# STATYSTYKA OPISOWA

**Jerzy Wierzbiński**

# STATYSTYKA OPISOWA

Jerzy Wierzbiński

Wydawnictwo Naukowe
Wydziału Zarządzania
Uniwersytetu Warszawskiego

Projekt okładki: Małgorzata Butkiewicz

Nadzór redakcyjny: Jerzy Jagodziński

© Copyright by Jerzy Wierzbiński, Warszawa 2008

Recenzent: Prof. dr hab. Jan B. Gajda

ISBN 978-83-61276-15-9

Opracowanie komputerowe, druk i oprawa:
Dom Wydawniczy ELIPSA,
ul. Inflancka 15/198, 00-189 Warszawa
tel./fax 022 635 03 01, 022 635 17 85,
e-mail: elipsa@elipsa.pl, www.elipsa.pl

# Spis treści


Wprowadzenie 5

Rozdział 1 **Podstawowe pojęcia** 7

1.1. Definicje statystyki . . . . . 7
1.2. Pojęcia wiążące się z definicją statystyki jako nauki . . . . . 8
1.3. Podział statystyki . . . . . 9
1.4. Podstawowe pojęcia statystyczne . . . . . 10
1.5. Pomiar statystyczny . . . . . 13
1.6. Badania statystyczne . . . . . 15

Rozdział 2 **Tabelaryczne formy prezentacji danych** 19

2.1. Formy prezentacji materiału statystycznego . . . . . 19
2.2. Tablice . . . . . 19
2.2.1. Tabelaryczne formy prezentacji cech jakościowych . . . . . 21
2.2.2. Tabelaryczne formy prezentacji cech ilościowych . . . . . 22
2.3. Zasady konstrukcji przedziałowych szeregów rozdzielczych . . . . . 25

Rozdział 3 **Graficzne formy prezentacji danych** 32

3.1. Typologia wykresów . . . . . 32
3.2. Wykresy powierzchniowe . . . . . 33
3.3. Wykresy w prostokątnym układzie współrzędnych . . . . . 36
3.4. Graficzne porównanie populacji . . . . . 40
3.5. Typy rozkładów empirycznych . . . . . 54

Rozdział 4 **Miary tendencji centralnej** 57

4.1. Średnia arytmetyczna . . . . . 57
4.2. Mediana i pozostałe kwantyle . . . . . 67
4.3. Dominanta (moda, wartość najczęstsza) . . . . . 79
4.4. Zależności $\bar{x}$, $M$, $D$ . . . . . 81
4.5. Średnia harmoniczna . . . . . 82
4.6. Średnia geometryczna . . . . . 84

Rozdział 5 **Miary zróżnicowania (zmienności, rozrzutu, dyspersji)** 88

5.1. Geneza miar . . . . . 88
5.2. Sposoby obliczania odchylenia standardowego . . . . . 93
5.3. Momenty statystyczne . . . . . 96
5.4. Ocena siły zróżnicowania . . . . . 97
5.5. Związki odchylenia standardowego z innymi miarami statystycznymi . . 99
5.6. Równość wariancyjna . . . . . 104

Rozdział 6 **Miary asymetrii (skośności)** 107

Rozdział 7 **Miary spłaszczenia (kurtozy) i miary koncentracji** 114

7.1. Miary spłaszczenia (wysmukłości, kurtozy) . . . . . 114
7.2. Miary koncentracji . . . . . 122

Rozdział 8 **Przykładowe zadania z rozwiązaniami** 128

8.1. Zadania kompleksowe dla szeregu przedziałowego . . . . . 128
8.2. Metody graficzne w analizie struktury zbiorowości . . . . . 146
8.3. Zadania na obliczanie średniej geometrycznej . . . . . 148

Rozdział 9 **Porównywanie zbiorowości statystycznych** 152

9.1. Porównania z użyciem narzędzi graficznych . . . . . 153
9.2. Porównania z użyciem miar statystycznych . . . . . 160

Rozdział 10 **Miary związku między zmiennymi ilościowymi: analiza regresji i współczynnik korelacji liniowej** 174

10.1. Związek liniowy między zmiennymi ilościowymi . . . . . 175
10.2. Regresja wielokrotna. Określanie związku zmiennej objaśnianej z więcej niż jednym predyktorem . . . . . 188

Bibliografia 200

# Wprowadzenie

*„Cała wiedza jest, w jej ostatecznej analizie, historią.*
*Wszystkie nauki przyrodnicze są, w ich najogólniejszej kwintesencji, matematyką.*
*Wszystkie sądy są, w ich podstawowych uzasadnieniach, statystyką..."*

C. Radhakrishna Rao, Statystyka i prawda, PWN, Warszawa 1994, s.12

Postęp wiedzy jest oparty na zdolności do wykrywania prawidłowości. Obserwacja pojedynczych zdarzeń uwarunkowanych wielozmiennowo jest obarczona dużym błędem losowym i nie pozwala na zrozumienie prawidłowości nimi rządzących.

Obserwacja zjawisk masowych daje nam poczucie chaosu, ale zastosowanie metod statystyki pozwala wyeliminować elementy nieistotne, ujawniając w tym „chaosie" prawidłowości.

Dzięki metodom statystyki opisowej, opisanym w tym podręczniku, potrafimy podsumować zbiór obserwacji, wychwycić tendencje ogólne abstrahując od szczegółów. Metody statystyczne znajdują zastosowanie przy opisie i analizie zjawisk przyrodniczych, społecznych i technicznych. Wykorzystuje się je m.in. formułując plany ekonomiczne i społeczne w skali makro, opracowując zasady statystycznej kontroli jakości, planując procesy produkcji, prognozując popyt na dobra i usługi, doskonaląc procesy diagnoz medycznych, przewidując pogodę, badając opinię publiczną, ustalając stawki w ubezpieczeniach na życie, używając dowodów statystycznych w sądzie itd.

Trudno sobie wyobrazić menedżera, ekonomistę, lekarza, prawnika, socjologa czy inżyniera nie rozumiejącego metod statystycznych. Nic więc dziwnego, że statystyki uczy się na prawie wszystkich kierunkach studiów.

Materiały zawarte w podręczniku są wykorzystywane (i udoskonalane) przeze mnie od ponad 20 lat podczas prowadzenia zajęć ze statystyki na Wydziale Zarządzania Uniwersytetu Warszawskiego. Ograniczona z konieczności liczba godzin, poświęconych przedmiotom ilościowym sprawia, że brakuje czasu na przedstawienie zagadnień, bez których zrozumienie wnioskowania statystycznego jest praktycznie

niemożliwe. W podręczniku zostały rozwinięte (w innych książkach dość marginalnie traktowane) zasady tworzenia tablic statystycznych, konstrukcja wykresów, interpretacja współzależności miar statystyki opisowej. Podane wzory pozwalają na wykonywanie wszystkich obliczeń za pomocą komputera. Zanim jednak to zrobimy warto rozwiązać przykładowe zadania „ręcznie". Ten trud — wbrew pozorom — bardzo się opłaci, ponieważ bezpośredni kontakt z liczbami pozwoli — po osiągnięciu pewnej wprawy — lepiej poczuć istotę wyliczanych miar. Warto więc czytać ten podręcznik aktywnie — z ołówkiem i kalkulatorem, sprawdzać wszystkie obliczenia po kolei. Tylko aktywność własna może przynieść efekty. Po latach dominacji metod dydaktycznych opartych na unikaniu obciążania uczących się „niepotrzebnym wysiłkiem", karierę robi japońska szkoła nauczania matematyki, która podstawową rolę przypisuje wyuczeniu pewnych umiejętności, uznając uzyskanie wglądu za wtórne. Do tego „japońskiego" podejścia zachęcam Czytelnika. Wykonanie przykładowych zadań zawartych w rozdziale 8 jest bardzo potrzebne. Inaczej będzie to jak oglądanie kasety z nauką jazdy samochodem. Oczywiście, że można ją obejrzeć, ale trzeba też usiąść za kierownicą. Rozwiązywanie zadań w podręczniku, to jak jazdy z instruktorem. Przygotują one Państwa do samodzielnego prowadzenia samochodu. Nawet jeżeli jesteśmy przekonani, że zrozumieliśmy jak się liczy miary tendencji centralnej, zróżnicowania, skośności, koncentracji, to prawdziwe ukorzenienie tej wiedzy nastąpi dopiero wtedy, gdy samodzielnie opiszemy kilkanaście zbiorów obserwacji. Możemy to potem już robić z wykorzystaniem pakietów statystycznych lub wprowadzając wzory podane w książce do Excela. Przykładowo, przedstawione wyniki analizy regresji w rozdziale 10 zostały wykonane za pomocą pakietu statystycznego *Statistical Package for Social Sciences* (SPSS).

Podręcznik jest rozwinięciem cieszącego się dużym zainteresowaniem, wyczerpanego już skryptu mojego autorstwa „Podstawy statystyki opisowej" (1998/2000). Oprócz udoskonalonych treści zawiera on nowe, bardzo ważne rozdziały takie jak porównywanie zbiorowości statystycznych czy wieloraką analizę regresji.

Podsumowując: statystyka jest samodzielną dyscypliną naukową zajmującą się metodami badań zjawisk masowych. Zjawiska masowe możemy opisywać za pomocą pojęć językowych, schematów, wykresów oraz liczb. Opis statystyczny jest opisem liczbowym. Tych, co jeszcze wątpią, czy warto uczyć się statystyki, mogę zapewnić, że nauczenie się metod statystyki opisowej przyczyni się nie tylko do podejmowania lepszych decyzji przez każdego z nas, ale także do lepszego zrozumienia otaczającej nas rzeczywistości. Ten wysiłek opłaca się stokrotnie.

*Jerzy Wierzbiński*

Warszawa, 14 listopad 2005

# Rozdział 1
# Podstawowe pojęcia

## 1.1. Definicje statystyki

Obecnie terminu „statystyka" używa się w kilku różnych znaczeniach, a mianowicie jako:

1) nazwy zbiorów danych dotyczących zjawisk gospodarczych, społecznych i przyrodniczych (w tym znaczeniu mówi się np. o statystyce ludności, statystyce przemysłu, statystyce wydobycia węgla, statystyce handlu zagranicznego, statystyce opadów);
2) nazwy czynności zbierania i opracowywania danych liczbowych (np. sporządzanie statystyki zatrudnienia to opracowywanie zestawień dotyczących poziomu zatrudnienia w określonym miejscu i czasie). W Polsce w jednolitym systemie ewidencji i statystyki gospodarki narodowej rozróżnia się:
   - ewidencję operatywną (bieżące notowanie zdarzeń),
   - księgowość (rejestrowanie ruchu wartości materialnych i pieniężnych),
   - statystykę (obliczanie uogólnionych charakterystyk liczbowych opartych na danych pochodzących z ewidencji operatywnej i księgowości);
3) nazwy charakterystyk liczbowych obliczonych z próby, np. średnią arytmetyczną i odchylenie standardowe wyznaczone dla próby nazwiemy ogólnie (wywoławczo) statystykami. Analogiczne miary, tj. średnią arytmetyczną i odchylenie standardowe policzone dla populacji generalnej nazywa się parametrami;
4) nauki o ilościowych metodach badania prawidłowości zjawisk masowych (najczęściej spotykana definicja w krajowych podręcznikach statystyki). Jedna z wielu bardziej opisowych definicji statystyki brzmi: nauka zajmująca się zbieraniem, przetwarzaniem, analizą i interpretacją danych liczbowych ze specjalnego typu zbiorowości. Statystyka ustala wielkości, za pomocą których można opisywać właściwości zjawisk masowych oraz podaje metody służące do ustalania związków między tymi wielkościami.

## 1.2. Pojęcia wiążące się z definicją statystyki jako nauki

1. Rzecz — fakt empiryczny podlegający obserwacji za pomocą dostępnych metod i środków, przejaw obiektywnej, poznawalnej rzeczywistości.
2. Zjawisko (zdarzenie) — utrata jakiejś cechy lub nabycie nowej cechy przez daną rzecz.
3. Proces — ciąg następujących po sobie i wzajemnie uzależnionych zjawisk.
4. Relacja zdarzenie — proces. Zdarzeniem nazywamy taką rozciągniętą w czasie obserwację, w której uwaga nasza koncentruje się jedynie na różnicy między stanem początkowym a stanem końcowym, podczas gdy w przypadku obserwacji procesu staramy się uchwycić pewien ciąg zdarzeń (z reguły więcej niż dwu) czy stanów przejściowych prowadzących od stanu początkowego do stanu końcowego.
5. Cecha — treść wrażenia wywołanego przez daną rzecz, inaczej właściwość jednostki pozwalająca ją odróżnić od innych jednostek.
6. Zjawiska masowe — zjawiska, które rozpatrywane w dużej masie odznaczają się prawidłowością nie dającą się ustalić na podstawie pojedynczej obserwacji. Mogą występować praktycznie nieograniczoną liczbę razy.
7. Prawidłowość — obiektywne, stale powtarzające się związki lub relacje stanów rzeczy. Atrybutami prawidłowości są: powtarzalność, obecność następstw i porządek w zjawiskach. Stanowisko w sprawie genezy prawidłowości statystycznych w uproszczeniu można sprowadzić do następujących tez:

   a) Prawidłowości statystyczne powstają w wyniku działania przyczyn głównych, tj. takich, które działają we wszystkich przypadkach oraz w obecności przyczyn ubocznych działających w indywidualnych, poszczególnych przypadkach.

   b) Gdyby działały wyłącznie przyczyny główne i gdyby każda jednostka w danej masie podlegała tylko wpływom przyczyn głównych, zawsze tych samych, prawidłowość byłaby absolutna. Dałaby się zauważyć w każdym poszczególnym przypadku.

   c) Działanie przyczyn ubocznych powoduje, że poszczególne zdarzenia mogą poważnie odbiegać od ogólnej prawidłowości. Przypadkowość jest skutkiem przyczyn nie powiązanych z systemem przyczyn głównych.

d) Przyczyny główne powodują, ze istnieje prawidłowość, natomiast przyczyny uboczne sprawiają, że prawidłowość ta jest tylko przybliżona, że wyraża się tylko w masie.

e) Łączenie zdarzeń w duże masy kompensuje działanie przyczyn ubocznych, wzajemne znoszenie się dodatniej i ujemnej (w stosunku do procesu głównego) przypadkowości.

W miarę zwiększania ilości zdarzeń konieczne związki bardziej uwypuklają się drogą wyrównywania się skutków przypadkowych.

## 1.3. Podział statystyki

1. Statystyka opisowa — podejście deterministyczne — dział zajmujący się metodami zbierania danych o zbiorowości, ich prezentacji oraz analizy:
   a) struktury statystycznej z punktu widzenia wybranych cech,
   b) występowania współzależności cech w danej zbiorowości,
   c) dynamiki zjawisk w zbiorowości.
2. Statystyka matematyczna — podejście probabilistyczne — dział statystyki zajmujący się metodami wnioskowania o właściwościach zbiorowości generalnej, jej parametrach oraz postaci rozkładu badanych cech, na podstawie danych z prób losowych pobranych z tej populacji.

<table>
<tr><td rowspan="2">KRYTERIUM ZAKRESOWO-PRZEDMIOTOWE</td><td colspan="2">KRYTERIUM FORMALNO-STATYSTYCZNE</td></tr>
<tr><td>OPIS</td><td>WNIOSKOWANIE</td></tr>
<tr><td rowspan="2">ANALIZA<br>STRUKTURY ZJAWISK</td><td colspan="2">MIARY TENDENCJI CENTRALNEJ<br>MIARY DYSPERSJI</td></tr>
<tr><td>MIARY ASYMETRII<br>MIARY SPŁASZCZENIA<br>MIARY KONCENTRACJI</td><td>————</td></tr>
<tr><td>ANALIZA WSPÓŁZALEŻNOŚCI<br>ZJAWISK</td><td colspan="2">MIARY REGRESJI I KORELACJI DWÓCH CECH<br>MIARY REGRESJI I KORELACJI WIELORAKIEJ</td></tr>
<tr><td rowspan="2">ANALIZA DYNAMIKI<br>ZJAWISK</td><td colspan="2">MIARY TENDENCJI ROZWOJOWEJ<br>MIARY OKRESOWOŚCI<br>MIARY WAHAŃ RESZTOWYCH</td></tr>
<tr><td>INDEKSY STAT.<br>— INDYWIDUALNE<br>— AGREGATOWE</td><td>————</td></tr>
</table>

Źródło: A. Luszniewicz, *Statystyka ogólna*, PWE, Warszawa 1987, s. 28.

# 1.4. Podstawowe pojęcia statystyczne

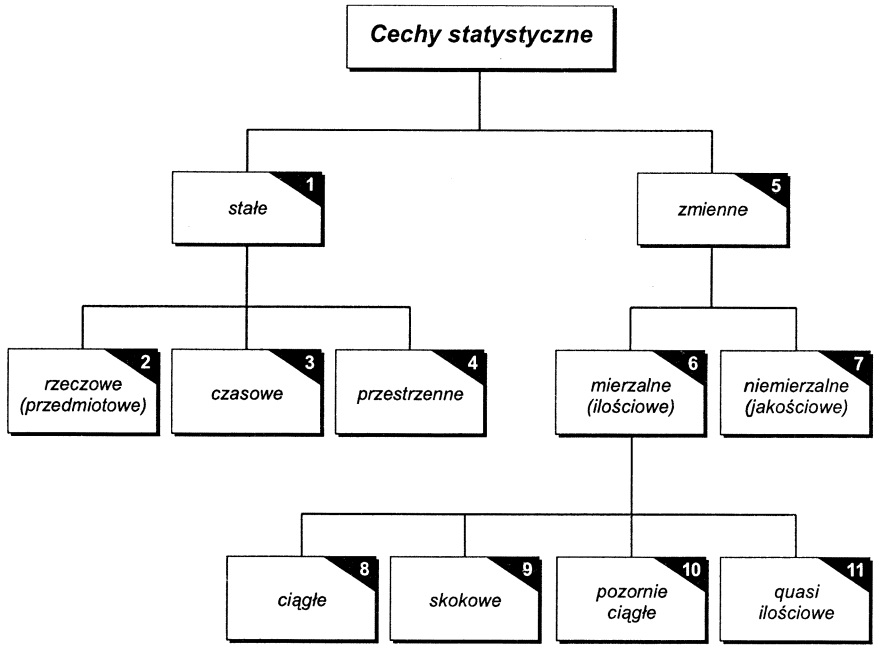
1. Zbiorowość (populacja, masa) statystyczna — ogół rzeczy jednoznacznie określonych i wyodrębnionych, posiadających przynajmniej jedną wspólną cechę różniącą się wartościami; inaczej zbiór elementów nieidentycznych z punktu widzenia badanej cechy.
2. Typy zbiorowości statystycznych:

| | | | |
|---|---|---|---|
| a1. | skończenie liczne, w sensie matematycznym przeliczalne<br>↙ ↘<br>skończone: np. ludność Polski<br>nieskończone: np. gwiazdy (jeżeli Wszechświat jest nieskończony); samochody, które mogą być wyprodukowane; wszystkie możliwe ciągi literowe | a2. | nieskończenie liczne, w sensie matematycznym nieprzeliczalne<br>np. kulki łożyskowe ze względu na możliwe masy czy średnice; drzewa ze względu na możliwą wysokość |
| b1. | jednowymiarowe<br>populacja badana ze względu na jedną cechę | b2. | wielowymiarowe<br>populacja badana ze względu na dwie lub więcej cech |
| c1. | statyczne<br>populacja badana na dany moment, np. ludności Polski w NSP | c2. | dynamiczne<br>badanie w przedziale czasu, najczęściej interesuje nas proces |
| d1. | względnie jednorodne<br>podstawowym kryterium różnicującym jednostki populacji jest ilość. Jednostki tego samego typu, rodzaju, gatunku, np. śledzie, sarny, studenci ze względu na wyniki w nauce, studenci uczelni wojskowych ze względu na zainteresowania, poborowi, księgowi | d2. | względnie niejednorodne<br>podstawowym kryterium różnicującym jednostki populacji jest jakość. W skład populacji wchodzą jakościowo odmienne jednostki lub grupy jednostek, np. studenci ze względu na wzrost czy wagę (kobiety i mężczyźni razem), akcje i obligacje, złowione ryby, wszystkie zwierzęta hodowlane |

3. Populacja generalna — obejmuje wszystkie elementy będące przedmiotem badania, w odniesieniu do których formułować będziemy wnioski ogólne.
4. Jednostka statystyczna — element zbiorowości statystycznej.

5. Cecha danej rzeczy — treść wrażenia wywołanego przez tę rzecz. Treść tę, jedynie myślowo, potrafimy rozłożyć na trzy składowe:
   - jakość — to, co pozwala odróżnić jedno wrażenie od drugiego. Podstawowe różnice jakości zachodzą między wrażeniami dostarczonymi przez różne zmysły,
   - intensywność (siłę) — to, co pozwala odróżnić wrażenia o tej samej jakości,
   - czas trwania.

   Jeżeli w ramach tej samej jakości, np. goryczy, rozważamy jedynie intensywność wrażenia, to skrótowo mówimy, że mamy do czynienia z cechą ilościową. Jeśli w danych warunkach nie interesują nas różnice intensywności wrażeń w ramach tej samej jakości, to skrótowo powiemy, że mamy do czynienia z cechą jakościową.




Wykres 1.1. Podział cech statystycznych

Ad 1. **Cechy stałe** — cechy wspólne wszystkim jednostkom danej zbiorowości, nie podlegają badaniu. Określają kogo lub co badamy (cechy rzeczowe 2), kiedy (cechy czasowe 3) oraz gdzie (cechy przestrzenne 4).

Ad 5. **Cechy zmienne** — właściwości, którymi różnią się jednostki poddane badaniu.

Ad 6. **Cechy mierzalne** — wyrażane za pomocą liczb.

Ad 7. **Cechy niemierzalne** — (jakościowe)wyrażane słownie lub za pomocą skal numerycznych. Istnienie lub nieistnienie danego wariantu cechy niemierzalnej można jedynie stwierdzić, ale nie można go zmierzyć.

Ad 8. **Cechy ciągłe** — w danym przedziale mogą przyjmować wszystkie wartości na skali liczbowej (zbiór ich wartości jest nieprzeliczalny).

Ad 9. **Cechy skokowe** — cechy, które mogą przyjmować jedynie wyróżnione wartości na osi liczbowej (w większości przypadków całkowitoliczbowe). Ich zbiór wartości jest przeliczalny.

Ad 10. **Cechy pozornie ciągłe** — cechy skokowe, które traktowane są jak cechy ciągłe. Czynimy tak wówczas, gdy różnice między kolejnymi możliwymi wartościami cechy skokowej są tak małe, że nie mają praktycznego znaczenia.

Ad 11. **Cechy quasi ilościowe** — cecha wyrażona opisowo. Opis określa natężenie cechy u poszczególnych jednostek. Podstawą wyróżnienia klas jest jednoznaczne kryterium ilościowe.

***Uwaga 1.*** Cechy stałe w jednym badaniu (np. bycie krową w badaniu mleczności) mogą być cechami zmiennymi (w badaniu stanu zwierząt gospodarskich).

***Uwaga 2.*** Cechy ciągłe przedstawiane są ze skończoną dokładnością (określona liczba miejsc po kropce dziesiętnej). Nie przestają mimo to być cechami ciągłymi.

Przykładowe cechy:

- **jakościowe**: płeć, wykształcenie, zawód, miejsce zamieszkania, rodzaj prowadzonej działalności gospodarczej, znajomość języków obcych, marka samochodu, kolor włosów, stosunek do programu uwłaszczenia obywateli;
- **mierzalne skokowe**: liczba akcji, które zmieniły właścicieli na sesji giełdy, wielkość rodziny, liczba dzieci w klasie, liczba sztuk wadliwych w partii wyrobów, liczba ziaren w kłosie, liczba zgłoszeń na ostrym dyżurze w szpitalu, liczba samochodów przekraczających przejścia graniczne w danym dniu, rozmiary obuwia;
- **mierzalne ciągłe**: długość ciętego elementu, masa tabliczki czekolady, objętość detergentu w pojemniku, droga pokonana przez pojazd na jednym litrze paliwa, temperatura w pomieszczeniu, czas wykonywania czynności, staż pracy, wiek, wysokość plonów;
- **mierzalne pozornie ciągłe**: wszystkie wyrażenia wartościowe, takie jak płace, zarobki, ceny, dywidendy.

***Przykład 1.1.***

| Populacja | Podpopulacja | Jednostka statystyczna | Cechy |
|---|---|---|---|
| Firmy mające siedzibę na danym terenie w określonym czasie | Firmy ubezpieczeniowe, Banki, Firmy elektroniczne, Sklepy spożywcze itp. | Pojedyncza firma Sklep chemiczny, restauracja Kuchcik, kantor wymiany Miliard, biuro pośrednictwa itp. | Rodzaj prowadzonej działalności, rok założenia, liczba zatrudnionych, obroty, roczna sprzedaż, zysk, płace. |

## 1.5. Pomiar statystyczny

1. Zmienną o danym zbiorze wartości, za pomocą której określamy jakąś rzecz, nazywamy miarą tej rzeczy.
2. Czynność odwzorowania danej rzeczy w jedną z wartości zmiennej nazywa się pomiarem tej rzeczy. Inaczej, pomiar to taki sposób manipulowania symbolami, w którym istotne relacje między symbolami pozostają takie same, jak relacje między przedstawionymi przez te symbole rzeczami.
3. Celem pomiaru jest przedstawienie treści obserwacji symbolami związanymi ze sobą tak samo jak są lub mogą być ze sobą połączone analizowane rzeczy, czy przedstawiające je pojęcia.
4. Użycie liczb jako symboli wynika stąd, że:
   a) związki pomiędzy liczbami zostały dobrze rozpoznane przez matematykę,
   b) pewne z tych związków mają swe odniesienie w relacjach między rzeczami.
5. Liczby mają wiele właściwości, które nie przysługują analizowanym rzeczom. Znajdując wartości charakterystyk statystycznych należy zastanowić się jakie z operacji matematycznych znajdują w konkretnym przypadku zastosowanie.

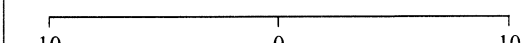
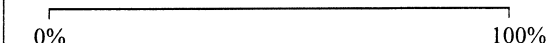
Skale pomiarowe:

1. Skala nominalna — najniższy typ skali pomiarowej. Na jej gruncie ma miejsce klasyfikowanie elementów. Dla dowolnej pary możemy jedynie powiedzieć, czy jest pod pewnym względem jednakowa, czy też nie. Liczby mogą być użyte wyłącznie jako nazwy klas, znacząc tyle samo, co numer na koszulce sportowca. Elementy w klasach mogą być zliczane. Miara statystyczna do wykorzystania: dominanta, czyli moda. Przykłady cech do przedstawienia w skali nominalnej: zawód, płeć, stan cywilny, nazwy firm, numery statystyczne przedsiębiorstw, kolor włosów.

2. Skala porządkowa (rangowa) — dotyczy również cech jakościowych, polega na uporządkowaniu elementów bez możliwości określenia odległości między nimi. Możliwe do użycia miary statystyczne (prócz przechodzących z poprzedniego poziomu): mediana, korelacje. Przykłady cech: poziom wykształcenia, ranga w strukturze organizacyjnej, jakość jakiegoś dobra, zdolność kredytowa firmy (duża, średnia, mała), stopnie wojskowe, pozycja społeczna.
3. Skala przedziałowa (interwałowa) — dotyczy cech mierzalnych. Zasadza się na przyjęciu stałej różnicy jako jednostki porównań między elementami w sytuacji, gdy nie istnieje (lub go nie znamy) naturalny początek. Zostają ustalone obiekty wzorcowe oraz wzorcowa różnica. Klasycznym przykładem jest temperatura. Raz przyjęto jako umowny punkt zerowy temperaturę czystej, zamarzającej wody, w innym przypadku temperaturę wody zmieszanej z salmiakiem itd. Dla cech wyrażonych w tej skali uprawnione są operacje dodawania i odejmowania (np. dzisiaj może być cieplej o 2 stopnie). Nie mają sensu działania mnożenia i dzielenia. Z tego, ze dzisiaj jest 8°C, a wczoraj były tylko 4°C nie wynika, że jest dwa razy cieplej. Miary statystyczne do wykorzystania: średnia arytmetyczna (obliczamy np. średnią temperaturę października) i odchylenie standardowe. Przykłady innych cech: czas kalendarzowy, iloraz inteligencji, wydajność uczenia się, zdolność zapamiętywania, wysokość nad poziom morza, zdolność kredytowa firmy (np. w skali 0–10).

***Przykład 1.2.***

Użycie różnych skal dla wyrażenia stosunku do słodyczy:

| | |
|---|---|
| 1. Skala nominalna | ● nie lubię słodyczy ● lubię słodycze |
| 2. Skala porządkowa (werbalna) | ● nie lubię ● lubię ● bardzo lubię |
| 2A. Skala porządkowa (graficzna) | zupełnie nie lubię; nie lubię; są mi obojętne; lubię; bardzo lubię |
| 3. Skala przedziałowa |  −10 0 10 |
| 4. Skala ilorazowa (graficzna) |  0% 100% |
| 4A. Skala ilorazowa: | ile tygodniowo wydaję na słodycze |

4. Skala stosunkowa (ilorazowa) — polega na uporządkowaniu z oznaczeniem odległości elementów od naturalnego początku. Możliwe jest wykonywanie wszystkich operacji arytmetycznych. Skala najsilniejsza. Odnosi się do wszystkich cech mierzonych za pomocą CGS i SI. Miara statystyczna: średnia geometryczna. Przykłady cech: długość, masa, objętość, wyrażenia wartościowe, czas (dojazdu do pracy).

## 1.6. Badania statystyczne

Badanie statystyczne jest to badanie poświęcone właściwościom zbiorowości.

1. Każde badanie realizowane jest w ramach określonych sił i środków.
2. Nie ma dwóch identycznych badań statystycznych.
3. Wspólne wszystkim badaniom są cztery etapy:
   a) projektowania (przygotowania, programowania),
   b) obserwacji statystycznej,
   c) opracowania wyników,
   d) analizy statystycznej.

<table>
<tr><td rowspan="13">1. PROGRAMOWANIE BADANIA STATYSTYCZNEGO</td><td colspan="4">Określenie celów badania statystycznego</td></tr>
<tr><td colspan="4">cele główne i szczegółowe, hipotezy badawcze</td></tr>
<tr><td colspan="4">Zakres podmiotowy badania<br>(definicja zbiorowości i jednostki statystycznej)</td></tr>
<tr><td colspan="2">rzeczowy</td><td>terytorialny</td><td>czasowy</td></tr>
<tr><td colspan="4">skończony / nieskończony<br>statyczny / dynamiczny<br>względnie jednorodny / względnie niejednorodny</td></tr>
<tr><td colspan="4">Zakres przedmiotowy badania (program obserwacji)</td></tr>
<tr><td colspan="2">zmienne zależne</td><td colspan="2">zmienne niezależne</td></tr>
<tr><td>ilościowe</td><td>jakościowe</td><td>ilościowe</td><td>jakościowe</td></tr>
<tr><td colspan="4">skale pomiarowe</td></tr>
<tr><td>nominalne</td><td>porządkowe</td><td>przedziałowe</td><td>ilorazowe</td></tr>
<tr><td colspan="4">określenie źródeł danych</td></tr>
<tr><td colspan="2">badanie pierwotne</td><td colspan="2">badanie wtórne</td></tr>
</table>

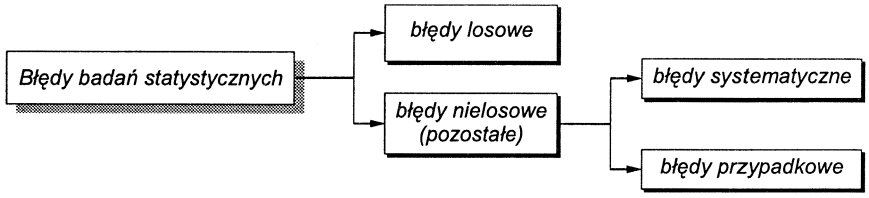
<table>
<tr><td rowspan="11">1. PROGRAMOWANIE BADANIA STATYSTYCZNEGO (c.d.)</td><td colspan="6">Metoda badania (metoda obserwacji empirycznej)</td></tr>
<tr><td rowspan="2">pełne</td><td colspan="3">częściowe</td><td colspan="2">szacunek</td></tr>
<tr><td>wybór celowy</td><td>wybór losowy</td><td>wybór celowo-losowy</td><td>interpolacja</td><td>ekstrapolacja</td></tr>
<tr><td colspan="6">Rodzaj i częstotliwość badania</td></tr>
<tr><td colspan="2">ciągłe</td><td colspan="2">okresowe</td><td colspan="2">doraźne</td></tr>
<tr><td colspan="6">Techniki obserwacji statystycznej</td></tr>
<tr><td colspan="6">eksperyment, spis, monografia, wywiad, ankieta, obserwacja, test, sprawozdawczość, badanie opinii, ewidencja, szacunki, wtórne wykorzystanie zbiorów danych</td></tr>
<tr><td colspan="6">Narzędzia badawcze, grupowania i przekroje stosowane w opisie</td></tr>
<tr><td colspan="6">Program przetwarzania danych</td></tr>
<tr><td colspan="6">wartości zmienne cech; makiety tablic wynikowych, algorytmy przetwarzania</td></tr>
<tr><td colspan="6">Harmonogram prac</td></tr>
<tr><td colspan="6">Pole zastosowań wyników badania</td></tr>
<tr><td colspan="2">użyteczność teoretyczna</td><td colspan="2">użyteczność metodologiczna</td><td colspan="2">użyteczność praktyczna</td></tr>
<tr><td rowspan="4">2. OBSERWACJA STATYSTYCZNA</td><td colspan="6">Badanie pilotażowe</td></tr>
<tr><td colspan="6">sprawdzenie terenu, sprawdzenie narzędzi badawczych</td></tr>
<tr><td colspan="6">Badanie zasadnicze</td></tr>
<tr><td colspan="3">bezpośrednie</td><td colspan="3">przez ankieterów</td></tr>
<tr><td rowspan="8">3. OPRACOWANIE WYNIKÓW BADAŃ</td><td colspan="6">Kontrola zebranego materiału</td></tr>
<tr><td colspan="3">formalna</td><td colspan="3">merytoryczna</td></tr>
<tr><td colspan="6">błędy pomiaru (przypadkowe i systematyczne)</td></tr>
<tr><td colspan="6">Grupowanie</td></tr>
<tr><td colspan="3">typologiczne</td><td colspan="3">wariancyjne</td></tr>
<tr><td colspan="6">Zliczanie danych</td></tr>
<tr><td colspan="6">Prezentacja materiału</td></tr>
<tr><td colspan="2">tablice, wykresy</td><td colspan="2">charakterystyki opisowe</td><td colspan="2">wzory analityczne</td></tr>
<tr><td rowspan="3">4. ANALIZA STATYSTYCZNA</td><td colspan="6">Analiza całej populacji (opis statystyczny)</td></tr>
<tr><td colspan="6">Analiza próby (wnioskowanie statystyczne)</td></tr>
<tr><td colspan="6">Analiza przedmiotowa i metodologiczna</td></tr>
</table>

4. Rodzaje badań statystycznych:

| Badania | Ciągłe | Okresowe | Doraźne |
|---|---|---|---|
| Całkowite | | | |
| Częściowe | | | |

Główne przyczyny określające potrzebę prowadzenia badań częściowych: brak czasu, koszty, niszczenie lub zmniejszenie wartości użytkowej, nieskończona liczba jednostek populacji, niedostępne dane, możliwość lepszego przygotowania badania, potrzebne są jedynie wyniki orientacyjne.

5. Błędy badań statystycznych:

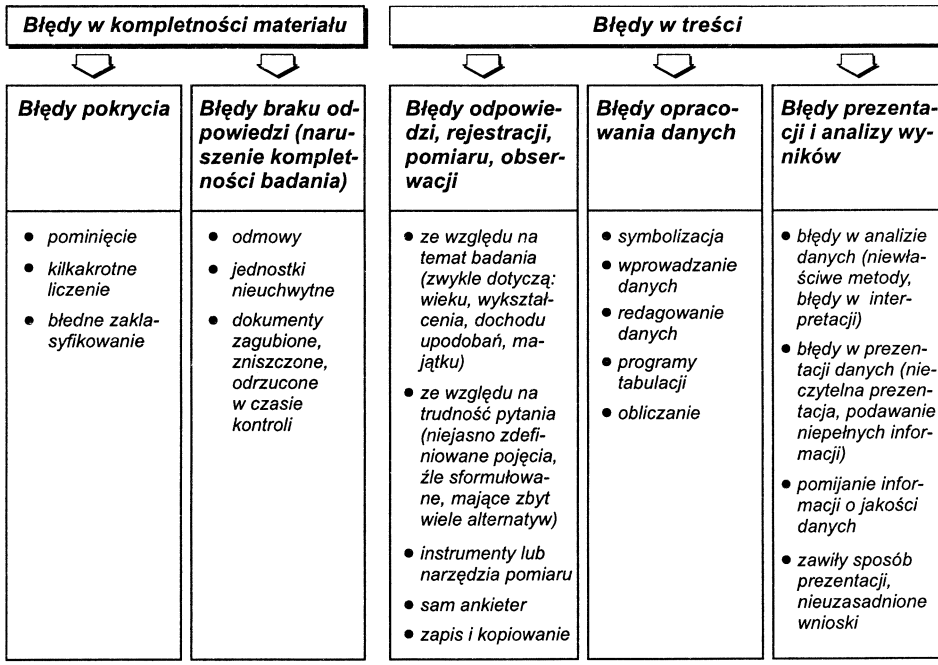



***Błąd losowy***

Błąd badania statystycznego w przypadku, gdy próba została wybrana metodą wyboru losowego. Ta część różnicy między wartościami z populacji, a ich oszacowaniami otrzymanymi na podstawie próby, która spowodowana jest korzystaniem z danych próby losowej, a nie z informacji o całej populacji generalnej (odchylenie jest efektem zdarzeń losowych). Rozmiary błędów losowych zależą od: liczebności próby, schematu losowania, metody szacowania parametru, zmienności badanych cech.

***Błędy nielosowe***

Ze względu na przyczynę, błędy nielosowe dzieli się na systematyczne i przypadkowe. Błędy systematyczne wynikają z jednokierunkowej tendencji do zniekształcania badanej rzeczywistości, błędy przypadkowe wywołane nieuwagą, niedbalstwem czy nieumiejętnością podania prawidłowej odpowiedzi powodują rozrzut wokół rzeczywistej wartości danej wielkości.




Wykres 1.2. Podział błędów nielosowych ze względu na miejsce ich powstawania.
Źródło: A. Kordos, *Jakość danych statystycznych*, PWE, Warszawa 1995.

# Rozdział 2
# Tabelaryczne formy prezentacji danych

## 2.1. Formy prezentacji materiału statystycznego

Formami prezentacji materiału statystycznego są:

1. Tablice
2. Wykresy
3. Charakterystyki opisowe
4. Wzory analityczne

## 2.2. Tablice

Tablica statystyczna to jeden lub więcej szeregów statystycznych. Szereg statystyczny (elementarna forma opisu badanej zbiorowości), to zbiór wyników obserwacji jednostek populacji według jednej cechy. Dane uzyskane po spisaniu z ankiet czy urządzeń pomiarowych nazywane są indywidualnym wykazem wartości lub szeregiem nieuporządkowanym.

***Przykład 2.1.***

Populacja: 64 dni pracy hurtowni Profi
Cecha: dzienna liczba sprzedanych palet

Indywidualny wykaz wartości

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | 16 | 22 | 9 | 22 | 12 | 39 | 19 | 14 | 23 | 6 | 24 | 16 | 18 | 7 | 17 |
| 20 | 25 | 28 | 18 | 10 | 24 | 20 | 21 | 10 | 7 | 18 | 28 | 24 | 20 | 14 | 23 |
| 25 | 34 | 22 | 5 | 33 | 23 | 26 | 29 | 13 | 36 | 11 | 26 | 11 | 37 | 30 | 13 |
| 8 | 15 | 22 | 21 | 32 | 21 | 31 | 17 | 16 | 23 | 12 | 9 | 15 | 27 | 17 | 21 |

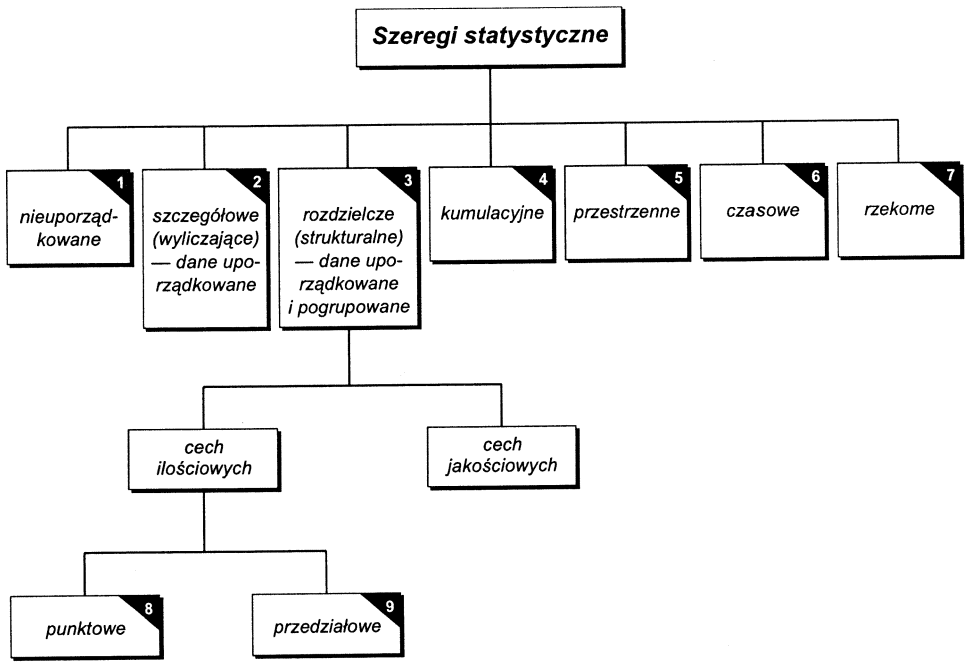
Dane te — po pierwsze — muszą zostać uporządkowane (uporządkowanie to wprowadzenie do zgromadzonych obserwacji ładu opartego na przyjętych kryteriach, najczęściej ustawienie jednostek od najmniejszej do największej).

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 6 | 7 | 7 | 8 | 9 | 9 | 10 | 10 | 11 | 11 | 12 | 12 | 13 | 13 | 14 |
| 14 | 15 | 15 | 16 | 16 | 16 | 17 | 17 | 17 | 18 | 18 | 18 | 19 | 19 | 20 | 20 |
| 20 | 21 | 21 | 21 | 21 | 22 | 22 | 22 | 22 | 23 | 23 | 23 | 23 | 24 | 24 | 24 |
| 25 | 25 | 26 | 26 | 27 | 28 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 36 | 37 | 39 |

Ta mało liczna populacja (tylko 64 elementy) mimo uporządkowania pozostaje mało czytelna. Uporządkowane dane poddaje się z kolei grupowaniu. Grupowanie to czynności związane z wyodrębnieniem jednorodnych lub względnie jednorodnych części zbiorowości statystycznej.

Rozróżnia się dwa rodzaje grupowań:

- typologiczne — prowadzi do rozróżnienia jakościowo odmiennych podzbiorowości,
- wariancyjne — podstawowym kryterium różnicującym klasy jest ilość.




Wykres 2.1. Typologia szeregów statystycznych

Ad 4. Szereg kumulacyjny to szereg zawierający liczebności lub częstości skumulowane (wyjaśnione na przykładzie na stronie 24).

Ad 7. Szereg rzekomy to szereg w którym liczebności nie są adekwatne względem wartości rozpatrywanej cechy.

Ad 8. Punktowemu ujęciu rozkładu jednej cechy mierzalnej odpowiadają jednostki zbiorowości statystycznej przyporządkowane konkretnym, pojedynczym wartościom wariantów cechy (najczęściej skokowej).

Ad 9. W ujęciu przedziałowym — jednostki zbiorowości przyporządkowane są wartościom cechy w granicach „od – do" tak dla cechy skokowej jak i cechy ciągłej.

### 2.2.1. Tabelaryczne formy prezentacji cech jakościowych

1. Notacja

$A$ — cecha niemierzalna,

$A_i$ — $i$-ty wariant cechy $A$, $i=1, 2, ..., k$,

$n_i$ — liczba elementów populacji posiadających $i$-ty wariant (typ),

$N$ — liczebność populacji, gdy mamy do czynienia z populacją generalną

$$N = \sum_{i=1}^{k} n_i$$

$n$ — liczebność populacji, gdy mamy do czynienia z próbą

$$n = \sum_{i=1}^{k} n_i$$

$w_i = \frac{n_i}{N}$ — częstość względna (wskaźnik struktury)

2. Gdy liczba wyróżnionych wariantów cechy wynosi dwa, szereg nazywa się dychotomicznym, gdy jest większa od dwóch — jest to szereg politomiczny.

| $A_i$ | $n_i$ | $w_i = \frac{n_i}{N}$ |
|---|---|---|
| $A_1$ | $n_1$ | $w_1$ |
| $A_2$ | $n_2$ | $w_2$ |
| ⋮ | ⋮ | ⋮ |
| $A_k$ | $n_k$ | $w_k$ |
| Razem | $N$ | 1 |

3. Przykład

Populacja: 20 pracowników

Cecha: stosunek do programu powszechnej prywatyzacji (A = bardzo pozytywny, B = pozytywny, C = negatywny, D = zdecydowanie negatywny)

Indywidualny wykaz wartości

B, B, A, B, C, B, B, C, D, A, B, D, C, C, D, B, B, B, D, B

**Szereg rozdzielczy**

| Poziom zadowolenia | $n_i$ | $w_i$ |
|---|---|---|
| A | 2 | 0,10 |
| B | 10 | 0,50 |
| C | 4 | 0,20 |
| D | 4 | 0,20 |
| Razem | 20 | 1,00 |

## 2.2.2. Tabelaryczne formy prezentacji cech ilościowych

1. Notacja

$X$ — cecha mierzalna,

$x_i$ — $i$-ty wariant badanej cechy $X$,

$n_i$ — liczebność odpowiadająca wariantowi $x_i$,

$k$ — liczba wyróżnionych klas,

$x_{0i}$ — dolna granica klasy $i$-tej,

$x_{1i}$ — górna granica klasy $i$-tej.

**Makieta szeregu punktowego**

| $x_i$ | $n_i$ |
|---|---|
| $x_1$ | $n_1$ |
| $x_2$ | $n_2$ |
| ⋮ | ⋮ |
| $x_k$ | $n_k$ |
| Razem | $N$ |

**Makieta szeregu przedziałowego**

| $x_{0i} - x_{1i}$ | $n_i$ |
|---|---|
| $x_{01} - x_{11}$ | $n_1$ |
| $x_{02} - x_{12}$ | $n_2$ |
| ⋮ | ⋮ |
| $x_{0k} - x_{1k}$ | $n_k$ |
| Razem | $N$ |

***Przykład 2.2.***

**Szereg punktowy dla danych o sprzedaży palet (w wersji horyzontalnej)**

| $x_i$ | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|
| $n_i$ | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 2 |

| $x_i$ | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
|---|---|---|---|---|---|---|---|---|---|
| $n_i$ | 2 | 2 | 3 | 3 | 3 | 2 | 3 | 4 | 4 |

| $x_i$ | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 |
|---|---|---|---|---|---|---|---|---|---|
| $n_i$ | 4 | 3 | 2 | 2 | 1 | 2 | 1 | 1 | 1 |

| $x_i$ | 32 | 33 | 34 | 36 | 37 | 39 | Razem |
|---|---|---|---|---|---|---|---|
| $n_i$ | 1 | 1 | 1 | 1 | 1 | 1 | 64 |

***Przykład 2.3.***

**Szereg przedziałowy dla danych o sprzedaży palet**

| $x_{0i} - x_{1i}$ | $n_i$ |
|---|---|
| 5–9 | 7 |
| 10–14 | 10 |
| 15–19 | 13 |
| 20–24 | 18 |
| 25–29 | 8 |
| 30–34 | 5 |
| 35–39 | 3 |
| Razem | 64 |

***Przykład 2.4.***

**Praktycznie spotykane przekształcenia kolumny liczebności bezwzględnych $n_i$**

| 1 | 2 | 3 | 4 | 5 | 6 | | 7 | | 8 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Kumulata | | Dekumulata | | |
| $x_{0i}$–$x_{1i}$ | $n_i$ | $w_i = \frac{n_i}{N}$ | $w_i$ | $w_i \cdot 100\%$ | $n_i(x \le x_{1i})$ | $w_i(x \le x_{1i})$ | $n_i(x \ge x_{0i})$ | $w_i(x \ge x_{0i})$ | $g_i = \frac{n_i}{h_i}$ |
| 5–9 | 7 | 0,109375 | 0,109 | 10,9 | 7 | 0,109 | 64 | 0,999 | 1,4 |
| 10–14 | 10 | 0,156250 | 0,156 | 15,6 | 17 | 0,265 | 57 | 0,890 | 2,0 |
| 15–19 | 13 | 0,203125 | 0,203 | 20,3 | 30 | 0,468 | 47 | 0,734 | 2,6 |
| 20–24 | 18 | 0,281250 | 0,281 | 28,1 | 48 | 0,749 | 34 | 0,531 | 3,6 |
| 25–29 | 8 | 0,125000 | 0,125 | 12,5 | 56 | 0,874 | 16 | 0,250 | 1,6 |
| 30–34 | 5 | 0,078125 | 0,078 | 7,8 | 61 | 0,952 | 8 | 0,125 | 1,0 |
| 35–39 | 3 | 0,046875 | 0,047 | 4,7 | 64 | 0,999 | 3 | 0,047 | 0,6 |
| Razem | 64 | 1,000000 | 0,999 | 99,9 | × | × | × | × | × |

***Uwaga 1.*** W kolumnie 4, gdzie zaokrąglenia spowodowały, że suma po kolumnie nie jest równa 1 możemy dodać brakujące 0,001 w wierszu gdzie najbliżej jest do zaokrąglenia w górę, tj. w wierszu pierwszym.

***Uwaga 2.*** Ostatnia, ósma kolumna to kolumna gęstości. Liczebności w klasach dzielone są przez długości klas ($h_i$). Długości (o nich dalej) klas w tej tablicy wynoszą każdorazowo 5. Gęstość mówi ile jednostek zbiorowości w danej klasie przypada na jednostkę długości przedziału. Gęstości możemy liczyć także dla częstości względnych $w_i$.

## 2.3. Zasady konstrukcji przedziałowych szeregów rozdzielczych

Znakomita większość badań statystycznych jest prezentowana za pomocą przedziałowych szeregów rozdzielczych. W oparciu o nie tworzy się następnie wykresy, znajduje charakterystyki liczbowe, wyciąga końcowe wnioski. Niepoprawnie zbudowana tablica statystyczna wprowadza do badania błąd systematyczny, a w sytuacjach krańcowych skazuje na porażkę wszelkie dalsze działania.

Szereg przedziałowy musi:

- spełniać kryteria formalne (takie jak rozłączność i zupełność klasyfikacji)
- zapewniać możliwość porównań w przestrzeni i czasie
- być maksymalnie dogodny do obliczeń i przedstawień graficznych.

Podstawowa dyrektywa jaką należy się kierować brzmi następująco: **należy tak dobrać przedziały klasowe, by szereg rozdzielczy odzwierciedlał możliwie szczegółowo, ale jednocześnie i przejrzyście obraz struktury badanej zbiorowości statystycznej.** Istotne jest sprawdzenie, czy jednostki populacji mają tendencję do skupiania się wokół jakichś wartości i czy jest to w miarę regularne.

W następnej kolejności brane są pod uwagę następujące reguły:

1. zupełność (klasyfikacja musi obejmować wszystkie dane, które przeszły kontrolę formalną i merytoryczną),
2. rozłączność (każdy element trafia tylko i wyłącznie do jednej klasy),
3. rzadko kiedy stosuje się mniej niż 6 i więcej niż 15 klas,
4. środki przedziałów klasowych powinny być bliskie środkom ciężkości klasy (inaczej w przybliżeniu równe średnim arytmetycznym elementów przedziału),
5. należy unikać tworzenia klas otwartych (tzn. bez określenia dolnej lub górnej granicy interwału),
6. w szeregu nie powinno być klas o liczebności zero,
7. jest pożądane by długości klas w szeregu były takie same (stałe),
8. ułatwi dalsze przekształcenia, jeśli długości klas są krotnościami pięciu,
9. należy dążyć by środki przedziałów były liczbami całkowitymi,
10. można się spotkać z wyróżnianiem przedziałów o największej liczebności (kosztem stałej długości klas).

Budując szereg rozdzielczy ustalamy:

A. liczbę przedziałów,

B. rozpiętości (długości) klas,

C. sposób określania granic przedziałów.

Ad. A. Na decyzję o liczbie klas wpływają:

1. liczebność zbiorowości (im liczniejsza populacja tym może być ich więcej),
2. rodzaj cechy — skokowa lub ciągła — oraz liczba wyróżnionych wariantów cechy,
3. cel badania (szczegółowsze badanie wymaga większej dokładności),
4. rozstęp (dla dłuższego obszaru zmienności cechy możliwość użycia większej liczby klas),
5. minimalna różnica między dwiema kolejnymi odmianami cechy.

Ad. B. Szereg rozdzielczy o takich samych (stałych) długościach klas jest znacznie dogodniejszy w obróbce statystycznej.

Ad. C. W dalszym ciągu przydatne będą następujące określenia:

- dolna granica klasy $i$-tej ($x_{0i}$) — najmniejsza wartość, jaką może przyjąć jednostka w danej klasie,
- górna granica klasy $i$-tej ($x_{1i}$) — największa wartość, jaką może przyjąć jednostka w klasie $i$-tej,
- środek przedziału $i$-tego ($\mathring{x}_i$) — punkt równo oddalony od granic klasy, inaczej średnia arytmetyczna granic

$$\mathring{x}_i = \frac{x_{0i} + x_{1i}}{2}$$

- szerokość przedziału $i$-tego — różnica między granicami klasy $i$-tej,
- rozpiętość przedziału $i$-tego ($h_i$) — różnica między środkami kolejnych przedziałów czy granicami (odpowiednio dolnymi lub górnymi) sąsiednich klas, a w przypadku, gdy dolna granica następnej klasy jest równa górnej granicy klasy poprzedniej — różnica między granicami klasy.

Sposoby określania granic przedziałów:

Podstawowa różnica w określaniu granic przedziałów sprowadza się do tego, czy:

dolna granica następnej klasy jest większa od górnej klasy poprzedniej ($x_{0,i+1} > x_{1i}$ $i = 1, ..., k-1$) — przykład 2.5.

dolna granica następnej klasy jest równa górnej granicy klasy poprzedniej ($x_{0,i+1} = x_{1i}$ $i = 1, ..., k-1$) — przykład 2.6.

***Przykład 2.5.***

**Sprzedaż w firmie Profi**

| $x_{0i} - x_{1i}$ | $n_i$ |
|---|---|
| 5–9 | 7 |
| 10–14 | 10 |
| 15–19 | 13 |
| 20–24 | 18 |
| 25–29 | 8 |
| 30–34 | 5 |
| 35-39 | 3 |
| Razem | 64 |

***Przykład 2.6.***

**Sprzedaż w firmie Profi**

| $x_{0i} - x_{1i}$ | $n_i$ |
|---|---|
| 5–10 | 7 |
| 10–15 | 10 |
| 15–20 | 13 |
| 20–25 | 18 |
| 25–30 | 8 |
| 30–35 | 5 |
| 35-40 | 3 |
| Razem | 64 |

Kwalifikowanie elementów do przedziałów klasowych może teoretycznie nastąpić zgodnie z jedną z czterech zasad:

a. element $x_j$ trafia do klasy $i$-tej, jeśli jest większy lub równy dolnej granicy i mniejszy lub równy górnej granicy klasy:

$$x_{0i} \leq x_j \leq x_{1i}$$

b. element xj trafia do klasy $i$-tej, jeśli jest większy od dolnej granicy i mniejszy lub równy górnej granicy klasy:

$$x_{0i} < x_j \leq x_{1i}$$

c. element $x_j$ trafia do klasy $i$-tej, jeśli jest większy lub równy dolnej granicy i mniejszy od górnej granicy klasy:

$$x_{0i} \leq x_j < x_{1i}$$

d. element $x_j$ trafia do klasy $i$-tej jeśli jest większy od dolnej granicy i mniejszy od górnej granicy klasy:

$$x_{0i} < x_j < x_{1i}$$

***Uwaga:*** Tam, gdzie nierówność jest ostra, mówimy o otwarciu klasy. Tam, gdzie nierówność jest nieostra, mówimy o zamknięciu (domknięciu) klasy.

Klasa według sposobu (a) jest obustronnie domknięta, według sposobu (b) jest otwarta z lewej strony i domknięta z prawej itd.

Nie każda z tych zasad spełnia właściwie swoje zadanie w konkretnej sytuacji

dla konwencji: $x_{0,i+1} > x_{1i}$ (granic klas: jak w przykładzie 2.5)

- stosując sposób (b) nie będziemy w stanie przypisać do żadnej z klas wartości równych dolnym granicom przedziałów,
- stosując sposób (c) będziemy mieli kłopoty z wartościami równymi górnym granicom klas,
- użycie sposobu (d) pozbawi przydziału wszystkie elementy równe granicom klas.

Jedyną zasadą, która tutaj potwierdza swą praktyczną przydatność jest (a), czyli $x_{0i} \leq x_j \leq x_{1i}$ (klasy z obu stron zamknięte).

dla konwencji: $x_{0,i+1} = x_{1i}$ (granic klas: jak w przykładzie 2.6)

- sposób (a) sprawi kłopoty ze wszystkimi elementami równymi granicom klas poza $x_{01}$ i $x_{1k}$,
- poza klasami pozostaną jednostki równe granicom przedziałów przydzielane zgodnie z zasadą (d).

Zakwalifikowanie elementu $x_j$ do konkretnej klasy jest możliwe na jeden z dwóch sposobów

$x_{0i} \leq x_j < x_{1i}$ (klasa zamknięta z lewej strony, otwarta z prawej).

$x_{0i} < x_j \leq x_{1i}$ (klasa otwarta z lewej strony, zamknięta z prawej).

Konwencję wykorzystuje się dla podkreślenia skokowości cechy (przy niewielkiej liczbie wariantów i co najwyżej dwucyfrowych wartościach cechy). Sprawia kłopoty przy prezentacji graficznej.

| $x_{0i} - x_{1i}$ | $n_i$ |
|---|---|
| 5–10 | 7 |
| 10–15 | 10 |
| 15–20 | 13 |
| 20–25 | 18 |
| 25–30 | 8 |
| 30–35 | 5 |
| 35–40 | 3 |
| Razem | 64 |

| $x_{0i} - x_{1i}$ | $n_i$ |
|---|---|
| 4–9 | 7 |
| 10–14 | 10 |
| 14–19 | 13 |
| 19–24 | 18 |
| 24–29 | 8 |
| 29–34 | 5 |
| 34–39 | 3 |
| Razem | 64 |

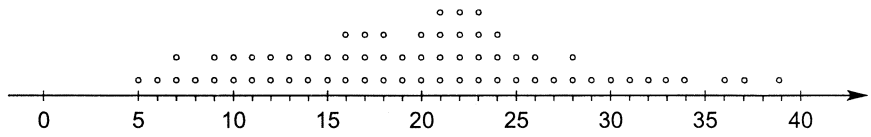
Tablice, gdzie klasy są z jednej strony otwarte, a z drugiej zamknięte buduje się zarówno dla cech skokowych jak i ciągłych.

Tablice, gdzie klasy są z jednej strony otwarte, a z drugiej zamknięte buduje się zarówno dla cech skokowych jak i ciągłych.

***Przykład 2.7.***

Budowa szeregu przedziałowego dla firmy Profi.

1. Dogodnym sposobem przekonania się o ewentualnych tendencjach do grupowania jednostek jest oś liczbowa:




***Wniosek:*** brak regularnych punktów skupień, nie ma przeciwwskazań dla stałych długości klas.

2. Populacja liczy 64 elementy. Jest mało liczna. liczba klas w dolnych granicach przedziału 6–15.

   Cecha — skokowa
   Cel badania — ogólny
   Rozstęp: $R = x_{max} - x_{min} = 39 - 5 = 34$ palety
   Minimalna różnica między kolejnymi wartościami cechy: jedna paleta.

***Wnioski:*** populacja nieliczna, cecha skokowa, nieduży rozstęp → liczba klas w dolnych granicach.

Pierwszy wariant rozwiązania

| $x_{0i} - x_{1i}$ | $n_i$ |
|---|---|
| 5–10 | 7 |
| 10–15 | 10 |
| 15–20 | 13 |
| 20–25 | 18 |
| 25–30 | 8 |
| 30–35 | 5 |
| 35–40 | 3 |
| Razem | 64 |

***Uwaga:*** klasy są zamknięte z lewej strony, otwarte z prawej.

Drugi wariant rozwiązania z podkreśleniem skokowości cechy

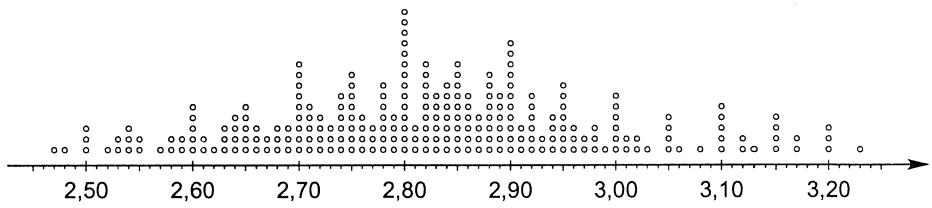
| $x_{0i} - x_{1i}$ | $n_i$ |
|---|---|
| 5–9 | 7 |
| 10–14 | 10 |
| 15–19 | 13 |
| 20–24 | 18 |
| 25–29 | 8 |
| 30–34 | 5 |
| 35–39 | 3 |
| Razem | 64 |

***Uwaga:*** klasy są zamknięte z obu stron.

***Przykład 2.8.***

**Godzinowe stawki płac w firmie XXC w marcu 1997 r. (w zł)**

| $x_i$ | $n_i$ | $x_i$ | $n_i$ | $x_i$ | $n_i$ |
|---|---|---|---|---|---|
| 2,47 | 1 | 2,75 | 8 | 3,00 | 6 |
| 2,48 | 1 | 2,76 | 4 | 3,01 | 2 |
| 2,50 | 3 | 2,77 | 3 | 3,02 | 2 |
| 2,52 | 1 | 2,78 | 7 | 3,03 | 1 |
| 2,53 | 2 | 2,79 | 4 | 3,05 | 4 |
| 2,54 | 3 | 2,80 | 14 | 3,06 | 1 |
| 2,55 | 2 | 2,81 | 3 | 3,07 | 1 |
| 2,57 | 1 | 2,82 | 9 | 3,10 | 5 |
| 2,58 | 2 | 2,83 | 6 | 3,12 | 2 |
| 2,59 | 2 | 2,84 | 7 | 3,13 | 1 |
| 2,60 | 5 | 2,85 | 9 | 3,15 | 4 |
| 2,61 | 2 | 2,86 | 6 | 3,17 | 2 |
| 2,62 | 1 | 2,87 | 4 | 3,20 | 3 |
| 2,63 | 3 | 2,88 | 8 | 3,23 | 1 |
| 2,64 | 4 | 2,89 | 6 | Razem | 250 |
| 2,65 | 5 | 2,90 | 11 | | |
| 2,66 | 3 | 2,91 | 3 | | |
| 2,67 | 2 | 2,92 | 5 | | |
| 2,68 | 3 | 2,93 | 2 | | |
| 2,69 | 3 | 2,94 | 4 | | |
| 2,70 | 9 | 2,95 | 7 | | |
| 2,71 | 5 | 2,96 | 3 | | |
| 2,72 | 4 | 2,97 | 2 | | |
| 2,73 | 3 | 2,98 | 3 | | |
| 2,74 | 6 | 2,99 | 1 | | |




***Wniosek:*** punktami skupień są „okrągłe" dziesięciogroszowe stawki. To one wyznaczają środki przedziałów jako środki ciężkości klas. Nic nie stoi na przeszkodzie zastosowania przedziałów o takich samych długościach.

Populacja liczy 250 elementów. Można ją uznać za średnio liczną. Cecha jest pozornie ciągła → można zbudować szereg w myśl zasady „dolna granica następnej klasy jest równa górnej granicy klasy poprzedniej". Cel badania — ogólny. Rozstęp: 3,23 – 2,47 = 0,76 zł. Minimalna różnica między dwiema kolejnymi wartościami cechy 0,01 zł.

Zalecany wariant rozwiązania

| $x_{0i}-x_{1i}$ | $n_i$ | $\mathring{x}_i$ |
|---|---|---|
| 2,45–2,55 | 11 | 2,5 |
| 2,55–2,65 | 22 | 2,6 |
| 2,65–2,75 | 43 | 2,7 |
| 2,75–2,85 | 65 | 2,8 |
| 2,85–2,95 | 58 | 2,9 |
| 2,95–3,05 | 27 | 3,0 |
| 3,05–3,15 | 14 | 3,1 |
| 3,15–3,25 | 10 | 3,2 |
| Razem | 250 | × |

***Uwaga:*** klasy są zamknięte z lewej, a otwarte z prawej strony.

Drugi wariant rozwiązania nawiązujący do pozornej ciągłości cechy

| $x_0-x_{1i}$ | $n_i$ | $\mathring{x}_i$ |
|---|---|---|
| 2,45–2,54 | 11 | 2,495 |
| 2,55–2,64 | 22 | 2,595 |
| 2,65–2,74 | 43 | 2,695 |
| 2,75–2,84 | 65 | 2,795 |
| 2,85–2,94 | 58 | 2,895 |
| 2,95–3,04 | 27 | 2,995 |
| 3,05–3,14 | 14 | 3,095 |
| 3,15–3,24 | 10 | 3,195 |
| Razem | 250 | × |

***Uwaga:*** klasy są obustronnie zamknięte.

# Rozdział 3

# Graficzne formy prezentacji danych

## 3.1. Typologia wykresów

Typologia prostych postaci wykresów statystycznych przedstawia się jak niżej:

- wykresy wymiarowe:
  - liniowe,
  - powierzchniowe (słupkowe, kołowe, pasowe),
  - objętościowe (bryłowe);
- wykresy ilościowe:
  - obrazkowe, czyli piktogramy,
  - punktowe;
- wykresy mapowe:
  - kartogram (przedstawia różnice w natężeniu zjawiska za pomocą skali barwnej),
  - kartodiagram (łączy mapę z wykresami obrazkowymi lub powierzchniowymi);
- wykresy specjalne np. harmonogramy, drzewa genealogiczne;
- kombinowane;
- wykresy sporządzone w prostokątnym układzie współrzędnych (histogram, wielobok liczebności, krzywa liczebności, wielobok liczebności skumulowanych, wykres pudełkowy (*box and whisker plot*)).

## 3.2. Wykresy powierzchniowe

W dalszym ciągu najwięcej uwagi poświęcimy stawiającym wysokie formalne wymagania wykresom w układzie kartezjańskim.

Istotne miejsce pośród wykresów statystycznych zajmują powierzchniowe wykresy wymiarowe. Dla wykresów powierzchniowych, jak sugeruje to ich nazwa, pola figur muszą być proporcjonalne do wielkości prezentowanych zjawisk.

***Przykład 3.1.***

Przedstawić w formie powierzchniowej długość sieci wodociągowej (bez magistralnej) w Polsce (stan w dniu 31.12.), dane GUS.

| Rok | Sieć w km |
|---|---|
| 1990 | 93187 |
| 1994 | 142179 |

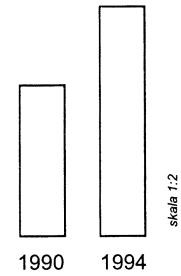
Relacja wielkości zjawisk $= \frac{142179}{93187} = 1{,}526$, z czego wynika, że sieć wodociągowa w 1994 r. była o prawie 53% dłuższa od sieci w 1990 r. Prawidłowo sporządzone wykresy powierzchniowe powinny być wykonane tak, aby pole figury dla roku 1994 było o 52,6% większe od pola figury dla roku 1990.

1. Najprościej wygląda sprawa dla prostokątów. Wystarczy, aby prostokąty miały podstawy takiej samej długości, a wysokości proporcjonalne do wielkości przedstawianych zjawisk, bo:

$$\frac{P_{1994}}{P_{1990}} = \frac{a \cdot h_2}{a \cdot h_1} = 1{,}526 \Rightarrow \frac{h_2}{h_1} = 1{,}526$$

$$h_2 = h_1 \cdot 1{,}526$$

jeśli $h_1 = 40\,\text{mm}$ to $h_2 = 40\,\text{mm} \cdot 1{,}526 = 61{,}04\,\text{mm}$



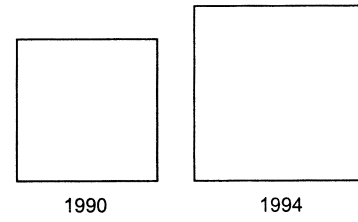

2. Dla kwadratów jak następuje:

$$\frac{P_{1994}}{P_{1990}} = \frac{b^2}{a^2} = 1{,}526 \quad \Rightarrow \quad b^2 = a^2 \cdot 1{,}526$$

jeśli np.

$a = 40\,\text{mm}$ to $b^2 = 40^2 \cdot 1{,}526 = 2441{,}6\,\text{mm}^2$

$$b = \sqrt{b^2} = \sqrt{2441{,}6\,\text{mm}^2} = 49{,}4\,\text{mm}$$



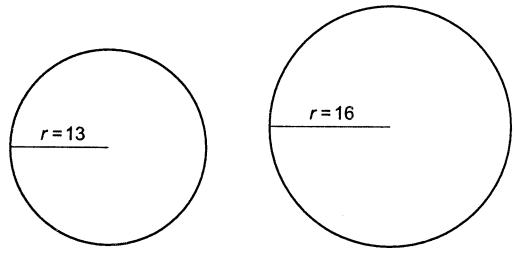
3. Dla wykresów kołowych:

$$\frac{P_{1994}}{P_{1990}} = \frac{\pi r_2^2}{\pi r_1^2} = 1{,}526 \quad \Rightarrow \quad \frac{r_2^2}{r_1^2} = 1{,}526$$

$$r_2^2 = r_1^2 \cdot 1{,}526$$

jeśli $r_1 = 13\,\text{mm}$ to $r_2^2 = 13^2 \cdot 1{,}526 = 169 \cdot 1{,}526 = 257{,}84\,\text{mm}^2$

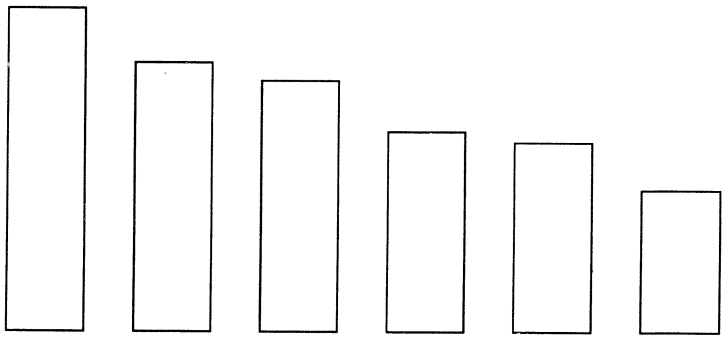
$$r_2 = \sqrt{r_2^2} = \sqrt{257{,}84\,\text{mm}^2} = 16{,}06\,\text{mm}$$




1990 1994

***Przykład 3.2.***

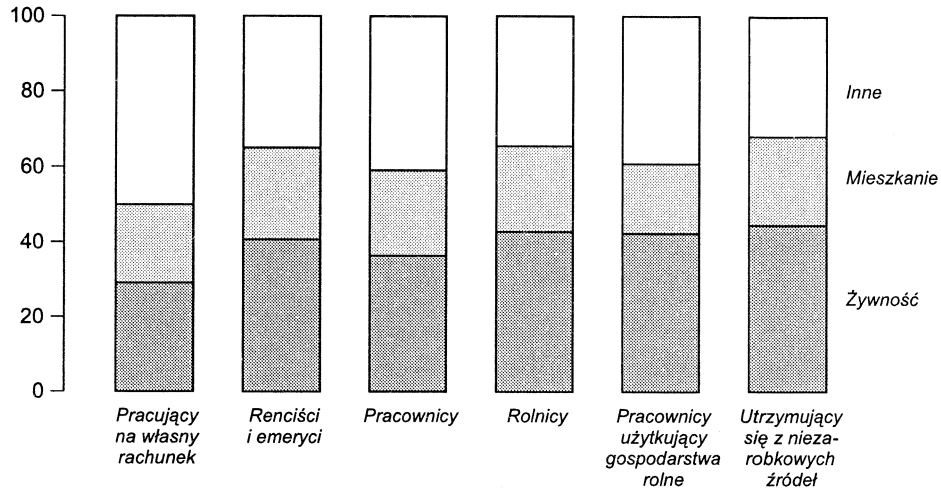
**Miesięczne dochody na osobę w gospodarstwach domowych w pierwszym półroczu 1996 r. (dane GUS)**




Porównując **strukturę** rozkładu cechy dla różnych populacji, przedstawiamy ją za pomocą figur o takich samych powierzchniach (traktowanych jako 100%).

***Przykład 3.3.***

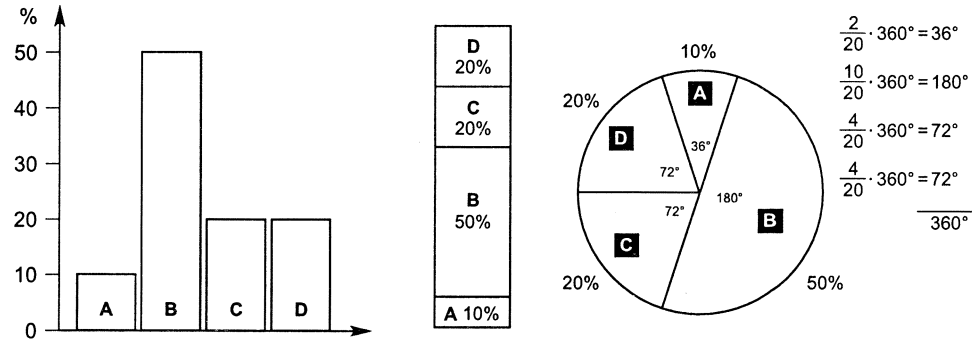
**Struktura wybranych przeciętnych wydatków na osobę w gospodarstwach domowych w pierwszym półroczu 1996 r. (w procentach)**




Najczęściej spotykanymi formami wykresów dla cech jakościowych są wykresy słupkowe, kołowe, pasowe oraz obrazkowe.

***Przykład 3.4.***

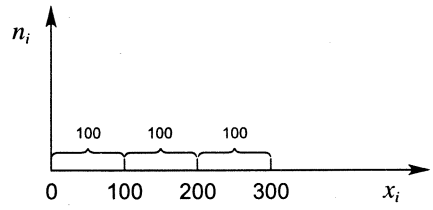
**Wykresy dla danych o preferencjach pracowniczych**



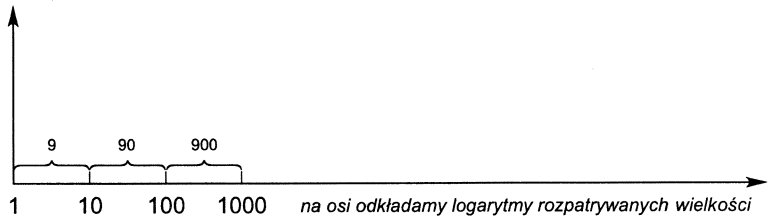

*Wykres słupkowy* *Złożony wykres słupkowy* *Wykres kołowy*

## 3.3. Wykresy w prostokątnym układzie współrzędnych

Wykresy te w większości przypadków realizowane są z zachowaniem zasady tzw. podziałki równomiernej. Zgodnie z nią jednakowym przedziałom graficznym odpowiadają jednakowe przedziały liczbowe:




Podstawową sytuacją, w której zasada ta jest łamana, jest użycie podziałki logarytmicznej na jednej lub na obu osiach:




Podziałka logarytmiczna jest wykorzystywana do przedstawienia szeregów czasowych, gdy występują duże rozpiętości wartości cech i/lub chcemy pokazać względne zmiany zjawiska.

***Histogram*** jest uporządkowanym zbiorem prostokątów, których podstawy spoczywają na osi odciętych i są wyznaczone przez odcinki odpowiadające kolejnym przedziałom wartości zmiennej, a których wysokości odpowiadają częstościom w kolejnych klasach. Pola prostokątów są proporcjonalne do liczebności w poszczególnych klasach, a pole histogramu jest proporcjonalne do liczebności populacji.

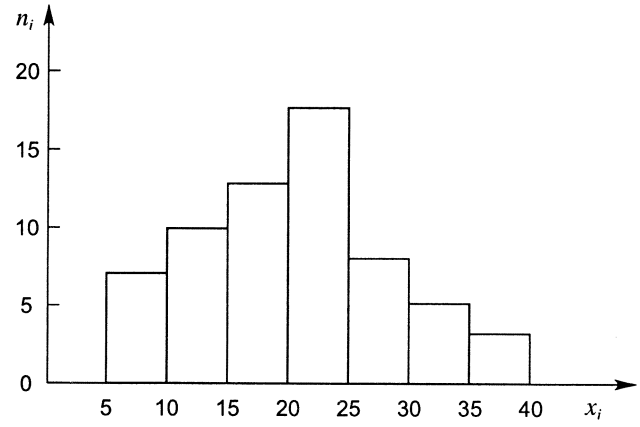
***Wielobok liczebności*** — powstaje z histogramu przez łączenie liniami prostymi punktów o współrzędnych: odcięta — środek przedziału, rzędna — liczebność w danej klasie. Skrajne punkty łączy się ze środkami dwóch sąsiadujących przedziałów na osi odciętych (o liczebnościach zero).

W zasadzie nie włącza się punktów zewnętrznych, gdy:

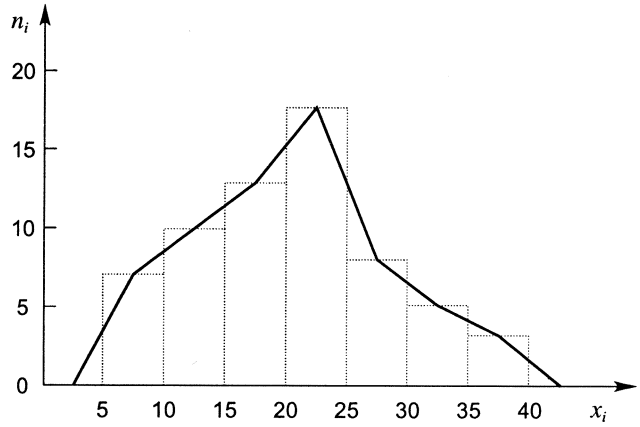
— przedziały są nierównej długości lub otwarte,

— liczebności w szeregu rozdzielczym nie mają tendencji spadkowej przy zbliżaniu się do końca obszaru zmienności rozpatrywanej cechy,

— przez ten zabieg wykres wkracza w przedział wartości sprzecznych z naturą zjawiska (ujemnych lub dodatnich).



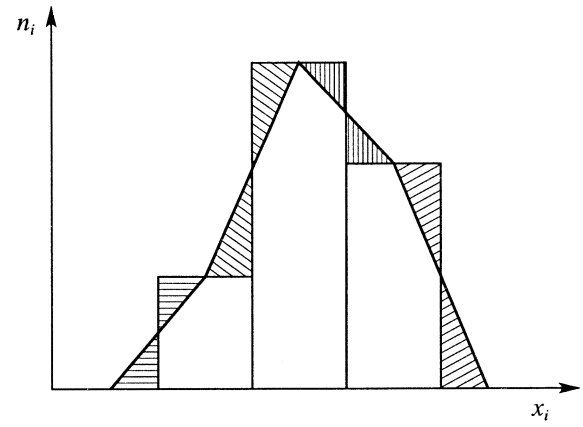

Wykres 3.1. Histogram do przykładu o paletach sprzedanych przez firmę Profi




Wykres 3.2. Wielobok liczebności dla przykładu o paletach

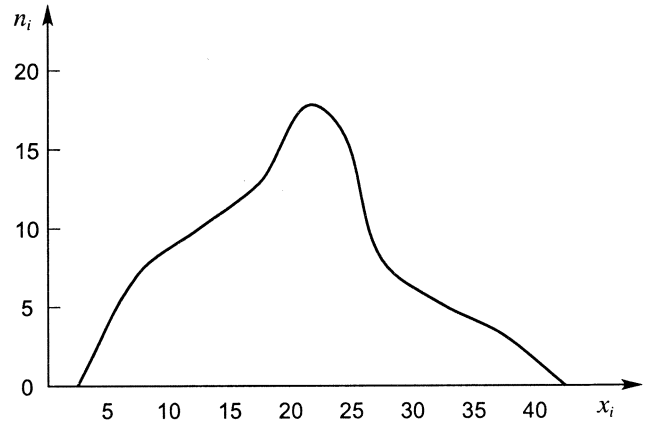
Celem łączenia skrajnych punktów wieloboku z osią odciętych jest osiągnięcie pola dla wieloboków liczebności takiego samego jak pole histogramu (patrz wykres poniżej) oraz lepsza czytelność kształtu rozkładu.

Zaletą wieloboku liczebności (w przeciwieństwie do histogramu) jest to, że jeden rysunek może pomieścić ich kilka, co jest szczególnie cenne przy porównaniach. Wadą — zniekształcenie proporcji liczebności w klasach. Obszar zmienności wieloboku jest powiększony o długość całego przedziału klasowego w porównaniu z histogramem.




Wykres 3.3. Równość pól histogramu i wieloboku liczebności

***Krzywa liczebności*** — uciąglony wielobok liczebności. Powstaje przez zmniejszenie rozpiętości przedziałów klasowych. Jest to w zasadzie możliwe dla zmiennych ciągłych i bardzo licznych zbiorowości, ale zastępowanie wieloboku przez krzywą jest powszechne, by zaprezentować kształt rozkładu cechy.




Wykres 3.4. Krzywa liczebności dla przykładu o paletach

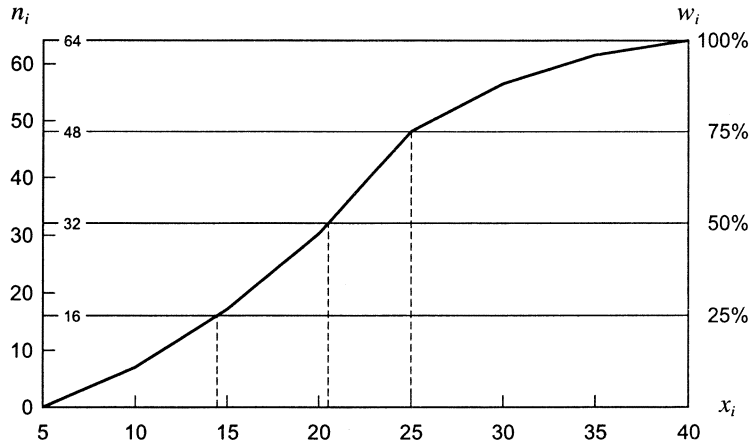
***Wielobok liczebności skumulowanych*** jest graficzną prezentacją szeregu kumulacyjnego. Specyfika jego tworzenia zasadza się na tym, że:

1. najlepiej jest go budować na planie prostokąta, którego podstawa jest określona przez obszar zmienności cechy, a wysokość przez liczebność populacji,

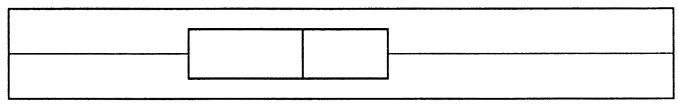
2. rzędne punktów są równe liczebnościom skumulowanym na granicach klas,
3. odciętymi są wartości granic przedziałów.

Tablica do wykreślenia kumulaty przy granicach określonych na zasadzie „dolna granica następnej klasy jest równa górnej granicy klasy poprzedniej".

| $x_{0i} - x_{1i}$ | $n_i$ | Kumulata |
|---|---|---|
| 5–10 | 7 | 7 |
| 10–15 | 10 | 17 |
| 15–20 | 13 | 30 |
| 20–25 | 18 | 48 |
| 25–30 | 8 | 56 |
| 30–35 | 5 | 61 |
| 35–40 | 3 | 64 |
| Razem | 64 | × |




Wykres 3.5. Wielobok liczebności skumulowanych dla przykładu o paletach



Wykres 3.6. *Box and whisker plot* dla przykładu o paletach

***Pudełko z wąsami (box and whisker plot)*** buduje się wykorzystując wykres kumulaty. Rysujemy pod nim kolejny prostokąt o identycznej podstawie. Jego wysokość jest uznaniowa. Na boku wykresu kumulaty znajdujemy punkty równe $^1/_4$, $^1/_2$ i $^3/_4$ liczebności populacji. Przez punkty te prowadzimy proste równoległe do podstawy. Z punktów przecięcia prostych z kumulatą opuszczamy prostopadłe do podstawy. Uzyskane punkty wyznaczają wartości miar statystycznych nazywanych kwartylami. Będzie o nich mowa dalej. Na razie interesuje nas tylko technika tworzenia wykresu. Wartości kwartyli odkładamy jako boki wewnętrznego prostokąta w pudełku. Na koniec łączymy boki wewnętrznego prostokąta z bokami prostokąta zewnętrznego. To wąsy, od których wywodzi się nazwa tej formy graficznej.

## 3.4. Graficzne porównanie populacji

Z zależności pola histogramu od liczebności w poszczególnych klasach wynika ważna konsekwencja praktyczna: porównanie będzie skazane na niepowodzenie, jeśli pola figur ograniczonych ramionami wieloboku liczebności nie będą takie same. Interesuje nas poznanie różnic struktury rozkładu badanej cechy, które nie uzewnętrzniają się bez sprowadzenia na wspólną płaszczyznę liczebności analizowanych zbiorowości.

Przed wyciągnięciem ostatecznych wniosków jak to należy robić, rozpatrzmy trzy przypadki szczególne:

a. wieloboki liczebności dla tej samej populacji, przedstawione za pomocą klas o różnych długościach,

b. wieloboki dla populacji o różnych liczebnościach, ale o takich samych długościach przedziałów klasowych,

c. wieloboki dla populacji o różnych długościach klas.

***Przykład 3.5.*** Wyniki uzyskane z kolokwium ze statystyki

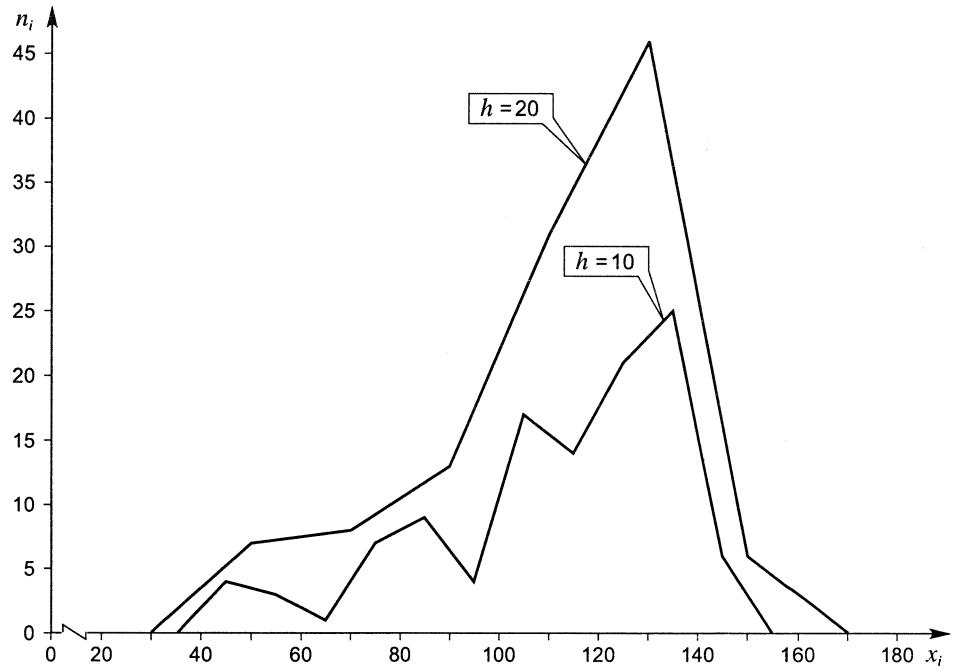
| $x_{0i} - x_{1i}$ | $n_i$ |
|---|---|
| 40–50 | 4 |
| 50–60 | 3 |
| 60–70 | 1 |
| 70–80 | 7 |
| 80–90 | 9 |
| 90–100 | 4 |
| 100–110 | 17 |
| 110–120 | 14 |
| 120–130 | 21 |
| 130–140 | 25 |
| 140–150 | 6 |
| Razem | 111 |

| $x_{0i} - x_{1i}$ | $n_i$ |
|---|---|
| 40–60 | 7 |
| 60–80 | 8 |
| 80–100 | 13 |
| 100–120 | 31 |
| 120–140 | 46 |
| 140–160 | 6 |
| Razem | 111 |

Źródło: dane umowne.

**Ad a.** Przykład dotyczy wyników kolokwium w punktach dla 111 osób.
Dysponujemy dla tej populacji dwoma szeregami rozdzielczymi o stałych granicach klas: w pierwszym przypadku przedziały mają długość 10, w drugim — 20.

Zobaczymy, jak będą wyglądały wieloboki liczebności dla tych szeregów na jednym wykresie.




Wykres 3.7. Wieloboki liczebności dla wyników kolokwium dla 111 osób (dla $h_i$)

Celem tego zestawienia było zwrócenie uwagi Czytelnika na następujące rzeczy:

1. wieloboki dla tej samej populacji mogą się znacznie od siebie różnić,
2. wielobok wykreślony dla szeregu o mniejszych długościach klas, jako zawierający więcej informacji jest bardziej pofałdowany (postrzępiony), wykres dla większych rozpiętości jest bardziej wygładzony, ogólny,
3. relacje pól obu figur są jak 2 : 1.
   Uzasadnieniem jest następujący rachunek. Pole wieloboku liczebności jest równe polu histogramu, a ono z kolei jest równe sumie pól prostokątów o podstawach równych długościom klas i wysokościom równym liczebnościom w przedziałach.

Pole dla wieloboku o rozpiętości 10 ($h=10$) wynosi:

$$P_{pk} = \sum h^{(10)} n_i = h^{(10)} \sum n_i = h^{(10)} N = 10 \cdot 111 = 1110 \text{ jp.}$$

Pole dla wieloboku o rozpiętości 20 ($h=20$) wynosi:

$$P_{pk} = \sum h^{(20)} n_i = h^{(20)} \sum n_i = h^{(20)} N = 20 \cdot 111 = 2220 \text{ jp.}$$

($P_{pk}$ — oznacza pole pod krzywą wieloboku).

W poprzednio poczynionych uwagach na temat porównywania struktur zbiorowości za pomocą wykresów powierzchniowych znalazła się myśl, że warunkiem porównywalności jest równość pól pod krzywymi. Oczywiście, nie została ona tutaj osiągnięta.

Tablice poniżej (za pomocą ostatnich wierszy opisujących pole pod krzywą dla różnych wariantów określania wysokości prostokątów), upewniają, że:

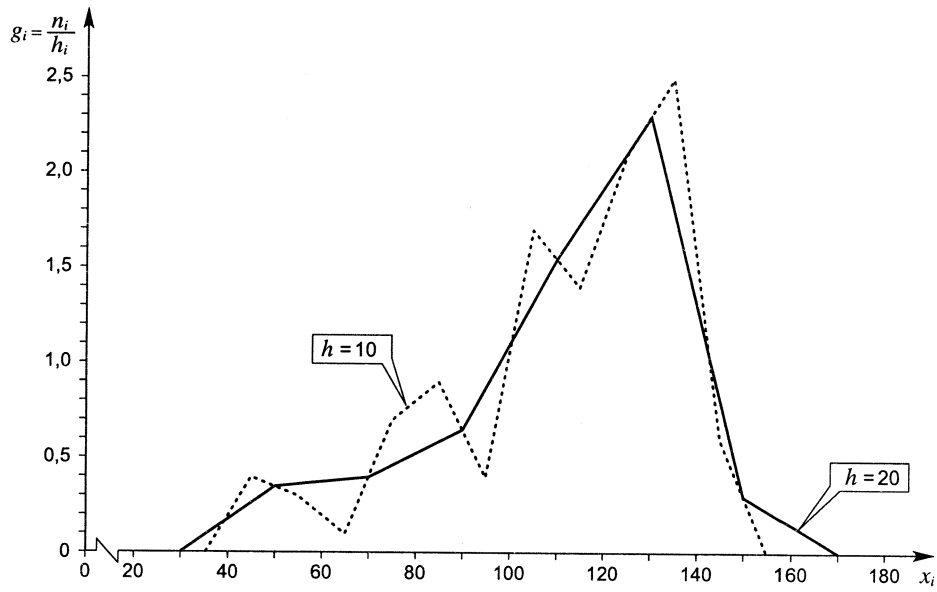
a) wieloboki wykreślone dla częstości względnych nadal będą miały pola w stosunku 2:1,

b) wieloboki wykreślone dla gęstości ($g_i = \frac{n_i}{h_i}$, gdzie $h_i$ jest długością $i$-tej klasy)

będą miały takie same pola pod krzywymi.

| $x_{0i} - x_{1i}$ | $n_i$ | $w_i$ | $g_i = \frac{n_i}{h_i}$ |
|---|---|---|---|
| 40–50 | 4 | 0,036 | 0,4 |
| 50–60 | 3 | 0,027 | 0,3 |
| 60–70 | 1 | 0,009 | 0,1 |
| 70–80 | 7 | 0,063 | 0,7 |
| 80–90 | 9 | 0,081 | 0,9 |
| 90–100 | 4 | 0,036 | 0,4 |
| 100–110 | 17 | 0,153 | 1,7 |
| 110–120 | 14 | 0,126 | 1,4 |
| 120–130 | 21 | 0,189 | 2,1 |
| 130–140 | 25 | 0,225 | 2,5 |
| 140–150 | 6 | 0,054 | 0,6 |
| Razem | 111 | 0,999 | 11,1 |
| Pole pod krzywą ($P_{pk}$) | 1110 | 10 | 111 |

| $x_{0i} - x_{1i}$ | $n_i$ | $w_i$ | $g_i = \frac{n_i}{h_i}$ |
|---|---|---|---|
| 40–60 | 7 | 0,063 | 0,35 |
| 60–80 | 8 | 0,072 | 0,40 |
| 80–100 | 13 | 0,117 | 0,65 |
| 100–120 | 31 | 0,279 | 1,55 |
| 120–140 | 46 | 0,414 | 2,30 |
| 140–160 | 6 | 0,054 | 0,30 |
| Razem | 111 | 0,999 | 5,55 |
| Pole pod krzywą ($P_{pk}$) | 2220 | 20 | 111 |




Wykres 3.8. Wielobok liczebności dla wyników kolokwium dla 111 osób (dla $g_i$).

Z dotychczasowych doświadczeń możemy wysnuć ogólniejszy wniosek:

**Warunkiem uzyskania równych pól pod krzywymi wieloboków dla populacji tak samo licznych, ale przedstawionych za pomocą klas o stałych, ale różnych dla zbiorowości długościach klas, jest przejście na gęstości $g_i$ (lub ich przypadek szczególny $l_i$, o którym będzie dalej mowa).**

**Ad b.** Dwie populacje $A$ i $B$, mają tym razem klasy o równych i stałych (takich samych dla populacji) długościach. W oparciu o jakie przekształcenia należy wykreślić wieloboki, aby pola pod krzywymi były takie same?

***Przykład 3.6.***

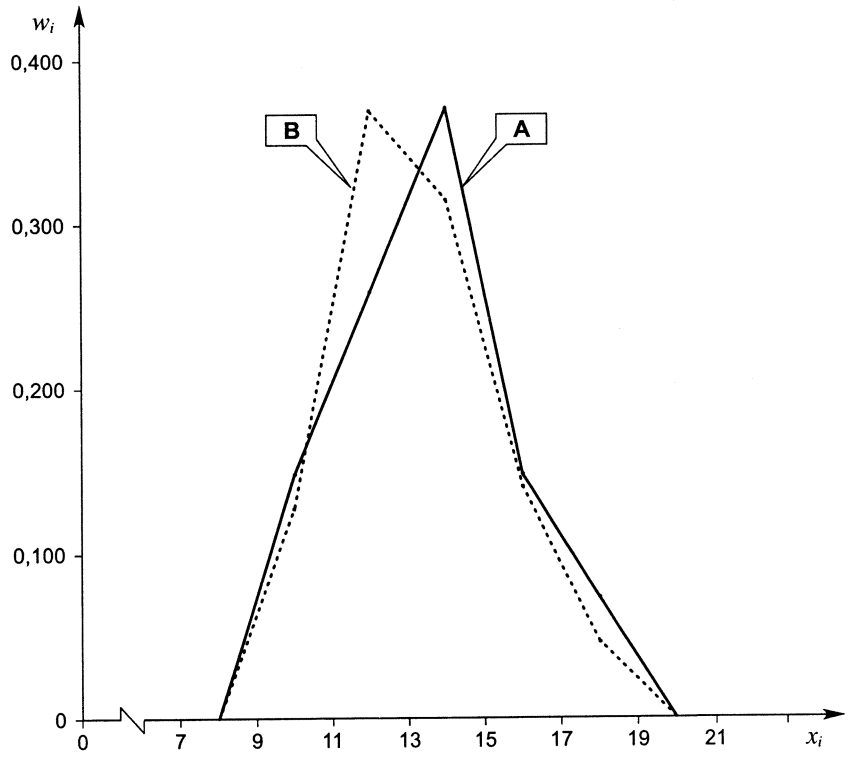
| $x_{0i} - x_{1i}$ | $n_i^A$ | $n_i^B$ |
|---|---|---|
| 9–11 | 4 | 19 |
| 11–13 | 7 | 55 |
| 13–15 | 10 | 47 |
| 15–17 | 4 | 21 |
| 17–19 | 2 | 7 |
| Razem | 27 | 149 |

Rachunek przeprowadzony na takiej samej zasadzie jak poprzednio pokazuje (wiersz $P_{pk}$), że figury wykreślone dla liczebności bezwzględnych $n_i$, ale także dla $g_i$, nie spełnią oczekiwań.

Rozwiązaniem jest przejście na częstości względne $w_i$:

| $x_{0i} - x_{1i}$ | $n_i^A$ | $n_i^B$ | $w_i^A$ | $w_i^B$ |
|---|---|---|---|---|
| 9–11 | 4 | 19 | 0,148 | 0,128 |
| 11–13 | 7 | 55 | 0,259 | 0,369 |
| 13–15 | 10 | 47 | 0,371 | 0,315 |
| 15–17 | 4 | 21 | 0,148 | 0,141 |
| 17–19 | 2 | 7 | 0,074 | 0,047 |
| Razem | 27 | 149 | 1,000 | 1,000 |
| $P_{pk}$ | 54 | 298 | 2,000 | 2,000 |

Potwierdzeniem jest wykres 3.9.




Wykres 3.9. Wieloboki liczebności dla różnolicznych populacji o stałych i równych długościach klas (wykres dla $w_i$)

**Wniosek: warunkiem uzyskania równych pól pod krzywymi liczebności dla populacji różniących się liczebnościami, ale o stałych i równych długościach klas jest zrobienie wykresu dla częstości względnych $w_i$.**

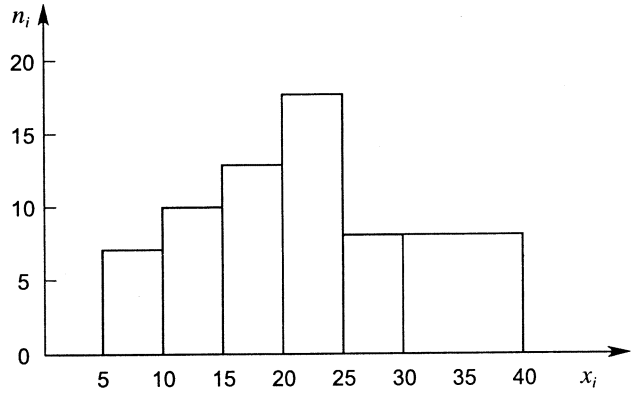
**Ad c.** Histogramy i wieloboki liczebności dla nierównych rozpiętości klas

***Przykład 3.7.***

W szeregu przedziałowym dla firmy Profi połączyliśmy dwa ostatnie przedziały:

| $x_{0i} - x_{1i}$ | $n_i$ |
|---|---|
| 5–10 | 7 |
| 10–15 | 10 |
| 15–20 | 13 |
| 20–25 | 18 |
| 25–30 | 8 |
| 30–40 | 8 |

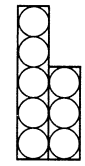
Jeśli wykreślimy histogram nie zwracając uwagi na różnice w długościach klas, otrzymamy:



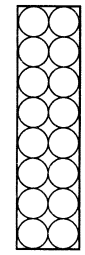

Wykres 3.10. Histogram bez uwzględniania różnic w długościach klas.
**Uwaga:** wykres jest wykonany niepoprawnie.

Pole tej figury jest równe: $\sum n_i \cdot h_i = 7 \cdot 5 + 10 \cdot 5 + 13 \cdot 5 + 18 \cdot 5 + 8 \cdot 5 + 8 \cdot 10 = 360$, czyli o 40 jednostek większe w porównaniu z wykresem 3.2. (Utworzenie tylko jednej klasy od 5 do 40 z liczebnością 64 zaowocowałoby prostokątem o boku 2240, siedem razy większym od figury wyjściowej). Różnice mają miejsce dla przedziału 30–40.

Wykres dla przypadku o równych rozpiętościach klas miał dla dwóch ostatnich przedziałów pola $5 \cdot 5 = 25$ i $5 \cdot 3 = 15$; razem 40.




Drugi wykres, po połączeniu tych klas, ma pole $8 \cdot 10 = 80$.




Stwarza błędne wrażenie, jakby w wyniku tego zabiegu liczba elementów o wartościach z przedziału 30–40 podwoiła się.

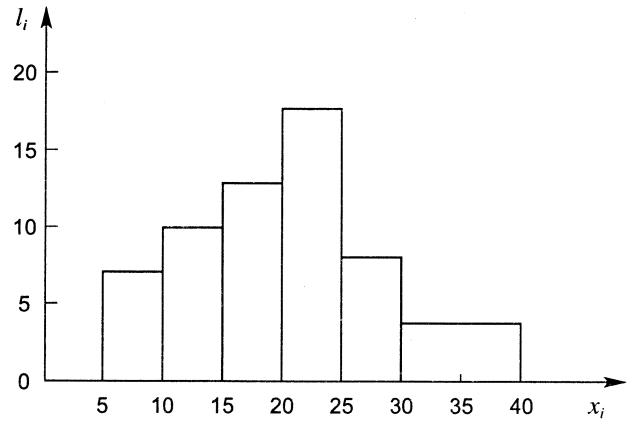
Poprawny, ze względu na pole pod krzywą, histogram lub wielobok liczebności dla populacji o różnych (niestałych) długościach klas musi być wykonany dla gęstości ($g_i$) lub natężeń liczebności ($l_i$):

$$l_i = \frac{n_i \,(\text{lub } w_i) \cdot h_{\text{bazowe}}}{h_i},$$

$h_{\text{bazowe}}$ to długość klasy uznanej za punkt odniesienia przy wykonywaniu wykresu. Wybiera się ją arbitralnie; jeśli będzie to długość w danej tablicy występująca najczęściej, liczba niezbędnych dodatkowych wyliczeń będzie najmniejsza.

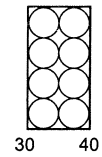
| $x_{0i} - x_{1i}$ | $n_i$ | $h_i$ | $l_i$ | $g_i$ |
|---|---|---|---|---|
| 5–10 | 7 | 5 | 7 | 1,4 |
| 10–15 | 10 | 5 | 10 | 2,0 |
| 15–20 | 13 | 5 | 13 | 2,6 |
| 20–25 | 18 | 5 | 18 | 3,6 |
| 25–30 | 8 | 5 | 8 | 1,6 |
| 30–40 | 8 | 10 | 4 | 0,8 |
| Razem | 64 | × | × | × |

$h_{\text{bazowe}} = 5$.




Wykres 3.11. Histogram dla natężeń liczebności

Pole tego wieloboku jest równe: $7 \cdot 5+10 \cdot 5+13 \cdot 5+ 18 \cdot 5 +8 \cdot 5 +4 \cdot 10=320$. Dzięki wysokości 4 w ostatniej klasie, liczba elementów w przedziale jest równa sumie elementów w obu klasach przed połączeniem.




**Wniosek: przedstawienie graficzne (za pomocą histogramu lub wieloboku liczebności) populacji o niestałych rozpiętościach przedziałów wymaga przejścia na gęstości (natężenia liczebności).**

Końcowe wnioski dla porównań za pomocą histogramów lub wieloboków liczebności:

**Porównanie populacji różniących się liczebnościami ($N_i$), o nierównych i niestałych rozpiętościach przedziałów wymaga:**

a) **przejścia z liczebności bezwzględnych $n_i$ na częstości względne $w_i$,**

b) **wyznaczenia gęstości (natężenia liczebności) w oparciu o częstości względne $w_i$ z punktu (a).**

Natężenia liczebności są szczególnym przypadkiem gęstości, gdy liczbę elementów w klasie odnosimy do długości klasy bazowej, przyjętej jako podstawa porównania. Związki pomiędzy $l_i$ a $g_i$ przedstawiają ostatnie kolumny tablicy na poprzedniej stronie. Pomnożenie każdego elementu kolumny gęstości przez 5 daje odpowiednie natężenie liczebności.

**Tablica 3.1. Budowanie histogramów i wieloboków liczebności w zależności od liczby populacji, ich liczebności oraz długość klas**

| Liczba populacji / Rozpiętości | Jedna | Kilka | |
|---|---|---|---|
| | | o takich samych liczebnościach | różniące się liczebnościami |
| stałe* i równe** | ①, ②, ③ | 1, ③ | 1, ③ |
| niestałe | ②, ③ | 3 | 3 |

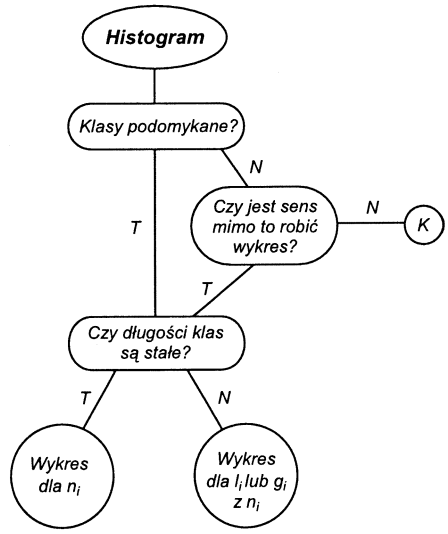
1 — linie pionowe (dla cech skokowych)
2 — histogramy
3 — wieloboki liczebności
◯ — forma zalecana

gęstości (natężenia liczebności)

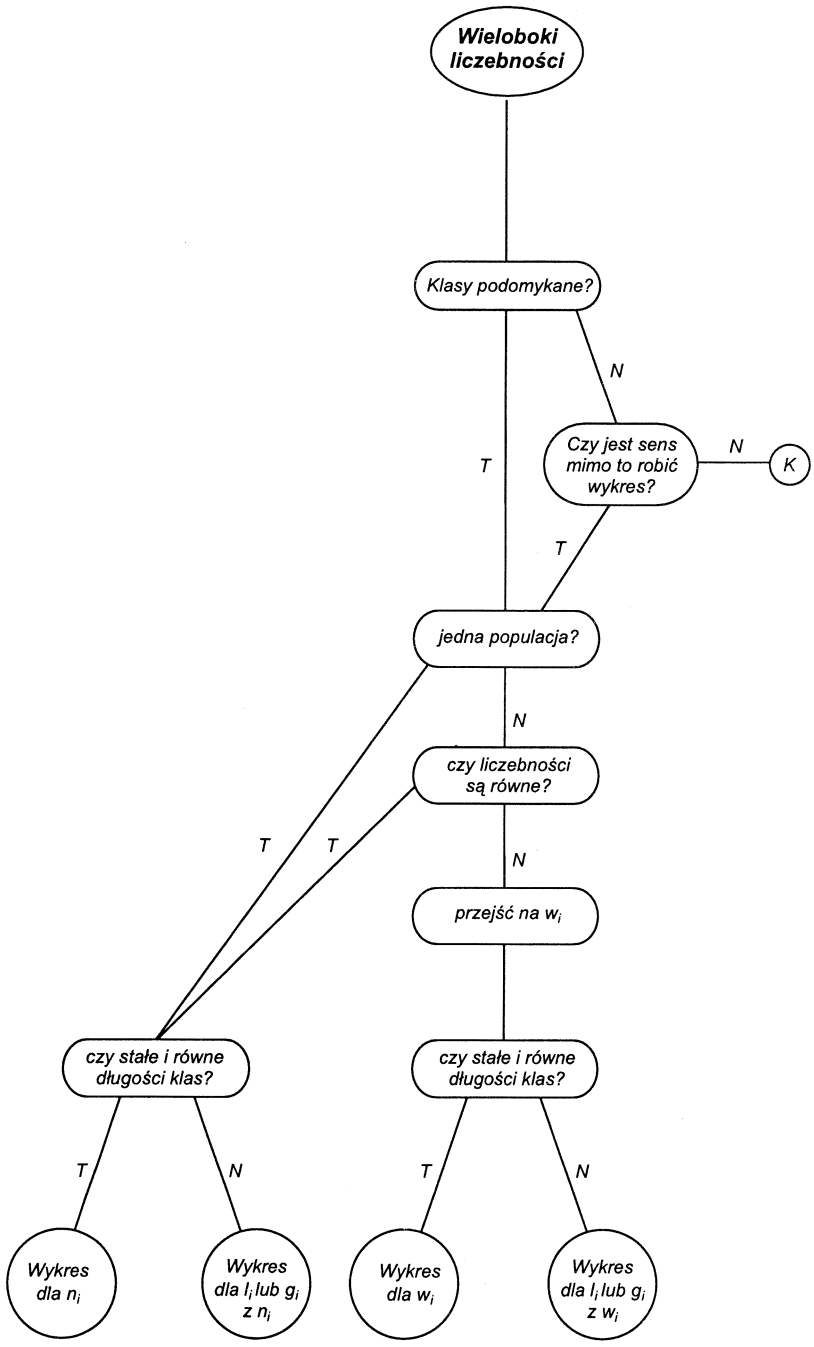
częstości względne

* stałe rozpiętości — wszystkie klasy w szeregu rozdzielczym mają takie same długości
** równe rozpiętości klas — porównywane populacje mają klasy takich samych długości

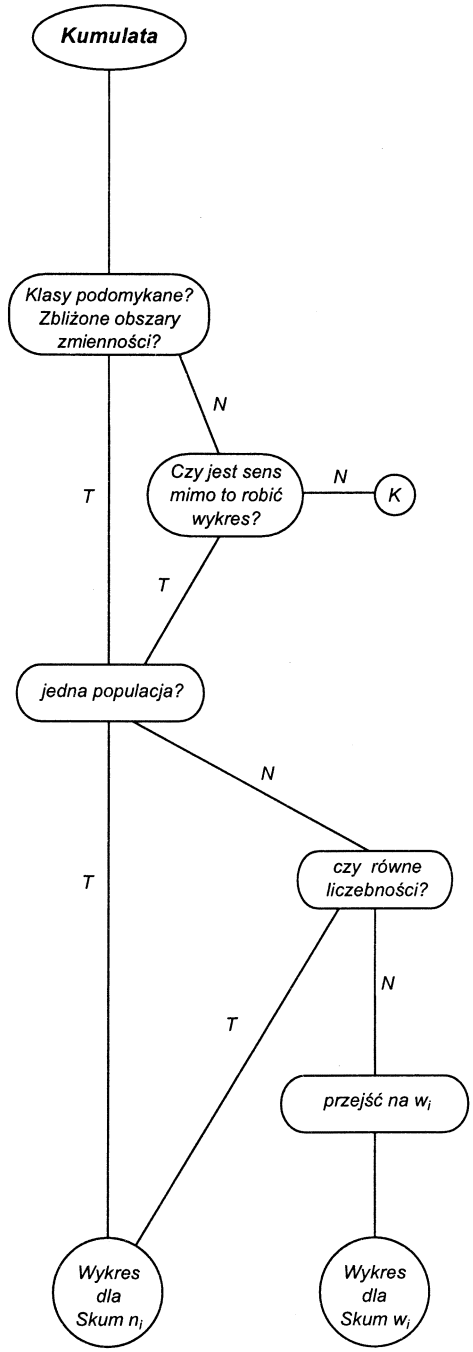
Czynności istotne dla wykreślania histogramów, wieloboków liczebności i kumulat przedstawiają schematy blokowe poniżej.




Wykres 3.12. Schemat blokowy budowy histogramu




Wykres 3.13. Schemat blokowy budowy wieloboków liczebności




Wykres 3.14. Schemat blokowy budowy kumulat

***Przykład 3.8.***

Metodami graficznymi porównać strukturę indywidualnych gospodarstw rolnych w województwach $K$ i $M$ w roku $X$ (dane hipotetyczne).

| Gospodarstwa o powierzchni ogólnej w ha | Liczba gospodarstw | |
|---|---|---|
| | $K$ | $M$ |
| 0,1–0,5 | 18179 | 3782 |
| 0,5–1 | 12856 | 3209 |
| 1–2 | 14170 | 4393 |
| 2–3 | 7258 | 3209 |
| 3–5 | 8225 | 3094 |
| 5–7 | 4355 | 3018 |
| 7–10 | 2696 | 8328 |
| 10–15 | 1037 | 7487 |
| 15–20 | 207 | 1337 |
| 20–30 | 138 | 344 |
| Razem | 69121 | 38201 |

Ze względu na różnice w liczebnościach zbiorowości i występowanie różnych długości przedziałów należy:

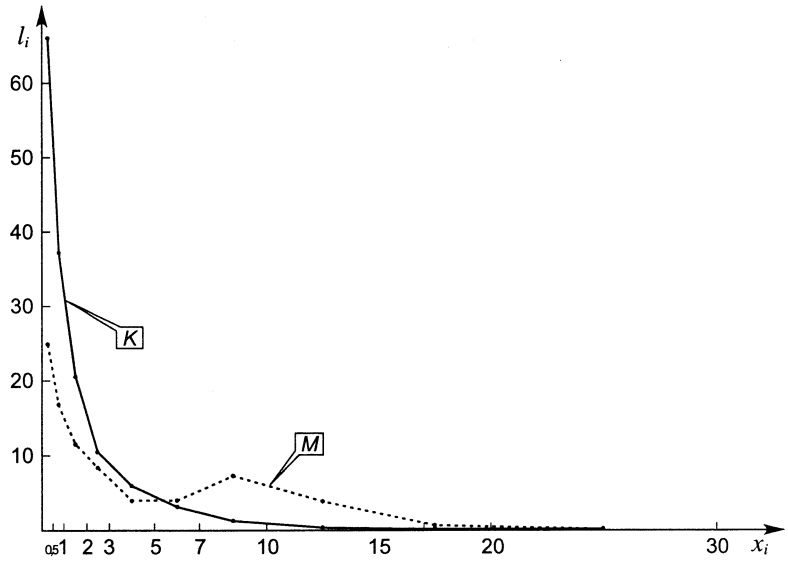
a) przejść z liczebności bezwzględnych $n_i$ na częstości względne $w_i$,

b) dla częstości względnych obliczyć gęstości $g_i$ lub natężenia liczebności $l_i$.

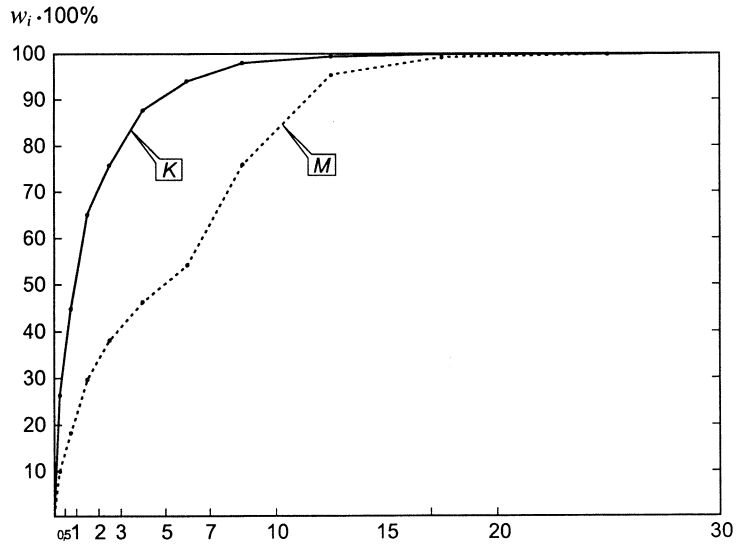
Tablica robocza dla sporządzenia wieloboków liczebności i wieloboków liczebności skumulowanych dla przykładu o gospodarstwach rolnych została zamieszczona następnej stronie.

**Tablica robocza**
**do sporządzenia wieloboków liczebności i wieloboków liczebności skumulowanych dla przykładu o gospodarstwach rolnych**

| $x_{0i} - x_{1i}$ | $h_i$ | $n_i$ dla $K$ | $n_i$ dla $M$ | $w_i \cdot 100\%$ dla $K$ | $w_i \cdot 100\%$ dla $M$ | $l_i$ dla $K^*$ | $l_i$ dla $M$ | Kumulaty dla $w_i \cdot 100\%$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | dla $K$ | dla $M$ |
| 0,1–0,5 | 0,4 | 18179 | 3782 | 26,3 | 9,9 | 65,75 | 24,75 | 26,3 | 9,9 |
| 0,5–1 | 0,5 | 12856 | 3209 | 18,6 | 8,4 | 37,20 | 16,80 | 44,9 | 18,3 |
| 1–2 | 1 | 14170 | 4393 | 20,5 | 11,5 | 20,50 | 11,50 | 65,4 | 29,8 |
| 2–3 | 1 | 7258 | 3209 | 10,5 | 8,4 | 10,50 | 8,40 | 75,9 | 38,2 |
| 3–5 | 2 | 8225 | 3094 | 11,9 | 8,1 | 5,95 | 4,05 | 87,8 | 46,3 |
| 5–7 | 2 | 4355 | 3018 | 6,3 | 7,9 | 3,15 | 3,95 | 94,1 | 54,2 |
| 7–10 | 3 | 2696 | 8328 | 3,9 | 21,8 | 1,30 | 7,30 | 98,0 | 76,0 |
| 10–15 | 5 | 1037 | 7487 | 1,5 | 19,6 | 0,30 | 3,92 | 99,5 | 95,6 |
| 15–20 | 5 | 207 | 1337 | 0,3 | 3,5 | 0,06 | 0,70 | 99,8 | 99,1 |
| 20–30 | 10 | 138 | 344 | 0,2 | 0,9 | 0,02 | 0,09 | 100,0 | 100,0 |
| Razem | × | 69121 | 38201 | 100,0 | 100,0 | × | × | × | × |



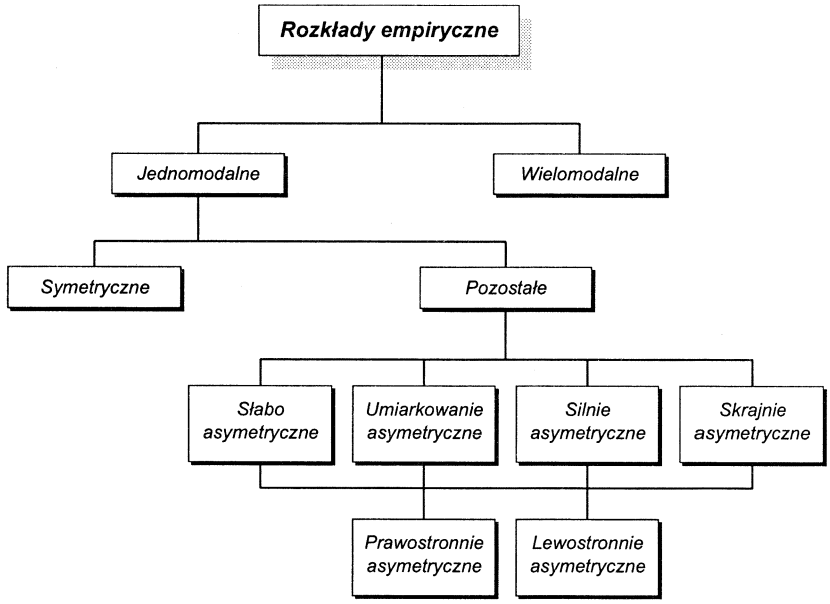

Wykres 3.13. Wieloboki liczebności dla gospodarstw indywidualnych




Wykres 3.14. Wieloboki liczebności skumulowanych dla gospodarstw indywidualnych

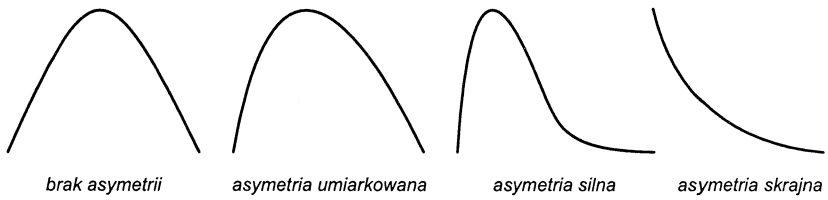
## 3.5. Typy rozkładów empirycznych

Z wykreślaniem wieloboków liczebności i krzywych gęstości wiąże się potrzeba wyróżnienia najważniejszych kształtów nakładów. Rozkładem empirycznym nazywa się przyporządkowanie kolejnym wartościom zmiennej $x_i$ odpowiadających im liczebności (w zależności od potrzeby $n_i$ lub $w_i$ lub $l_i$ lub $g_i$) na podstawie konkretnych obserwacji. Rozkładem teoretycznym nazywa się takie przyporządkowanie w oparciu o wyidealizowany model (np. rozkład dwumianowy, trójkątny, normalny).




Wykres 3.17. Rodzaje rozkładów empirycznych

Rozkłady jednomodalne to rozkłady z jednym, wyraźnie zaznaczonym ośrodkiem dominującym (posiadające ekstremum właściwe); z tego wynika, że umieszczenie wśród nich rozkładów skrajnie asymetrycznych jest naciągane, ale świadome. Asy-




Wykres 3.18. Siła asymetrii rozkładu

metria rozkładu jest tym silniejsza, im dłuższe jest jedno z ramion wykresu w stosunku do drugiego.

W zależności od tego, które z ramion wykresu jest dłuższe, mówi się o asymetrii lewo lub prawostronnej.

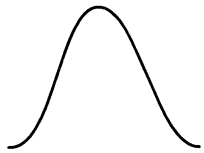
Jaki jest sens wyróżniania kształtów rozkładów?

**Po pierwsze:** liczba kształtów, jakie teoretycznie mogą się pojawić jest ogromna. Rozpatrzmy przykład. Populacja liczy 100 elementów, z których każdy może trafić do jednej z siedmiu klas. Sto elementów w siedmiu klasach można rozmieścić na $\binom{106}{100}$ sposobów. Daje to 1.705.904.746 kombinacji. Gdyby każdej z nich poświęcić tylko jedną sekundę, zajęłoby to przeszło 50 lat! Większość z nich nie jest dla nas interesująca, choćby ze względu na wielomodalność. Warto więc operować kilkunastoma najczęściej spotykanymi rozkładami, by móc stwierdzić, do którego z nich najbardziej podobny jest rozkład analizowanej zbiorowości.

**Po drugie:** dla populacji **jednorodnych**, gdzie różnice między jednostkami są wywołane jedynie losowymi przyczynami ubocznymi (mamy z taką sytuacją do czynienia tylko w przypadku kontrolowanych eksperymentów) otrzymujemy rozkład symetryczny zbliżony do rozkładu normalnego (w kształcie dzwonu).
Wraz ze wzrostem niejednorodności populacji odkształcenia rozkładu od postaci symetrycznej stają się coraz wyraźniejsze. Dla zbiorowości **względnie** jednorodnych obserwowana asymetria rozkładu może być dość silna.
W populacjach niejednorodnych wpływ przyczyn głównych na elementy zbiorowości maleje, rośnie oddziaływanie przyczyn ubocznych. Rozkłady stają się nieregularne, przybierając postać wielomodalną, siodłową czy skrajnie asymetryczną.
Statystyka nie wypracowała narzędzi pozwalających dobrze charakteryzować zbiorowości niejednorodne. Należy je uprzednio rozbić na podpopulacje o bardziej zbliżonym do siebie składzie.

**Zatem:** „... kształt rozkładów empirycznych nie tylko warunkuje dobór właściwych w danej sytuacji metod statystycznych, lecz także określa zakres dopuszczalnego wnioskowania." [A. Luszniewicz, *op. cit.*, s. 38]

**Wniosek: zanim przystąpimy do liczenia miar, analizujemy wykres!**



***rozkład symetryczny***

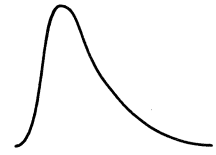
*np. uchylający się od płacenia alimentów wg wieku; rozkład wzrostu poborowych; rozkład częstości błędów przypadkowych*

***umiarkowana asymetria prawostronna (dodatnia)***

*zarobki; zgony wywołane gruźlicą wg. wieku; wiek osób skazanych za udzielanie łapówek*

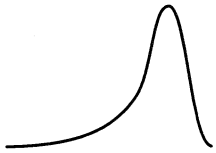
***umiarkowana asymetria lewostronna (ujemna)***

*zmarli na choroby nowotworowe wg wieku; rodziny robotnicze wg dochodu na osobę*

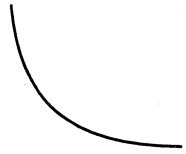


***silna prawa asymetria***

*zatrudnienie pracowników fizycznych w rolnictwie wg płacy miesięcznej; wiek sprawców wypadków drogowych; liczba kobiet wg liczby urodzonych dzieci*

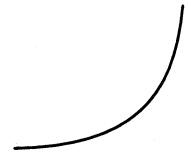


***asymetria silna lewostronna***



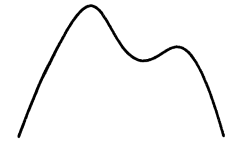
***skrajna prawa asymetria***

*męższczyźni w wieku produkcyjnym bierni zawodowo ze względu na naukę; rozkład czasu przeżycia chorego na raka po radykalnej operacji; sprawcy rozbojów, gwałtów wg wieku; indywidualne gospodarstwa rolne wg powierzchni*



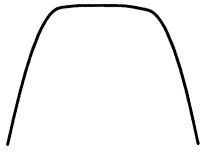
***skrajna lewa asymetria***

*mężczyźni w wieku produkcyjnym bierni zawodowo ze względu na otrzymywanie emerytury lub renty; straty wartości budynków drewnianych spowodowane pożarem*



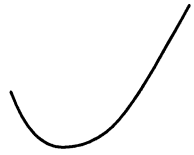
***rozkład bimodalny***

*rozkład wieku osób zmarłych na grypę; rozkład wieku zabójców, mankowiczów*



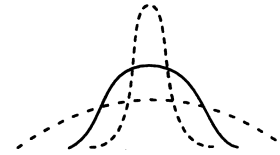
***rozkład bez wyraźnie zaznaczonego ośrodka dominującego***

*uczniowie szkoły podstawowej wg wieku; wyniki testu nie uwzględniającego różnic w poziomie nauczania; wiek skazanych za paserstwo*



***rozkład siodłowy (antymodalny)***

*mężczyźni w wieku produkcyjnym bierni zawodowo; zachmurzenie nieba w Greenwich*



***rozkład mezokurtyczny wraz z rozkładami symetrycznymi: lepto- i platokurtycznym***

Wykres 3.17. Najczęściej spotykane w praktyce kształty rozkładów

# Rozdział 4
# Miary tendencji centralnej

Miary tendencji centralnej określają przeciętny poziom cechy. Rozróżnia się:

przeciętne wielkości klasyczne — stanowią one wypadkową wartości cechy wszystkich badanych jednostek zbiorowości (średnia arytmetyczna, geometryczna, harmoniczna),

przeciętne wielkości pozycyjne — są wyznaczone pozycją niektórych jednostek (mediana) lub grup jednostek (dominanta).

## 4.1. Średnia arytmetyczna

***Średnia arytmetyczna*** — suma wartości cechy mierzalnej podzielona przez liczbę jednostek zbiorowości statystycznej. Wynika stąd interpretacja tej miary → jaka wartość cechy przypadałaby na jednostkę populacji, gdyby suma wartości cechy została rozłożona po równo.

$\bar{x}$ — symbol średniej arytmetycznej dla próby (statystyka),

$\mu$ — symbol średniej arytmetycznej dla populacji generalnej (parametr).

Obliczenia średniej arytmetycznej ($\bar{x}$) przeprowadza się według wzorów:

— dla indywidualnego wykazu wartości

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n} \quad \left( \mu = \frac{\sum x_i}{N} \right),$$

— dla szeregu punktowego $\bar{x} = \frac{\sum x_i n_i}{\sum n_i}$,

— dla szeregu przedziałowego $\bar{x} \cong \frac{\sum \mathring{x}_i n_i}{\sum n_i}$,

— dla szeregu przedziałowego z użyciem częstości względnych $\bar{x} \cong \frac{\sum \mathring{x}_i w_i}{\sum w_i} = \sum \mathring{x}_i w_i$.

Zalety średniej arytmetycznej:

a) jest znana większości ludzi,
b) zawsze istnieje,
c) jest zawsze tylko jedna,
d) nadaje się do dalszych czynności statystycznych (średnie ze średnich),
e) jest stosunkowo wiarygodna (średnie z prób tej samej populacji nie różnią się zbytnio).

Wady średniej arytmetycznej: podatna na wpływ wielkości skrajnych.

$\bar{x}$ jest wielkością abstrakcyjną, jej wartość poznawcza i zakres zastosowań zależą od typu rozkładu empirycznego. Prawidłowo charakteryzuje średni poziom w przypadku rozkładów symetrycznych lub o umiarkowanej sile asymetrii.

Nie należy jej obliczać
- dla rozkładów skrajnie asymetrycznych, bimodalnych lub wielomodalnych,
- dla szeregu o otwartych przedziałach klasowych (z wyjątkiem bardzo małych liczebności w tych przedziałach).

Podstawowe własności $\bar{x}$

a. $x_{\min} < \bar{x} < x_{\max}$;

b. $\sum(\mathring{x}_i - \bar{x})n_i = 0$;

c. $\overline{(x+y)} = \bar{x} + \bar{y}$;

d. jeśli wszystkie wartości zmiennej powiększymy (pomniejszymy lub pomnożymy) o pewną stałą, to średnia arytmetyczna będzie równa sumie (różnicy, iloczynowi) średniej arytmetycznej wyjściowych zmiennych i tej stałej;

$$\frac{\sum(x_i + c)n_i}{n} = \frac{\sum x_i n_i + \sum cn_i}{n} = \bar{x} + \frac{c\sum n_i}{n} = \bar{x} + c$$

$$\frac{\sum(x_i - c)n_i}{n} = \frac{\sum x_i n_i - \sum cn_i}{n} = \bar{x} - \frac{c\sum n_i}{n} = \bar{x} - c$$

$$\frac{\sum(x_i \cdot c)n_i}{n} = \frac{c\sum x_i n_i}{n} = c\bar{x}$$

$$\frac{\sum\left(\frac{x_i}{c}\right)n_i}{n} = \frac{\frac{1}{c}\sum x_i n_i}{n} = \frac{\bar{x}}{c}$$

e. wartość $\bar{x}$ zależy nie od absolutnych wielkości wag, ale od ich proporcji.

$$\frac{\sum x_i(cn_i)}{\sum(cn_i)} = \frac{c\sum x_i n_i}{c\sum n_i} = \bar{x}$$

$$\frac{\sum x_i\left(\frac{n_i}{c}\right)}{\sum\left(\frac{n_i}{c}\right)} = \frac{\frac{1}{c}\sum x_i n_i}{\frac{1}{c}\sum n_i} = \bar{x}$$

***Przykłady na liczenie średniej arytmetycznej***

***Przykład 4.1.***

W czasie sesji uzyskałem na egzaminach następujące oceny: 3; 3,5; 4; 5. Jaka była średnia ocena?

$$\bar{x} = \frac{3 + 3{,}5 + 4 + 5}{4} = \frac{15{,}5}{4} = 3{,}875$$

$$\text{wzór } \bar{x} = \frac{\sum x_i}{N}$$

***Uwaga:*** W statystyce opisowej każdą badaną populację uznajemy za populację generalną. Operowanie symbolem μ nie jest wygodne i dlatego w dalszym ciągu stosowana będzie powyższa, hybrydowa postać wzoru.

***Przykład 4.2.***

Jaką średnią uzyskali studenci na egzaminie z ekonomii? Zdawały 523 osoby, 47 dostało piątki, 64 — czwórki z plusem, 113 — czwórki, 92 — trójki z plusem, 165 — trójki, pozostali nie zdali.

Liczba osób, które dostały ocenę niedostateczną:

$$523-(47+64+113+92+165)=523-481=42$$

$$\bar{x}=\frac{47\cdot 5+65\cdot 4{,}5+113\cdot 4+92\cdot 3{,}5+165\cdot 3+42\cdot 2}{523}=$$

$$=\frac{235+288+452+322+495+84}{523}=\frac{1876}{523}=3{,}587\cong 3{,}59$$

$$\text{wzór } \bar{x}=\frac{\sum x_i n_i}{N}$$

Wygodniejsza forma zapisu i obliczeń:

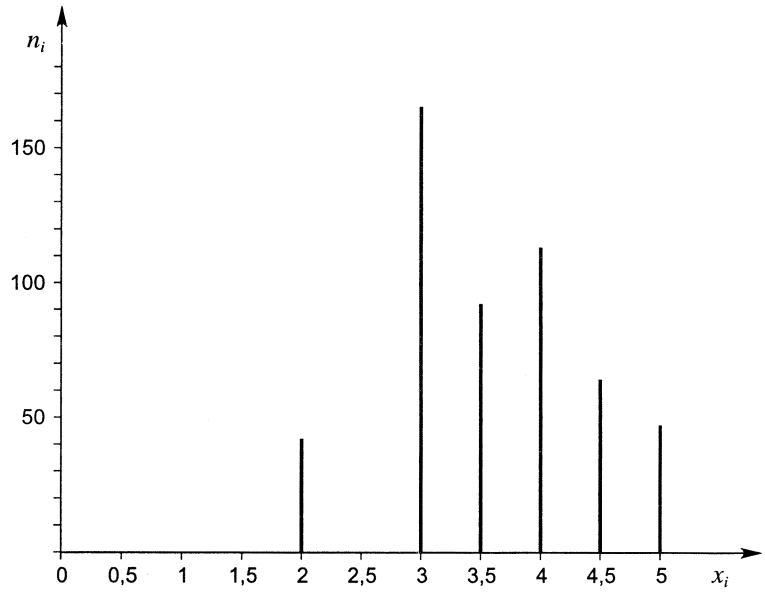
| $x_i$ | $n_i$ | $x_i n_i$ |
|---|---|---|
| 2 | 42 | 84 |
| 3 | 165 | 495 |
| 3,5 | 92 | 322 |
| 4 | 113 | 452 |
| 4,5 | 64 | 288 |
| 5 | 47 | 235 |
| Razem | 523 | 1876 |

***Przykład 4.3.***

Moja dotychczasowa średnia wynosi 4,11. Na ostatniej sesji dostałem: 4; 4,5; 5; 5. Ile wynosi zaktualizowana średnia?

a) Zgodnie z regułą liczenia średniej arytmetycznej, w liczniku musi się znaleźć suma ocen ze wszystkich egzaminów. Dane, którymi dysponujemy umożliwiają jedynie wyznaczenie średniej dla **ostatniej** sesji. Bez dodatkowych informacji polecenie jest niewykonalne!

b) Dowiadujemy się, że 4,11 jest średnią z 23 egzaminów.




Wykres 4.1. Wykres linii pionowych dla rozkładu z przykładu 4.2 (cecha skokowa, mała liczba wyróżnionych wartości zmiennej)

$$\bar{x} = \frac{\sum_{1}^{27} x_i}{27} = \frac{\sum_{1}^{23} x_i + x_{24} + x_{25} + x_{26} + x_{27}}{27}$$

$$\sum_{1}^{23} x_i = ? \qquad \bar{x} = \frac{\sum x_i}{N} \rightarrow \sum x_i = \bar{x} \cdot N$$

$$4{,}11 = \frac{\sum_{1}^{23} x_i}{23} \rightarrow \sum_{1}^{23} x_i = \overbrace{4{,}11 \cdot 23}^{94{,}53} = \bar{x} \cdot N$$

$$\bar{x} = \frac{94{,}53 + 4 + 4{,}5 + 5 + 5}{27} = \frac{113{,}03}{27} = 4{,}186$$

Wzór na średnią ze średnich:

$$\bar{X} = \frac{\bar{x}_1 \cdot N_1 + \bar{x}_2 \cdot N_2 + \ldots + \bar{x}_n \cdot N_n}{N_1 + N_2 + \ldots + N_n}$$

***Przykład 4.4.***

Ile wynosiła średnia miesięczna płaca w wojsku w roku 1994? Wiemy, że:

- wariant a.

| Wyszczególnienie | Liczba zatrudnionych | Średnia miesięczna płaca brutto |
|---|---|---|
| Mundurowi | 146151 | 669,00 |
| Cywilni | 35629 | 391,14 |
| Razem | 181780 | × |

Źródło: dane GUS.

$$\overline{X} = \frac{\overline{x}_z \cdot N_z + \overline{x}_c \cdot N_c}{N_z + N_c} = \frac{669 \cdot 146151 + 391{,}14 \cdot 35629}{181780} =$$

$$= \frac{97775019 + 13935927{,}06}{181780} =$$

$$= \frac{111710946{,}06}{181780} = 614{,}539 \cong 614{,}54$$

- wariant b.

| Wyszczególnienie | Odsetek zatrudnionych $w_i \cdot 100\%$ | Średnia miesięczna płaca brutto $\overline{x}_i$ |
|---|---|---|
| Mundurowi | 80,4 | 669,00 |
| Cywilni | 19,6 | 391,14 |
| Razem | 100,0 | × |

$$\overline{X} = \frac{\sum \overline{x}_i w_i}{\sum w_i} = \frac{\sum \overline{x}_i \cdot w_i \cdot 100}{\sum w_i \cdot 100} = \frac{669 \cdot 80{,}4 + 391{,}14 \cdot 19{,}6}{100} =$$

$$= \frac{53787{,}6 + 7666{,}344}{100} = \frac{61453{,}944}{100} = 614{,}539$$

***Przykład 4.5.***

Zbadano 20 studentów ze względu na liczbę lat poświęconych dotychczas na studia. Okazało się, że dwóch z nich studiuje mniej niż dwa lata, dziesięciu — krócej niż 4 lata, a osiemnastu mniej niż 6 lat. Za pomocą średniej arytmetycznej proszę ustalić, od ilu lat średnio studiują badani studenci (maksymalny okres studiowania w badanej grupie nie przekracza 8 lat).

***Uwaga.*** Specyfika zadania polega na podaniu danych w formie kumulaty.

***Wniosek:*** należy odtworzyć szereg rozdzielczy.

| $x_{0i} - x_{1i}$ | Cum |
|---|---|
| 0–2 | 2 |
| 2–4 | 10 |
| 4–6 | 18 |
| 6–8 | 20 |
| Razem | × |

| $x_{0i} - x_{1i}$ | $n_i$ | $\mathring{x}_i$ | $\mathring{x}_i n_i$ |
|---|---|---|---|
| 0–2 | 2 | 1 | 2 |
| 2–4 | 8 | 3 | 24 |
| 4–6 | 8 | 5 | 40 |
| 6–8 | 2 | 7 | 14 |
| Razem | 20 | × | 80 |

$$\bar{x} \cong \frac{\sum \mathring{x}_i n_i}{N} = \frac{80}{20} = 4{,}0$$

***Przykład 4.6.***

Za pomocą średniej arytmetycznej ocenić przeciętną wielkość gospodarstwa rolnego, jeżeli dane tworzą następujący szereg:

| Wielkość gospodarstwa w ha | poniżej 2 ha | 2–4 | 4–7 | 7–10 | 10 i więcej | Razem |
|---|---|---|---|---|---|---|
| Liczba gospodarstw | 2 | 4 | 6 | 5 | 3 | 20 |

Powyższy szereg rozdzielczy zawiera dwie otwarte klasy. Przede wszystkim powinniśmy spróbować je domknąć. Za dolną granicę pierwszej klasy można uznać zero. Gorzej wygląda ostatni przedział. Nie ma żadnych przesłanek by uznać, że górna granica wynosi 20,50 czy 1000 ha.

***Uwaga.*** W praktyce obliczeń przyjmuje się, że jeśli w tablicy występuje otwarta klasa, nie wiemy jak ją domknąć, liczebność w przedziale nie przekracza ok. 5% liczebności populacji i możemy dodatkowo przyjąć, że pominięcie tej klasy nie wpłynie w istotny sposób na wyniki badania to dopuszcza się je **odrzucenie** (liczymy miary dla **zredukowanej** tablicy).

***Wniosek:*** Ponieważ w ostatniej klasie znajduje się aż 15% populacji, bez dodatkowych informacji średniej arytmetycznej nie policzymy.

Uzyskujemy dodatkowe dane:

a) średnia powierzchni dwóch najmniejszych gospodarstw wynosi 1,5 ha;

b) gospodarstwa liczące 10 i więcej ha zajmują razem powierzchnię 51 ha.

Tablica robocza przed uzyskaniem dodatkowych danych:

| $\mathring{x}_i$ | . | 3 | 5,5 | 8,5 | . | × |
|---|---|---|---|---|---|---|
| $x_{0i} - x_{1i}$ | 2 ha i mniej | 2–4 | 4–7 | 7–10 | 10 i więcej | Razem |
| $n_i$ | 2 | 4 | 6 | 5 | 3 | 20 |
| $\mathring{x}_i n_i$ | . | 12 | 33 | 42,5 | . | . |

Tablica robocza z wykorzystaniem dodatkowych danych:

| $\mathring{x}_i$ | (1,5) | 3 | 5,5 | 8,5 | (17) | × |
|---|---|---|---|---|---|---|
| $x_{0i} - x_{1i}$ | 1–2 | 2–4 | 4–7 | 7–10 | 10–24 | Razem |
| $n_i$ | 2 | 4 | 6 | 5 | 3 | 20 |
| $\mathring{x}_i n_i$ | 3 | 12 | 33 | 42,5 | 51 | 141,5 |

$$\bar{x} \cong \frac{\sum \mathring{x}_i n_i}{\sum n_i} = \frac{141,5}{20} = 7,075 \text{ ha}$$

$$\bar{x} \cong 7,1 \text{ ha}$$

***Przykład 4.7.***

Wyznaczmy różnymi sposobami średnią arytmetyczną dla przykładu o firmie Profi

- dla indywidualnego wykazu wartości:

$$\bar{x} = \frac{\sum x_i}{N} = \frac{19+16+22+9+22+\ldots+27+17+21}{64} =$$

$$\frac{1274}{64} = 19,90625 \approx 19,9$$

- dla szeregu punktowego:

| $x_i$ | $n_i$ |
|---|---|
| 5 | 1 |
| 6 | 1 |
| 7 | 2 |
| 8 | 1 |
| ⋮ | ⋮ |
| 36 | 1 |
| 37 | 1 |
| 39 | 1 |
| Razem | 64 |

$$\bar{x} = \frac{\sum x_i n_i}{N} = \frac{5 \cdot 1 + 6 \cdot 1 + 7 \cdot 2 + 8 \cdot 1 + \ldots + 36 \cdot 1 + 37 \cdot 1 + 39 \cdot 1}{64} = \frac{1274}{64} = 19{,}9$$

- dla szeregu przedziałowego — częstości względne $w_i$:

| $x_{0i} - x_{1i}$ | $w_i$ | $\mathring{x}_i$ | $\mathring{x}_i w_i$ |
|---|---|---|---|
| 5–10 | 0,109 | 7,5 | 0,8175 |
| 10–15 | 0,156 | 12,5 | 1,9500 |
| 15–20 | 0,203 | 17,5 | 3,5525 |
| 20–25 | 0,281 | 22,5 | 6,3225 |
| 25–30 | 0,125 | 27,5 | 3,4375 |
| 30–35 | 0,078 | 32,5 | 2,5350 |
| 35–40 | 0,047 | 37,5 | 1,7625 |
| Razem | 0,999 | × | 20,3775 |

$$\bar{x} = \sum \mathring{x}_i w_i = 20{,}3775 \cong 20{,}38$$

- dla szeregu przedziałowego — liczebności bezwzględne $n_i$:

| $x_{0i} - x_{1i}$ | $n_i$ | $\mathring{x}_i$ | $\mathring{x}_i n_i$ |
|---|---|---|---|
| 5–10 | 7 | 7,5 | 52,5 |
| 10–15 | 10 | 12,5 | 125,0 |
| 15–20 | 13 | 17,5 | 227,5 |
| 20–25 | 18 | 22,5 | 405,0 |
| 25–30 | 8 | 27,5 | 220,0 |
| 30–35 | 5 | 32,5 | 162,5 |
| 35–40 | 3 | 37,5 | 112,5 |
| Razem | 64 | × | 1305,0 |

$$\bar{x} \cong \frac{\sum \mathring{x}_i n_i}{N} = \frac{1305{,}0}{64} = 20{,}390625 \cong 20{,}4$$

Różnica między średnią dokładną dla indywidualnego wykazu wartości (także szeregu punktowego) wynoszącą 19,90625, a średnią dla szeregu przedziałowego równą 20,390625 bierze się z zastąpienia danych indywidualnych ich reprezentantami (środkami przedziałów). Nazywa się ją błędem grupowania.

Podobnie jak dla przykładu o pensjach pracowników MON otrzymaliśmy praktycznie takie same wyniki:

raz: dla kolumny liczebności bezwzględnych $n_i$,
dwa: dla kolumny częstości względnych $w_i$,

wynika to z *e*-tej własności średniej arytmetycznej:

**Wartość $\bar{x}$ zależy nie od absolutnych wielkości wag, ale od ich proporcji.**

Aby wykazać niektóre z własności średniej arytmetycznej z punktów d. oraz e. przekształcamy poprzednią tabelę:

| $x_{0i} - x_{1i}$ | $n_i$ | $\mathring{x}_i$ | $\mathring{x}_i n_i$ |
|---|---|---|---|
| 17–22 | 70 | 19,5 | 1365 |
| 22–27 | 100 | 24,5 | 2450 |
| 27–32 | 130 | 29,5 | 3835 |
| 32–37 | 180 | 34,5 | 6210 |
| 37–42 | 80 | 39,5 | 3160 |
| 42–47 | 50 | 44,5 | 2225 |
| 47–52 | 30 | 49,5 | 1485 |
| Razem | 640 | × | 20730 |

1. dodając do każdej wartości $x_i$ stałą $c = 12$,
2. mnożąc każdą z liczebności przez stałą $k = 10$.

Zgodnie z własnościami średniej wynikowa średnia musi być o 12 większa od średniej dla poprzedniej tabeli.

$$\bar{x}^* = \frac{\sum(\mathring{x}_i + c)\cdot(n_i k)}{\sum n_i k} = \frac{\sum \mathring{x}_i n_i k + \sum c n_i k}{k\sum n_i} = \bar{x} + c$$

$$\bar{x}^* = \frac{20730}{640} = 32{,}390625 = 20{,}390625 + 12 = \bar{x} + c$$

## 4.2. Mediana i pozostałe kwantyle

**Mediana** (wartość środkowa, kwartyl drugi) to wartość cechy $X$ w rozkładzie uporządkowanych wariantów, poniżej oraz powyżej której znajduje się połowa jednostek zbiorowości (dzieli populację na połowy).

Może być wyznaczana metodą graficzną lub rachunkową. Dla rozkładów punktowych odczytujemy wartość elementu środkowego (gdy $N$ nieparzyste) lub znajdujemy średnią arytmetyczną z dwóch centralnych wartości (gdy $N$ parzyste). W przypadku rozkładów przedziałowych wartość mediany trzeba interpolować.

Własności mediany:

a) zawsze istnieje,

b) zawsze jest tylko jedna,

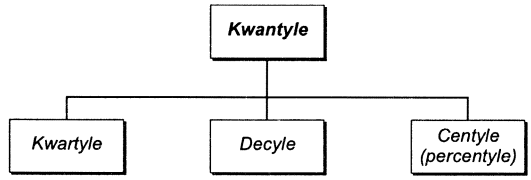
c) nie ulega wpływom wartości skrajnych,

d) może służyć do opisania populacji, których nie można określić liczbowo,

e) bardziej podatna na przypadkowe odchylenia od średniej (mniej wiarygodna dla próby od $\bar{x}$),

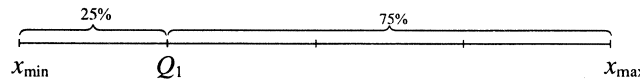
f) $\sum|x_i - M| < \sum|x_i - T|$ dla $M \neq T$.

Walory poznawcze mediany rosną w miarę zwiększania liczebności populacji (jeśli liczba obserwacji jest niewielka jej użycie nie jest wskazane). Jest jedyną średnią, którą wyznacza się dla rozkładów skrajnie asymetrycznych. Jest też jedną z miar statystycznych nazwanych kwantylami. Są to miary pozycyjne dzielące populację statystyczną w określonej proporcji.

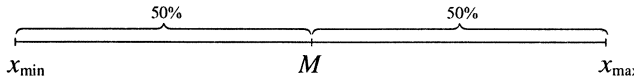



Kwartyle ($Q$) dzielą populację z dokładnością do ćwiartek, decyle ($D$) dzielą populację z dokładnością do dziesiątek, centyle ($C$) dzielą populację z dokładnością do pojedynczych procentów.

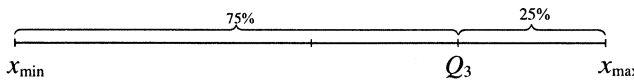
Pierwszy kwartyl ($Q_1$) odcina 25% elementów populacji o najmniejszych wartościach od pozostałych 75% jednostek:



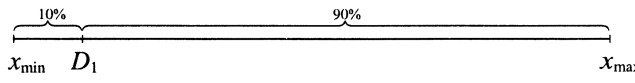
Mediana (kwartyl drugi) symbolizowana jako $M$ lub $Q_2$ dzieli populację na połowy:



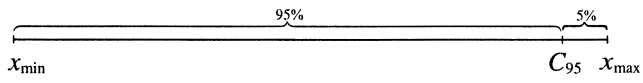

Trzeci kwartyl ($Q_3$) oddziela 75% populacji od 25% jednostek o największych wartościach:




Przykładowo decyl pierwszy ($D_1$) dzieli elementy populacji w proporcji 10%–90%:




Dziewięćdziesiąty piąty centyl ($C_{95}$) oddziela 95% populacji od górnych 5%:




Odsetek populacji odcinany przez dany kwantyl na lewo od niego nazywamy indeksem kwantyla. Indeks mediany wynosi 0,5; pierwszego kwartyla 0,25; czwartego decyla 0,4 itd.

Tryb postępowania przy wyznaczaniu kwantyli jest następujący:

krok pierwszy: określenie numeru elementu, który dzieli populację w zadanej dla kwantyla proporcji;

krok drugi: znalezienie wartości elementu o numerze ustalonym w pierwszym kroku.

***Wyznaczanie kwantyli dla szeregu uporządkowanego***

1. numer kwantyla znajdujemy ze wzoru:

$$\begin{array}{ccccc} \text{numer kwantyla} & = & \text{indeks kwantyla} & \cdot & (N+1) \\ \downarrow & & \downarrow & & \\ n_{kw} & = & i_{kw} & \cdot & (N+1) \end{array}$$

2. z uporządkowanego wykazu wartości odczytuje się wartość konkretnego elementu.

***Uwaga.*** 1 dodaje się w celu prawidłowego ustalenia pozycji mediany niezależnie od tego czy liczba elementów populacji jest parzysta czy nieparzysta.

| 1 | 2 | 3 |
|---|---|---|
| 6 | 8 | 9 |

$$n_M = (3+1)\cdot 0{,}5 = 2$$

wartością drugiego elementu jest tutaj 8, a więc $M = 8$.

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| 6 | 8 | 9 | 10 |

$$n_M = (4+1)\cdot 0{,}5 = 2{,}5$$

wynik 2,5 interpretujemy: wartością mediany jest wartość elementu drugiego i połowy różnicy między elementem następnym, czyli w tym wypadku trzecim a poprzednim, drugim

$$M = 8 + 0{,}5(9-8) = 8{,}5$$

inaczej jest to średnia elementów drugiego i trzeciego.

$$M = \frac{x_2 + x_3}{2}$$

***Przykład 4.9.***

Firma Profi

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

Znajdujemy kwartyle:

$n_{Q_1} = (64+1)\cdot 0{,}25 = 16{,}25$ wynik oznacza, że pierwszy kwartyl jest równy sumie wartości elementu 16. i 0,25 różnicy między elementami o numerach 17 i 16.

$$Q_1 = 14 + 0{,}25(14-14) = 14$$

Ponieważ pierwszy kwartyl dzieli populację w proporcji 25%–$Q_1$–75% wynik zinterpretujemy: w 25% przypadków wielkość sprzedaży była nie większa niż 14 palet, w pozostałych 75% przypadków sprzedaż była nie mniejsza od 14 palet.

$$n_{Q_2} = 65\cdot 0{,}5 = 32{,}5$$

$$Q_2 = M = \frac{20+20}{2} = 20 \quad \text{lub} \quad M = 20 + 0{,}5(20-20) = 20$$

W połowie przypadków dzienna sprzedaż była nie większa od 20 palet, w pozostałej połowie przypadków dzienna sprzedaż była nie mniejsza od 20 palet.

$$n_{Q_3} = 65\cdot 0{,}75 = 48{,}75$$

$$Q_3 = 24 + 0{,}75(25-24) = 24{,}75$$

W 75% przypadków dzienna sprzedaż była nie większa od 24,75 palety, w pozostałych 25% przypadkach była nie mniejsza od 24,75 palety.

***Przykład 4.10.***

Wyniki kolokwium dla 111 osób:

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 43 | 44 | 47 | 49 | 50 | 56 | 58 | 61 | 70 | 74 | 75 | 77 | 77 | 77 | 78 |
| 80 | 81 | 82 | 82 | 83 | 85 | 86 | 86 | 86 | 91 | 92 | 96 | 97 | 100 | 100 |
| 100 | 100 | 101 | 104 | 104 | 105 | 106 | 106 | 106 | 107 | 108 | 108 | 109 | 109 | 109 |
| 110 | 110 | 111 | 112 | 113 | 114 | 115 | 115 | 117 | 118 | 118 | 118 | 119 | 119 | 120 |
| 122 | 122 | 122 | 122 | 123 | 123 | 123 | 125 | 125 | 125 | 126 | 126 | 127 | 127 | 128 |
| 129 | 129 | 129 | 129 | 129 | 130 | 130 | 130 | 130 | 130 | 130 | 130 | 131 | 131 | 131 |
| 132 | 132 | 132 | 133 | 134 | 135 | 135 | 136 | 136 | 136 | 137 | 138 | 138 | 138 | 139 |
| 140 | 141 | 141 | 142 | 143 | 145 | | | | | | | | | |

Znajdowanie wartości $D_1$, $Q_1$, $Q_2$, $C_{65}$:

$n_{D_1} = (1 + 111) \cdot 0{,}1 = 11{,}2$ — wynik ten interpretuje się następująco: pierwszy decyl jest równy wartości jedenastego elementu i dwóch dziesiątych różnicy pomiędzy elementem następnym, tj. dwunastym a jedenastym.

$D_1 = 75 + 0{,}2 \cdot (77 - 75) = 75{,}4$

$n_{Q_1} = (1 + 111) \cdot 0{,}25 = 28$ $\quad Q_1 = 97$

$n_{Q_2} = (1 + 111) \cdot 0{,}5 = 56$ $\quad Q_2 = 118$

$n_{C_{65}} = (1 + 111) \cdot 0{,}65 = 72{,}8$ $\quad C_{65} = 126 + 0{,}8\,(127 - 126) = 126{,}8$

***Wyznaczanie kwantyli dla szeregu punktowego***

Znajdowanie kwantyli dla szeregu punktowego umożliwia kumulata.

| $x_i$ | $n_i$ | Cum |
|---|---|---|
| 5 | 1 | 1 |
| 6 | 1 | 2 |
| 7 | 2 | 4 |
| 8 | 1 | 5 |
| 9 | 2 | 7 |
| 10 | 2 | 9 |
| 11 | 2 | 11 |
| 12 | 2 | 13 |
| 13 | 2 | 15 |
| 14 | 2 | 17 |
| 15 | 2 | 19 |

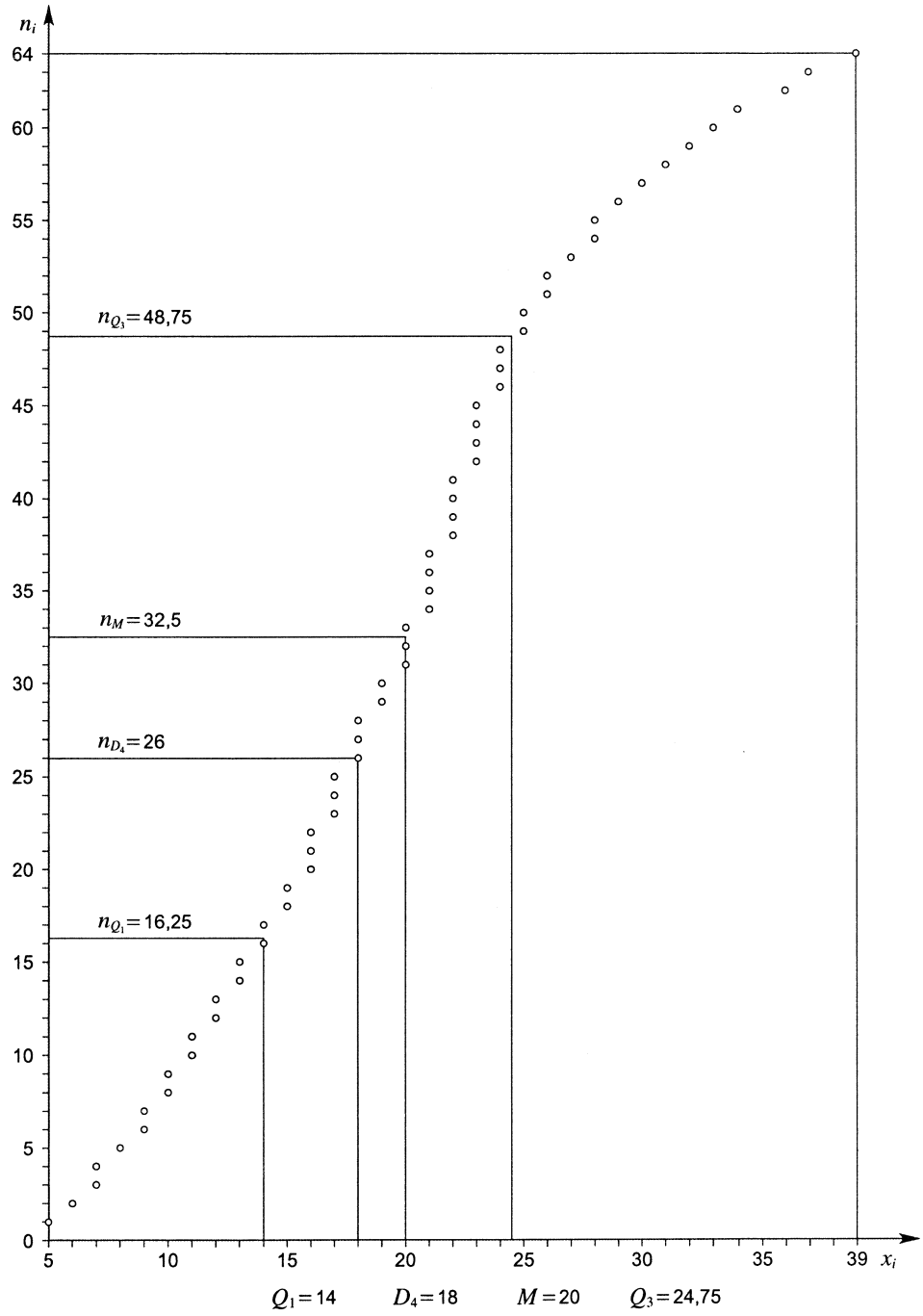
| $x_i$ | $n_i$ | Cum |
|---|---|---|
| 16 | 3 | 22 |
| 17 | 3 | 25 |
| 18 | 3 | 28 |
| 19 | 2 | 30 |
| 20 | 3 | 33 |
| 21 | 4 | 37 |
| 22 | 4 | 41 |
| 23 | 4 | 45 |
| 24 | 3 | 48 |
| 25 | 2 | 50 |
| 26 | 2 | 52 |

| $x_i$ | $n_i$ | Cum | |
|---|---|---|---|
| 27 | 1 | 53 | |
| 28 | 2 | 55 | |
| 29 | 1 | 56 | |
| 30 | 1 | 57 | |
| 31 | 1 | 58 | $\leftarrow M$ |
| 32 | 1 | 59 | |
| 33 | 1 | 60 | |
| 34 | 1 | 61 | |
| 36 | 1 | 62 | |
| 37 | 1 | 63 | |
| 39 | 1 | 64 | |
| Razem | 64 | × | |

Ile wynosi mediana?

$$n_M = (64 + 1) \cdot 0{,}5 = 32{,}5$$

Kumulata dla szeregu punktowego pomaga ustalić wartości elementów 32. i 33. w szeregu uporządkowanym. W wierszu 16. znajdujemy, że są 33 elementy o wartoś-




Wykres 4.1. Graficzne wyznaczanie kwantyli dla kumulaty dokładnej

ciach mniejszych/równych 20. W wierszu 15. znajduje się informacja o 28 elementach populacji o wartościach mniejszych/równych 18. Wynika stąd, że elementy o numerach 31, 32 i 33 mają wartość 20. Mediana, jako średnia elementów o numerach 32 i 33 równa jest

$$M = \frac{20+20}{2} = 20 \quad \text{lub} \quad M = 20 + 0{,}5(20-20) = 20$$

Znajdujemy $Q_3$.

$$n_{Q_3} = (64+1)\cdot 0{,}75 = 48{,}75$$

Musimy znaleźć wartości elementów o numerze 48 i 49 w uporządkowanym wykazie wartości. Dzięki kumulacie wiemy, że 48. element ma wartość 24 (wiersz 20), a elementy 49. i 50. wynoszą po 25 (wiersz 21).

$$Q_3 = 24 + 0{,}75(25-24) = 24{,}75$$

Znajdujemy $C_{95}$.

$$n_{C_{95}} = (64+1)\cdot 0{,}95 = 61{,}75$$

Element 61. ma wartość 34, element 62. ma wartość 36.

$$C_{95} = 34 + 0{,}75(36-34) = 34 + 0{,}75\cdot 2 = 35{,}5$$

***Graficzne wyznaczanie kwantyli dla kumulaty dokładnej***

Kwantyle z wykresu kumulaty znajdujemy prowadząc przez punkt na osi rzędnych odpowiadający numerowi danego kwantyla prostą równoległą do podstawy do miejsca spotkania z kumulatą, opuszczając stąd prostopadłą do podstawy i uznając odczytaną na osi odciętych liczbę za wartość kwantyla (wykres 4.1).

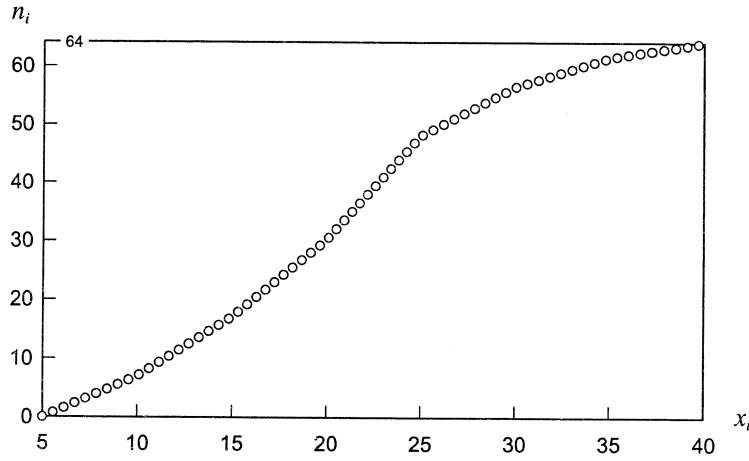
Oczywiście w praktyce nie wykreśla się kumulat dokładnych. Sporządzenie takiego wykresu dla populacji liczącej choćby kilkaset jednostek nie jest celowe.

Sytuację upraszczają:

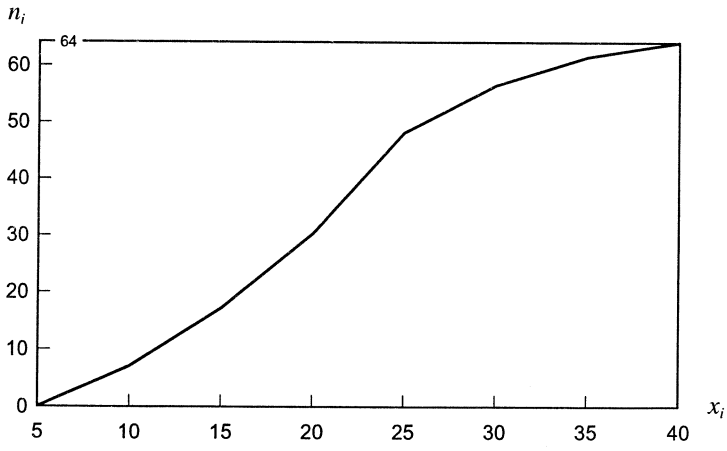
a. przejście na szereg przedziałowy,

b. przyjęcie, że zmienność wartości elementów w klasie jest liniowa.

Efekt tych założeń zawiera wykres poniżej:



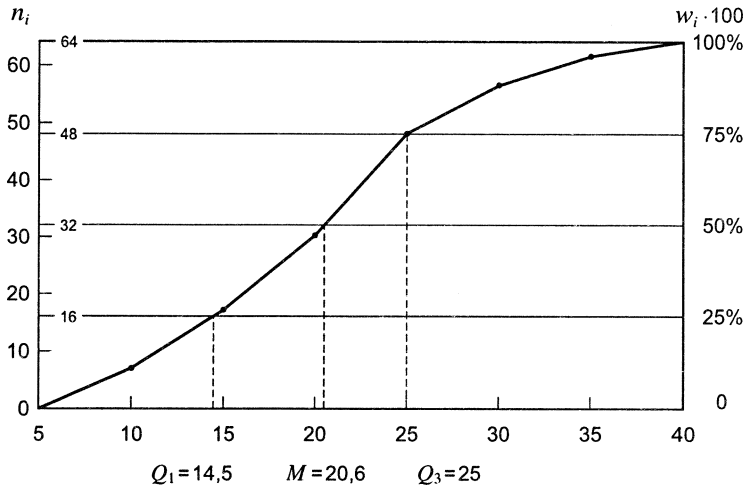

Ostatecznie kumulatę dla szeregu przedziałowego przedstawia się w jeszcze większym uproszczeniu za pomocą odcinków, rezygnując z prezentowania poszczególnych jednostek, skoro ich rzeczywisty rozkład jest inny. Jest to postać, która wyprzedzająco znalazła się na stronie 39. Przypomnijmy ją:




***Graficzne wyznaczenie kwantyli dla szeregu przedziałowego***

Numer kwantyla dla szeregu przedziałowego, ze względu na stosowane przybliżenia, zarówno dla metody graficznej, jak i analitycznej, obliczamy mnożąc liczebność populacji przez indeks kwantyla (bez dodawania jedynki).

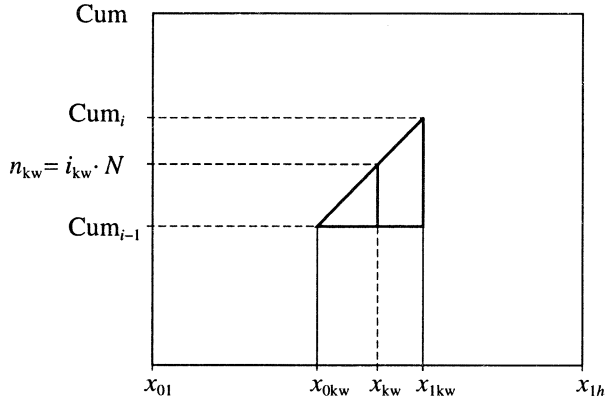
$$n_{Q_1} = 64 \cdot 0{,}25 = 16 \qquad n_M = 64 \cdot 0{,}5 = 32 \qquad n_{Q_3} = 64 \cdot 0{,}75 = 48$$




Wykres 4.1. Graficzne wyznaczanie kwartyli (przykład dla firmy Profi)

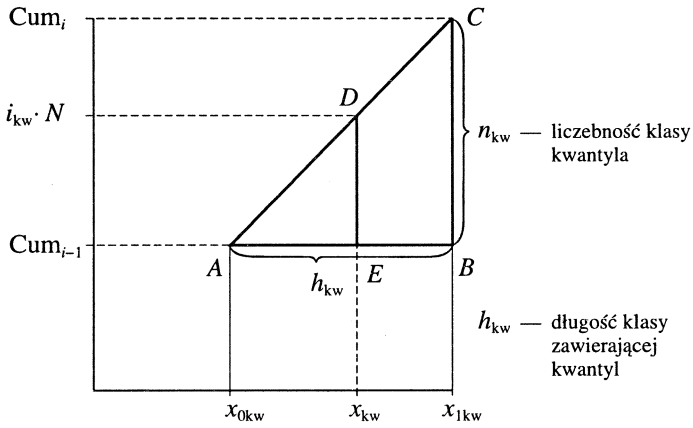
***Wyznaczanie kwantyli dla szeregu przedziałowego***

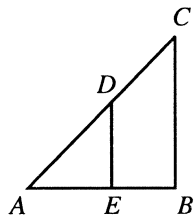
Szczegóły procedury interpolacyjnej, ważne dla wyprowadzenia wzoru na liczenie kwantyli, obrazuje wycinek wykresu kumulaty ograniczony do przedziału, w którym mieści się kwantyl. W wyznaczaniu wartości kwantyli uczestniczą:




gdzie:

$\text{Cum}_{i-1} = n(x < x_{0kw})$ – kumulata na dolnej granicy klasy,
$\text{Cum}_i$ – kumulata na górnej granicy klasy,
$x_{0kw}$ – dolna granica klasy, w której mieści się kwantyl,
$x_{1kw}$ – górna granica klasy, w której mieści się kwantyl,
$x_{kw}$ – wartość kwantyla.




Z podobieństwa trójątów

$$\frac{AE}{AB} = \frac{DE}{CB}$$

$$AE = x_{kw} - x_{0kw}$$

$$AB = h_{kw}$$

$$DE = i_{kw} \cdot N - \text{Cum}_{i-1}$$

$$CB = n_{kw}$$

$$\frac{x_{kw} - x_{0kw}}{h_{kw}} = \frac{i_{kw} \cdot N - \text{Cum}_{i-1}}{n_{kw}} \quad / \cdot h_{kw}$$

$$x_{kw} - x_{0kw} = \frac{i_{kw} \cdot N - \text{Cum}_{i-1}}{n_{kw}} \cdot h_{kw}$$

kwantyl = dolna granica przedziału w którym znajduje się kwantyl + (indeks kwantyla · $N$ − dystrybuanta do klasy zawierającej kwantyl) / liczebność przedziału, w którym znajduje się kwantyl · długość przedziału kwantyla

***Przykład 4.11.***

Sprzedaż w firmie Profi

| Nr | $x_{0i} - x_{1i}$ | $n_i$ | Kumulata | |
|---|---|---|---|---|
| 1 | 5–10 | 7 | 7 | |
| 2 | 10–15 | 10 | 17 | $\leftarrow Q_1$ |
| 3 | 15–20 | 13 | 30 | |
| 4 | 20–25 | 18 | 48 | $\leftarrow M, Q_3$ |
| 5 | 25–30 | 8 | 56 | |
| 6 | 30–35 | 5 | 61 | |
| 7 | 35–40 | 3 | 64 | |
| × | Razem | 64 | × | |

$$n_{Q_1} = 0{,}25 \cdot 64 = 16$$

Dzięki kumulacie wiemy, że 16. element znajduje się w klasie drugiej. Znane są zatem wartości wyrażeń, które musimy podstawić do wzoru.

$$Q_1 = 10 + \frac{16-7}{10} \cdot 5 = 10 + \frac{9 \cdot 5}{10} = 10 + 4{,}5 = 14{,}5$$

Wartości kwantyli dla szeregu punktowego (dokładna) i dla szeregu przedziałowego (przybliżona) różnią się z tych samych powodów co wartości średnich arytmetycznych.

$$n_M = 0{,}5 \cdot 64 = 32$$

Element o numerze 32 znajduje się w klasie czwartej.

$$M = 20 + \frac{32-30}{18} \cdot 5 = 20 + \frac{10}{18} = 20 + 0{,}5(5) \cong 20{,}6$$

$$n_{Q_3} = 0{,}75 \cdot 64 = 48$$

Element 48. jest ostatni w klasie czwartej, a więc równy górnej granicy klasy.

$$Q_3 = 20 + \frac{48-30}{18} \cdot 5 = 20 + \frac{18 \cdot 5}{18} = 20 + 5 = 25$$

***Przykład 4.12.*** Wyniki kolokwium 111 osób

| Nr | $x_{0i} - x_{1i}$ | $n_i$ | Kumulata dla $n_i$ | $w_i$ | Kumulata dla $w_i$ | |
|---|---|---|---|---|---|---|
| 1 | 40–50 | 4 | 4 | 0,036 | 0,036 | |
| 2 | 50–60 | 3 | 7 | 0,027 | 0,063 | |
| 3 | 60–70 | 1 | 8 | 0,009 | 0,072 | |
| 4 | 70–80 | 7 | 15 | 0,063 | 0,135 | ← $D_1$ |
| 5 | 80–90 | 9 | 24 | 0,081 | 0,216 | |
| 6 | 90–100 | 4 | 28 | 0,036 | 0,252 | ← $Q_1$ |
| 7 | 100–110 | 17 | 45 | 0,153 | 0,405 | |
| 8 | 110–120 | 14 | 59 | 0,126 | 0,531 | ← $Q_2$ |
| 9 | 120–130 | 21 | 80 | 0,189 | 0,720 | ← $C_{65}$ |
| 10 | 130–140 | 25 | 105 | 0,225 | 0,945 | ← przedział dominanty |
| 11 | 140–150 | 6 | 111 | 0,054 | 0,999 | |
| | Razem | 111 | × | 0,999 | × | |

Kolumny dla liczebności bezwzględnych i częstości względnych są opcjonalne (wyniki są identyczne).

Znajdowanie wartości $D_1$, $Q_1$, $Q_2$, $C_{65}$:

ad $D_1$. Najpierw lokalizuje się pozycję zajmowaną przez pierwszy decyl $(i_{D_1} \cdot N) = 0{,}1 \cdot 111 = 11{,}1$. W kolumnie kumulat znajdujemy numer wiersza, w którym znajdują się elementy jedenasty i dwunasty, a dokładnie jedenasty i jedna dziesiąta różnicy pomiędzy elementami dwunastym i jedenastym. Jest to wiersz czwarty o granicach 70–80. W tym momencie znamy wszystkie wartości, które trzeba podstawić do wzoru.

$$D_1 = 70 + \frac{11{,}1 - 8}{7} \cdot 10 = 70 + 4{,}4 = 74{,}4 \quad \text{lub dla części względnych}$$

$$D_1 = 70 + \frac{0{,}1 - 0{,}072}{0{,}063} \cdot 10 = 74{,}4$$

ad $Q_1$. $Q_1 = 90 + \frac{27{,}75 - 24}{4} \cdot 10 = 99{,}4$ lub $Q_1 = 90 + \frac{0{,}25 - 0{,}216}{0{,}036} \cdot 10 = 99{,}4$

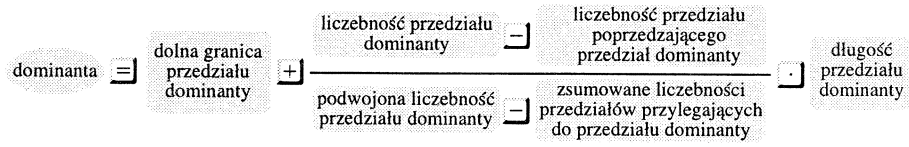
ad $Q_2$. $Q_2 = 110 + \frac{55{,}5 - 45}{14} \cdot 10 = 117{,}5$

ad $C_{65}$. $C_{65} = 120 + \frac{72{,}15 - 59}{21} \cdot 10 = 126{,}3$

## 4.3. Dominanta (moda, wartość najczęstsza)

**Dla indywidualnego wykazu wartości** — wartość najczęściej spotykana. W przykładzie dotyczącym kolokwium najczęściej występującym wynikiem jest 130 punktów. Tyle zatem wynosi dominanta.

W populacji może być kilka dominant, gdy różne wartości występują tyle samo razy (mimo to, populacja może być względnie jednorodna!); zdarza się też, że dominanta nie istnieje, zachodzi to wtedy, gdy żadna z wartości się nie powtarza. Przypadek kilku dominant obserwujemy dla przykładów o firmie Profi. Wartościami najczęstszymi są: 21, 22 i 23. Zatem $D^{(1)}=21$, $D^{(2)}=22$, $D^{(3)}=23$.




**Dla szeregu rozdzielczego**, aby określić wartość dominanty stosuje się wzór:

w skrócie: $D = X_{OD} + \dfrac{n_D - n_{D-1}}{2 \cdot n_D - (n_{D-1} + n_{D+1})} \cdot h_D$

| $x_{0i} - x_{1i}$ | $n_i$ | |
|---|---|---|
| 5–10 | 7 | |
| 10–15 | 10 | |
| 15–20 | 13 | |
| 20–25 | 18 | ← $D$ |
| 25–30 | 8 | |
| 30–35 | 5 | |
| 35–40 | 3 | |
| Razem | 64 | |

***Przykład 4.13.*** Firma Profi

$$D = 20 + \frac{18-13}{36-(13+8)} \cdot 5 = 20 + \frac{5 \cdot 5}{15} = 20 + 1{,}6(6) \cong 21{,}7$$

Najczęściej występującą dzienną wielkością sprzedaży było 21,7 palet.

Dominanta dla szeregu przedziałowego o wynikach kolokwium:

$$D = 130 + \frac{25-21}{50-(21+6)} \cdot 10 \cong 131{,}7$$

Dla szeregów rozdzielczych o nierównych rozpiętościach przedziałów w uzasadnionych przypadkach, tj. wtedy, gdy uważamy, że potrafimy zlokalizować przedział dominanty, do obliczeń wykorzystujemy gęstości (natężenia liczebności).

***Przykład 4.14.***

| Odchylenie od normy w mm | Liczba wyrobów | Gęstości | |
|---|---|---|---|
| do 1 | 8 | 8 | |
| 1–3 | 20 | 10 | |
| 3–5 | 50 | 25 | |
| 5–7 | 200 | 100 | ← $D$ |
| 7–11 | 180 | 45 | |
| 11–15 | 90 | 22,5 | |
| 15 i więcej | 30 | – | |

$$D = 5 + \frac{100 - 25}{200 - (25 + 45)} \cdot 2 \cong 6{,}2$$

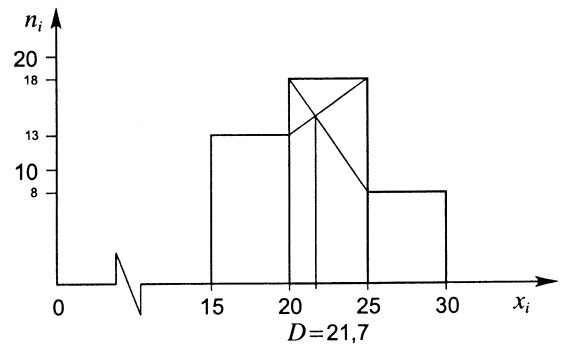
### *Graficzne wyznaczanie dominanty*

Wykreślamy część histogramu ograniczoną do klasy najliczniejszej i jej dwóch najbliższych sąsiadów. Następnie łączymy odcinkami punkty o współrzędnych:

— górna granica klasy poprzedzającej najliczniejszą i częstość w tej klasie z górną granicą przedziału dominanty i częstością w tej klasie,

— dolna granica klasy dominanty i częstość w tej klasie z dolną granicą przedziału następującego po dominancie i częstością w tej klasie.

Z punktu przecięcia prowadzimy prostą prostopadłą do podstawy i odczytujemy interpolowaną wartość dominanty.

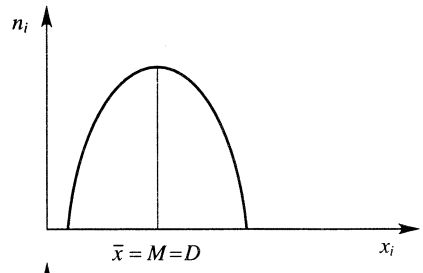



Wykres 4.3. Graficzne wyznaczanie dominanty (przykład dla firmy Profi)

# 4.4. Zależności $\bar{x}$, $M$, $D$

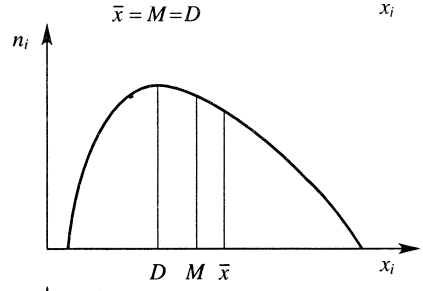
a. dla rozkładów symetrycznych

$$\bar{x} = M = D$$



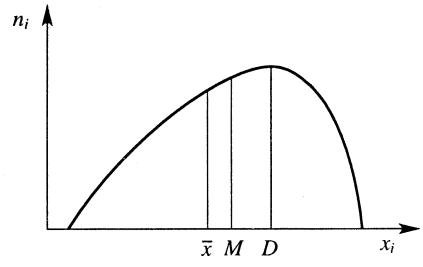

b. dla rozkładów prawostronnie asymetrycznych

$$D < M < \bar{x}$$



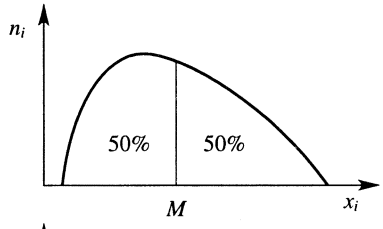

c. dla rozkładów lewostronnie asymetrycznych

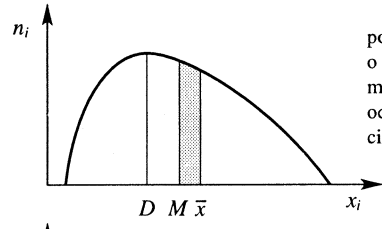
$$\bar{x} < M < D$$



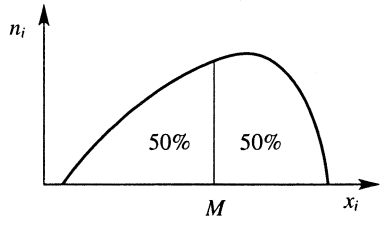

***Uwaga:*** nie zawsze mediana leży wewnątrz przedziału $\bar{x} - D$!

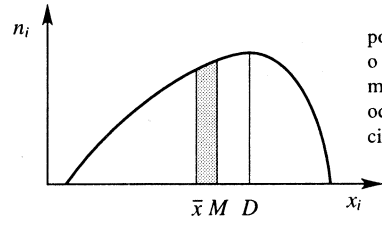
Ponieważ mediana dzieli pole pod krzywą liczebności na połowy to:



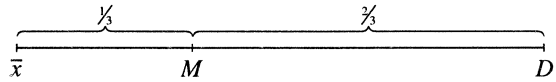

ponad 50% populacji o asymetrii dodatniej ma wartości mniejsze od średniej (o pole zaciemnione)




ponad 50% populacji o asymetrii ujemnej ma wartości większe od średniej (o pole zaciemnione)

Dla rozkładów umiarkowanie asymetrycznych mediana leży w około $^1/_3$ odległości między średnią a dominantą, licząc od średniej.




Zależność tę wyraża wzór Pearsona, pozwalając wyznaczyć nieznaną z triady, jeśli znamy wartość dwójki pozostałych.

$$D \cong 3M - 2\bar{x}$$

Zalety i wady stosowania $\bar{x}$, $M$, $D$

$\bar{x}$ — jest stosowana do celów ogólnych, łatwo ją obliczyć, niemal zawsze daje się wyznaczyć, szczególnie łatwe są jej algebraiczne przekształcenia, błędy losowe wywierają na nią mniejszy wpływ niż na medianę,

$M$ — łatwiejsza do obliczenia przy danym rozkładzie liczebności niż średnia, z reguły bardziej lub nie mniej stabilna niż średnia, może stwarzać kłopoty z interpretacją dla zmiennych losowych skokowych (że dzieli pole pod krzywą liczebności na połowy) w szczególności w sytuacji, gdy wiele elementów zbiorowości ma wartości równe medianie, algebraiczne przekształcenia trudne lub niemożliwe,

$D$ — mało przydatna w zastosowaniach elementarnych w związku z trudnością jej obliczania, niemożliwe algebraiczne przekształcenia, często jednak stanowi cenną charakterystykę zmiennej.

## 4.5. Średnia harmoniczna

Jest w swej treści poznawczej identyczna ze średnią arytmetyczną i jest stosowana wtedy, gdy wartości zmiennej podane są w formie odwrotności. Formuła średniej harmonicznej dla indywidualnego wykazu wartości $H = \dfrac{N}{\sum \dfrac{1}{x_i}}$ wyraźnie pokazuje, że jest to odwrotność średniej arytmetycznej z odwrotności wartości zmiennych. Wzór stosowany dla szeregu przedziałowego: $H = \dfrac{\sum n_i}{\sum \dfrac{n_i}{\mathring{x}_i}}$.

Średnią harmoniczną stosuje się wówczas, kiedy miana wartości zmiennych podane są w jednostkach względnych („łamanych"), np. km/godz., kg/osobę, sztuka/prac., wagi zaś — w jednostkach liczników tych jednostek względnych. Przykładowo, zmienne:

- prędkość pojazdu (miano w km/godz.), waga — liczba przejechanych kilometrów,
- gęstość zaludnienia (miano mieszkańcy/km$^2$), waga — liczba mieszkańców,
- spożycie artykułu „$x$" na osobę (miano w kg/osobę), waga — ilość kg.

Relację między średnią arytmetyczną a harmoniczną obrazuje przykład 4.15.

***Przykład 4.15.***

Należy obliczyć średnią wydajność pracy.

| Wydajność pracy w szt./rob. $x_{0i} - x_{1i}$ | Liczba robotników $n_i$ |
|---|---|
| 8–12 | 17 |
| 12–16 | 42 |
| 16–20 | 25 |
| 20–24 | 16 |
| 24–28 | 7 |
| Razem | 107 |

Jest to klasyczna sytuacja, gdy liczymy $\bar{x}$. Obliczamy więc kolumnę →

| Liczba sztuk $\mathring{x}_i n_i$ |
|---|
| 170 |
| 588 |
| 450 |
| 352 |
| 182 |
| 1742 |

$$\rightarrow \bar{x} = \frac{\sum \mathring{x}_i n_i}{\sum n_i} = \frac{1472}{107} \cong 16{,}28$$

Natomiast, gdy będziemy dysponowali danymi w poniższym układzie:

| Wydajność pracy w szt./rob. $x_{0i} - x_{1i}$ | Liczba sztuk $\mathring{x}_i n_i$ |
|---|---|
| 8–12 | 170 |
| 12–16 | 588 |
| 16–20 | 450 |
| 20–24 | 352 |
| 24–28 | 182 |
| Razem | 1742 |

Zastosujemy wzór na średnią harmoniczną, gdzie „odtwarzamy" dane dotyczące liczb w poszczególnych klasach wydajności

↓

$$H = \frac{n}{\sum \frac{n_i}{\mathring{x}_i}} = \frac{1472}{\frac{170}{10} + \frac{588}{14} + \frac{352}{22} + \frac{182}{26}} = \frac{1472}{107} \left[ \frac{\text{sztuki}}{\frac{\text{szt}}{\text{rob}}} \right]$$

***Przykład 4.16.***

Klasyczny przykład operowania średnią harmoniczną dotyczy znajdowania przeciętnej prędkości. Ktoś pokonał 30 km z prędkością 90 km/godz.; 50 z prędkością 100 km/godz. i 20 km z prędkością 50 km/godz.

$$H = \frac{100}{\frac{30}{90} + \frac{50}{100} + \frac{20}{50}} = \frac{100}{\frac{1}{3} + \frac{1}{2} + \frac{2}{5}} = \frac{100}{\frac{37}{30}} = \frac{100 \cdot 30}{37} \cong 81{,}081 \left[ \frac{\text{km}}{\frac{\text{km} \cdot \text{godz}}{\text{km}}} \right]$$

## 4.6. Średnia geometryczna

Średnia geometryczna nieważona

$$G = \sqrt[n]{x_1 \cdot x_1 \cdot \ldots \cdot x_n}$$

Średnia geometryczna ważona (liczebnościami)

$$G = \sqrt[\Sigma n_i]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_n^{n_n}}$$

Znajduje zastosowanie w:

1. badaniach szeregów czasowych przy określaniu przeciętnego tempa zmiany danego zjawiska,
2. opisie tendencji rozwojowej badanego zjawiska przy zastosowaniu wykładniczej funkcji trendu.

Średniej geometrycznej używa się dla tzw. zmiennych multiplikatywnych, tj. takich, które ze względu na swoją naturę mnoży się przez siebie, aby osiągnąć wartość wynikową, w odróżnieniu od zmiennych addytywnych, które się sumuje (przypadek średniej arytmetycznej i harmonicznej).

Przyjmijmy, że początkowa wielkość pewnego zjawiska wynosi $L_0$. Jeśli zjawisko to zmienia się z okresu na okres o stałą wielkość względną $p$, to jego wartość będzie wynosiła odpowiednio w chwili:

$$L_1 = L_0 + L_0 \cdot p = L_0(1+p)$$

$$L_2 = L_1 + L_1 \cdot p = L_0 + L_0 \cdot p + L_0 \cdot p + L_0 \cdot p^2 = L_0(1+2p+p^2) = L_0(1+p)^2$$

$$L_3 = L_2 + L_2 \cdot p = L_0(1+p)^2 + L_0(1+p)^2 \cdot p = L_0(1+p)^2(1+p) = L_0(1+p)^3$$

$$\vdots$$

$$L_n = L_0(1+p)^n$$

$$(1+p)^n = \frac{L_n}{L_0}$$

$$1+p = \sqrt[n]{\frac{L_n}{L_0}} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \sqrt[n]{(1+p_1)(1+p_2)\cdot \ldots \cdot (1+p_n)} = \sqrt[n]{(1+p)^n}$$

$$p = \sqrt[n]{\frac{L_n}{L_0}} - 1$$

oznaczenia:

| | |
|---|---|
| $p$ | — ułamek dziesiętny, |
| $s = p \cdot 100\%$, | — średnie tempo zmiany wyrażone w %, |
| $x = 1 + p$ | — średni indeks łańcuchowy, |
| $x_i = 1 + p_i$ | — indeks łańcuchowy (iloraz fizycznej wielkości zjawiska w danym roku przez wielkość w roku poprzednim), |
| $L_0$ | — fizyczna wielkość zjawiska w momencie początkowym, |
| $L_n$ | — fizyczna wielkość zjawiska w momencie końcowym. |

***Przykład 4.17.***

| Lata | Obroty firmy w mln zł | Indeksy łańcuchowe $x_i = (1 + p_i)$ | Wielkości obrotów przy założeniu stopy $s$ |
|---|---|---|---|
| 1990 | 2000 | . | 2000 |
| 1991 | 2500 | $1{,}25 = \frac{2500}{2000}$ | $2000 \cdot 1{,}5137 = 3027{,}4$ |
| 1992 | 5000 | 2,00 ⋮ | $3027{,}4 \cdot 1{,}5137 = 4582{,}6$ |
| 1993 | 7500 | $1{,}50 = \frac{7500}{5000}$ | $4582{,}6 \cdot 1{,}5137 = 6936{,}6$ |
| 1994 | 10500 | 1,40 ⋮ | $6936{,}6 \cdot 1{,}5137 = 10499{,}9$ |

Aby obliczyć średnie tempo wzrostu obrotów obliczamy pierwiastek stopnia równego liczbie indeksów wyrażenia podpierwiastkowego z ilorazu wartości końcowej przez początkową i od otrzymanego wyniku odejmujemy jednostkę (otrzymując $p$), a następnie mnożymy go przez 100%, otrzymując $s$. Zatem:

$$p = \sqrt[4]{\frac{10500}{2000}} - 1 = \sqrt[4]{5{,}25} - 1 = 1{,}5137 - 1 = 0{,}5137$$

$s = 51{,}37\%$ (obroty rosły po 51,37% rocznie).

Zwróćmy uwagę, że:

$$1 + p = \underbrace{\sqrt[4]{\frac{10500}{2000}}}_{\text{cz‡on a}} = \underbrace{\sqrt[4]{\frac{2500}{2000} \cdot \frac{5000}{2500} \cdot \frac{7500}{5000} \cdot \frac{10500}{7500}}}_{\text{cz‡on b}} = \underbrace{\sqrt[4]{1{,}25 \cdot 2{,}00 \cdot 1{,}50 \cdot 1{,}40}}_{\text{cz‡on c}}$$

Oczywiście, wyznaczanie wartości członów $b$ i $c$ jest zbędne, wyrażenia zostały zapisane głównie po to, aby na przykładzie pokazać multiplikatywność indeksów łańcuchowych.

***Przykład 4.18.***

Ludność Polski (stan w dniu 31.12) w mln wynosiła:

| 1970 | 1980 | 1985 | 1990 | 1994 |
|---|---|---|---|---|
| 32,7 | 35,7 | 37,3 | 38,2 | 38,6 |

1. Proszę obliczyć tempo zmiany dla lat 1970–1994,
2. ekstrapolować liczbę ludności Polski w 1998 r.,
3. w którym roku ludność Polski osiągnie 40 mln, zakładając utrzymanie tendencji z lat 1990–1994?

ad 1. $$p_{94/70} = \sqrt[24]{\frac{38,6}{32,7}} - 1 = \sqrt[24]{1,1804281} - 1 = 1,0069355 - 1 \cong 0,0069$$

$$s = p \cdot 100\% = 0,0069 \cdot 100\% = 0,69\% \cong 0,7\%$$

W okresie 1970–1994 liczba ludności Polski przyrastała w tempie 0,7% rocznie.

ad 2. Tempo przyrostu liczby ludności znacznie spadło w ostatnich latach, nierealistyczne byłoby zatem posługiwanie się tempem dla lat 1970–1994.

$$p = \sqrt[4]{\frac{38,6}{38,2}} - 1 = 1,0026 - 1 = 0,0026$$

Zakładając utrzymanie się niskiego tempa wzrostu z lat 1990–1994:

$$L_{1998} = L_{1994}\,(1+p)^4 = 38,6 \cdot 1,0026^4 = 39,0.$$

Szacowana liczebność populacji w 1998 roku wynosi 39 mln.

ad 3.

$$40 = L_{1994}(1+p)^x$$

$$40 = 38,6 \cdot 1,0026^x$$

$$\frac{40}{38,6} = 1,0026^x$$

$$1,0362694 = 1,0026^x$$

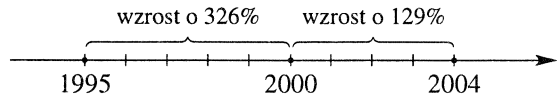
$$\ln 1,0362694 = x \ln 1,0026$$

$$x = \frac{\ln 1{,}0362694}{\ln 1{,}0026} = \frac{0{,}0356272}{0{,}0025966} = 13{,}72$$

Przy tym założeniu nastąpi to po niespełna czternastu latach, a więc w roku 2008.

***Przykład 4.19.***

W firmie Niebieski Ocean, w latach 1995–2000, obroty wartościowo wzrosły o 326%, w ciągu następnych czterech lat obroty wzrosły o 129%. O ile średniorocznie rosły obroty tej firmy pomiędzy 1995 i 2004 rokiem?




$$\frac{L_{2000}}{L_{1995}} = 1 + p_{2000/1995} = 1 + 3{,}26 = 4{,}26 = (1 + p_{śr})^5$$

$$\frac{L_{2004}}{L_{2000}} = 1 + p_{2004/2000} = 1 + 1{,}29 = 2{,}29 = (1 + p_{śr})^4$$

$$p_{śr} = \sqrt[9]{\frac{L_{2004}}{L_{1995}}} - 1 = \sqrt[9]{\frac{L_{2000}}{L_{1995}} \cdot \frac{L_{2004}}{L_{2000}}} - 1 =$$

$$= \sqrt[9]{4{,}26 \cdot 2{,}29} - 1 = \sqrt[9]{9{,}7554} - 1 = 1{,}288 - 1 = 0{,}288$$

$$s = 0{,}288 \cdot 100\% = 28{,}8\%.$$

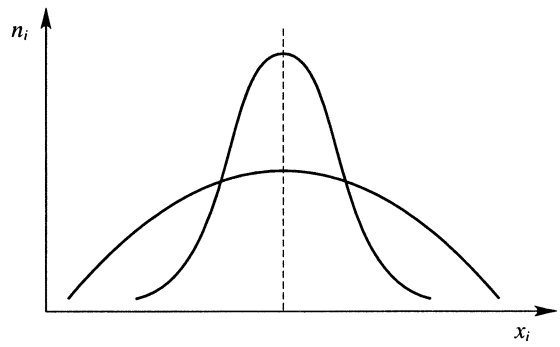
Średnio w ciągu roku obroty firmy rosły po 28,8%.

# Rozdział 5

# Miary zróżnicowania (zmienności, rozrzutu, dyspersji)

## 5.1. Geneza miar

Potrzebę operowania innymi, poza miarami tendencji centralnej, charakterystykami wyraża wykres:




Wykres 5.1. Rozkłady o różnej zmienności

Obie populacje (o wspólnej osi symetrii) mają identyczne średnie arytmetyczne, medianty i dominanty. Mimo to rozkłady nie są identyczne.

Rysunek nasuwa koncepcję najprostszego rozwiązania: im zbiorowość jest bardziej skupiona, smuklejsza, tym jednostki są bardziej do siebie podobne, mniej się między sobą różnią. Im populacja bardziej spłaszczona, jednostki rozrzucone są na większym obszarze zmienności, zróżnicowanie wzrasta. Miarą może więc być rozstęp:

$$R = x_{\max} - x_{\min}.$$

Rozstęp wyraża długość przedziału, w jakim realizuje się zmienna. Taki sam rozstęp mają np. populacje:

| | | | |
|---|---|---|---|
| populacja $A$ | $x_{max}=52$ | $x_{min}=2$ | $R_A=52-2=50$ |
| populacja $B$ | $x_{max}=10050$ | $x_{min}=10000$ | $R_B=10050-10000=50$ |

Zaletami rozstępu jako miary są prostota koncepcji i łatwość obliczenia. Wadą: podatność na wpływ wielkości skrajnych.

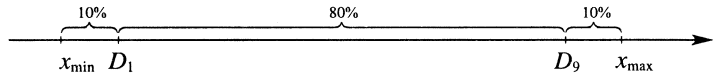
***Przykład 5.1.***

Dane są dwie populacje. Różnią się jedynie wartościami ostatniego elementu:

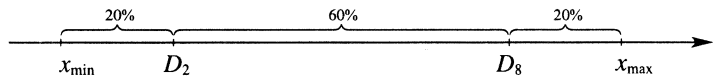
| | | |
|---|---|---|
| populacja $C$ | 52, 52, ..., 86, 91 | $R=91-52=39$ |
| populacja $D$ | 52, 52, ..., 86, 244 | $R=244-52=192$ |

Znając jedynie wartości rozstępów możemy upierać się, że populacja $D$ mocno różni się od populacji $C$, co w istocie nie ma miejsca. Próbą zaradzenia temu są następujące propozycje:

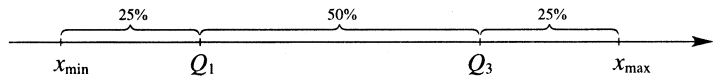
$R_D=D_9-D_1$, czyli rozstęp decylowy




$R_{D'}=D_8-D_2$, również rozstęp decylowy




$R_Q=Q_3-Q_1$, rozstęp międzykwartylowy




i połowa $R_Q$, odchylenie ćwiartkowe

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

W efekcie osiągamy charakterystyki, które nie mogą pretendować do oceny zmienności wszystkich jednostek analizowanej zbiorowości. Odchylenie ćwiartkowe, jako połowa długości przedziału pozostałego po odrzuceniu pierwszej i ostatniej ćwiartki zbiorowości, jest wykorzystywane w analizach jako miara zróżnicowania wewnętrznych 50% elementów. $Q$ interpretuje się jako średnią rozpiętość cechy

w dwóch środkowych ćwiartkach zbiorowości. Założenia, wiodące ku miarom klasycznym są następujące:

1. przyjąć jakiś punkt odniesienia,
2. obliczyć jak odchylają się poszczególne jednostki populacji od punktu odniesienia,
3. znaleźć średnią arytmetyczną tych odchyleń.

ad 1. niech punktem odniesienia będzie średnia arytmetyczna, więc:

$$\frac{\sum(x_i - \bar{x})n_i}{N}$$

niestety:

$$\frac{\sum(x_i - \bar{x})n_i}{N} = \frac{\sum(x_i n_i - \bar{x}n_i)}{N} = \frac{\sum x_i n_i}{N} - \bar{x}\frac{\sum n_i}{N} = \bar{x} - \bar{x} = 0$$

Inne możliwości

a. policzyć średnią arytmetyczną modułów odchyleń wartości cechy od średniej arytmetycznej:

$$d = \frac{\sum|x_i - \bar{x}|n_i}{N}$$

Miarę tę nazywa się odchyleniem przeciętnym.

b. policzyć średnią arytmetyczną modułów odchyleń wartości cechy od mediany (własność (f) ze str. 66)

$$d_M = \frac{\sum|x_i - M|n_i}{n}$$

Miary bazujące na modułach nie sprawdziły się w praktyce obliczeniowej. Ich użycie wiązało się z wieloma błędami, w szczególności łatwo był mylony nawias z symbolem modułu. Ponadto wyrażenia tej postaci nie są podatne na przekształcenia algebraiczne. Przedstawionych wad nie posiada średnia arytmetyczna kwadratów odchyleń wartości cechy od średniej arytmetycznej, nazywana wariancją:

$$Var = \frac{\sum(x_i - \bar{x})^2 n_i}{N}$$

Jej wadą są miana w kwadratach, a więc przykładowo: osoby$^2$, złote$^2$ itd. Niedogodność tę obchodzi się wyciągając pierwiastek kwadratowy z wariancji. Powstałą miarę nazywa się odchyleniem standardowym i symbolizuje się przez $\sigma$.

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 n_i}{N}}$$

Nie do końca poprawnie, ale zgodnie z ideą formuły liczenia odchylenia standardowego, wygodnie jest interpretować $\sigma$ jako określające, o ile wszystkie jednostki danej zbiorowości różnią się średnio, ze względu na wartość badanej zmiennej, od średniej arytmetycznej tej zmiennej. Średnia arytmetyczna i odchylenie standardowe to dwie najważniejsze miary w statystyce.

Miary zróżnicowania (i dalej miary asymetrii) wygodnie jest zestawić w tabeli o następującej budowie:

| | miary absolutne ↓ | miary stosunkowe ↓ |
|---|---|---|
| miary klasyczne → | | |
| miary pozycyjne → | | |

Określenia miar klasycznych i pozycyjnych zostały wprowadzone w rozdziale 4.

Miary absolutne to miary mianowane, najczęściej w tych samych mianach, co analizowana cecha (w tabeli miar zróżnicowania wyjątkiem jest wariancja).

Miary stosunkowe to miary bezmianowe, czasem zwyczajowo wyrażane w procentach. Jako ilorazy stosownych wielkości służą do określenia siły.

**Tablica 5.1. Miary zróżnicowania**

| Miary | Absolutne | Stosunkowe |
|---|---|---|
| Klasyczne | $d = \frac{\sum \lvert x_i - \bar{x} \rvert n_i}{N}$<br>odchylenie przeciętne<br>$d_{Q_2} = \frac{\sum \lvert x_i - Q_2 \rvert n_i}{N}$<br>odchylenie medialne<br>$Var = \frac{\sum (x_i - \bar{x})^2}{N} = \overline{x^2} - (\bar{x})^2$<br>wariancja<br>$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 n_i}{N}}$<br>odchylenie standardowe | $V_{zk} = \frac{\sigma}{\bar{x}} \cdot 100\%$<br>klasyczny współczynnik zmienności |
| Pozycyjne* | $R = x_{max} - x_{min}$<br>rozstęp<br>$R_D = D_9 - D_1$<br>rozstęp decylowy<br>$R_Q = Q_3 - Q_1$<br>rozstęp międzykwartylowy<br>$Q = \frac{Q_3 - Q_1}{2}$<br>odchylenie ćwiartkowe (połówkowy rozstęp wewnątrzkwartylowy) | $V_{zQ} = \frac{Q}{M} \cdot 100\%$<br>pozycyjny współczynnik zmienności<br>$V_{zD} = \frac{D_9}{D_1}$<br>wskaźnik zróżnicowania decylowego<br>$V_R = \frac{R}{\bar{X}}$<br>względna wartość obszaru zmienności<br>$V_Q = \frac{Q_3 - Q_1}{Q_3 + Q_1}$<br>współczynnik zmienności kwartylowej |

* Im zbiorowość mniej jednorodna tym większa wartość poznawcza tych miar.

Do miar praktycznie wykorzystywanych należą: odchylenie standardowe, klasyczny współczynnik zmienności, odchylenie ćwiartkowe, pozycyjny współczynnik zmienności (stosunkowe odchylenie ćwiartkowe) oraz rozstęp.

Bardzo przybliżona ocena związku tych miar w populacjach symetrycznych i umiarkowanie asymetrycznych:

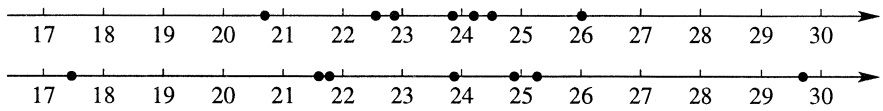
$$R \cong 6\sigma \cong 7{,}5d \cong 9Q$$

## 5.2. Sposoby obliczania odchylenia standardowego

***Przykład 5.2.***

Należy porównać płace dyrektorów regionalnych w dwóch organizacjach:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| płace w org. 1 | 20700 | 22500 | 22900 | 23800 | 24100 | 24500 | 26000 |
| płace w org. 2 | 17500 | 21600 | 21700 | 23800 | 24900 | 25300 | 29700 |




Płace zaznaczone na osi liczbowej

Proste obliczenia wykazują, że średnie arytmetyczne płac w obu przypadkach wynoszą po 23 500; identyczne są mediany, po 23 800; dominant brak. Z wykresów można stwierdzić, że płace w organizacji 1. są mniej zróżnicowane od płac w organizacji 2.

***Obliczanie odchylenia standardowego dla danych indywidualnych***

$$\text{wzór: } \sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$$

Odchylenie standardowe dla populacji 1:

$$\sigma^{(1)} = \sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}} =$$

$$= \sqrt{\frac{(20700-23500)^2+(22500-23500)^2+(22900-23500)^2+(23800-23500)^2+}{7}}$$

$$\frac{+(24100-23500)^2+(24500-23500)^2+(26000-23500)^2}{7} =$$

$$= \sqrt{\frac{(-2800)^2+(-1000)^2+(-600)^2+(300)^2+(600)^2+(1000)^2+(2500)^2}{7}} =$$

$$= \sqrt{\frac{7840000+1000000+360000+90000+360000+1000000+6250000}{7}} =$$

$$= \sqrt{\frac{16900000}{7}} = \sqrt{2414285,7} = 1553,7971 \approx 1553,8$$

Znajdowanie miar dla danych indywidualnych, dosyć uciążliwe w analizie tendencji centralnej, dla odchylenia standardowego staje się wyjątkowo żmudne. Ułatwieniem będzie przejście na szereg punktowy.

***Obliczanie odchylenia standardowego dla szeregu punktowego***

$$\text{wzór: } \sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 n_i}{N}}$$

**Tablica 5.2. Odchylenie standardowe dla populacji 2.**

| $x_i$ | $n_i$ | $x_i n_i$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ | $(x_i - \bar{x})^2 n_i$ |
|---|---|---|---|---|---|
| 17500 | 1 | 17500 | –6000 | 36000000 | 36000000 |
| 21600 | 1 | 21600 | –1900 | 3610000 | 3610000 |
| 21700 | 1 | 21700 | –1800 | 3240000 | 3240000 |
| 23800 | 1 | 23800 | 300 | 90000 | 90000 |
| 24900 | 1 | 24900 | 1400 | 1960000 | 1960000 |
| 25300 | 1 | 25300 | 1800 | 3240000 | 3240000 |
| 29700 | 1 | 29700 | 6200 | 38440000 | 38440000 |
| Razem | 7 | 164500 | × | 86580000 | 86580000 |

$$\sigma^{(2)} = \sqrt{\frac{\sum (x_i - \bar{x})^2 n_i}{N}} = \sqrt{\frac{86580000}{7}} = \sqrt{12368571} = 3516{,}8979 \cong 3516{,}9$$

***Obliczanie odchylenia standardowego dla szeregu przedziałowego***

$$\text{wzór: } \sigma = \sqrt{\frac{\sum (\mathring{x}_i - \bar{x})^2 n_i}{N}}$$

**Tablica 5.3. Odchylenie standardowe wielkości sprzedaży w firmie Profi**

| $x_{0i} - x_{1i}$ | $n_i$ | $\mathring{x}_i$ | $\mathring{x}_i n_i$ | $\mathring{x}_i - \bar{x}$ | $(\mathring{x}_i - \bar{x})^2$ | $(\mathring{x}_i - \bar{x})^2 n_i$ |
|---|---|---|---|---|---|---|
| 5–10 | 7 | 7,5 | 52,5 | –12,9 | 166,41 | 1164,87 |
| 10–15 | 10 | 12,5 | 125,0 | –7,9 | 62,41 | 624,10 |
| 15–20 | 13 | 17,5 | 227,5 | –2,9 | 8,41 | 109,33 |
| 20–25 | 18 | 22,5 | 405,0 | 2,1 | 4,41 | 79,38 |
| 25–30 | 8 | 27,5 | 220,0 | 7,1 | 50,41 | 403,28 |
| 30–35 | 5 | 32,5 | 162,5 | 12,1 | 146,41 | 732,05 |
| 35–40 | 3 | 37,5 | 112,5 | 17,1 | 292,41 | 877,23 |
| Razem | 64 | × | 1305,0 | × | × | 3990,24 |

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 n_i}{N}} = \sqrt{\frac{3990{,}24}{64}} = \sqrt{62{,}3475} \cong 7{,}9 \text{ palety}$$

***Alternatywny sposób obliczania wariancji i odchylenia standardowego***

Ponieważ:

$$\frac{\sum (\mathring{x}_i - \bar{x})^2 n_i}{N} = \frac{\sum (\mathring{x}_i^2 - 2\mathring{x}_i \bar{x} + \bar{x}^2) n_i}{N} = \frac{\sum \mathring{x}_i^2 n_i - \sum 2\mathring{x}_i \bar{x}\, n_i + \sum \bar{x}^2 n_i}{N} =$$

$$= \frac{\sum \mathring{x}_i^2 n_i}{N} - \frac{2\bar{x}\sum \mathring{x}_i n_i}{N} + \frac{\bar{x}^2 \sum n_i}{N} = \overline{x^2} - 2 \cdot \bar{x} \cdot \bar{x} + \bar{x}^2 =$$

$$= \overline{x^2} - 2\bar{x}^2 + \bar{x}^2 = \overline{x^2} - \bar{x}^2$$

Zatem wariancja, czyli średnia arytmetyczna kwadratów odchyleń wartości cechy od ich średniej arytmetycznej jest równa średniej kwadratów minus kwadrat średniej arytmetycznej.

**Tablica 5.4. Wykorzystanie momentów zwykłych do liczenia wariancji**

| $x_{0i} - x_{1i}$ | $n_i$ | $\mathring{x}_i$ | $\mathring{x}_i n_i$ | $\mathring{x}_i^2 n_i$ |
|---|---|---|---|---|
| 5–10 | 7 | 7,5 | 52,5 | 393,75 |
| 10–15 | 10 | 12,5 | 125,0 | 1562,50 |
| 15–20 | 13 | 17,5 | 227,5 | 3981,25 |
| 20–25 | 18 | 22,5 | 405,0 | 9181,25 |
| 25–30 | 8 | 27,5 | 220,0 | 9112,50 |
| 30–35 | 5 | 32,5 | 162,5 | 6050,00 |
| 35–40 | 3 | 37,5 | 112,5 | 5281,25 |
| Razem | 64 | × | 1305,0 | 30600,00 |

Średnia kwadratów $\overline{x^2} = \frac{\sum \mathring{x}_i^2 n_i}{N} = \frac{30600}{64} = 478{,}125 \text{ palet}^2$

Kwadrat średniej arytmetycznej $\left(\frac{\sum \mathring{x}_i n_i}{N}\right)^2 = \left(\frac{1305}{64}\right)^2 = 20{,}39^2 = 415{,}752 \text{ palet}^2$

Wariancja obliczona drugim sposobem:

$$Var = \overline{x^2} - \bar{x}^2 = 478{,}125 - 415{,}752 = 62{,}373 \text{ palety}^2$$

$$\sigma = \sqrt{62{,}373 \text{ palety}^2} = 7{,}9 \text{ palety}$$

## 5.3. Momenty statystyczne

Dwie drogi prowadzące do wyznaczania wariancji:

$$\sqrt{\frac{\sum (x_i - \bar{x})^2 n_i}{N}} = \frac{\sum x_i^2 n_i}{N} - \left(\frac{\sum x_i n_i}{N}\right)^2 = \overline{x^2} - \bar{x}^2$$

dające się słownie wyrazić, że wariancja, czyli średnia arytmetyczna kwadratów odchyleń wartości cechy od średniej arytmetycznej jest równa średniej arytmetycznej kwadratów wartości cechy minus kwadrat średniej arytmetycznej, znajdują odniesienie przy wyliczaniu innych miar klasycznych. Średnie arytmetyczne *k*-tych potęg odchyleń wartości cechy od średniej arytmetycznej nazywa się momentami centralnymi i symbolizuje $e_k$:

$$e_k = \frac{\sum (x_i - \bar{x})^k n_i}{N}$$

Średnie arytmetyczne *k*-tych potęg wartości cechy nazywa się momentami zwykłymi i symbolizuje $m_k$:

$$m_k = \frac{\sum x_i^k n_i}{N}$$

Pierwszy moment zwykły to średnia arytmetyczna, pierwszy moment centralny jako średnia arytmetyczna odchyleń wartości cechy od średniej arytmetycznej jest z definicji równy zero, drugi moment centralny to wariancja, trzeci moment centralny jest miarą asymetrii, czwarty moment centralny jest miarą spłaszczenia.

Zależności między momentami są następujące:

$$e_2 = m_2 - m_1^2$$
$$e_3 = m_3 - 3m_1m_2 + 2m_1^3$$
$$e_4 = m_4 - 4m_1m_3 + 6m_1^2m_2 - 3m_1^4$$

**Tablica 5.5. Momenty statystyczne**

| Rząd momentów $k$ | Moment zwykły $m_k$ | Moment centralny $e_k$ |
|---|---|---|
| 1 | $m_1 = \frac{\sum x_i n_i}{N} = \bar{x}$ | $e_1 = \frac{\sum (x_i - \bar{x}) n_i}{N} = 0$ |
| 2 | $m_2 = \frac{\sum x_i^2 n_i}{N}$ | $e_2 = \frac{\sum (x_i - \bar{x})^2 n_i}{N} = Var$ |
| 3 | $m_3 = \frac{\sum x_i^3 n_i}{N}$ | $e_3 = \frac{\sum (x_i - \bar{x})^3 n_i}{N}$ |
| ⋮ | ⋮ | ⋮ |
| $k$ | $m_k = \frac{\sum x_i^k n_i}{N}$ | $e_k = \frac{\sum (x_i - \bar{x})^k n_i}{N}$ |

## 5.4. Ocena siły zróżnicowania

Jest pożądane, aby opisując populacje móc wypowiadać się na temat siły zróżnicowania. Brak jest podstaw do ustalenia takiego kryterium w sposób obiektywny. Na bazie doświadczeń praktycznych można jednak dojść do pewnych ustaleń.

Za punkt wyjścia weźmy dziewięć hipotetycznych populacji, przedstawionych w poniższej tablicy.

**Tablica 5.6. Liczebności analizowanych populacji**

| $x_{0i} - x_{1i}$ | $n_i^1$ | $n_i^2$ | $n_i^3$ | $n_i^4$ | $n_i^5$ | $n_i^6$ | $n_i^7$ | $n_i^8$ | $n_i^9$ |
|---|---|---|---|---|---|---|---|---|---|
| 300–700 | 1 | 5 | 5 | 5 | 6 | 8 | 13 | 14 | 8 |
| 700–1100 | 1 | 10 | 10 | 11 | 10 | 10 | 14 | 14 | 20 |
| 1100–1500 | 1 | 15 | 16 | 15 | 15 | 15 | 15 | 14 | 25 |
| 1500–1900 | 94 | 40 | 39 | 39 | 39 | 37 | 16 | 16 | 29 |
| 1900–2300 | 1 | 15 | 15 | 15 | 15 | 15 | 15 | 14 | 10 |
| 2300–2700 | 1 | 10 | 10 | 10 | 10 | 10 | 14 | 14 | 5 |
| 2700–3100 | 1 | 5 | 5 | 5 | 5 | 5 | 13 | 14 | 3 |
| Razem | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

Populacja o numerze 1 zawiera aż 94 elementy ze 100 (czyli 94%) w środkowym przedziale klasowym. Populacja 2., podobnie jak populacja poprzednia symetryczna, zawiera 40% jednostek w środkowej klasie, po 15% elementów w przedziałach sąsiadujących ze środkowym. Populacje 3., 4., 5. i 6. różnią się od populacji 2. tym, że ze środkowej klasy nieliczne elementy zostały przerzucone do którejś z sąsiednich

klas. Populacje 7. i 8., to rozkłady silnie spłaszczone, kształtem zbliżone do rozkładu jednostajnego (prostokątnego). Populacja 9. jest podobnie spłaszczona jak populacja 2., przy umiarkowanej prawostronnej asymetrii.

Wartości średnich arytmetycznych, odchyleń standardowych i klasycznych współczynników zmienności dla populacji:

**Tablica 5.7. Klasyczne miary zróżnicowania w porównywanych populacjach**

| populacja | $\bar{x}$ | $\sigma$ | $V_{zk}$ |
|---|---|---|---|
| 1 | 1700 | 211,662 | 12,45% |
| 2 | 1700 | 565,686 | 33,28% |
| 3 | 1696 | 567,10 | 33,44% |
| 4 | 1692 | 571,26 | 33,76% |
| 5 | 1688 | 578,15 | 34,25% |
| 6 | 1664 | 601,59 | 36,15% |
| 7 | 1700 | 775,63 | 45,63% |
| 8 | 1700 | 791,96 | 46,59% |
| 9 | 1460 | 562,85 | 38,6% |

I tak dla populacji 1., bardzo silnie skupionej (wysmukłej) klasyczny współczynnik zmienności wynosi około 12,5%, dla populacji 2., w kształcie dzwonu, klasyczny współczynnik zmienności wynosi około 33,3%, dla populacji 8., silnie spłaszczonej, klasyczny współczynnik zmienności rośnie do 46,6%.

Proponowane przedziały do oceny siły zróżnicowania cechy w całej populacji:

$V_{zk} < 20\%$ zróżnicowanie słabe

$20\% < V_{zk} < 40\%$ zróżnicowanie wyraźne

$V_{zk} > 40\%$ zróżnicowanie silne

Wartości median, odchyleń ćwiartkowych i pozycyjnych współczynników zmienności dla tych populacji:

**Tablica 5.8. Odchylenia ćwiartkowe i pozycyjne wsp. zmienności w przykł. populacjach**

| populacja | $M$ | $Q$ | $V_{zQ}$ |
|---|---|---|---|
| 1 | 1700,0 | 106,38 | 6,26% |
| 2 | 1700,0 | 333,33 | 19,61% |
| 3 | 1694,9 | 341,67 | 20,16% |
| 4 | 1694,9 | 346,67 | 20,45% |
| 5 | 1694,9 | 346,67 | 20,45% |
| 6 | 1683,8 | 373,33 | 22,17% |
| 7 | 1700,0 | 657,14 | 38,66% |
| 8 | 1700,0 | 685,71 | 40,34% |
| 9 | 1452,0 | 381,72 | 26,29% |

Osiągnięcie dużych (w kategoriach bezwzględnych) wartości współczynników zmienności dla wewnętrznych 50% populacji jest znacznie trudniejsze niż dla miar liczonych dla całej populacji, gdzie elementy skrajne bardzo silnie wpływają na wynik.

Proponowane przedziały do oceny zróżnicowania cechy wewnętrznych 50% populacji:

$V_{zQ} < 10\%$ zróżnicowanie słabe

$10\% < V_{zQ} < 20\%$ zróżnicowanie wyraźne

$V_{zQ} > 20\%$ zróżnicowanie silne

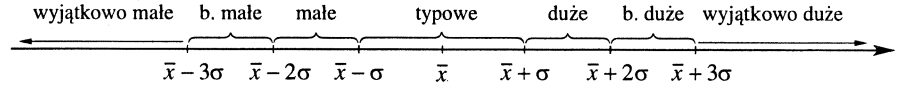
## 5.5. Związki odchylenia standardowego z innymi miarami statystycznymi

a. Iloraz odchylenia standardowego i średniej arytmetycznej wyrażony w procentach, to omówiony w poprzednim punkcie klasyczny współczynnik zmienności, mówiący o sile zróżnicowania (im większy procent średniej stanowi odchylenie standardowe, tym, oczywiście, zróżnicowanie silniejsze).

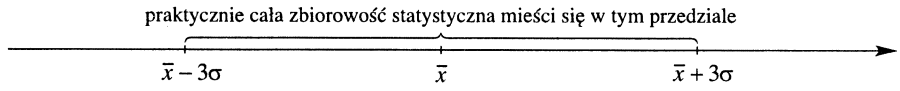
b. Obszar (przedział) typowy. Za typowe w statystyce uważa się te elementy zbiorowości, które różnią się od średniej o wartość nie więcej niż jednego odchylenia standardowego.

$$\bar{x} - \sigma \leq x_{typ} \leq \bar{x} + \sigma$$

c. Z pojęciem przedziału typowego wiąże się dalsza klasyfikacja elementów populacji, w zależności od tego, o ile odchyleń standardowych odbiegają od średniej populacji.



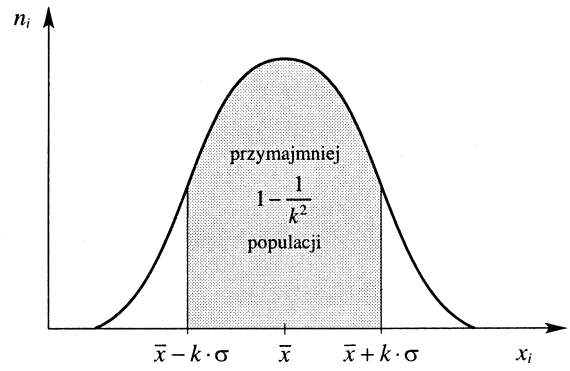

d. Prawo $3\sigma$ wywodzi się z obserwacji praktycznych, że bardzo nieliczne jednostki zbiorowości różnią się od średniej o więcej niż trzy odchylenia standardowe.




e. Standaryzacja. Jest to transformacja pozwalająca określić o ile odchyleń standardowych odbiega wartość danego elementu od średniej arytmetycznej.

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

f. Twierdzenie Czebyszewa. Mówi ono: w przypadku każdej populacji statystycznej można być pewnym, że część elementów, które różnią się od średniej arytmetycznej o **mniej** niż $k$ odchyleń standardowych, stanowi ułamek **nie mniejszy** niż $1-\frac{1}{k^2}$.




Wykres 5.2. Graficzny wyraz twierdzenia Czebyszewa

Na przykład: co najmniej 75% danych każdego zbioru musi mieć wartości różniące się od średniej arytmetycznej o co najwyżej dwa odchylenia standardowe ($k=2$); analogicznie, co najmniej 96% danych musi się mieścić w granicach ± pięciu odchyleń standardowych od średniej (poza tym przedziałem będzie więc **co najwyżej** 4% populacji!).

Twierdzenie Czebyszewa jako znajdujące zastosowanie do dowolnego kształtu zbiorowości daje wyniki bardzo ogólne (zgrubne). Jeśli wiemy z jakim konkretnym typem rozkładu mamy do czynienia (np. dwumianowym, normalnym, wykładniczym), stosowanie twierdzenia Czebyszewa mija się z celem, co obrazuje zestawienie:

**Tablica 5.9. Porównanie wskazań twierdzenia Czebyszewa z rozkładem normalnym**

| $k$ | minimalny procent populacji w przedziale ± $k$ odchyleń standardowych od średniej według twierdzenia Czebyszewa | procent populacji w przedziale ± $k$ odchyleń standardowych od średniej dla rozkładu normalnego |
|---|---|---|
| 1 | 0* | 68,27% |
| 2 | 75% | 95,45% |
| 3 | 88,9% | 99,73% |
| 4 | 93,75% | 99,994% |
| 5 | 96% | 99,99994% |

*Wartość 0 jako wynik wyrażenia $1-\frac{1}{1^2}$ nie ma żadnej wartości poznawczej. Przyjmuje się więc, że:

a) dla $k=1$ twierdzenia Czebyszewa nie stosuje się lub

b) w przedziale ± jednego odchylenia standardowego od średniej mieści się co najmniej jeden element populacji.

g. Oszacowanie Snedecora. Snedecor zbudowal tabelę wiążącą liczebność populacji ($N$), rozstęp ($R$) i odchylenie standardowe ($\sigma$):

**Tablica 5.10. Szacowanie wartości $\sigma$ według Snedecora**

| $N$ | $\sigma$ |
|---|---|
| 5 | $R/2$ |
| 10 | $R/3$ |
| 25 | $R/4$ |
| 50 | $R/4{,}5$ |
| 100 | $R/5$ |
| 200 | $R/5{,}5$ |
| 500 i więcej | $R/6$ |

Tabela umożliwia znalezienie **przybliżonej** wartości odchylenia standardowego, jeśli znamy jedynie liczebność populacji i rozstęp. Dla bardzo mało licznych zbiorowości (trzy pierwsze wiersze) do wyników należy podchodzić z wielką ostrożnością, dla licznych (wiersz ostatni) — rozstęp jest w przybliżeniu równy sześciu odchyleniom standardowym, co nawiązuje do prawa $3\sigma$.

Warto również, po wyznaczeniu odchylenia standardowego tradycyjnymi metodami, podzielić rozstęp przez odchylenie i sprawdzić, czy relacja jest zgodna z oszacowaniem Snedecora. Procedura ta często pozwala prawidłowo zinterpretować na pozór podejrzane wskazania miar.

***Przykłady***

ad a. W zadaniu o pensjach dyrektorów regionalnych mieliśmy:

$$\bar{x}^{(1)} = 23500 \text{ zł} \quad \text{i} \quad \sigma^{(1)} = 1553{,}8 \text{ zł}$$

$$\bar{x}^{(2)} = 23500 \text{ zł} \quad \text{i} \quad \sigma^{(2)} = 3519{,}9 \text{ zł}$$

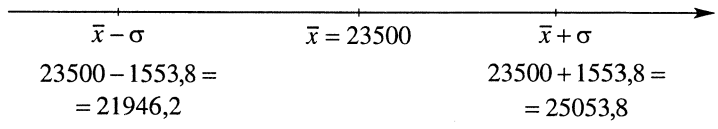
zatem:

$$V_{zk}^{(1)} = \frac{\sigma^{(1)}}{\bar{x}^{(1)}} \cdot 100\% = \frac{1553{,}8 \text{ zł}}{23500 \text{ zł}} \cdot 100\% = 6{,}6\%$$

$$V_{zk}^{(2)} = \frac{\sigma^{(2)}}{\bar{x}^{(2)}} \cdot 100\% = \frac{3516{,}9 \text{ zł}}{23500 \text{ zł}} \cdot 100\% \cong 15\%$$

W obu przypadkach współczynnik zmienności jest mniejszy od 20%, zatem zróżnicowanie płac jest słabe, ale w organizacji 2 dwukrotnie silniejsze.

ad b. Przedział typowy dla płac dyrektorów w organizacji 1 wynosi:




$$21946{,}2 \le x_{typ} \le 25053{,}8$$

Typowa płaca dyrektora organizacji 1 należy do przedziału $\langle 21946{,}2;\ 25053{,}8\rangle$.

ad b. Przedział typowy dla wielkości sprzedaży w firmie Profi jest następujący $\langle 12{,}5;\ 28{,}3\rangle$. Jaka jest średnia dzienna wielkość sprzedaży i jej odchylenie standardowe?

Ponieważ średnia arytmetyczna leży na środku przedziału typowego, więc:

$$\bar{x} = \frac{12{,}5 + 28{,}3}{2} = \frac{40{,}8}{2} = 20{,}4$$

Górna granica przedziału typowego jest równa $\bar{x} + \sigma$.
Zatem:

$$\bar{x} + \sigma = 28{,}3 \quad i \quad \bar{x} = 20{,}4$$

$$\sigma = 28{,}3 - \bar{x} = 28{,}3 - 20{,}4 = 7{,}9$$

ad c, e. Zapłaciliśmy za koszulę 70 zł, za garnitur 750 zł. Z danych opublikowanych przez organizację konsumentów dowiadujemy się, że średnia cena takiej koszuli wynosi 55 zł z odchyleniem standardowym 10 zł; średnia cena takiego garnituru wynosi 675 zł z odchyleniem standardowym 125 zł. W którym przypadku bardziej przepłaciliśmy? Które ceny są bardziej zróżnicowane: koszul czy garniturów?

$$z_k = \frac{x_k - \bar{x}_k}{\sigma_k} = \frac{70 - 55}{10} = 1{,}5$$

$$z_G = \frac{x_G - \bar{x}_G}{\sigma_G} = \frac{750 - 675}{125} = 0{,}6$$

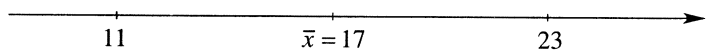
Cena zapłacona za koszulę różni się od średniej o 1,5 odchylenia standardowego. Jest to cena wysoka (należy do przedziału $\bar{x}+\sigma; \bar{x}+2\sigma$ ). Cena zapłacona za garnitur różni się od średniej o 0,6 odchylenia standardowego. Jest typowa (należy do przedziału typowego). O większą liczbę odchyleń standardowych od średniej odchyla się cena zapłacona za koszulę.

$$V_{zk}^{(k)} = \frac{\sigma^k}{\bar{x}^k} \cdot 100\% = \frac{10}{55} \cdot 100\% = 18{,}2\%$$

$$V_{zk}^{(G)} = \frac{\sigma^G}{\bar{x}^G} \cdot 100\% = \frac{125}{675} \cdot 100\% = 18{,}5\%$$

Nieznacznie większy jest klasyczny współczynnik zmienności dla cen garniturów, a więc ceny garniturów są silniej zróżnicowane. W obu przypadkach zróżnicowanie jest słabe, choć w górnych granicach przedziału (0–20%).

ad f. Średnia dzienna frekwencja w salonie fryzjerskim wynosi 17 osób z odchyleniem standardowym 3,6 osoby. Właściciel lokalu wie, że granicą ekonomicznej opłacalności jest 11 klientów dziennie. Jaki procent dni roboczych przynosi straty?
Informacje o rozkładzie są bardzo ogólne, w związku z czym stosujemy twierdzenie Czebyszewa. Dolna granica przedziału jest oddalona od średniej o 6 jednostek, górna granica przedziału wynosi zatem $17+6=23$.

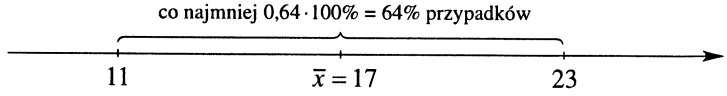



Dolna granica przedziału jest oddalona od średniej o nieznaną liczbę $k$ odchyleń standardowych (tak samo, jak górna granica).

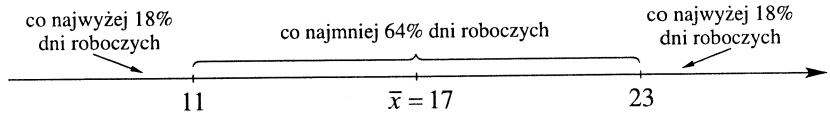
$$k = \frac{23-17}{\sigma} = \frac{23-17}{3{,}6} = \frac{6}{3{,}6} = 1{,}67$$

Twierdzenie Czebyszewa mówi, że z przedziałem (11; 23) wiąże się co najmniej $1-\frac{1}{k^2}$ przypadków:

$$1-\frac{1}{k^2} = 1-\frac{1}{1{,}67^2} = 1-0{,}36 = 0{,}64$$



Frekwencja mniejsza od 11 lub większa od 23 zdarza się zatem w co najwyżej 100% – 64% = 36% przypadków. Odpowiadając możemy bardzo ostrożnie stwierdzić, że właściciel firmy ponosi straty w co najwyżej 36% przypadków. Nic jednak nie stoi na przeszkodzie przyjęciu założenia o symetrycznym rozkładzie populacji na obu jej końcach, więc pozostałe co najwyżej 36% dzielimy na dwie części i podajemy końcową wersję: właściciel ponosi straty w co najwyżej 18% dni roboczych.




## 5.6. Równość wariancyjna

Wariancja jest ważną statystycznie miarą nie tylko dlatego, że w oparciu o nią wyznacza się wartość odchylenia standardowego. Dzięki swoim ciekawym własnościom jest np. wykorzystywana w analizie współzależności zjawisk masowych. Jedną z tych własności jest możliwość dekompozycji wariancji na dwa addytywne składniki: wariancję wewnątrzgrupową i wariancję międzygrupową. Mianowicie, jeśli badaną zbiorowość statystyczną podzielimy na $k$ podpopulacji, znając liczebności tych populacji, a także dysponując wartościami średnich arytmetycznych i wariancji to przy oznaczeniach:

$k$ — liczba podzbiorowości (podpopulacji),

$n_i$ — liczebność $i$-tej podzbiorowości,

$N = \sum_{i=1}^{k} n_i$ — liczebność populacji,

$Var_i$ — wariancja cechy w w $i$-tej podzbiorowości,

$\bar{x}_i$ — średnia arytmetyczna cechy w $i$-tej podzbiorowości,

$\bar{\bar{x}} = \frac{\sum_{1}^{k} \bar{x}_i n_i}{N}$ — średnia arytmetyczna dla całej populacji.

$$Var = \overline{Var_i} + Var(\bar{x}_i)$$

$\overline{Var_i} = \frac{\sum_1^k Var_i n_i}{N}$ — to wariancja wewnątrzgrupowa wyrażająca zmienność wewnątrzgrupową,

$Var(\bar{x}_i) = \frac{\sum_1^k (\bar{x}_i - \bar{\bar{x}})^2 n_i}{N}$ — to wariancja międzygrupowa wyrażająca zmienność międzygrupową.

W dwóch krańcowych, teoretycznie możliwych, przypadkach:

a. gdy wszystkie elementy w podpopulacji są sobie równe dla wymienionych grup wtedy wariancja wewnątrzgrupowa wynosi zero. Wariancja ogólna równa jest wariancji międzygrupowej.
b. gdy przy różniących się wartościami elementach w podpopulacjach średnie arytmetyczne dla grup są takie same, wariancja międzygrupowa jest równa zero. Wariancja ogólna równa jest wariancji wewnątrzgrupowej.

Równość wariancyjna umożliwia obliczenie wariancji ogólnej, gdy znane są liczebności podpopulacji, średnie arytmetyczne i wariancje cząstkowe. Dzięki równości wariancyjnej dowiadujemy się także, który ze składników miał większy udział w wariancji ogólnej.

Ponieważ równości wariancyjnej nie daje się przekształcić na równość odchyleń standardowych, ze względu na miana podniesione do kwadratu, składników równości nie interpretuje się merytorycznie.

***Przykład 5.3.***

Należy zbadać zróżnicowanie ogólne i źródła zróżnicowania wyników kolokwiów (te same zadania, miejsce i czas pisania) dla studentów różnych toków studiów. Dane rzeczywiste. Średnia arytmetyczna (w punktach) dla pierwszej grupy wyniosła 23,95; odchylenie standardowe 19,59 pkt.; liczba zdających 95. W drugiej grupie znalazły się 133 osoby, średnia arytmetyczna wyników była równa 58,03 pkt., a odchylenie standardowe 23,63 pkt.

**Tablica robocza dla wyznaczania składników równości wariacyjnej**

| Nr grupy | $n_i$ | $\bar{x}_i$ | $Var_i$ | $n_i\bar{x}_i$ | $Var_i n_i$ | $\bar{x}_i - \bar{\bar{x}}$ | $(\bar{x}_i - \bar{\bar{x}})^2$ | $(\bar{x}_i - \bar{\bar{x}})^2 n_i$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 95 | 23,95 | 383,7681 | 2275,25 | 36457,9695 | –19,88 | 395,2144 | 37545,368 |
| 2 | 133 | 58,03 | 558,3769 | 7717,99 | 74264,1277 | 14,20 | 201,64 | 26818,120 |
| Razem | 228 | — | — | 9993,24 | 110722,0972 | — | — | 64363,488 |

$$\bar{\bar{x}} = \frac{\sum_1^k \bar{x}_i n_i}{N} = \frac{9993,24}{228} = 43,83$$

$$\overline{Var_i} = \frac{\sum_1^k (\bar{x}_i - \bar{\bar{x}})^2 n_i}{N} = \frac{64363,448}{228} = 282,296$$

$$Var = \overline{Var_i} + Var(\bar{x}_i) = 485,623 + 282,296 = 767,919$$

ponieważ

$$\frac{Var}{Var} = \frac{\overline{Var_i}}{Var} + \frac{Var(\bar{x}_i)}{Var}$$

$$1 = 0,632 + 0,368$$

Na zróżnicowanie wyników kolokwium wyraźnie silniejszy wpływ miało wewnątrzgrupowe zróżnicowanie wyników niż międzygrupowe zróżnicowanie wyników. Odchylenie standardowe wyników dla wszystkich piszących wyniosło:

$$\sqrt{767,919 \text{ pkt.}^2} = 27,71 \text{ pkt.}$$

Klasyczny współczynnik zmienności dla wszystkich piszących

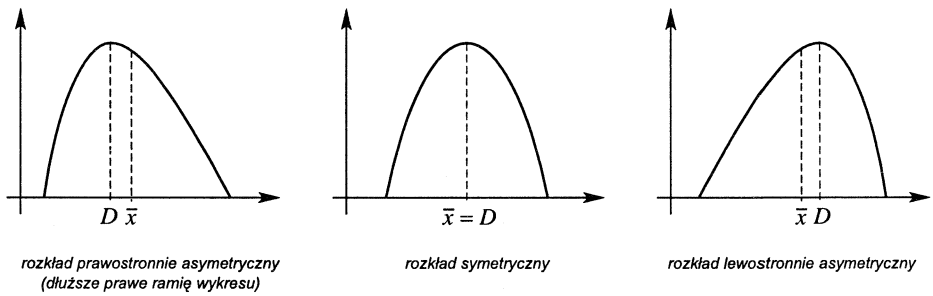
$$V_{zk} = \frac{27,71}{43,83} \cdot 100\% = 63,22\%$$

# Rozdział 6

# Miary asymetrii (skośności)

Celem miar asymetrii jest określenie, w którym kierunku dany rozkład odchyla się od rozkładu symetrycznego i jak silne jest to odchylenie. Sprowadzając zagadnienie do najbardziej zewnętrznych oznak zauważmy, że rozkład symetryczny ma ramiona równych długości, w rozkładach asymetrycznych jedno z ramion jest dłuższe niż drugie (aż do całkowitego zaniku drugiego w przypadku rozkładów skrajnie asymetrycznych). Przedstawiają to wykresy:

Wykres 6.1 pokazuje, że dla oceny asymetrii rozkładu można wykorzystać położenie średniej arytmetycznej i dominanty.




Wykres 6.1. Rozkład symetryczny i asymetryczny

Jeśli:

$\bar{x} - D = 0$ — to brak asymetrii,

$\bar{x} - D > 0$ — to asymetria prawostronna, ze względu na znak wskazania nazywana inaczej dodatnią,

$\bar{x} - D < 0$ — to asymetria lewostronna, nazywana inaczej ujemną.

Ta prosta w koncepcji i liczeniu miara nosi nazwę wskaźnika asymetrii i jest symbolizowana jako $w_{as}$. Ma dwie niedogodności: nic nie mówi o sile skośności oraz skupiając się na położeniu jedynie dwóch wartości może wręcz wprowadzać w błąd (pamiętajmy o interpolacyjnym wzorze na dominantę, który dla szeregów rozdzielczych dla tej samej populacji z wariantowymi długościami klas da różne wartości).

Pierwszą wadę przezwycięża się dzieląc wskaźnik asymetrii przez odchylenie standardowe. Miara ta nosi nazwę współczynnika asymetrii Pearsona. Zgodnie ze swoją konstrukcją mówi o ile odchyleń standardowych oddalona jest w populacji średnia od dominanty — im wartość współczynnika większa, tym silniejsza asymetria rozkładu. Praktycznie nie przekracza przedziału $(-1, +1)$, osiągając wartości zbliżone do jedności dla spłaszczonych rozkładów, bez wyraźnie zaznaczonego ośrodka dominującego.

Trudniej jest znaleźć remedium na drugie zastrzeżenie. Wiedząc, że wskazania miary mogą być mylące dla rozkładów spłaszczonych, w kształcie rekiniej płetwy, również takich, gdzie znaczny odsetek zbiorowości plasuje się blisko krańców obszaru zmienności cechy, stosujemy ją z zachowaniem zasady ograniczonego zaufania i konfrontujemy z miarami klasycznymi.

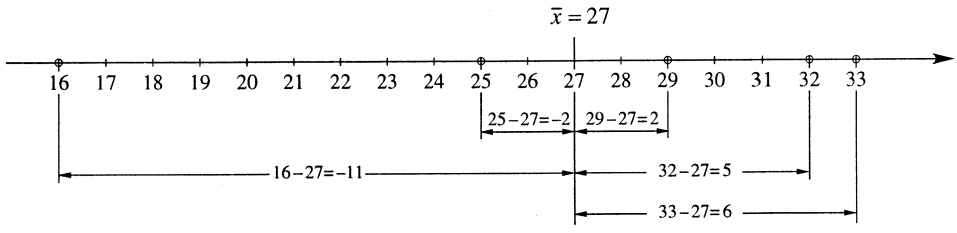
Koncepcja klasycznej miary asymetrii jest kontynuacją rozważań, które doprowadziły nas do wariancji. Sumy odchyleń wartości cechy od średniej arytmetycznej podniesione do trzeciej potęgi $\sum(x_i - \bar{x})^3 n_i$ wynoszą zero dla rozkładów symetrycznych, przyjmują wartości większe od zera dla rozkładów prawostronnie asymetrycznych, są ujemne dla rozkładów lewostronnie asymetrycznych. Podobnie jak przy wariancji, uśredniamy te sumy odchyleń uzyskując:

$$\frac{\sum(x_i - \bar{x})^3 n_i}{N}$$

Jest to wzór na trzeci moment centralny ($e_3$). Istotę miary oddaje przykład 6.1.

***Przykład 6.1.***

Wyznaczyć $e_3$ dla liczb 16, 25, 29, 32 i 33.




Odchylenia wartości cechy od średniej arytmetycznej

Po doświadczeniach z liczeniem wariancji, lepiej jest potraktować ten indywidualny wykaz wartości jak szereg punktowy (każda z wartości o liczebności jeden). Zatem:

| $x_i$ | $n_i$ | $x_i n_i$ | $x-\bar{x}$ | $(x-\bar{x})^2$ | $(x-\bar{x})^3$ | $(x-\bar{x})^3 n_i$ |
|---|---|---|---|---|---|---|
| 16 | 1 | 16 | –11 | 121 | –1331 | –1331 |
| 25 | 1 | 25 | –2 | 4 | –8 | –8 |
| 29 | 1 | 29 | 2 | 4 | 8 | 8 |
| 32 | 1 | 32 | 5 | 25 | 125 | 125 |
| 33 | 1 | 33 | 6 | 36 | 216 | 216 |
| Razem | 5 | 135 | 0 | × | × | –990 |

$$\bar{x} = \frac{135}{5} = 27$$

Suma sześcianów odchyleń wartości cechy od średniej jest mniejsza od zera (–990), co świadczy o asymetrii lewostronnej.

$$e_3 = \frac{\sum (x_i - \bar{x})^3 n_i}{N} = \frac{-990}{5} = -198$$

Średnia arytmetyczna sześcianów odchyleń wartości cechy od średniej asymetrycznej wynosi –198.

Klasyczną absolutną miarą skośności jest zatem $e_3$. Do oceny siły asymetrii potrzebna jest miara stosunkowa. Otrzymujemy ją dzieląc trzeci moment centralny przez odchylenie standardowe podniesione do trzeciej potęgi. Powstała miara (trzeci moment centralny standaryzowany) jest dzięki temu:

a) bezmianowa,

b) „oczyszczona" z wpływu zróżnicowania.

ad b. Wyjaśnienie jest następujące. Asymetria idzie w parze ze zróżnicowaniem w tym sensie, że im rozkład jest bardziej asymetryczny, tym zróżnicowanie cechy jest silniejsze (ale niekoniecznie na odwrót). Podzielenie miary opisującej łącznie zróżnicowanie i asymetrię przez wielkość zróżnicowania z populacji daje „czystą" miarę asymetrii.

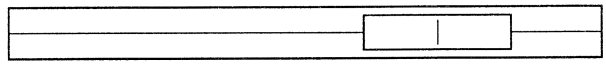
Odpowiednikiem współczynnika asymetrii Pearsona, w którym klasyczne składowe (tylko, gdy nie możemy ich wyznaczyć) zastępujemy ich pozycyjnymi zamiennikami, jest współczynnik asymetrii:

$$V_{as} = \frac{M - D}{Q}$$

Do oceny kierunku i siły asymetrii tej części populacji, która dotyczy wewnętrznie położonych 50% elementów służy wzór:

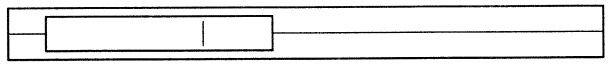
$$V_{asp} = \frac{Q_1 + Q_3 - 2M}{2Q}$$

Jego interpretacja staje się prosta, gdy odwołamy się do *box and whisker plot.* Wzór określa, która z podstaw wewnętrznego prostokąta wykresu (na lewo, czy na prawo od mediany) jest dłuższa i jak silna jest ta relacja. Długości podstaw reprezentują długości ramion wykresu dla wewnętrznych ćwiartek. Jeśli równe są obszary zmienności dla obu wewnętrznych ćwiartek, jak niżej:



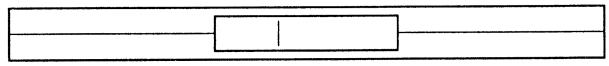
to powiemy, że wewnętrzne 50% populacji jest symetryczne.

Jeśli obszar zmienności dla drugiej ćwiartki jest większy niż dla trzeciej, jak np.:



to powiemy, że wewnętrzne 50% populacji jest asymetrycznie lewostronne.

Jeśli obszar zmienności dla trzeciej ćwiartki jest dłuższy niż dla drugiej ćwiartki, jak:



to wewnętrzne 50% populacji jest prawostronnie asymetryczne.

Przedstawione miary asymetrii (tab. 6.1) są tak skonstruowane, by dla populacji:

a) symetrycznych: ich wskazania były równe zero,

b) prawostronnie asymetrycznych: ich wskazania były dodatnie, stąd alternatywna nazwa dla populacji tego kształtu — dodatnio asymetryczne,

c) lewostronnie asymetrycznych: ich wskazania były ujemne — populacje odpowiednio nazywamy ujemnie asymetrycznymi.

**Tablica 6.1. Miary asymetrii**

| Miary | Absolutne | Stosunkowe | |
|---|---|---|---|
| Klasyczne | $e_3 = \frac{\sum(x_i - \bar{x})^3 n_i}{\sum n_i} =$ $= m_3 - 3m_2 \cdot m_1 + 2m_1^3$ trzeci moment centralny | $\gamma_3 = \frac{e_3}{\sigma^3}$ | trzeci moment centralny standaryzowany bardzo rzadko przekracza (–2,2) |
| Pozycyjne | $w_{as} = \bar{x} - D$ wskaźnik asymetrii | $V_{as} = \frac{\bar{x} - D}{\sigma}$ | praktycznie nie wykracza poza przedział (–1,1) |
| | | $V_{as} = \frac{3(M - D)}{2\sigma}$ | Pearsonowskie odmiany poprzedniego wzoru |
| | | $V_{as} = \frac{3(\bar{x} - M)}{\sigma}$ | |
| | | $V_{as} = \frac{M - D}{Q}$ | |
| | | $V_{asp} = \frac{Q_1 + Q_3 - 2M}{2Q}$ [–1,1] | współczynnik asymetrii wewnętrznych 50% |

Propozycja przedziałów dla oceny siły asymetrii z użyciem trzeciego momentu centralnego standaryzowanego:

$\gamma_3 < -1$ — silna ujema
$-1 < \gamma_3 < -0{,}5$ — wyraźna ujemna
$-0{,}5 < \gamma_3 < 0$ — słaba ujemna
$\gamma_3 = 0$ — brak asymetrii
$0 < \gamma_3 < 0{,}5$ — słaba dodatnia
$0{,}5 < \gamma_3 < 1$ — wyraźna dodatnia
$\gamma_3 > 1$ — silna dodatnia

***Przykład 6.2.***

Miary asymetrii charakteryzujące sprzedaż w firmie Profi.

*Miary klasyczne*

Suma po ostatniej kolumnie $\left(\sum(\mathring{x} - \bar{x})^3 n_i\right)$ jest dodatnia, a więc rozkład jest prawostronnie asymetryczny.

$$e_3 = \frac{\sum(\mathring{x} - \bar{x})^3 n_i}{N} = \frac{6614{,}154}{64} = 103{,}35$$

**Tablica robocza**

| $x_{01}-x_{1i}$ | $n_i$ | $\mathring{x}_i$ | $\mathring{x}_i n_i$ | $\mathring{x}_i-\bar{x}$ | $(\mathring{x}_i-\bar{x})^2$ | $(\mathring{x}_i-\bar{x})^2 n_i$ | $(\mathring{x}_i-\bar{x})^3 n_i$ |
|---|---|---|---|---|---|---|---|
| 5–10 | 7 | 7,5 | 52,5 | –12,9 | 166,41 | 1164,87 | –15026,823 |
| 10–15 | 10 | 12,5 | 125,0 | –7,9 | 62,41 | 624,10 | –4930,390 |
| 15–20 | 13 | 17,5 | 227,5 | –2,9 | 8,41 | 109,33 | –317,057 |
| 20–25 | 18 | 22,5 | 405,0 | 2,1 | 4,41 | 79,38 | 166,698 |
| 25–30 | 8 | 27,5 | 220,0 | 7,1 | 50,41 | 403,28 | 2863,288 |
| 30–35 | 5 | 32,5 | 162,5 | 12,1 | 146,41 | 732,05 | 8857,805 |
| 35–40 | 3 | 37,5 | 112,5 | 17,1 | 292,41 | 877,23 | 15000,633 |
| Razem | 64 | × | 1305,0 | × | × | 3990,24 | 6614,154 |

Przypadające na jednostkę odchylenie od średniej do trzeciej potęgi wynosi 103,35 palety$^3$.

Trzeci moment centralny standaryzowany

$$\gamma_3 = \frac{e_3}{\sigma^3} = \frac{103{,}35\,\text{palety}^3}{(7{,}9\,\text{palety})^3} = \frac{103{,}35\,\text{palety}^3}{493{,}039\,\text{palety}^3} \cong 0{,}21$$

Asymetria cechy jest słaba dodatnia.

*Miary pozycyjne*

Wskaźnik asymetrii

$$w_{as} = \bar{x} - D = 20{,}4 - 21{,}7 = -1{,}3$$

Wskaźnik asymetrii sugeruje asymetrię lewostronną, choć przed chwilą $\gamma_3$ mówił o słabej asymetrii dodatniej i nie został popełniony żaden błąd w obliczeniach. Z czego wynikają sprzeczne wskazania miar i jak w takich sytuacjach powinniśmy się zachować?

1. Warto przypomnieć, że każda z analizowanych miar podchodzi do skośności ze swego punktu widzenia: $\gamma_3$ uwzględnia odchylenia wszystkich jednostek populacji od średniej, wskaźnik asymetrii interesuje się tylko wartościami średniej i dominanty.
2. Wskazania miar pozycyjnych, gdy z wykresu nie wynikają w sposób oczywisty wnioski odnośnie kierunku asymetrii, traktujemy z dużą dozą ostrożności, szczególnie wtedy, gdy $\bar{x}$ i $D$ leżą bardzo blisko siebie.
3. Wyżej cenimy zawsze miary klasyczne (gdy istnieją merytoryczne podstawy, by je wyliczać), bo formuła uwzględnia wszystkie elementy populacji.

Współczynnik asymetrii Pearsona

$$V_{as} = \frac{\bar{x} - D}{\sigma}$$

W tym konkretnym przypadku, po zastrzeżeniach do wskaźnika asymetrii, interpretowanie go jako miary asymetrii całej populacji nie jest zasadne.

Współczynnik asymetrii wewnętrznych 50%

$$V_{asp} = \frac{Q_1 + Q_3 - 2M}{2Q} = \frac{14{,}5 + 25 - 41{,}2}{10{,}5} = -0{,}162$$

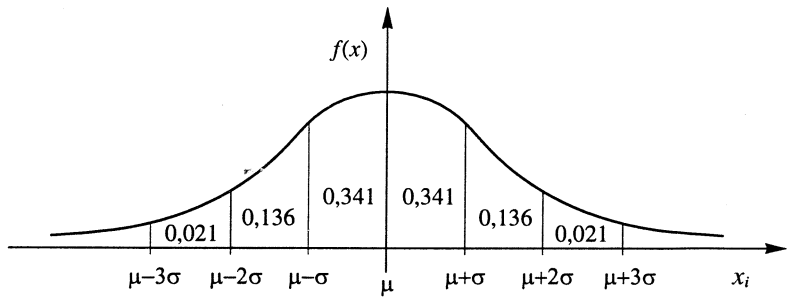
Dwie wewnętrzne ćwiartki populacji wykazują słabą asymetrię lewostronną. Sprzeczne wskazania $w_{as}$ i $\gamma_3$ otrzymujemy m.in. gdy sytuacja jest właśnie taka: kierunek asymetrii całej populacji jest inny niż wewnętrznych 50%.

# Rozdział 7

# Miary spłaszczenia (kurtozy) i miary koncentracji

## 7.1. Miary spłaszczenia (wysmukłości, kurtozy)

Rozkład normalny charakteryzuje się tym, że niezależnie od konkretnych wartości parametrów $\mu$ i $\sigma$, w przedziałach określonej liczby odchyleń standardowych od średniej znajduje się zawsze taki sam odsetek populacji.




Wykres 7.1. Własności rozkładu normalnego

Oznacza to, że jednostki zbiorowości statystycznej o rozkładzie normalnym w bardzo określony sposób rozkładają się w obszarze zmienności, co przekłada się na konkretną koncentrację elementów wokół średniej arytmetycznej.

Za miarę koncentracji elementów populacji wokół wartości średniej można uznać formułę bazującą na czwartym momencie centralnym, który dla rozkładu normalnego wynosi $3\sigma^4$.

Można go wyliczyć ze wzoru na parzyste momenty centralne dla rozkładu normalnego o postaci ogólnej:

$$e_{2n} = \frac{\sum (x_i - \bar{x})^{2n} n_i}{N} = \frac{(2n)!}{2^n n!} \sigma^4$$

skąd dla $n=2$

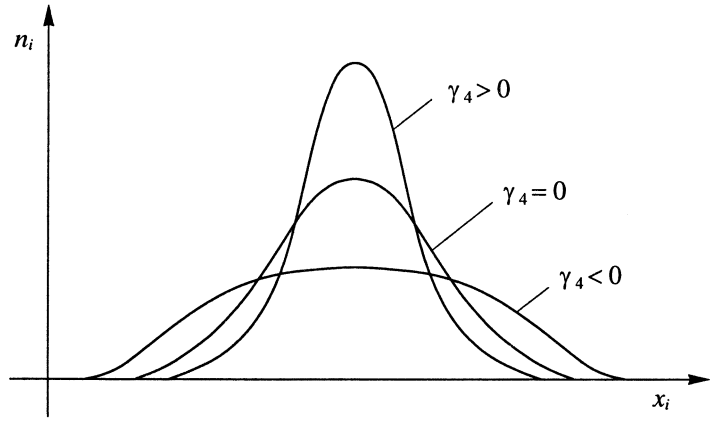
$$e_{2\cdot 2} = e_4 = \frac{\sum (x_i - \bar{x})^4 n_i}{N} = \frac{(2\cdot 2)!}{2^2 \cdot 2!} \sigma^{2\cdot 2} = \frac{4!}{4\cdot 2} \sigma^4 = 3\sigma^4$$

Jest to miara mianowana o mianach w czwartej potędze. Jeśli ją zestandaryzujemy i od wyników odejmiemy 3 to otrzymamy miarę

$$\gamma_4 = \frac{e_4}{\sigma^3} - 3$$

nazywaną czwartym momentem centralnym standaryzowanym lub współczynnikiem ekscesu, która przyjmuje wartość zero jeśli skupienie elementów wokół średniej arytmetycznej jest dla badanej populacji o konkretnej średniej i odchyleniu standardowym takie samo jak dla rozkładu normalnego o identycznej średniej i odchyleniu standardowym. Jeśli skupienie elementów wokół średniej jest dla danej populacji większe niż dla rozkładu normalnego (o takiej samej średniej i odchyleniu standardowym) to $\gamma_4$ jest większe od zera i tym większe im silniejsza jest koncentracja.

Jeśli badany rozkład jest spłaszczony w stosunku do rozkładu normalnego to $\gamma_4$ jest mniejsza od zera. Rozkłady zgodne z rozkładem normalnym nazywa się mezokur-




Wykres 7.2. Rozkłady wypukłe, spłaszczone i mezokurtyczne

tycznymi, rozkłady wysmukłe w stosunku do rozkładu normalnego nazywa się leptokurtycznymi, rozkłady spłaszczone w stosunku do rozkładu normalnego nazywa się platokurtycznymi lub platykurtycznymi (od greckiego kurtos — garbaty).

Zobaczmy, jak w kilku przypadkach, rozkłady liczebności przekładają się ma kształty rozkładów i wartości czwartego momentu centralnego standaryzowanego.

Populacja 1. charakteryzuje się wyjątkowo silnym skupieniem jednostek wokół średniej arytmetycznej co skutkuje niespotykaną w praktyce badań statystycznych wartością współczynnika ekscesu wynoszącą aż 22. Rozkład 3. przedstawiony jako wielobok liczebności jest bardzo zbliżony kształtem do rozkładu normalnego. Dla niego $\gamma_4 = -0{,}02$ co interpretujemy jako minimalnie spłaszczony rozkład mezokurtyczny.

**Tablica 7.1. Liczebności przykładowych populacji**

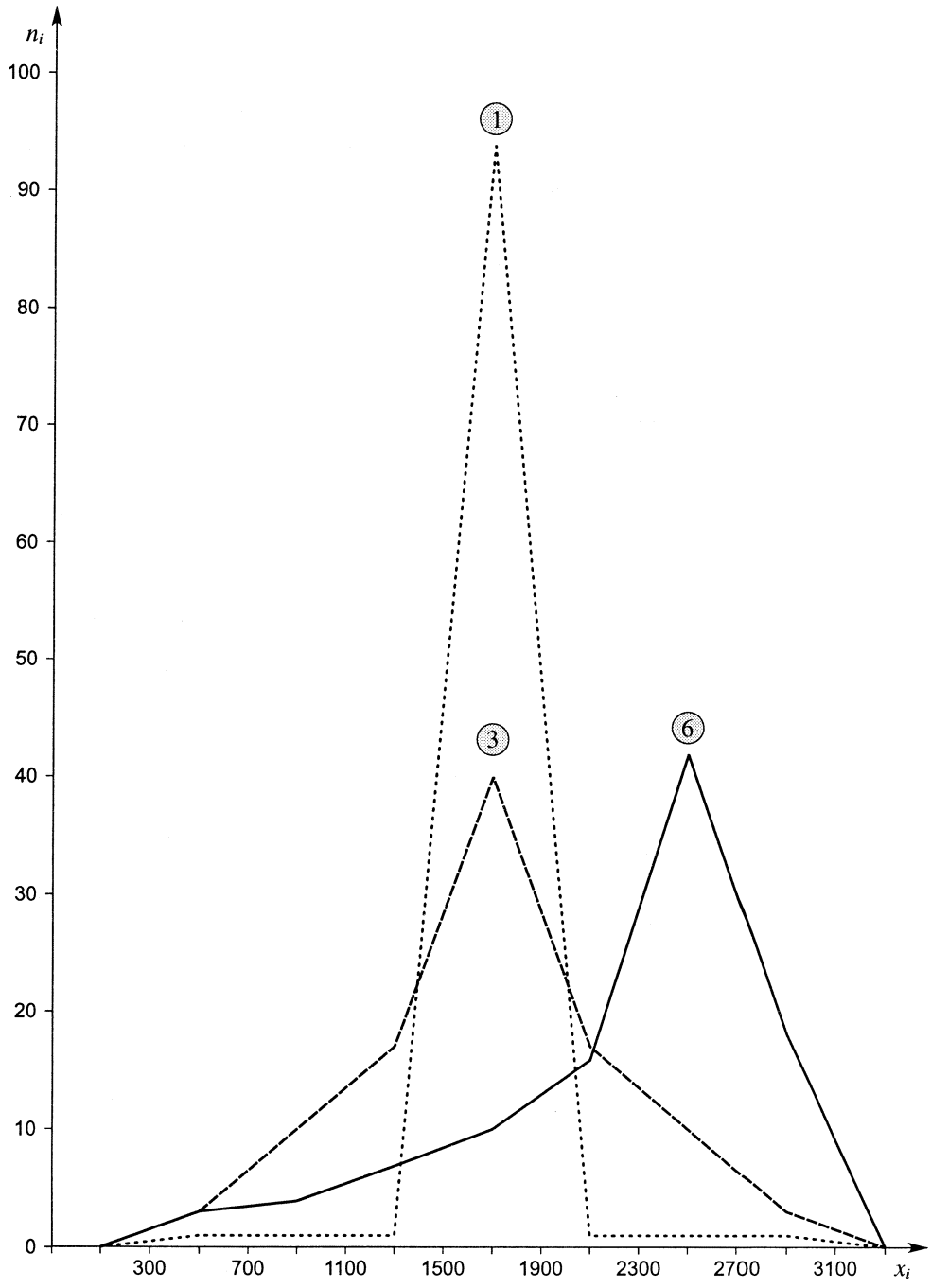
| $x_{0i}-x_{1i}$ | $n_i^1$ | $n_i^2$ | $n_i^3$ | $n_i^4$ | $n_i^5$ | $n_i^6$ | $n_i^7$ |
|---|---|---|---|---|---|---|---|
| 300–700 | 1 | 1 | 3 | 14 | 50 | 3 | 18 |
| 700–1100 | 1 | 4 | 10 | 14 | 0 | 4 | 42 |
| 1100–1500 | 1 | 10 | 17 | 14 | 0 | 7 | 16 |
| 1500–1900 | 94 | 70 | 40 | 16 | 0 | 10 | 10 |
| 1900–2300 | 1 | 10 | 17 | 14 | 0 | 16 | 7 |
| 2300–2700 | 1 | 4 | 10 | 14 | 0 | 42 | 4 |
| 2700–3100 | 1 | 1 | 3 | 14 | 50 | 18 | 3 |
| Razem | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

Źródło: dane umowne.

**Tablica 7.2. Miary klasyczne w porównywanych populacjach**

| Nr populacji | $\bar{x}$ | $\sigma$ | $V_{zk}$ | $\gamma_3$ | $\gamma_4$ |
|---|---|---|---|---|---|
| 1 | 1700 | 211,66 | 12,45% | 0 | 22,01 |
| 2 | 1700 | 334,66 | 19,69% | 0 | 3,33 |
| 3 | 1700 | 518,46 | 30,50% | 0 | –0,02 |
| 4 | 1700 | 791,96 | 46,59% | 0 | –1,22 |
| 5 | 1700 | 1200,00 | 70,59% | 0 | –4,00 |
| 6 | 2220 | 602,66 | 27,15% | –1,13 | 0,66 |
| 7 | 1180 | 602,66 | 51,07% | 1,13 | 0,66 |

Źródło: opracowanie własne na podstawie tablicy 7.1.




Wykres 7.3. Kształty wyróżnionych wieloboków liczebności ze względu na wysmukłość.
Źródło: opracowanie własne na podstawie tablicy 7.1.

Rozkład 4. bardzo nieznacznie różni się od rozkładu prostokątnego. Czwarty moment centralny standaryzowany równy −1,22 potwierdza silne spłaszczenie wykresu w stosunku do wzorcowego rozkładu normalnego. Zauważmy także, że rozkłady 6. i 7. jako swoje lustrzane odbicia mają identyczne wartości miary wysmukłości.

Analityczna ocena stopnia spłaszczenia szeregu rozdzielczego i odpowiadającej mu krzywej liczebności wiąże się z dosyć uciążliwymi obliczeniami (jeśli wykonujemy je osobiście) niezależnie od tego czy posłużymy się formułą momentów centralnych czy momentów zwykłych.

$$\gamma_4 = \frac{e_4}{\sigma^4} - 3 = \frac{\frac{\sum(\mathring{x}_i - \bar{x})^4 n_i}{N}}{\sigma^4} - 3 = \frac{m_4 - 4m_1m_3 + 6m_1^2m_2 - 3m_1^4}{\sigma^4} - 3$$

Poniższe tablice robocze przedstawiają sposoby obliczania $\gamma_4$ w obu wariantach dla wielkości w firmie Profi.

**Tablica robocza wykorzystująca momenty zwykłe do liczenia miar klasycznych**

| $x_{0i} - x_{1i}$ | $n_i$ | $\mathring{x}_i$ | $\mathring{x}_i n_i$ | $\mathring{x}_i^2 n_i$ | $\mathring{x}_i^3 n_i$ | $\mathring{x}_i^4 n_i$ |
|---|---|---|---|---|---|---|
| 5–10 | 7 | 7,5 | 52,5 | 393,75 | 2953,125 | 22148,44 |
| 10–15 | 10 | 12,5 | 125,0 | 1562,50 | 19531,250 | 244140,63 |
| 15–20 | 13 | 17,5 | 227,5 | 3981,25 | 69671,875 | 1219257,81 |
| 20–25 | 18 | 22,5 | 405,0 | 9112,50 | 205031,250 | 4613203,13 |
| 25–30 | 8 | 27,5 | 220,0 | 6050,00 | 166375,000 | 4575312,50 |
| 30–35 | 5 | 32,5 | 162,5 | 5281,25 | 171640,625 | 5578320,31 |
| 35–40 | 3 | 37,5 | 112,5 | 4218,75 | 158203,125 | 5932617,19 |
| Razem | 64 | × | 1305,0 | 30600,00 | 793406,250 | 22185000,01 |

Źródło: obliczenia własne.

$$e_4 = \frac{\sum(\mathring{x}_i - \bar{x})^4 n_i}{N} = \frac{618085{,}23}{64} = 9657{,}58$$

$$\gamma_4 = \frac{e_4}{\sigma^4} - 3 = \frac{9657{,}58}{7{,}9^4} - 3 = 2{,}48 - 3 = -0{,}52$$

Wynik świadczy o dość wyraźnym spłaszczeniu rozkładu.

**Tablica robocza wykorzystująca momenty centralne do liczenia miar klasycznych**

| $x_{0i} - x_{1i}$ | $n_i$ | $\mathring{x}_i$ | $\mathring{x}_i n_i$ | $\mathring{x}_i - \bar{x}$ | $(\mathring{x}_i - \bar{x})^2$ | $(\mathring{x}_i - \bar{x})^2 n_i$ | $(\mathring{x}_i - \bar{x})^3 n_i$ | $(\mathring{x}_i - \bar{x})^4 n_i$ | Cum | |
|---|---|---|---|---|---|---|---|---|---|---|
| 5–10 | 7 | 7,5 | 52,5 | –12,9 | 166,41 | 1164,87 | –15026,823 | 193846,01 | 7 | |
| 10–15 | 10 | 12,5 | 125,0 | –7,9 | 62,41 | 624,10 | –4930,390 | 38950,08 | 17 | ← $Q_1$ |
| 15–20 | 13 | 17,5 | 227,5 | –2,9 | 8,41 | 109,33 | –317,057 | 919,47 | 30 | |
| 20–25 | 18 | 22,5 | 405,0 | 2,1 | 4,41 | 79,38 | 166,698 | 350,07 | 48 | ← $Q_1$, $Q_2$, $D$ |
| 25–30 | 8 | 27,5 | 220,0 | 7,1 | 50,41 | 403,28 | 2863,288 | 20329,34 | 56 | |
| 30–35 | 5 | 32,5 | 162,5 | 12,1 | 146,41 | 732,05 | 8857,805 | 107179,44 | 62 | |
| 35–40 | 3 | 37,5 | 112,5 | 17,1 | 292,41 | 877,23 | 150000,633 | 256510,82 | 64 | |
| Razem | 64 | × | 1305,0 | × | × | 3990,24 | 6614,154 | 618085,23 | × | |

Źródło: obliczenia własne.

$$m_1 = \frac{\sum \mathring{x}_i n_i}{N} = \frac{1305}{64} = 20{,}391$$

$$m_2 = \frac{\sum \mathring{x}_i^2 n_i}{N} = \frac{30600}{64} = 478{,}125$$

$$m_3 = \frac{\sum \mathring{x}_i^3 n_i}{N} = \frac{793406{,}25}{64} = 12396{,}973$$

$$m_4 = \frac{\sum \mathring{x}_i^4 n_i}{N} = \frac{22185000{,}01}{64} = 12396{,}973$$

$$e_4 = m_4 - 4m_1m_3 + 6m_1^2m_2 - 3m_1^4 = 9648{,}59$$

$$\gamma_4 = \frac{9648{,}59}{7{,}9^4} - 3 = 2{,}48 - 3 = -0{,}52$$

Pozostałe momenty centralne:

$$e_2 = Var = m_2 - m_1^2 = \frac{30600}{64} - \left(\frac{1305}{64}\right)^2 = 62{,}347$$

$$\sigma(x) = 7{,}896 \cong 7{,}9$$

$$e_3 = m_3 - 3m_2m_1 + 2m_1^3 = 12396{,}973 - 29248{,}341 + 16956{,}865 = 105{,}497$$

$$\gamma_3 = \frac{105{,}497}{7{,}9^3} = 0{,}21$$

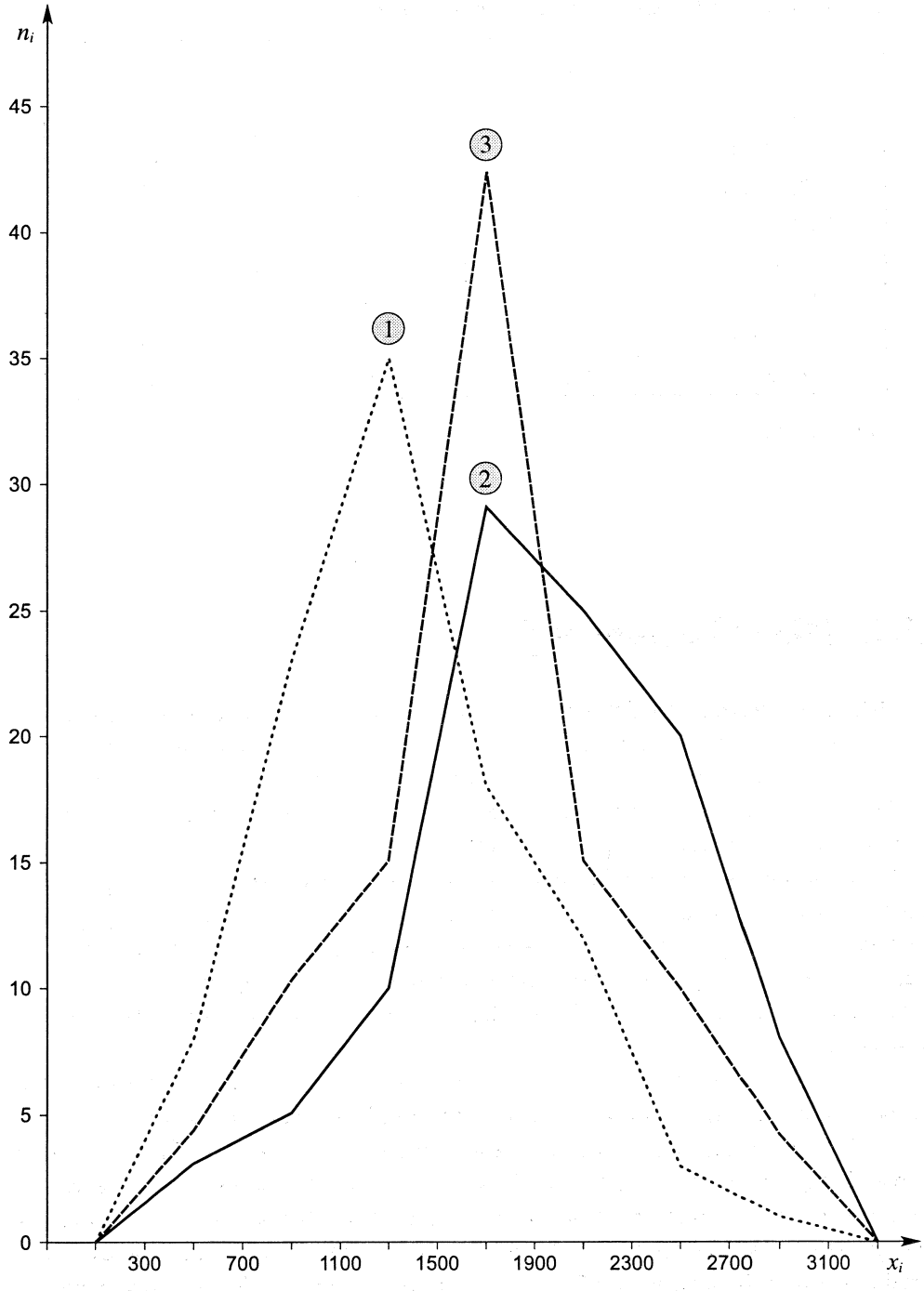
Czwarty moment centralny standaryzowany nie jest miarą łatwą w interpretacji, szczególnie dla rozkładów asymetrycznych.

Przykładowo, dla trzech rozkładów przedstawionych w tablicy 7.3, z kształtami wieloboków pokazanymi na wykresie 7.4, wszystkie współczynniki ekscesu wskazują na rozkłady bliskie mezokurtycznemu i odpowiednio wynoszą:

$$\gamma_4^{(1)} = -0{,}03,$$

$$\gamma_4^{(2)} = -0{,}08,$$

$$\gamma_4^{(3)} = -0{,}01.$$




Wykres 7.4. Wieloboki liczebności dla danych z tablicy 7.3. Źródło: opracowanie własne.

**Tablica 7.3. Liczebności analizowanych populacji**

| $x_{0i} - x_{1i}$ | $n_i^1$ | $n_i^2$ | $n_i^3$ |
|---|---|---|---|
| 300–700 | 8 | 3 | 4 |
| 700–1100 | 23 | 5 | 10 |
| 1100–1500 | 35 | 10 | 15 |
| 1500–1900 | 18 | 29 | 42 |
| 1900–2300 | 12 | 25 | 15 |
| 2300–2700 | 3 | 20 | 10 |
| 2700–3100 | 1 | 8 | 4 |
| Razem | 100 | 100 | 100 |

Źródło: dane umowne.

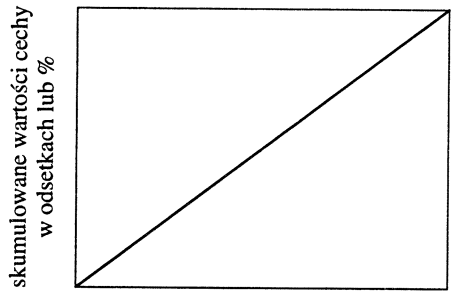
## 7.2. Miary koncentracji

Ostatnim aspektem analizy struktury zbiorowości statystycznej jest koncentracja rozumiana jako nierównomierne rozłożenie sumy wartości cechy mierzalnej pomiędzy jednostki zbiorowości. Bada się koncentrację kapitałów, produkcji, ziemi, oszczędności, dochodów czy zanieczyszczeń. Miary koncentracji stosowane są także w odniesieniu do procesów epidemicznych.

Brak koncentracji występuje wówczas, gdy na każdą jednostkę populacji przypada taka sama część ogólnej sumy wartości cechy. Z tego wynika, że 10% populacji posiada 10% sumy wartości (np. dochodów z pracy), 35% populacji posiada 35% sumy wartości cechy, 75% populacji dysponuje 75% ogólnej sumy wartości etc. Koncentracja jest absolutna, jeśli na jeden element zbiorowości przypada cała wartość cechy.

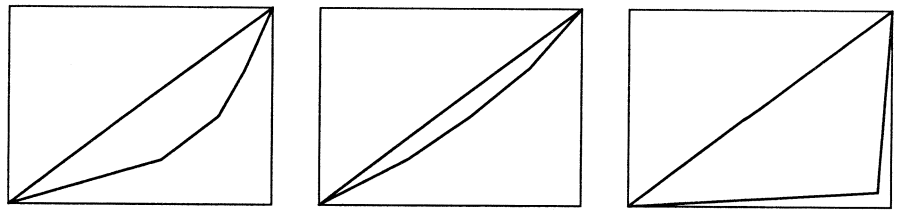
Do opisu zjawiska koncentracji stosuje się metody graficzną i analityczną. Obie wykorzystują kumulaty (dystrybuanty) liczebności i sum wartości cechy.

Istotą koncepcji jest tzw. krzywa (koncentracji) Lorentza. Na planie kwadratu (prostokąta) na podstawie odkłada się skumulowane liczebności, na boku figury odkłada się skumulowane wartości cechy. Krzywa Lorentza powstaje przez połączenie punktów o współrzędnych: odcięta – częstość skumulowana na granicy klasy, rzędna – skumulowana wartość cechy (wyrażona w odsetkach lub procentach) na granicy klasy.




Przekątna figury to tzw. linia równomiernego rozkładu opowiadająca sytuacji całkowitego braku koncentracji, gdzie każdemu elementowi populacji przypisana jest taka sama część sumy wartości.

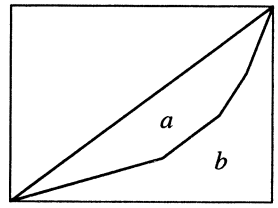
W praktyce badań empirycznych nie zdarza się, by krzywa Lorentza pokrywała się z linią równomiernego rozkładu. Konstrukcja tej formy wykresu sprawia, że krzywa koncentracji przebiega poniżej linii równomiernego rozkładu.



Rozkłady różniące się stopniem koncentracji

Im dalej od linii równomiernego rozkładu plasuje się krzywa Lorentza, tym cecha jest silniej skoncentrowana.

Jeśli przez $a$ oznaczymy pole zawarte między krzywą koncentracji, a linią równomiernego rozkładu, a przez $b$ pole leżące pod krzywą Lorentza:



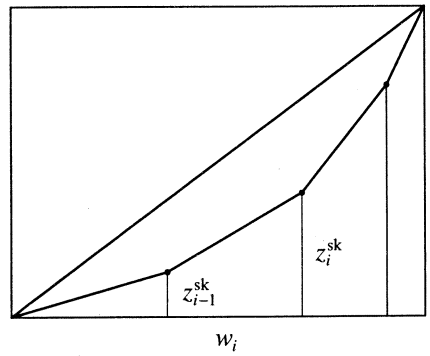

to do oceny siły koncentracji zjawiska może służyć iloraz $\frac{a}{a+b}$.

Miarę tę nazywa się współczynnikiem koncentracji Lorentza i symbolizuje przez $K_L$

$$K_L = \frac{a}{a+b} \qquad 0 \le K_L \le 1$$

Wartość 0 (brak koncentracji) i 1 (całkowita koncentracja) są jedynie teoretycznie możliwe.

Wyznaczenie, w sposób analityczny, pola pod krzywą koncentracji wymaga znajomości pól trójkąta i trapezów tworzących krzywą.




Wzór na współczynnik koncentracji Lorentza uwzględniający znajdowanie pola (-*b*-) figury pod krzywą koncentracji przedstawia się następująco:

$$K_L = \frac{0{,}5 - \sum\left(\frac{z_i^{sk} - z_{i-1}^{sk}}{2}\right) w_i}{0{,}5} = 1 - \sum\left(z_i^{sk} - z_{i-1}^{sk}\right) w_i$$

Wyznaczenie wartości $K_L$ wymaga odwołania się do tablicy roboczej o następujących składowych:

| $x_{0i} - x_i$ | $n_i$ | $\mathring{x}_i$ | $w_i$ | $m_i = \mathring{x}_i w_i$ | $z_i = \frac{m_i}{\sum m_i}$ | $z_i^{sk}$ | $w_i^{sk}$ | $\left(z_i^{sk} - z_{i-1}^{sk}\right) w_i$ |
|---|---|---|---|---|---|---|---|---|

***Przykład 7.1.***

Przedstawić graficznie i zbadać analitycznie stopień koncentracji rent rodzinnych w Polsce według wysokości świadczeń w marcu 2003 r.

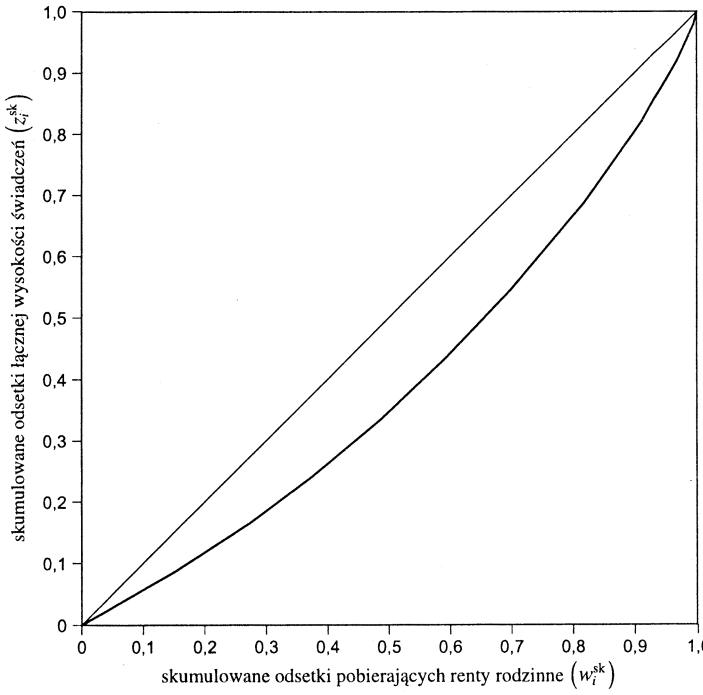
Dane o wysokości i liczbie pobieranych świadczeń możemy znaleźć w Roczniku Statystycznym Rzeczpospolitej Polskiej 2004 na str. 278, tab. 5 (199).

W zagregowanej formie przedstawia się ona jak niżej:

**Tablica 7.4. Pobierający renty rodzinne według wysokości świadczeń w marcu 2003 r.**

| Wysokość świadczenia brutto w PLN | Pobierający renty rodzinne w tys. |
|---|---|
| 550–600 | 192,3 |
| 600–700 | 149,5 |
| 700–800 | 128,2 |
| 800–900 | 140,9 |
| 900–1000 | 135,7 |
| 1000–1100 | 130,7 |
| 1100–1300 | 152,1 |
| 1300–1600 | 115,7 |
| 1600–2000 | 70,4 |
| 2000–2400 | 33,9 |
| 2400 i więcej | 7,5 |
| Razem | 1256,9 |

Źródło: obliczenia własne na podstawie tablicy 5 (199), RSRP 204, s. 278.




Wykres 7.5. Krzywa koncentracji Lorentza dla pobierających renty rodzinne w Polsce w marcu 2003 r.

**Tablica robocza**

| Wysokość świadczenia brutto w PLN | Liczba pobierających rentę | Odsetek pobierających rentę | Środek klasy | Łączna wysokość świadczenia brutto | Odsetek łącznej wysokości świadczenia | Skumulowane odsetki | | Pole wg wzoru |
|---|---|---|---|---|---|---|---|---|
| | | | | | | łącznej wysokości świadczeń | liczby pobierających renty | |
| $x_{0i} - x_{1i}$ | $n_i$ | $w_i$ | $\mathring{x}_i$ | $m_i = \mathring{x}_i w_i$ | $z_i = \frac{m_i}{\sum m_i}$ | $z_i^{sk}$ | $w_i^{sk}$ | $\left(z_i^{sk} + z_{i-1}^{sk}\right) w_i$ |
| 550–600 | 192,3 | 0,153 | 575 | 87,975 | 0,087 | 0,087 | 0,153 | 0,013 |
| 600–700 | 149,5 | 0,119 | 650 | 77,350 | 0,077 | 0,164 | 0,272 | 0,030 |
| 700–800 | 128,2 | 0,102 | 750 | 76,500 | 0,076 | 0,240 | 0,374 | 0,041 |
| 800–900 | 140,9 | 0,112 | 850 | 95,200 | 0,094 | 0,334 | 0,486 | 0,064 |
| 900–1000 | 135,7 | 0,108 | 950 | 102,600 | 0,102 | 0,436 | 0,594 | 0,083 |
| 1000–1100 | 130,7 | 0,104 | 1050 | 109,200 | 0,108 | 0,544 | 0,698 | 0,102 |
| 1100–1300 | 152,1 | 0,121 | 1200 | 145,200 | 0,144 | 0,688 | 0,819 | 0,149 |
| 1300–1600 | 115,7 | 0,092 | 1450 | 133,400 | 0,132 | 0,820 | 0,911 | 0,139 |
| 1600–2000 | 70,4 | 0,056 | 1800 | 100,800 | 0,100 | 0,920 | 0,967 | 0,097 |
| 2000–2400 | 33,9 | 0,027 | 2200 | 59,400 | 0,059 | 0,979 | 0,994 | 0,051 |
| 2400 i więcej | 7,5 | 0,006 | 3782,5 | 22,695 | 0,020 | 1,000 | 1,000 | 0,012 |
| Razem | 1256,9 | 1,000 | × | 1010,32 | 1,000 | × | × | 0,781 |

Źródło: obliczenia własne na podstawie tablicy 7.4.

W oparciu o wykres możemy stwierdzić słabą koncentrację rent rodzinnych wypłacanych w marcu 2003 roku. Przy takim przebiegu krzywej koncentracji mówi się, że analizowany rozkład jest zbliżony do rozkładu równomiernego.

Współczynnik koncentracji $K_L$

$$K_L = 1 - 0{,}781 = 0{,}219$$

potwierdza w pełni wniosek wyciągnięty na podstawie wykresu, że rozkład rent rodzinnych jest słabo skoncentrowany.

# Rozdział 8

# Przykładowe zadania z rozwiązaniami

## 8.1. Zadania kompleksowe dla szeregu przedziałowego

***Przykład 8.1.***

Proszę scharakteryzować rozkład liczby palet sprzedanych przez firmę Profi w ciągu badanych 64 dni.

| Dzienna liczba sprzedanych palet $x_{0i} - x_{1i}$ | Liczba dni o danej wysokości sprzedaży $n_i$ |
|---|---|
| 5–10 | 7 |
| 10–15 | 10 |
| 15–20 | 13 |
| 20–25 | 18 |
| 25–30 | 8 |
| 30–35 | 5 |
| 35–40 | 3 |
| Razem | 64 |

Z czego powinno składać się rozwiązanie:

A. wykresu,

B. wniosków z wykresu,

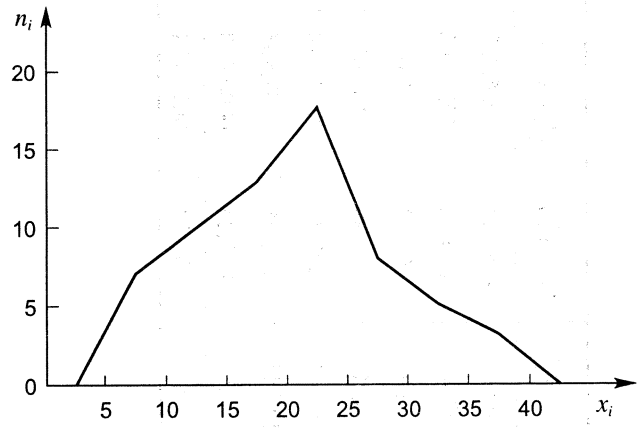
C. miar i ich interpretacji,

D. końcowych wniosków.

1. (dotyczy A i B)

Zaczynamy od wykresu, aby ustalić czy populacja jest względnie jednorodna czy względnie niejednorodna. Jeśli okaże się wielomodalna nie możemy nic sensownego zrobić przed uprzednim rozbiciem jej na względnie jednorodne części, do czego potrzebne jest nowe badanie. Jeśli jest jednomodalna oceniamy przybliżone zróżnicowanie, asymetrię i ewentualne spłaszczenie (w celu porównania z później wyznaczonymi miarami) oraz decydujemy się na liczenie charakterystyk, które w danej sytuacji najlepiej opiszą strukturę badanego zjawiska.

Formą graficzną która umożliwia odpowiedź na postawione pytania i prostą w wykonaniu jest wielobok liczebności.




Wykres 8.1. Wielobok liczebności dla przykładu o paletach

Wnioski z wykresu:

a) populacja jest jednomodalna,

b) wyraźnie zróżnicowana, co wiąże się ze spłaszczeniem rozkładu; asymetria wygląda na prawostronną (słabą), ale kształt wykresu postaci ⌒ może się wiązać ze **sprzecznymi** wskazaniami miar asymetrii,

c) nie ma przeciwwskazań dla wyznaczenia miar klasycznych i pozycyjnych (domknięte przedziały, umiarkowana asymetria). Końcowe wnioski mogą być dzięki temu pełniejsze.

**Tablica robocza**

| $x_{0i} - x_{1i}$ | $n_i$ | $\mathring{x}_i$ | $\mathring{x}_i n_i$ | $\mathring{x}_i - \bar{x}$ | $(\mathring{x}_i - \bar{x})^2$ | $(\mathring{x}_i - \bar{x})^2 n_i$ | $(\mathring{x}_i - \bar{x})^3 n_i$ | $(\mathring{x}_i - \bar{x})^4 n_i$ | Cum | |
|---|---|---|---|---|---|---|---|---|---|---|
| 5–10 | 7 | 7,5 | 52,5 | –12,9 | 166,41 | 1164,87 | –15026,823 | 193846,01 | 7 | |
| 10–15 | 10 | 12,5 | 125,0 | –7,9 | 62,41 | 624,10 | –4930,390 | 38950,08 | 17 | ← $Q_1$ |
| 15–20 | 13 | 17,5 | 227,5 | –2,9 | 8,41 | 109,33 | –317,057 | 919,47 | 30 | |
| 20–25 | 18 | 22,5 | 405,0 | 2,1 | 4,41 | 79,38 | 166,698 | 350,07 | 48 | ← $Q_1$, $Q_2$, $D$ |
| 25–30 | 8 | 27,5 | 220,0 | 7,1 | 50,41 | 403,28 | 2863,288 | 20329,34 | 56 | |
| 30–35 | 5 | 32,5 | 162,5 | 12,1 | 146,41 | 732,05 | 8857,805 | 107179,44 | 62 | |
| 35–40 | 3 | 37,5 | 112,5 | 17,1 | 292,41 | 877,23 | 150000,633 | 256510,82 | 64 | |
| Razem | 64 | × | 1305,0 | × | × | 3990,24 | 6614,154 | 618085,23 | × | |

2. (dotyczy C)

## Miary klasyczne

***Średnia arytmetyczna*** (klasyczna absolutna miara tendencji centralnej):

$$\bar{x} = \frac{\sum x_i^{\circ} n_i}{N} = \frac{1305}{64} = 20{,}390625 \cong 20{,}4$$

Przeciętnie, w badanym okresie, firma sprzedawała 20,4 palet dziennie.

***Wariancja*** (klasyczna absolutna miara zróżnicowania)

$$Var = \frac{\sum \left(x_i^{\circ} - \bar{x}\right)^2 n_i}{N} = \frac{3990{,}24}{64} = 62{,}3475$$

(bez interpretacji, ze względu na miana w kwadratach).

***Odchylenie standardowe*** (klasyczna absolutna miara zróżnicowania):

$$\sigma = \sqrt{Var} = \sqrt{62{,}3475} = 7{,}896 \cong 7{,}9$$

Liczba sprzedanych palet różniła się od średniej (20,4) o 7,9 palet.

***Współczynnik zmienności*** (klasyczna względna miara zróżnicowania)

$$V_{zk} = \frac{\sigma}{\bar{x}} \cdot 100\% = \frac{7{,}9}{20{,}4} \cdot 100\% = 38{,}73\%$$

Odchylenie standardowe stanowi niemal 40% średniej arytmetycznej. Zróżnicowanie cechy na pograniczu wyraźnego i silnego.

***Trzeci moment centralny*** (klasyczna absolutna miara asymetrii):

$$e_3 = \frac{\sum \left(x_i^{\circ} - \bar{x}\right)^3 n_i}{N} = \frac{6614{,}154}{64} = 103{,}346 \cong 103{,}3$$

Ponieważ $e_3 > 0$ (odchylenia na plus wartości cechy od $\bar{x}$ podniesione do sześcianu przeważają nad odchyleniami na minus) rozkład jest prawostronnie (dodatnio) asymetryczny.

***Trzeci moment centralny standaryzowany*** (klasyczna względna miara asymetrii):

$$\gamma_3 = \frac{e_3}{\sigma^3} = \frac{103{,}3}{7{,}9^3} = \frac{103{,}3}{493{,}039} = 0{,}2095 \cong 0{,}21$$

$\gamma_3 < 0{,}5$. Słaba asymetria prawostronna (rozkład niewiele odchyla się od rozkładu symetrycznego).

***Czwarty moment centralny*** (klasyczna absolutna miara spłaszczenia):

$$e_4 = \frac{\sum\left(\mathring{x}_i - \bar{x}\right)^4 n_i}{N} = \frac{618085{,}23}{64} = 9657{,}58$$

(bez interpretacji, miana w czwartej potędze).

***Czwarty moment centralny standaryzowany*** (klasyczna względna miara spłaszczenia):

$$\gamma_4 = \frac{e_4}{\sigma^4} - 3 = \frac{9657{,}58}{7{,}9^4} - 3 = \frac{9657{,}58}{3895{,}0081} - 3 = 2{,}479 - 3 = -0{,}521$$

$\gamma_4 < -0{,}5$. Dość wyraźne spłaszczenie rozkładu.

## Miary pozycyjne

Obliczenia pomocnicze do wyznaczenia kwantyli (określamy numery elementów, których wartości będziemy następnie wyznaczali).

$$n_{Q_1} = 0{,}25 \cdot 64 = 16$$
$$n_{Q_2} = 0{,}5 \cdot 64 = 32$$
$$n_{Q_3} = 0{,}75 \cdot 64 = 48$$

***Pierwszy kwartyl*** (pozycyjna absolutna miara położenia):

$$Q_1 = 10 + \frac{16 - 7}{10} \cdot 5 = 10 + 4{,}5 = 14{,}5$$

W $^1/_4$ przypadków (dni analizowanej działalności) wielkość sprzedaży była nie większa (mniejsza, równa) niż 14,5 palety.

***Mediana = drugi kwartyl*** (pozycyjna absolutna miara tendencji centralnej):

$$Q_2 = M = 20 + \frac{32-30}{18} \cdot 5 = 20 + 0{,}5(5) \cong 20{,}6$$

W połowie przypadków wielkość sprzedaży była nie większa niż 20,6 palety.

***Trzeci kwartyl*** (pozycyjna absolutna miara położenia)

$$Q_3 = 20 + \frac{48-30}{18} \cdot 5 = 20 + 5 = 25$$

(wartość trzeciego kwartyla jest równa górnej granicy klasy, bo jest to w tym przypadku wartość ostatniego elementu w przedziale).

W 75% przypadków wielkość sprzedaży była nie większa niż 25 palet.

***Dominanta (moda)*** (pozycyjna absolutna miara tendencji centralnej):

$$D = 20 + \frac{18-13}{18-13+18-8} \cdot 5 = 20 + 1{,}6(6) \cong 21{,}7$$

Najczęściej spotykaną dzienną wielkością sprzedaży było 21,7 palet.

$$w_{as} = \bar{x} - D = 20{,}4 - 21{,}7 = -1{,}3$$

$w_{as} < 0$ (średnia poprzedza dominantę na wykresie), co mówi o ujemnej (lewostronnej) asymetrii rozkładu cechy. Porządek miar na osi liczbowej $\bar{x} < M < D$ jest charakterystyczny dla rozkładów ujemnie asymetrycznych. Wynik jest sprzeczny ze wskazaniem trzeciego momentu centralnego ($e_3 > 0$!) co wiąże się z przypominającym płetwę rekina kształtem rozkładu — wniosek b. z punktu 1. Oba wyniki są poprawne i cenne poznawczo. Charakteryzują różne aspekty skośności analizowanego rozkładu.

***Współczynnik asymetrii*** (pozycyjna względna miara asymetrii):

$$V_{as} = \frac{\bar{x} - D}{\sigma} = -\frac{1{,}3}{7{,}9} = -0{,}165$$

Średnia znajduje się na lewo od dominanty w odległości niespełna 0,2 odchylenia standardowego.

***Odchylenie ćwiartkowe*** (pozycyjna absolutna miara zróżnicowania):

$$Q = \frac{Q_3 - Q_1}{2} = \frac{25 - 14{,}5}{2} = \frac{10{,}5}{2} = 5{,}25$$

Średnie zróżnicowanie w połowie obszaru zmienności (wyznaczonej przez pierwszy i trzeci kwartyl) wynosi 5,3 palety; inaczej: zróżnicowanie elementów środkowych 50% populacji względem mediany wynosi 5,3 palety.

***Współczynnik zmienności pozycyjnej*** (pozycyjna względna miara zróżnicowania):

$$V_{zQ} = \frac{Q}{M} \cdot 100\% = \frac{5{,}25}{20{,}6} \cdot 100\% = 25{,}49\% \cong 25{,}5\%$$

Odchylenie ćwiartkowe stanowi 25,5% mediany. Ponieważ $V_{zQ} > 20\%$ zróżnicowanie środkowych 50% populacji określamy jako silne.

***Współczynnik asymetrii wewnętrznych 50%*** (pozycyjna względna miara asymetrii):

$$V_{asp} = \frac{Q_1 + Q_3 - 2M}{2Q} = \frac{14{,}5 + 25 - 2 \cdot 20{,}6}{2 \cdot 5{,}25} = -\frac{1{,}7}{10{,}5} = -0{,}162$$

Wewnętrzne 50% populacji ma słabą asymetrię ujemną (lewostronną).

3. (dotyczy D)

## *Syntetyczne wnioski*

| Wyszczególnienie | Cała populacja | Wewnętrznie położone 50% populacji |
|---|---|---|
| Zróżnicowanie | na pograniczu wyraźnego i silnego ($V_{zk} = 38{,}7\%$) | silne ($V_{zQ} > 25{,}5\%$) |
| Asymetria | słaba asymetria dodatnia ($\gamma_3 = 0{,}21$) | słaba asymetria ujemna ($V_{asp} = -0{,}16$) |

Ponadto warto zwrócić uwagę na:

a. przeciwstawne wskazania miar asymetrii tłumaczone szczególnym kształtem rozkładu,

b. rozstęp pokrywany jest przez 4,4 odchylenia standardowego:

$$\frac{R}{\sigma} = \frac{x_{max} = x_{min}}{\sigma} = \frac{40 - 5}{7{,}9} = 4{,}4,$$

co jest zgodne z oszacowaniem Snedecora,

c. rozkład jest wyraźnie spłaszczony ($\gamma_4 = -0{,}52$).

***Przykład 8.2.***

W teście na zużycie paliwa uczestniczyło 170 pojazdów. Dane o trasie w km pokonanej na tej samej ilości paliwa przedstawia tablica poniżej. Proszę scharakteryzować rozkład tej cechy.

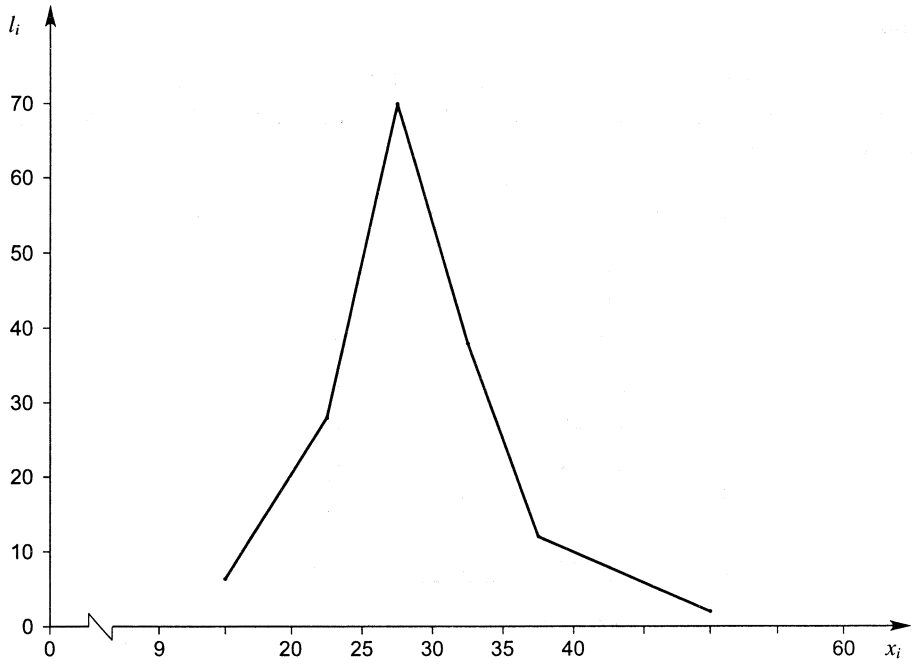
| $x_{0i}-x_{1i}$ | $n_i$ |
|---|---|
| 9–20 | 14 |
| 20–25 | 28 |
| 25–30 | 70 |
| 30–35 | 38 |
| 35–40 | 12 |
| 40–60 | 8 |
| Razem | 170 |

Źródło: dane umowne.

Wielobok liczebności wobec klas o różnych długościach wykonujemy dla natężeń liczebności lub gęstości.

| $x_{0i}-x_{1i}$ | $n_i$ | $h_i$ | $\mathring{x}_i$ | $l_i$ | | $g_i$ |
|---|---|---|---|---|---|---|
| 9–20 | 14 | 11 | 14,5 | 6,4 | | 1,3 |
| 20–25 | 28 | 5 | 22,5 | 28,0 | | 5,6 |
| 25–30 | 70 | 5 | 27,5 | 70,0 | | 14,0 |
| 30–35 | 38 | 5 | 32,5 | 38,0 | lub | 7,6 |
| 35–40 | 12 | 5 | 37,5 | 12,0 | | 2,4 |
| 40–60 | 8 | 20 | 50,0 | 2,0 | | 0,4 |
| Razem | 170 | × | × | × | | × |

$h_{\text{bazowe}} = 5$




Wnioski z wykresu:

— populacja jest jednomodalna,
— asymetria jest prawostronna,
— rozkład wygląda na mocno wysmukły,
— zróżnicowanie niezbyt silne.

Warto liczyć miary klasyczne i pozycyjne (obliczenia pomocnicze w tablicy roboczej).

## Miary klasyczne

### Średnia arytmetyczna

$$\bar{x} = \frac{\sum \mathring{x}_i n_i}{N} = \frac{4843}{170} = 28{,}488 \approx 28{,}5 \text{ km}$$

Gdyby każdy z pojazdów uczestniczących w teście pokonał taką samą trasę byłoby to 28,5 km.

### Wariancja

$$Var = \frac{\sum (\mathring{x}_i - \bar{x}) n_i}{N} = \frac{9100}{170} = 53{,}529 \text{ km}^2$$

**Tablica robocza**

| $x_{0i}-x_{1i}$ | $n_i$ | $\mathring{x}_i$ | $\mathring{x}_i n_i$ | $\mathring{x}_i-\bar{x}$ | $(\mathring{x}_i-\bar{x})^2$ | $(\mathring{x}_i-\bar{x})^2 n_i$ | $(\mathring{x}_i-\bar{x})^3 n_i$ | $(\mathring{x}_i-\bar{x})^4 n_i$ | Cum | |
|---|---|---|---|---|---|---|---|---|---|---|
| 9–20 | 14 | 14,5 | 203 | –14 | 196 | 2744 | –38416 | 537824,0 | 14 | $\leftarrow Q_3$ |
| 20–25 | 28 | 22,5 | 630 | –6 | 36 | 1008 | –6048 | 36288,0 | 42 | |
| 25–30 | 70 | 27,5 | 1925 | –1 | 1 | 70 | –70 | 70,0 | 112 | $\leftarrow Q_1, M, D$ |
| 30–35 | 38 | 32,5 | 1235 | 4 | 16 | 608 | 2432 | 9728,0 | 150 | |
| 35–40 | 12 | 37,5 | 450 | 9 | 81 | 972 | 8748 | 78732,0 | 162 | |
| 40–60 | 8 | 50,0 | 400 | 21,5 | 462,25 | 3698 | 79507 | 1709400,5 | 170 | |
| Razem | 170 | × | 4843 | × | × | 9100 | 46153 | 2372042,5 | × | |

$$\bar{x} = \frac{\sum \mathring{x}_i n_i}{N} = \frac{4843}{170} = 28{,}488 \cong 28{,}5$$

***Odchylenie standardowe***

$$\sigma = \sqrt{Var} = \sqrt{53{,}529} = 7{,}316 \text{ km}$$

Trasa pokonana przez każdy z pojazdów różni się od średniej przeciętnie o 7,3 km.

***Klasyczny współczynnik zmienności***

$$V_{zk} = \frac{\sigma}{\bar{x}} \cdot 100\% = \frac{7{,}3}{28{,}5} \cdot 100\% = 25{,}6\%$$

Wyraźne zróżnicowanie cechy choć na pograniczu ze słabym.

***Trzeci moment centralny***

$$e_3 = \frac{\sum(\mathring{x}_i - \bar{x})^3 n_i}{N} = \frac{46153}{170} = 271{,}488 \text{ km}^3$$

Dodatnia asymetria rozkładu.

***Trzeci moment centralny standardyzowany***

$$\gamma_3 = \frac{e_3}{\sigma^3} = \frac{271{,}488}{7{,}3^3} = \frac{271{,}488}{389{,}017} = 0{,}698$$

Wyraźna asymetria prawostronna rozkładu.

***Czwarty moment centralny***

$$e_4 = \frac{\sum(\mathring{x}_i - \bar{x})^4 n_i}{N} = \frac{2372042{,}5}{170} = 13953{,}191 \text{ km}^4$$

***Czwarty moment centralny standaryzowany***

$$\gamma_4 = \frac{e_4}{\sigma^4} - 3 = \frac{13953{,}191}{7{,}3^4} - 3 = \frac{13953{,}191}{2839{,}824} - 3 = 4{,}91 - 3 = 1{,}91$$

Wyraźna smukłość rozkładu (jednostki populacji mocno skupione wokół średniej asymetrycznej).

## Miary pozycyjne

### Kwartyle

$$n_{Q_1} = 0{,}25 \cdot 170 = 42{,}5$$
$$n_{Q_2} = 0{,}5 \cdot 170 = 85$$
$$n_{Q_3} = 0{,}75 \cdot 170 = 127{,}5$$

— ***Pierwszy kwartyl***

$$Q_1 = 25 + \frac{42{,}5 - 42}{70} \cdot 5 = 25 + \frac{2{,}5}{70} = 25{,}04 \text{ km}$$

Jedna czwarta pojazdów uczestniczących w teście pokonała trasę nie większą niż 25,04 km, pozostałe trzy czwarte przejechały trasę nie mniejszą od 25,04 km.

— ***Mediana (drugi kwartyl)***

$$M = 25 + \frac{85 - 42}{70} \cdot 5 = 25 + \frac{215}{70} = 25 + 3{,}07 = 28{,}07 \text{ km}$$

Połowa pojazdów przejechała nie więcej niż 28,07 km, pozostała połowa przejechała nie mniej niż 28,07 km.

— ***Trzeci kwartyl***

$$Q_3 = 30 + \frac{127{,}5 - 112}{38} \cdot 5 = 30 + \frac{77{,}5}{38} = 30 + 2{,}04 = 32{,}04 \text{ km}$$

75% pojazdów przejechało co najwyżej 32,04 km, 25% nie mniej niż 32,04 km.

### Odchylenie ćwiartkowe

$$Q = \frac{Q_3 - Q_1}{2} = \frac{32{,}04 - 25{,}04}{2} = \frac{7}{2} = 3{,}5 \text{ km}$$

Połowa długości przedziału zajmowanego przez wewnętrznie położone 50% populacji wynosi 3,5 km, inaczej: zróżnicowanie elementów należących do wewnętrznych 50% populacji względem mediany wynosi 3,5 km.

### Pozycyjny współczynnik zmienności

$$V_{zQ} = \frac{Q}{M} \cdot 100\% = \frac{3{,}5}{28{,}07} \cdot 100\% = 12{,}47\% \cong 12{,}5\%$$

Wyraźne zróżnicowanie jednostek należących do centralnych ćwiartek populacji.

***Pozycyjny współczynnik asymetrii***

$$V_{asp} = \frac{Q_1 + Q_3 - 2M}{2Q} = \frac{25{,}04 + 32{,}04 - 2 \cdot 28{,}07}{7} = \frac{0{,}94}{7} = 0{,}134$$

Wewnętrzne 50% zbiorowości jest słabo prawostronnie asymetryczne.

***Dominanta***

Przy różnej długości klasach musi być liczona dla natężeń lub gęstości:

$$D = 25 + \frac{70 - 28}{70 - 28 + 70 - 38} \cdot 5 = 25 + \frac{210}{74} = 27{,}84 \text{ km}$$

Najczęściej pokonywaną trasą w tym teście było 27,8 km.

***Wskaźnik asymetrii***

$$w_{as} = \bar{x} - D = 28{,}5 - 27{,}8 = 0{,}7 \text{ km}$$

Średnia jest większa od dominanty, rozkład prawostronny.

***Współczynnik asymetrii***

$$V_{as} = \frac{\bar{x} - D}{\sigma} = \frac{0{,}7}{7{,}3} = 0{,}096 \cong 0{,}1$$

Średnia leży na prawo od dominanty w odległości 0,1 odchylenia standardowego, a więc blisko.

## *Syntetyczne wnioski*

| Wyszczególnienie | Cała populacja | Wewnętrznie położone 50% |
|---|---|---|
| Zróżnicowanie | wyraźne, na pograniczu ze słabym ($V_{zk} = 25{,}6\%$) | wyraźno-słabe ($V_{zQ} = 12{,}5\%$) |
| Asymetria | wyraźna prawostronna ($\gamma_3 = 0{,}7$) | słaba prawostronna ($V_{asp} = 0{,}134$) |

Rozkład jest mocno wysmukły ($\gamma_4 = 1{,}9$) i co logiczne, niezbyt zróżnicowany ($V_{zk} = 25{,}6\%$), ale podejrzanie wyraźnie asymetryczny ($\gamma_3 = 0{,}7$). Sytuację tłumaczy fakt, że rozstęp jest równy aż siedmiu odchyleniom standardowym $\left(\frac{R}{\sigma} = \frac{51}{7{,}3} \cong 7\right)$, co zgodnie z prawem trzech sigm nie przytrafia się często rozkładom empirycznym. Asymetria prawostronna mówi, że ponad 50% pojazdów miało przebiegi mniejsze od średniej asymetrycznej.

W największym skrócie: rozkład jest mocno wysmukły i położony na aż siedmiu odchyleniach standardowych.

***Przykład 8.3.*** (dla cechy skokowej)

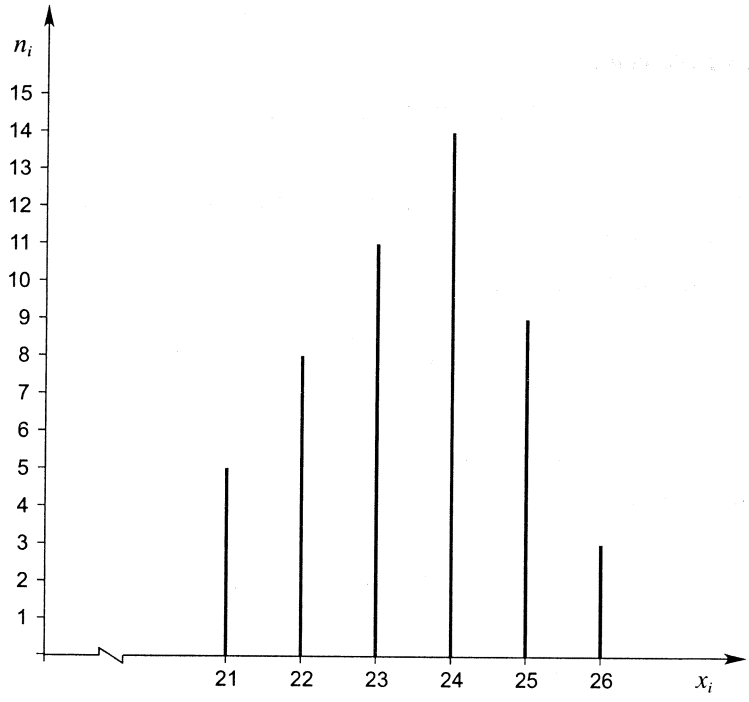
Dane o liczbie braków wyprodukowanych przez pracowników pewnego oddziału w grudniu 1999 roku przedstawiają się następująco:

| Liczba braków | 21 | 22 | 23 | 24 | 25 | 26 |
|---|---|---|---|---|---|---|
| Liczba robotników | 5 | 8 | 11 | 14 | 9 | 3 |

Źródło: dane umowne.

Scharakteryzować rozkład liczby braków.

Cecha jest skokowa, liczba wariantów cechy niewielka, stąd rozkład możemy przedstawić za pomocą wykresu linii pionowych.




Wnioski z wykresu:

— populacja jest jednomodalna,
— asymetria rozkładu jest lewostronna,
— rozkład jest spłaszczony,
— populacja jest skupiona na niewielkim obszarze zmienności, co będzie miało wpływ na niezbyt silne zróżnicowanie.

Liczymy miary klasyczne i pozycyjne (obliczenia pomocnicze w tablicy roboczej).

## Miary klasyczne

### Średnia arytmetyczna

$$\bar{x} = \frac{\sum x_i n_i}{N} = \frac{1173}{50} = 23{,}46 \cong 23{,}5 \text{ braków}$$

Gdyby każdy z pracowników oddziału wyprodukował taką samą liczbę braków byłoby to 23,5.

### Wariancja

$$Var = \frac{\sum (x_i - \bar{x})^2 n_i}{N} = \frac{94{,}50}{50} = 1{,}89 \text{ braków}^2$$

### Odchylenie standardowe

$$\sigma = \sqrt{Var} = \sqrt{1{,}89} = 1{,}375 \cong 1{,}38 \text{ braków}$$

Liczba braków przypadających na pracownika przeciętnie różni się od średniej arytmetycznej o 1,4 braku.

### Klasyczny współczynnik zmienności

$$V_{zk} = \frac{\sigma}{\bar{x}} \cdot 100\% = \frac{1{,}38}{23{,}46} \cdot 100\% = 5{,}9\%$$

Bardzo słabe zróżnicowanie cechy.

### Trzeci moment centralny

$$e_3 = \frac{\sum (x_i - \bar{x})^3 n_i}{N} = \frac{-27{,}5}{50} = -0{,}55 \text{ braków}^3$$

Ujemna asymetria rozkładu.

### Trzeci moment centralny standaryzowany

$$\gamma^3 = \frac{e_3}{\sigma^3} = \frac{-0{,}55}{1{,}38^3} = -0{,}21$$

Słaba asymetria lewostronna.

**Tablica robocza**

| $x_i$ | $n_i$ | $x_i n_i$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ | $(x_i - \bar{x})^2 n_i$ | $(x_i - \bar{x})^3 n_i$ | $(x_i - \bar{x})^4 n_i$ | Cum | |
|---|---|---|---|---|---|---|---|---|---|
| 21 | 5 | 105 | −2,5 | 6,25 | 31,25 | −78,125 | 195,3125 | 5 | |
| 22 | 8 | 176 | −1,5 | 2,25 | 18,00 | −27,000 | 40,5000 | 13 | $\leftarrow Q_1$ |
| 23 | 11 | 253 | −0,5 | 0,25 | 2,75 | −1,375 | 0,6875 | 24 | |
| 24 | 14 | 336 | 0,5 | 0,25 | 3,50 | 1,750 | 0,8750 | 38 | $\leftarrow M$ |
| 25 | 9 | 225 | 1,5 | 2,25 | 20,25 | 30,375 | 45,5625 | 47 | $\leftarrow Q_3$ |
| 26 | 3 | 78 | 2,5 | 6,25 | 18,75 | 46,875 | 117,1875 | 50 | |
| Razem | 50 | 1173 | × | × | 94,50 | −27,500 | 400,1250 | × | |

***Czwarty moment centralny***

$$e_4 = \frac{\sum (x_i - \bar{x})^4 n_i}{N} = \frac{400{,}125}{50} = 8{,}0025 \text{ braków}^4$$

***Czwarty moment centralny standaryzowany***

$$\gamma_4 = \frac{e_4}{\sigma^4} - 3 = \frac{8{,}0025}{1{,}384} - 3 = 2{,}21 - 3 = -0{,}79$$

Wyraźne spłaszczenie rozkładu.

## *Miary pozycyjne*

### *Kwartyle*

$$n_{Q_1} = 0{,}25 \cdot (N+1) = 0{,}25 \cdot 51 = 12{,}75$$

$$n_{Q_2} = 0{,}5 \cdot 51 = 25{,}5$$

$$n_{Q_3} = 0{,}75 \cdot 51 = 38{,}25$$

— ***Pierwszy kwaryl***

Dla szeregu punktowego możemy odczytać dokładne wartości kwantyli. Z kolumny kumulaty wynika, że elementy 12. i 13. uporządkowanego szeregu wynoszą po 22, więc

$$Q_1 = 22 + 0{,}75(22 - 22) = 22$$

25% pracowników wyprodukowało po nie więcej niż 22 braki, pozostałych 75% pracowników wykonało po nie mniej niż 22 braki.

— ***Mediana***

Elementy 25. i 26. w uporządkowanym szeregu wynoszą po 24.

$$M = 24 + 0{,}5(24 - 24) = 24$$

Połowa pracowników wyprodukowała w grudniu po co najwyżej 24 braki.

— ***Trzeci kwartyl***

Element 38. w uporządkowanym szeregu ma wartość 24, element 39. ma wartość 25, a $n_{Q_3} = 38{,}25$.

$$Q_3 = 24 + 0{,}25(25 - 24) = 24{,}25$$

75% robotników wyprodukowało po nie więcej niż 24,25 braków, pozostałych 25% po nie mniej niż 24,25 braków.

***Odchylenie ćwiartkowe***

$$Q = \frac{Q_3 - Q_1}{2} = \frac{24{,}25 - 22}{2} = \frac{2{,}25}{2} = 1{,}125$$

Pracownicy należący do wewnętrznych 50% populacji średnio różnią się od mediany ze względu na liczbę braków o 1,1 braku.

***Pozycyjny współczynnik zmienności***

$$V_{zQ} = \frac{Q}{M} \cdot 100\% = \frac{1{,}125}{24} \cdot 100\% = 4{,}7\%$$

Bardzo słabe zróżnicowanie wewnętrznych 50% populacji.

***Pozycyjny współczynnik asymetrii***

$$V_{asp} = \frac{Q_1 + Q_3 - 2M}{2Q} = \frac{22 + 24{,}25 - 2 \cdot 24}{2{,}25} = \frac{-1{,}75}{2{,}25} = -0{,}78$$

Silna asymetria ujemna wewnętrznych 50% populacji.

***Dominanta***

Najczęściej występującą wartość cechy odczytujemy bezpośrednio z szeregu punktowego.

$$D = 24.$$

Najwięcej robotników wyprodukowało po 24 braki.

***Wskaźnik asymetrii***

$$w_{as} = \bar{x} - D = 23{,}5 - 24 = -0{,}5$$

Potwierdza lewostronną asymetrie rozkładu.

***Współczynnik asymetrii***

$$V_{as} = \frac{\bar{x} - D}{\sigma} = \frac{-0{,}5}{1{,}38} = -0{,}36 \cong -0{,}4$$

Średnia asymetryczna leży na lewo od dominanty w odległości niespełna 0,4 odchylenia standardowego.

***Syntetyczne wnioski***

| Wyszczególnienie | Cała populacja | Wewnętrznie położone 50% |
|---|---|---|
| Zróżnicowanie | bardzo słabe ($V_{zk} = 5{,}9\%$) | bardzo słabe ($V_{zQ} = 4{,}7\%$) |
| Asymetria | słaba lewostronna ($\gamma_3 = -0{,}21$) | silna lewostronna ($V_{asp} = -0{,}78$) |

Rozkład jest bardzo słabo zróżnicowany, lewostronnie asymetryczny, wyraźnie spłaszczony. Mimo spłaszczenia zróżnicowanie jest słabe dzięki temu, że populacja leży tylko na 3,6 odchylenia standardowego $\left(\frac{R}{\sigma} = \frac{5}{1{,}38} = 3{,}6\right)$. Wewnętrzne 50% jest silnie ujemnie asymetryczne.

## 8.2. Metody graficzne w analizie struktury zbiorowości

***Przykład 8.4.***

Metodami graficznymi scharakteryzować rozkład liczby palet sprzedanych przez firmę Profi.

Polecenie realizujemy wykreślając podstawowe wykresy w prostokątnym układzie współrzędnych — wielobok liczebności, kumulatę, *box and whisker plot* (wykres pudełkowy) — i formułując wynikające z nich wnioski. Oczywiście, jeżeli na podstawie jednej określonej formy graficznej uzyskamy wystarczającą wiedzę o rozkładzie cechy, wykonywanie innych rysunków jest zbędne.

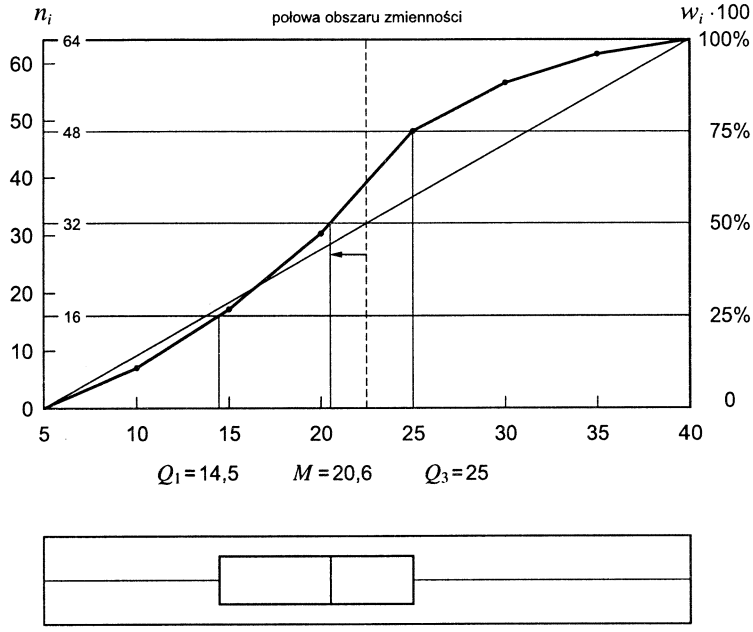
Część pierwsza rozwiązania będzie identyczna z punktem 1. przykładu 8.1.

Wykres kumulaty umożliwi:

a) precyzyjniejszą ocenę kierunku i siły asymetrii rozkładu,

b) precyzyjniejszą ocenę zróżnicowania rozkładu,

c) przez graficzne wyznaczenie wartości kwartyli sporządzenie *box and whisker plotu* i dalsze wnioski szczegółowe.




Wnioski z kumulaty:

1. Asymetria rozkładu jest słaba dodatnia (przecięcie kumulaty z prostą przechodzącą przez punkt równy połowie liczebności populacji następuje na lewo od środka obszaru zmienności — asymetria dodatnia; i niezbyt daleko od niego — asymetria słaba.
2. Zróżnicowanie populacji jest wyraźno-silne. Kumulata jest bliska przekątnej prostokąta, co oznacza bliski równomiernemu rozkład cechy, a im bardziej wysmukły jest rozkład, tym zróżnicowanie cechy jest mniejsze.

Wnioski z *box and whisker plotu*:

1. Asymetria środkowych 50% populacji jest słaba ujemna (bok lewego wewnętrznego prostokąta jest dłuższy od boku prawego — asymetria lewostronna; różnice długości nie są duże — asymetria słaba).
2. Zróżnicowanie całej populacji jest silne (wewnętrzny prostokąt stanowi około $^1/_3$ rozstępu; im ten prostokat jest mniejszy, tym zróżnicowanie słabsze).
3. Zróżnicowanie środkowych 50% populacji jest silne (z wykresu znajdujemy przybliżone wartości $Q \cong 5, M = 20 \rightarrow \frac{Q}{M} \cdot 100\% \cong 25\%$).

## 8.3. Zadania na obliczanie średniej geometrycznej

***Przykład 8.5.***

**Zasoby mieszkaniowe w Polsce. Mieszkania w tysiącach. Stan w dniu 31.12.**

| 1950 | 1970 | 1990 | 2000 | 2001 | 2002 | 2003 | 2004 |
|---|---|---|---|---|---|---|---|
| 5851 | 8081 | 11022 | 11845 | 11946 | 12438 | 12596 | 12683 |

Źródło: opracowanie własne w oparciu o dane GUS, MRS Polski 2005.

1. W jakim tempie rosła liczba mieszkań pomiędzy 1950 a 2000 rokiem?
2. W jakim tempie rosła liczba mieszkań w Polsce w okresie 2000–2004?
3. W ciągu ilu lat liczba mieszkań uległa by podwojeniu gdyby udało się utrzymać tempo z roku 2002?
4. Proszę ekstrapolować liczbę mieszkań w Polsce na koniec 2006 roku.

ad 1.

$$p = \sqrt[50]{\frac{11845}{5851}} - 1 = 1{,}0142 - 1 = 0{,}0142$$

$$s = 1{,}42\%$$

ad 2.

$$p = \sqrt[4]{\frac{12683}{11845}} - 1 = 1{,}0172 - 1 = 0{,}0172$$

$$s = 1{,}72\%$$

ad 3.

$$p = \frac{12438}{11946} - 1 = 1{,}0412 - 1 = 0{,}0412$$

$$s = 4{,}12\%$$

W 2002 r. liczba mieszkań zwiększyła się o 4,12% w porównaniu z rokiem 2001.

$$L_n = L_0(1+p)^n$$

$$L_n = 2L_0$$

$$2L_0 = L_0(1+0{,}0412)^n$$

$$2 = 1{,}0412^n$$

$$\log 2 = n \log 1{,}0412$$

$$n = \frac{\log 2}{\log 1{,}0412} = \frac{0{,}30103}{0{,}0175342} = 17{,}168 \cong 17{,}2$$

Gdyby liczba mieszkań zwiększyła się z roku na rok o 4,12%, ich zasób uległ by podwojeniu po 17,2 roku.

ad 4.

Prosta ekstrapolacja na liniowym przedłużeniu dotychczasowej tendencji, powinna znaleźć potwierdzenie w przyszłych zdarzeniach czy procesach gdy dane statystyczne są powiązane ze sobą sposób deterministyczny (np. charakteryzują je jednokierunkowe zmiany w czasie i mała zmienność). Jednoroczne tempa dla okresu 2000–2004 wynoszą odpowiednio 0,85%, 4,12%, 1,27% i 0,69%, a więc wahania są spore. W zależności od przeczuć co do stanu przyszłej koniunktury gospodarczej, znajomości rynku budownictwa mieszkaniowego można przyjmować różne założenia.

Pesymistyczna prognoza będzie oparta na tempie dla 2004 roku, wynoszącym zaledwie 0,69%.

$$L_{2006} = L_{2004} \cdot (1+0{,}0069)^2 = 12683 \cdot 1{,}0069^2 = 12858{,}6$$

Średnie tempo dla dwóch ostatnich lat szeregu czasowego: 0,98% pozwoli wyznaczyć wariant bardziej optymistyczny.

$$L_{2006} = L_{2004} \cdot (1+0{,}0098)^2 = 12683 \cdot 1{,}0098^2 = 12932{,}8$$

***Przykład 8.6.***

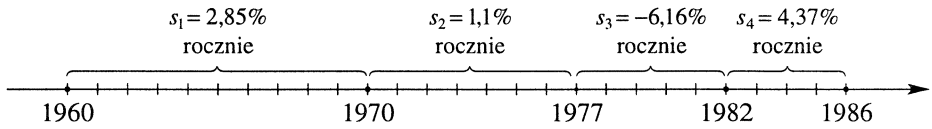
Dysponujemy danymi o produkcji tkanin bawełnianych w Polsce: w roku 1960 wyprodukowano 665 tys. km, do roku 1970 produkcja rosła po 2,85% rocznie, w latach 1970–1977 rosła w tempie 1,1%, w latach 1977–1982 malała średniorocznie po 6,16%, by dla okresu 1982–1986 ponownie rosnąć po 4,37% rocznie. Proszę obliczyć:

a. tempo zmiany dla lat 1960–1986,

b. fizyczną wielkość produkcji w roku 1986,

c. ekstrapolować wielkość produkcji w roku 1989 (przyjmijmy, że nie możemy po te dane sięgnąć do rocznika statystycznego).

Dane na osi czasu można przedstawić jak niżej




$L_{1960} = 665$ tys. km

ad a.

$$p = \sqrt[26]{(1+0{,}285)^{10} \cdot (1+0{,}011)^{7} \cdot (1-0{,}0616)^{5} \cdot (1+0{,}0437)^{4}} - 1$$

$$p = \sqrt[26]{1{,}0285^{10} \cdot 1{,}011^{7} \cdot 0{,}9384^{5} \cdot 1{,}0437^{4}} - 1 = \sqrt[26]{1{,}2346499} - 1 = 0{,}00841402$$

$$s = p \cdot 100 = 0{,}814\%$$

ad b.

$$L_{1986} = L_{1960}(1+p)^{26} = 665 \cdot (1+0{,}0081402)^{26} 665 \cdot 1{,}2346499 = 821{,}0$$

ad c.

Koniec lat 80. i początek 90. to czas kryzysu gospodarczego, utraty rynków wschodnich. Wykres 8.4 przedstawia jak kształtowała się produkcja tkanin bawełnianych i bawełnopodobnych w latach 1960–1991.

Przewidzenie z niewielkim błędem w roku 1986 wielkości produkcji w roku 1989 (wiemy teraz, że wyniosła ona 760 tys. km), a tym bardziej dla roku 1990 czy 1991 (286 tys. km) było zadaniem praktycznie niemożliwym.

Prześledźmy jednak konsekwencje kilku wariantów obliczeń.

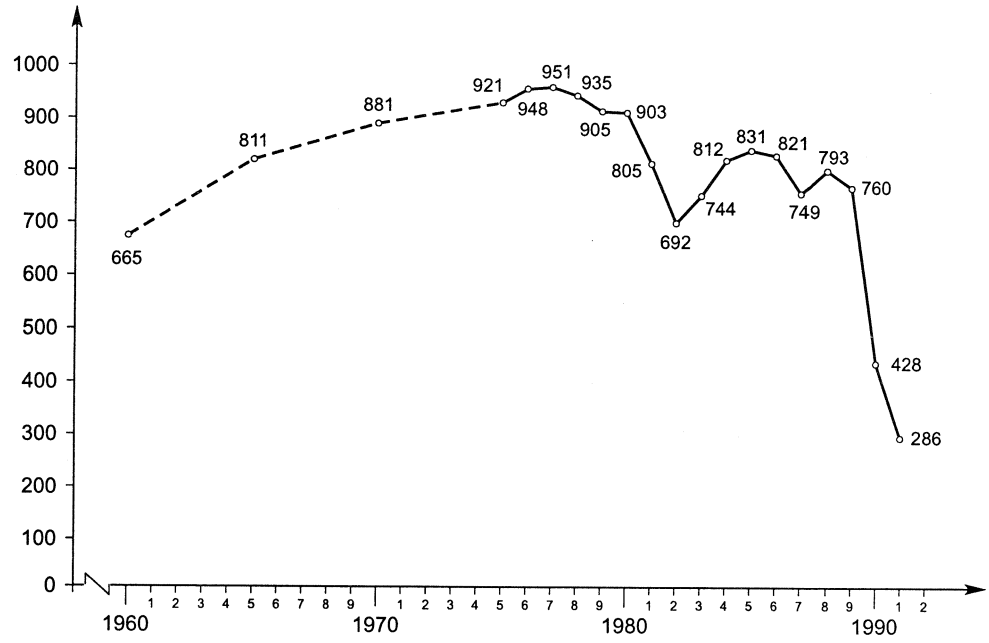
Ekstrapolacja w oparciu o średnią dla lat 1977–1986:

$$p = \sqrt[9]{(1-0{,}0616)^{5} \cdot (1+0{,}0437)^{4}} - 1 = \sqrt[9]{0{,}9384^{5} \cdot 1{,}0437^{4}} - 1 =$$

$$= \sqrt[9]{0{,}863461} - 1 = 0{,}9838 - 1 = -0{,}0162$$

$$s = -1{,}62\%$$

$$L_{1989} = L_{1986} \cdot (1+p)^{3} = 821 \cdot 0{,}9838^{3} = 781{,}7$$




Wykres 8.4. Produkcja tkanin bawełnianych i bawełnopodobnych (w tys. km) w Polsce w latach 1960–1991. Źródło: opracowanie własne na podstawie danych GUS.

Ekstrapolacja w oparciu o średnią dla lat 1960–1986 (całkowicie w tej sytuacji nieuprawniona):

$$L_{1989} = L_{1986} \cdot 1{,}00814^3 = 821 \cdot 1{,}00814^3 = 841{,}2$$

Przykład ten, podobnie jak poprzedni, miał za zadanie dobitnie pokazać, że liniową ekstrapolację można stosować w bardzo ograniczonym zakresie, nie daje ona żadnych rezultatów gdy warunki *ceteris paribus* nie są spełnione w wystarczającym stopniu.

# Rozdział 9

# Porównywanie zbiorowości statystycznych

Badaniem populacji, zarówno pod kątem dokonywania porównań jak i analizowania współzależności zjawisk, zajmują się w sposób komplementarny statystyka opisowa i statystyka matematyczna. Porównując populacje — w sposób kompleksowy — w zakresie analizy struktury należy odpowiedzieć na trzy pytania:

1. Czy porównywane populacje są do siebie podobne, czy też się różnią?
2. Na czym polegają różnice?
3. Jak duże są różnice?

Obok już zaprezentowanych narzędzi graficznych i analitycznych użyteczne będą metody statystyki matematycznej, wszędzie tam, gdzie badane populacje będą mogły być potraktowane jako pochodzące z większych całości (nadsystemów). Powyższa uwaga stosuje się zwłaszcza do pytania trzeciego i pytania pierwszego. W tym miejscu, z oczywistych, ograniczamy się do metod deskryptywnych i analitycznych statystyki opisowej.

Porównywanie populacji stwarza nowe problemy w stosunku do tych, które były omawiane w kontekście kompleksowej analizy struktury jednej zbiorowości statystycznej. Problemami tymi są w szczególności:

1. ograniczenia w stosowaniu niektórych spośród dotychczas poznanych narzędzi, ze względu na niejednoznaczność ich interpretacji;
2. potrzeba posługiwania się specyficznymi instrumentami właściwymi dla dokonywania porównań;

3. pilne zapotrzebowania ze strony pochodzących z praktyki użytkowników metod statystycznych na syntetyczny opis danych, możliwość wielostronnego spojrzenia na dane, wychwytywanie obserwacji nieprzystających do pozostałych etc., najlepiej w formie graficznej.

## 9.1. Porównania z użyciem narzędzi graficznych

***Przykład 9.1.***

Korzystając z danych umownych przedstawionych w tablicy 9.1 przeprowadzimy metodami graficznymi porównanie struktury dwóch grup pracowników ze względu na czas wykonania prostego zadania, w zależności od tego czy robotnik został przeszkolony teoretycznie i praktycznie (grupa I) czy tylko praktycznie (grupa II).

**Tablica 9.1. Rozkład czasu wykonania zadania w dwóch grupach robotników**

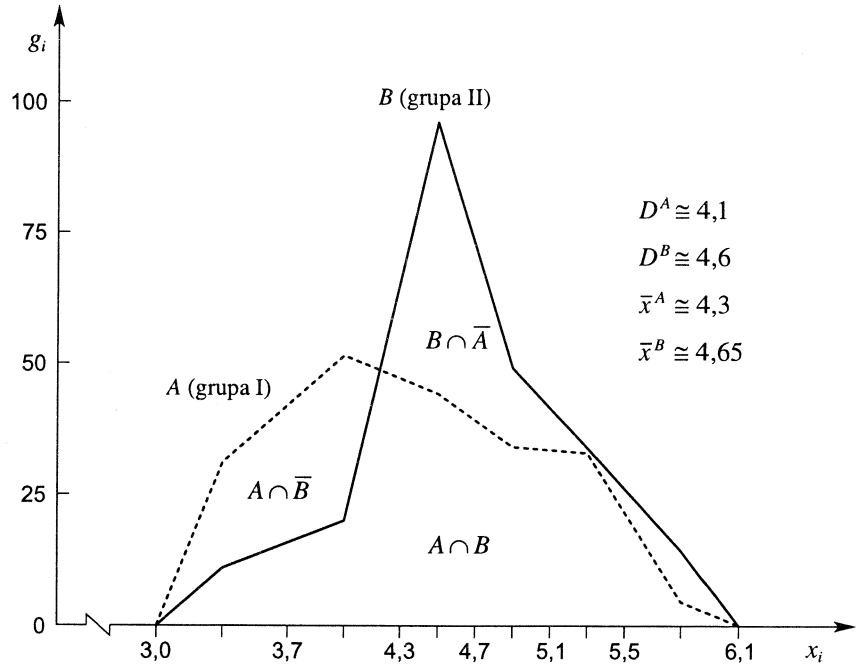
| Czas wykonania w min | Liczba robotników | |
|---|---|---|
| | Grupa I | Grupa II |
| 3,0–3,7 | 48 | 9 |
| 3,7–4,3 | 68 | 14 |
| 4,3–4,7 | 39 | 45 |
| 4,7–5,1 | 30 | 23 |
| 5,1–5,5 | 29 | 16 |
| 5,5–6,1 | 6 | 10 |
| Razem | 220 | 117 |

Źródło: dane umowne.

**Tablica 9.2. Obliczenia pomocnicze potrzebne do wykonania porównań graficznych**

| $x_{0i} - x_{1i}$ | $n_i^A$ | $n_i^B$ | $w_i^A \cdot 100\%$ | $w_i^B \cdot 100\%$ | $g_i^A$ | $g_i^B$ | Cum$w_i^A$ | Cum$w_i^B$ | $h_i$ |
|---|---|---|---|---|---|---|---|---|---|
| 3,0–3,7 | 48 | 9 | 21,82 | 7,69 | 31,2 | 11,0 | 21,82 | 7,69 | 0,7 |
| 3,7–4,3 | 68 | 14 | 30,91 | 11,97 | 51,5 | 20,0 | 52,73 | 19,66 | 0,6 |
| 4,3–4,7 | 39 | 45 | 17,73 | 38,46 | 44,3 | 96,2 | 70,46 | 58,12 | 0,4 |
| 4,7–5,1 | 30 | 23 | 13,64 | 19,66 | 34,1 | 49,2 | 84,10 | 77,78 | 0,4 |
| 5,1–5,5 | 29 | 16 | 13,18 | 13,68 | 33,0 | 34,0 | 97,28 | 91,46 | 0,4 |
| 5,5–6,1 | 6 | 10 | 2,73 | 8,55 | 4,6 | 14,3 | 100,01 | 100,01 | 0,6 |
| Razem | 220 | 117 | 100,1 | 100,1 | × | × | × | × | × |

Źródło: Obliczenia własne na podstawie tablicy 9.1.




Wykres 9.1. Wielobok liczebności dla czasu wykonania w grupach I i II

***Uwaga!*** Dzięki połączeniu skrajnych punktów wieloboków odpowiednio z dolną granicą pierwszej klasy i z górną granicą klasy ostatniej kształty rozkładów są łatwiejsze do odczytania. Przybliżone wartości średnich arytmetycznych i dominant zostały odczytane z wykresu.

Symbole $A \cap B$, $A \cap \overline{B}$, $B \cap \overline{A}$ odnoszą się odpowiednio do tych elementów w populacjach $A$ i $B$, które dla pola:

— $A \cap B$ — ze względu na wartości analizowanej cechy znajdują swoje odpowiedniki w drugiej zbiorowości,

— $A \cap \overline{B}$ — ze względu na wartości analizowanej cechy mają reprezentantów w populacji $A$, bez odpowiedników w populacji $B$,

— $B \cap \overline{A}$ — ze względu na wartości analizowanej cechy mają reprezentantów w populacji $B$, bez odpowiedników w populacji $A$.

Wnioski płynące z analizy wieloboków liczebności w odniesieniu do pierwszego z podstawowych pytań dla dokonywania porównań („Czy porównywane populacje są do siebie podobne, czy też się różnią?") są jednoznaczne. Rozkłady znacznie się

między sobą różnią. Rozkład $A$ jest bez wątpienia silnie spłaszczony, $B$ raczej zbliżony do mezokurtycznego (w każdym razie jest smuklejszy od rozkładu $A$). Oznacza to, że odchylenie standardowe dla $A$ jest większe od odchylenia standardowego dla $B$. Choć średnia arytmetyczna dla $B$ (jako środek ciężkości rozkładu) musi być większa od średniej arytmetycznej dla $A$ należy oczekiwać, że klasyczne współczynniki zmienności wykażą silniejszą zmienność populacji $A$. Ponadto rozkład $A$ jest rozkładem prawostronnie asymetrycznym (słabo), a rozkład $B$ charakteryzuje się asymetrią lewostronną, również słabą. Rozkład $B$ przypomina kształtem płetwę rekina, a więc asymetria wewnętrznie położonych 50% populacji może być dodatnia. Pole $A \cap \overline{B}$ pokazuje, że znacznie większy procent pracowników w grupie szkolonej teoretycznie i praktycznie wykonał zadanie w czasie poniżej 4,3 minuty niż procent ich odpowiedników poddanych tylko treningowi praktycznemu.

Nawet gdybyśmy nie mieli czasu na wykonanie dalszych wykresów ani policzenie żadnych miar, odpowiedź na pytanie: który ze sposobów szkolenia okazał się bardziej skuteczny? jest bardzo prosta. Dla takiej cechy jak czas wykonania zadania, której wielkość chcemy minimalizować, przy takim samym obszarze zmienności rozkładem korzystniejszym od lewostronnie asymetrycznego jest rozkład prawostronnie asymetryczny. Gdyby oba były prawostronnie asymetryczne wybralibyśmy ten o silniejszej asymetrii, przy zbliżonych wartościach $\gamma_3$ — rozkład bardziej wysmukły (mniej zróżnicowany).

W sposób analityczny podobieństwo struktur można zmierzyć posługując się tzw. wskaźnikiem podobieństwa struktur. Mierzy on część wspólną porównywanych populacji (dla omawianego przykładu pole $A \cap B$) w oparciu o wzór

$$w_{ps} = \sum_{i=1}^{k} \min\left(w_i^A, w_i^B\right)$$

(dla każdego wiersza tablicy, poza wierszem sumy, z kolumn częstości względnych $w_i^A$ lub $w_i^A \cdot 100\%$ wybieramy wielkość najmniejszą i poddajemy sumowaniu)

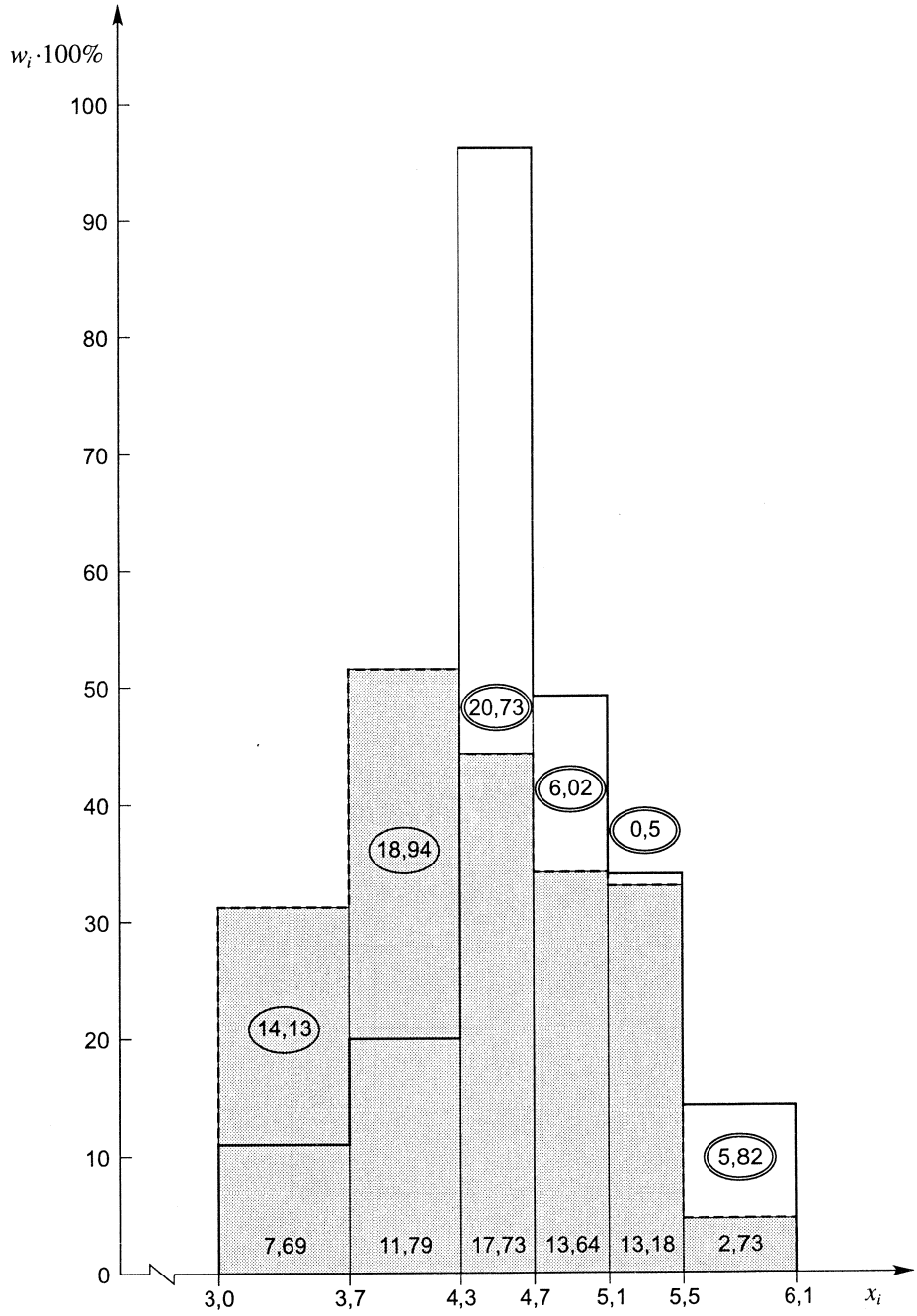
$$\begin{aligned} &\text{dla } w_i && 0 \le w_{ps} \le 1, \\ &\text{dla } w_i \cdot 100\% && 0 \le w_{ps} \le 100. \end{aligned}$$

Dla identycznych struktur wartość wskaźnika wynosi 100.

W sposób graficzny ideę wskaźnika podobieństwa struktur wyraża wykres 9.2. Dla przykładu 9.1 wskaźnik podobieństwa struktur (dla $w_i \cdot 100\%$) wynosi

$$w_{ps} = 7{,}69 + 11{,}97 + 17{,}73 + 13{,}64 + 13{,}18 + 2{,}73 = 67{,}94\%$$

$$\text{Pole } A \cap B = \text{Pole } \overline{A} \cap B = 100 - w_{ps} = 100 - 67{,}94 = 32{,}06$$




Wykres 9.2. Graficzna interpretacja wskaźnika podobieństwa struktur. Źródło: tablica 9.2.

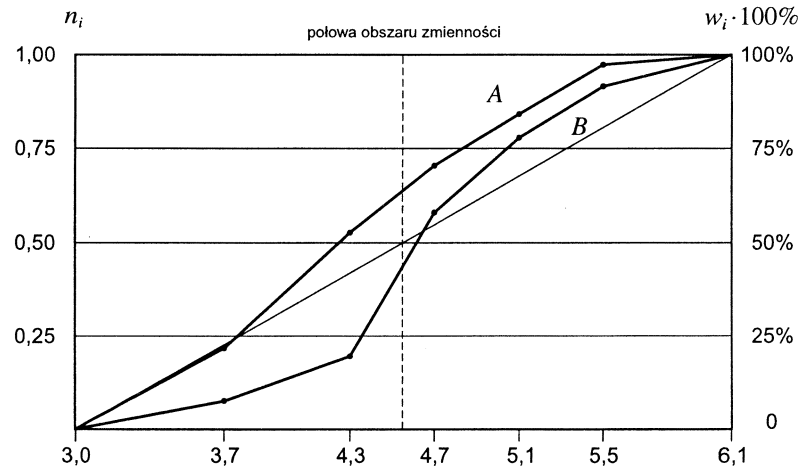
32,06% populacji $A$, ze względu na wartości badanej cechy, nie ma swoich odpowiedników w populacji $B$; 32,06% populacji $B$ nie ma swoich odpowiedników w populacji $A$; 67,94% populacji $A$ ma swoje odpowiedniki w populacji $B$. Mimo, że wskaźnik podobieństwa struktur przyjmuje wartość stosunkowo dużą, wobec charakterów różnic, w istocie możemy mówić jedynie o dość umiarkowanym podobieństwie badanych struktur.

Struktury oceniamy jako podobne gdy $w_{ps}$ przekracza 80%. To nie jest jedyna wada wskaźnika podobieństwa struktur. Z konstrukcji wzoru wynika, że porównywane populacje muszą mieć takie same granice klas, choćby po dokonaniu pewnych agregacji. Zauważmy, że w krańcowym przypadku, dwa rozkłady o identycznym kształcie, uplasowane w różnych miejscach na osi liczbowej będą miały $w_{ps}$ równy zero.

Wszystkie te zastrzeżenia powodują niezbyt częste sięganie po tę miarę w praktyce porównań statystycznych.

Drugim elementem porównań graficznych są wykresy kumulat (wykres 9.3). Z wykresu 9.3 dowiadujemy się m.in., że:

A. Kumulata dla populacji $A$ leży powyżej kumulaty dla populacji $B$. Wynika stąd, iż dla dowolnego czasu $x_i$ realizacji zadania — poza $x_{min}$ i $x_{max}$ — z wykonania zadania wywiązuje się większy odsetek pracowników szkolonych teoretycznie i praktycznie niż odsetek robotników szkolonych praktycznie. Największą różnicę obserwujemy dla czasu 4,3 minuty. Aż 52,73% pracowników grupy I wykonuje w czasie poniżej 4,3 minuty, podczas gdy w grupie II było to zaledwie 19,66% robotników. Przebieg kumulat przemawia na korzyść szkolenia teoretycznego i praktycznego.




Wykres 9.3. Czas wykonania zadania w dwóch grupach robotników — kumulaty (wieloboki liczebności skumulowanych). Źródło: tablica 9.2.

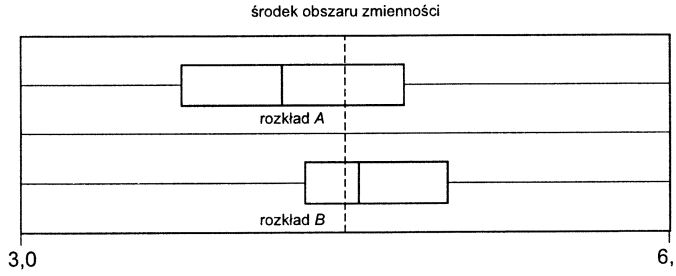
B. Asymetria rozkładu $A$ (oceniana na podstawie mediany) jest słaba dodatnia. Ponad 50% pracowników z grupy I wykonało zadanie w czasie krótszym niż średnia arytmetyczna.

C. Asymetria rozkładu $B$ (oceniana na podstawie położenia od mediany) jest słaba ujemna. Ponad 50% pracowników grupy II wykonało zadanie w czasie dłuższym niż średnia arytmetyczna. Asymetria rozkładu $B$ jest słabsza niż asymetria rozkładu $A$.

D. Kumulata dla rozkładu $A$ leży blisko przekątnej prostokąta wykresu. Potwierdza to wniosek o spłaszczeniu rozkładu i o silniejszym zróżnicowaniu rozkładu $A$ niż rozkładu $B$.

Z kumulaty odczytujemy wartości kwartyli dla porównywanych populacji i tworzymy wykres pudełkowy.

$$Q_1^A = 3{,}75 \quad M^A = 4{,}25 \quad Q_3^A = 4{,}85$$
$$Q_1^B = 4{,}35 \quad M^B = 4{,}60 \quad Q_3^B = 5{,}05$$




Wykres 9.4. Czas wykonania zadania w dwóch grupach robotników — wykres pudełkowy (*box plot*)

W porównaniu z wielobokami liczebności i kumulatami wykresy pudełkowe wzbogacają naszą wiedzę o asymetrii i zróżnicowaniu wewnętrznie położonych 50% populacji, ale też podkreśla się ich użyteczność dla porównana miedzy zbiorowościami (asymetrie i zróżnicowania całych populacji, długości obszarów zmienności zajmowanych przez ćwiartki populacji).

Aczel przypomina, że Tukey i współautorzy pracy Graphical Methods of Data Analysis byli zdania, że pewne odmiany wykresów pudełkowych można wykorzystać przy sprawdzaniu hipotezy o równości mediany w dwóch populacjach[1]: Inne

[1] A. Aczel, *Statystyka w zarządzaniu*, PWN, Warszawa 2000, s. 52.

formy wykresów pudełkowych pozwalają identyfikować dane podejrzane o nietypowość (*suspected outliers*) i dane nietypowe (*outliers*).

Na podstawie wykresu 9.4 uzyskujemy potwierdzenie prawostronnej asymetrii dla grupy I, lewostronnej asymetrii dla grupy II, silniejszej asymetrii czasu wykonania zadania w grupie I. Wewnętrzne 50% populacji jest silniej skupione w grupie II (mniejsza długość wewnętrznego prostokąta wykresu), co przekłada się na silniejsze zróżnicowanie w grupie I.

Jak widać z wykresu, w obu badanych populacjach, ma miejsce prawostronna asymetria wewnętrznych 50% populacji, silniej asymetryczne jest wnętrze grupy II, choć asymetria w obu przypadkach jest słaba.

Wykres pudełkowy zwraca naszą uwagę na to, że $Q_1^B$ jest większy od $M^A$, a więc o ile w grupie I 50% badanych wykonało zadanie w czasie nie dłuższym niż 4,25 minuty to w grupie II aż 75% badanych na realizację zadania potrzebowało nie mniej niż 4,35 minuty. Stosując metody graficzne o sile zróżnicowania wewnętrznie położonych 50% populacji możemy się wypowiedzieć podstawiając do odpowiednich wzorów wartości odczytane z wykresów.

Dla grupy I mamy przybliżone wartości

$$Q^A = \frac{4{,}85 - 3{,}75}{2} = \frac{1{,}1}{2} = 0{,}55 \qquad V_{zQ}^A = \frac{0{,}55}{4{,}25} \cdot 100\% = 12{,}9\%$$

Dla grupy II

$$Q^B = \frac{5{,}05 - 4{,}35}{2} = \frac{0{,}7}{2} = 0{,}35 \qquad V_{zQ}^B = \frac{0{,}35}{4{,}60} \cdot 100\% = 7{,}6\%$$

Stwierdzamy, że wewnętrzne 50% populacji w grupie II było zróżnicowane słabo (dzięki większej smukłości wykresu), natomiast wewnętrzne 50% populacji w grupie I było zróżnicowane wyraźnie (wykres spłaszczony).

Przybliżone wartości miar asymetrii wewnętrznych 50% populacji są następujące

$$V_{asp}^A = \frac{Q_1 + Q_3 - 2M}{2Q} = \frac{0{,}1}{1{,}1} \cong 0{,}09$$

$$V_{asp}^B = \frac{0{,}2}{0{,}7} \cong 0{,}29$$

Wnioski wyciągnięte z wykresów w syntetycznej formie przedstawiają się następująco:

| Wyszczególnienie | Cała populacja | Wewnętrznie położone 50% populacji |
|---|---|---|
| Zróżnicowanie | Trudne do jednoznacznego określenia, raczej na pograniczu słabego i wyraźnego. Gr. II słabiej zróżnicowana | Gr. I — zróżnicowanie wyraźne. Gr. II — zróżnicowanie słabe. |
| Asymetria | Gr. I — słaba asymetria dodatnia. Gr. II — słaba asymetria ujemna. Silniejsza asymetria w Gr. I | Gr. I — bardzo słaba asymetria dodatnia. Gr. II — słaba asymetria dodatnia. |

- Populacja I jest silnie spłaszczona (silniej zróżnicowana niż populacja II).
- Populacja II jest zbliżona w kształtem do rozkładu normalnego czyli mezokurtyczna.
- W przypadku populacji II stwierdzamy, że cała populacja jest ujemnie asymetryczna, natomiast wewnętrznie położone 50% populacji jest słabo dodatnio asymetryczne.

Mimo silnego spłaszczenia rozkładu czasu pracy w grupie pracowników szkolonych teoretycznie i praktycznie i związanego z nim większego zróżnicowania wyników grupa I wypadła w badaniu znacznie lepiej od osób, które odbyły tylko praktyczne przeszkolenie.

## 9.2. Porównania z użyciem miar statystycznych

***Przykład 9.2.***

Należy dokonać wszechstronnej analizy porównawczej dwóch grup pracowników pod względem czasu realizacji zadania zgodnie z tablicą 9.1. Wieloboki liczebności zostały przedstawione na wykresie 9.1.

Wnioski z wykresu są następujące:

a) obie populacje są jednomodalne,

b) rozkład *A* jest spłaszczony, lekko prawostronnie asymetryczny,

c) rozkład *B* wygląda na mezokurtyczny, jest lewostronnie asymetryczny. Kształt rozkładu nasuwa podejrzenie, że miary asymetrii dla całej populacji i dla wewnętrznie położonych 50% populacji mogą mieć przeciwne znaki,

d) siła zróżnicowania obu rozkładów jest trudna do oceny,

e) nie ma przeciwwskazań dla wyznaczenia miar klasycznych i pozycyjnych.

Miary statystyczne dla grupy I (symbolizowanej również jako *A*)

$$\bar{x} = \frac{\sum \mathring{x}_i n_i}{N} = \frac{943{,}8}{220} = 4{,}29\,\text{min}$$

Gdyby wszyscy robotnicy z grupy I wykonywali zadanie tak samo długo, byłoby to 4,29 minuty.

$$Var = \frac{\sum\left(\mathring{x}_i - \bar{x}\right)}{N} = \frac{104{,}278}{220} = 0{,}474\ \text{min}^2$$

$$\sigma = \sqrt{Var} = \sqrt{0{,}474} = 0{,}688 \cong 0{,}69\ \text{min}$$

Czas wykonania zadania przez każdego z pracowników różnił się przeciętnie od średniej arytmetycznej o 0,69 minuty.

$$V_{zk} = \frac{\sigma}{\bar{x}} \cdot 100 = \frac{0{,}69}{4{,}29} \cdot 100 = 16{,}1\%$$

Słabe zróżnicowanie cechy, choć bliskie wyraźnemu.

Tablica 9.3 przedstawia obliczenia pomocnicze potrzebne do scharakteryzowania czasu wykonania zadania przez osoby szkolone teoretycznie i praktycznie (grupa I).

$$e_3 = \frac{\sum\left(\mathring{x}_i - \bar{x}\right)^3}{N} = \frac{16{,}18}{220} = 0{,}0735$$

Dodatnia asymetria rozkładu.

$$\gamma_3 = \frac{e_3}{\sigma^3} = 0{,}225$$

Rozkład czasu wykonania zadania jest słabo prawostronnie asymetryczny.

$$e_4 = \frac{\sum\left(\mathring{x}_i - \bar{x}\right)^4}{N} = \frac{75{,}561}{220} = 0{,}343\ \text{min}^2$$

$$\gamma_4 = \frac{0343}{0{,}474^2} - 3 = 1{,}53 - 3 = -1{,}47$$

**Tablica 9.3. Obliczenia pomocnicze potrzebne do scharakteryzowania czasu wykonania zadania przez osoby szkolone teoretycznie i praktycznie (grupa I)**

| $h_i$ | $x_{0i}-x_{1i}$ | $n_i^A$ | $\mathring{x}_i$ | $\mathring{x}_i n_i$ | $\mathring{x}_i-\bar{x}$ | $(\mathring{x}_i-\bar{x})^2$ | $(\mathring{x}_i-\bar{x})^2 n_i$ | $(\mathring{x}_i-\bar{x})^3 n_i$ | $(\mathring{x}_i-\bar{x})^4 n_i$ | Cum | $g_i$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0,7 | 3,0–3,7 | 48 | 3,35 | 160,8 | –0,94 | 0,8836 | 42,4128 | –39,868 | 34,48 | 48 | 68,6 |
| 0,6 | 3,7–4,3 | 68 | 4,0 | 272,0 | –0,29 | 0,0841 | 5,7188 | –1,658 | 0,48 | 116 | 113,3 |
| 0,4 | 4,3–4,7 | 39 | 4,5 | 175,5 | 0,21 | 0,0441 | 1,7199 | 0,361 | 0,076 | 155 | 97,5 |
| 0,4 | 4,7–5,1 | 30 | 4,9 | 147,0 | 0,61 | 0,3721 | 11,163 | 6,809 | 4,154 | 185 | 75,0 |
| 0,4 | 5,1–5,5 | 29 | 5,3 | 153,7 | 1,01 | 1,0201 | 29,5829 | 29,879 | 30,178 | 214 | 72,5 |
| 0,6 | 5,5–6,1 | 6 | 5,8 | 34,8 | 1,51 | 2,2801 | 13,6806 | 20,658 | 31,193 | 220 | 10 |
| × | Razem | 220 | × | 943,8 | × | × | 104,278 | 16,18056 | 75,561 | × | × |

Źródło: Obliczenia własne na podstawie tablicy 9.1.

Rozkład silnie spłaszczony (platokurtyczny) w stosunku do rozkładu normalnego.

$$n_{Q_1} = 0{,}25 \cdot 220 = 55$$

$$n_M = 0{,}5 \cdot 220 = 110$$

$$n_{Q_3} = 0{,}75 \cdot 220 = 165$$

$$Q_1 = 3{,}7 + \frac{55-48}{68} \cdot 0{,}6 = 3{,}76 \text{ min}$$

25% robotników wykonało zadanie w czasie nie większym niż 3,76 minuty.

$$M = 3{,}7 + \frac{110-48}{68} \cdot 0{,}6 = 4{,}25 \text{ min}$$

50% robotników wykonało zadanie w czasie nie większym niż 4,25 minuty.

$$Q_3 = 4{,}7 + \frac{165-155}{30} \cdot 0{,}4 = 4{,}83 \text{ min}$$

75% robotników wykonało zadanie w czasie niż większym niż 4,83 minuty.

$$Q = \frac{Q_3 - Q_1}{2} = \frac{4{,}83 - 3{,}76}{2} = \frac{1{,}07}{2} = 0{,}535 \text{ min}$$

Czas wykonania zadania przez robotników należących do wewnętrznych 50% populacji różnił się średnio od mediany o 0,54 minuty.

$$V_{zQ} = \frac{Q}{M} \cdot 100 = \frac{0{,}535}{4{,}25} \cdot 100 = 12{,}6\%$$

Zróżnicowanie czasu w centralnych ćwiartkach populacji jest wyraźne.

$$V_{asp} = \frac{Q_1 + Q_3 - 2M}{2Q} = \frac{3{,}76 + 4{,}83 - 2 \cdot 4{,}25}{1{,}07} = \frac{0{,}09}{1{,}07} = 0{,}08$$

Wewnętrzne 50% zbiorowości jest słabo prawostronnie asymetryczne.

$$D = 3{,}7 + \frac{113{,}3 - 68{,}6}{113{,}3 - 68{,}6 + 113{,}3 - 97{,}5} \cdot 0{,}6 = 3{,}7 + \frac{44{,}7}{60{,}5} \cdot 0{,}6 = 4{,}14 \text{ min}$$

Najczęstszym czasem wykonania zadania było 4,14 minuty.

$$w_{as} = \bar{x} - D = 4{,}29 - 4{,}14 = 0{,}15 \text{ min}$$

Średnia arytmetyczna jest większa od dominanty. Wynik potwierdza prawostronną asymetrię rozkładu.

$$V_{as} = \frac{\bar{x} - D}{\sigma} = \frac{0{,}15}{0{,}69} = 22$$

Średnia arytmetyczna leży na prawo od dominanty w odległości 0,22 odchylenia standardowego, czyli dość blisko.

$$R = x_{\max} - x_{\min} = 6{,}1 - 3{,}0 = 3{,}1 \text{ min}$$

Różnica między najdłużej, a najkrócej wykonanym zadaniem wyniosła 3,1 minuty.

$$\frac{R}{\sigma} = \frac{3{,}1}{0{,}69} = 4{,}5$$

Rozkład leży na 4,5 odchylenia standardowego, a więc jest bardziej skupiony niż wynika to z oszacowania Snedcora (dla $N \cong 200$, $\frac{R}{\sigma} \cong 5{,}5$). Duże skupienie elementów populacji tłumaczy dlaczego przy silnym spłaszczeniu zróżnicowanie populacji jest słabo/wyraźne.

$$D < M < \bar{x}$$

Porządek miar tendencji centralnej jest właściwy dla rozkładu prawostronnie asymetrycznego.

Miary statystyczne dla grupy II

$$\bar{x} = \frac{\sum \mathring{x}_i n_i}{N} = \frac{544{,}15}{117} = 4{,}65 \text{ min}$$

Gdyby wszyscy robotnicy z grupy II wykonywali zadanie tak samo długo, byłoby to 4,65 minuty.

$$Var = \frac{\sum \left(\mathring{x}_i - \bar{x}\right)^2 n_i}{N} = \frac{43{,}56}{117} = 0{,}372 \text{ min}^2$$

$$\sigma = \sqrt{Var} = \sqrt{0{,}372} = 0{,}61 \text{ min}$$

Czas wykonania zadania przez każdego z pracowników różnił się przeciętnie od średniej arytmetycznej o 0,61 minuty.

**Tablica 9.4. Obliczenia pomocnicze potrzebne do scharakteryzowania czasu wykonania zadania przez robotników szkolonych praktycznie (grupa II)**

| $h_i$ | $x_{0i} - x_{1i}$ | $n_i$ | $\mathring{x}_i$ | $\mathring{x}_i n_i$ | $\mathring{x}_i - \bar{x}$ | $(\mathring{x}_i - \bar{x})^2$ | $(\mathring{x}_i - \bar{x})^2 n_i$ | $(\mathring{x}_i - \bar{x})^3 n_i$ | $(\mathring{x}_i - \bar{x})^4 n_i$ | Cum | $g_i$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0,7 | 3,0–3,7 | 9 | 3,35 | 30,15 | –1,3 | 1,69 | 15,21 | –19,773 | 25,7049 | 9 | 12,86 |
| 0,6 | 3,7–4,3 | 14 | 4,0 | 56,00 | –0,65 | 0,4225 | 5,915 | –3,84475 | 2,4991 | 23 | 23,33 |
| 0,4 | 4,3–4,7 | 45 | 4,5 | 202,50 | –0,15 | 0,0225 | 1,0125 | –0,151875 | 0,0228 | 68 | 112,50 |
| 0,4 | 4,7–5,1 | 23 | 4,9 | 112,70 | 0,25 | 0,0625 | 1,4375 | 0,359375 | 0,0898 | 91 | 57,50 |
| 0,4 | 5,1–5,5 | 16 | 5,3 | 84,80 | 0,65 | 0,4225 | 6,76 | 4,394 | 2,8561 | 107 | 40,00 |
| 0,6 | 5,5–6,1 | 10 | 5,8 | 58,00 | 1,15 | 1,3225 | 13,225 | 15,20875 | 17,4901 | 117 | 16,67 |
| × | Razem | 117 | × | 544,15 | × | × | 43,56 | –3,8075 | 48,6628 | × | × |

Źródło: Obliczenia własne na podstawie tablicy 9.1.

$$V_{zk} = \frac{\sigma}{\bar{x}} \cdot 100 = \frac{0,61}{4,65} \cdot 100 = 13,1\%$$

Czas wykonania zadania w tej grupie jest słabo zróżnicowany.

$$e_3 = \frac{\sum\left(\mathring{x}_i - \bar{x}\right)^3 n_i}{N} = \frac{-3,8075}{117} = -0,0325$$

Asymetria rozkładu jest lewostronna.

$$\gamma_3 = \frac{e_3}{\sigma^3} = \frac{-0,0325}{0,61^3} = -0,143$$

Rozkład czasu wykonania zadania jest słabo lewostronnie asymetryczny.

$$e_4 = \frac{\sum\left(\mathring{x}_i - \bar{x}\right)^4 n_i}{N} = \frac{48,6628}{117} = 0,4159 \text{ min}^4$$

$$\gamma_4 = \frac{e_4}{\sigma^4} = \frac{0,4159}{0,372^2} - 3 = 3,005 - 3 = 0,005$$

Rozkład jest praktycznie mezokurtyczny.

$$n_{Q_1} = 0,25 \cdot 117 = 29,25$$

$$n_M = 0,5 \cdot 117 = 58,5$$

$$n_{Q_3} = 0,75 \cdot 117 = 87,75$$

$$Q_1 = 4,3 + \frac{29,25 - 23}{45} \cdot 0,4 = 4,36 \text{ min}$$

25% robotników wykonało zadanie w czasie nie większym niż 4,36 minuty.

$$M = 4,3 + \frac{58,5}{45} \cdot 0,4 = 4,62 \text{ min}$$

50% robotników wykonało zadanie w czasie nie większym niż 4,62 minuty.

$$Q_3 = 4{,}7 + \frac{87{,}75 - 68}{23} \cdot 0{,}4 = 5{,}04 \text{ min}$$

75% robotników wykonało zadanie w czasie niż większym niż 5,04 minuty.

$$Q = \frac{Q_3 - Q_1}{2} = \frac{5{,}04 - 4{,}36}{2} = \frac{0{,}68}{2} = 0{,}34 \text{ min}$$

Czas wykonania zadania przez robotników należących do wewnętrznych 50% populacji różnił się średnio od mediany o 0,34 minuty.

$$V_{zQ} = \frac{0{,}34}{4{,}62} \cdot 100 = 7{,}36\%$$

Zróżnicowanie czasu wykonania zadania w centralnych ćwiartkach populacji jest słabe.

$$V_{asp} = \frac{Q_1 + Q_3 - 2M}{2Q} = \frac{4{,}36 + 5{,}04 - 2 \cdot 4{,}62}{0{,}68} = \frac{0{,}16}{0{,}68} = 0{,}24$$

Słaba prawostronna asymetria w wewnętrznych 50% populacji.

$$D = 4{,}3 + \frac{112{,}5 - 23{,}33}{112{,}5 - 23{,}33 + 112{,}5 - 57{,}5} \cdot 0{,}4 = 4{,}3 + \frac{89{,}17}{144{,}17} \cdot 0{,}4 = 4{,}55 \text{ min}$$

Najczęściej występującym czasem wykonania zadania w grupie II było 4,55 minuty.

$$w_{as} = \bar{x} - D = 4{,}65 - 4{,}55 = 0{,}10 \text{ min}$$

Średnia arytmetyczna jest większa od dominanty. Miara sugeruje prawostronną asymetrię rozkładu, co jest sprzeczne ze wskaźnikiem $\gamma_3$.

$$V_{as} = \frac{w_{as}}{\sigma} = \frac{0{,}1}{0{,}61} = 0{,}16$$

Średnia arytmetyczna leży na prawo od dominanty w odległości 0,16 odchylenia standardowego, a więc blisko.

$$R = x_{max} - x_{min} = 6{,}1 - 3{,}0 = 3{,}1 \text{ min}$$

Różnica między najdłużej, a najkrócej wykonanym zadaniem wyniosła 3,1 minuty.

$$\frac{R}{\sigma} = \frac{3,1}{0,61} = 5,1$$

Rozkład leży na 5,1 odchylenia standardowego. Wynik zbliżony do oszacowania Snedecora (dla $N \cong 100$). Duże skupienie elementów populacji tłumaczy, dlaczego przy silnym spłaszczeniu zróżnicowanie populacji jest słabo/wyraźne.

$$D < M < \bar{x}$$

Porządek miar tendencji centralnej jest właściwy dla rozkładu prawostronnie asymetrycznego, choć rozkład jest lewostronnie asymetryczny (ujemny $\gamma_3$). Różnice we wskaźnikach miar asymetrii tłumaczone są specyficznym kształtem wykresu.

**Tablica 9.5. Zestawienie miar statystycznych dla dwóch grup robotników pod względem czasu wykonania**

| Miara statystyczna | Grupa I | Grupa II |
|---|---|---|
| $\bar{x}$ | 4,29 | 4,65 |
| $\sigma$ | 0,474 | 0,372 |
| $Var$ | 0,688 | 0,61 |
| $V_{zk}$ | 16,05% | 13,12% |
| $e_3$ | 0,074 | –0,033 |
| $\gamma_3$ | 0,226 | –0,143 |
| $e_4$ | 0,343 | 0,416 |
| $\gamma_4$ | –1,74 | 0,001 |
| $\frac{R}{\sigma}$ | 4,5 | 5,1 |
| $Q_1$ | 3,76 | 4,36 |
| $M$ | 4,25 | 4,62 |
| $Q_3$ | 4,83 | 5,04 |
| $Q$ | 0,535 | 0,4 |
| $V_{zQ}$ | 12,6% | 7,4% |
| $w_{asp}$ | 0,08 | 0,24 |
| $D$ | 4,14 | 4,55 |
| $w_{as}$ | 0,15 | 0,10 |
| $\frac{\bar{x} - D}{\sigma}$ | 0,22 | 0,16 |

Znajomość charakterystyk opisowych pozwala uściślić wnioski wyciągnięte uprzednio w oparciu o metody graficzne. W szczególności możemy ocenić siłę zróżnicowania cechy w badanych populacjach.

**Syntetyczne wnioski**

| Wyszczególnienie | Całe populacje | Wewnętrznie położone 50% populacji |
|---|---|---|
| Zróżnicowanie | Silniejsze dla gr. I (16,1%) ale na pograniczu słabego i wyraźnego, $V_{zk}$ dla gr. II wyniosło 13,1%. | Wyraźne w gr. I (12,6%). Słabe w gr. II (7,4%). |
| Asymetria | Słaba dodatnia w gr. I (0,23). Słaba ujemna w gr. II (–0,14). Asymetria w gr. II słabsza niż w gr. II. | Bardzo słaba dodatnia w gr. I. Słaba dodatnia w gr. II. Asymetria w gr. II silniejsza od asymetrii w gr. I. |

Źródło: opracowanie własne.

Wszystkie miary tendencji centralnej są wyższe dla grupy II

$$\bar{x}^{I} < \bar{x}^{II}, \quad M^{I} < M^{II}, \quad D^{I} < D^{II},$$

a więc przemawiają za pierwszym sposobem szkolenia. Również kierunek asymetrii (dodatni) jest korzystny dla grupy I.

Spłaszczenie rozkładu I, wobec silnego skupienia elementów na 4,5 odchylenia standardowego, nie przemawia również za drugim sposobem szkolenia.

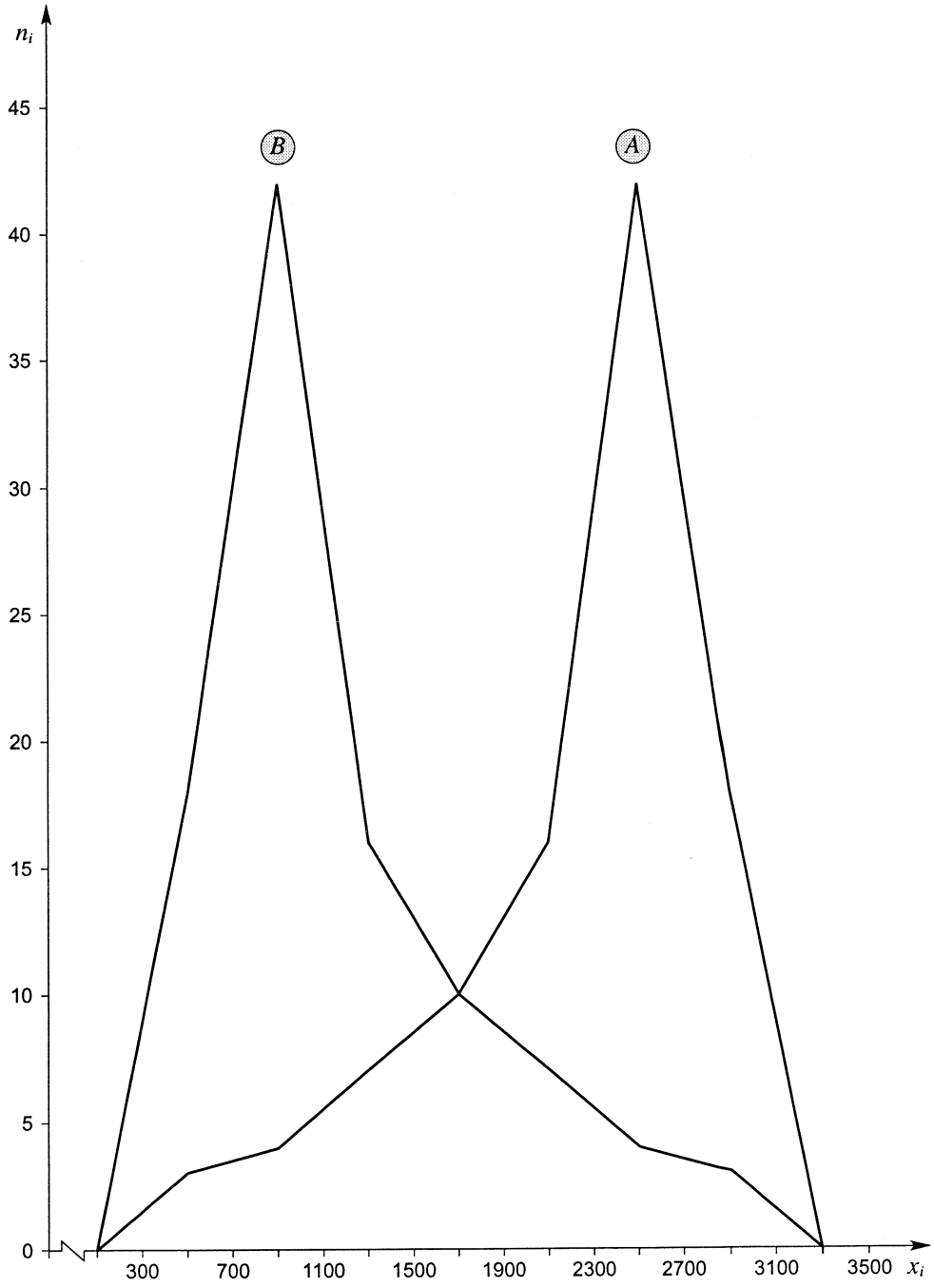
Wskaźnik podobieństwa struktur ($W_{ps} = 67{,}94$) sugeruje większe podobieństwo porównanych struktur niż ma to miejsce w rzeczywistości.

W przedstawionych przykładach porównań między zbiorowościami statystycznymi, przy okazji omawiania wskaźnika podobieństwa struktur, pojawił się problem trudności jakie sprawiają w analizie różne obszary zmienności badanej cechy. Trzeba jednak mieć świadomość, że nawet przy takich samych obszarach zmienności mogą się pojawiać się wyniki przynajmniej na pierwszy rzut oka zastanawiające.

***Przykład 9.3.***

Metodą graficzną i rachunkową zaprezentujemy specyficzne problemy jakie występują przy dokonywaniu porównań między populacjami o takim samym kształcie rozkładu leżących w różnych miejscach na osi liczbowej.

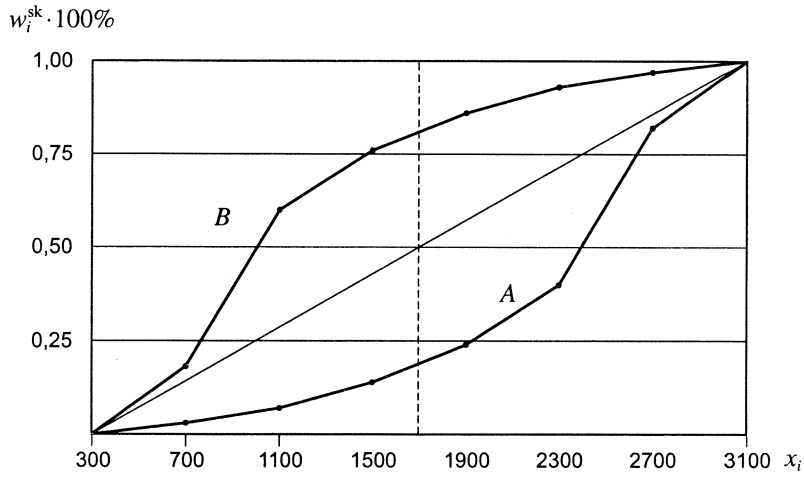
Rozpatrzymy przykład dwóch populacji, które są swoimi lustrzanymi odbiciami na osi liczbowej.



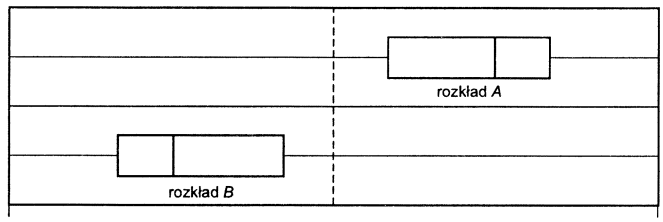

Wykres 9.5. Wieloboki liczebności do przykładu 9.3

**Tablica 9.6. Obliczenia pomocnicze do wykreślenia kumulat**

| $x_{0i} - x_{1i}$ | $n_i^A$ | $n_i^B$ | $\text{Cum}^A$ | $\text{Cum}^B$ |
|---|---|---|---|---|
| 300–700 | 3 | 18 | 3 | 18 |
| 700–1100 | 4 | 42 | 7 | 60 |
| 1100–1500 | 7 | 16 | 14 | 76 |
| 1500–1900 | 10 | 10 | 24 | 86 |
| 1900–2300 | 16 | 7 | 40 | 93 |
| 2300–2700 | 42 | 4 | 82 | 97 |
| 2700–3100 | 18 | 3 | 100 | 100 |
| Razem | 100 | 100 | × | × |




Wykres 9.6. Kumulaty do przykładu 9.3




Wykres 9.7. Wykresy pudełkowe do przykładu 9.3

Wnioski z wykresów:

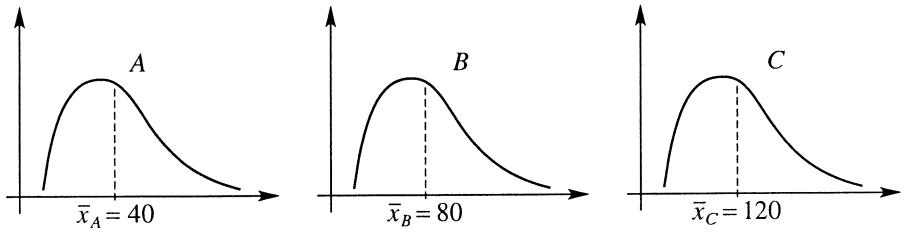
- Populacje mają taki sam obszar zmienności.
- Rozkłady są względem siebie symetryczne.
- Pamiętamy z analizy $\gamma_4$, że rozkłady względem siebie symetryczne są tak samo wysmukłe.
- Kierunki asymetrii (zarówno całych populacji jak i wewnętrznych 50%) są przeciwne, ale siła asymetrii jest taka sama.
- Skoro wieloboki są swoimi lustrzanymi odbiciami, odchylenia standardowe w obu populacjach muszą być takie same.

Miary różne dla badanych rozkładów:

- dominanty,
- średnie arytmetyczne,
- klasyczne współczynniki zmienności (bo różne $\bar{x}$, przy takich samych $\sigma$),
- pozycyjne współczynniki zmienności (bo różne $M$ przy takich samych $Q$),
- kwartyle.

Otrzymane wyniki w pełni potwierdzają oczekiwania sformułowane na podstawie analizy wykresów. Rzecz jasna, do tych samych wniosków możemy dojść wychodząc od matematycznej postaci wzorów i warunku wzajemnej symetrii rozkładów.

Warto było rozpatrzeć powyższy przykład, aby się przekonać jak ostrożnie należy podchodzić do interpretacji tak ważnej miary jaką jest $V_{zk}$. Klasyczny współczynnik zmienności — uznawany przez wielu za narzędzie oceny jednorodności rozkładu — będzie przyjmował dla takiego samego odchylenia standardowego bardzo różne wartości w zależności od tego, jaka jest wartość średniej arytmetycznej dla danej populacji.




$$\sigma_A = \sigma_B = \sigma_C = 6{,}4$$

$$V_{zk}^A = \frac{6{,}4}{40} \cdot 100 = 16\%, \quad V_{zk}^B = \frac{6{,}4}{80} \cdot 100 = 8\%, \quad V_{zk}^A = \frac{6{,}4}{120} \cdot 100 = 5{,}3\%$$

**Tablica 9.7. Zestawienie charakterystyk opisowych dla populacji z przykładu 9.3**

| Miara statystyczna | Populacja *A* | Populacja *B* |
|---|---|---|
| $\bar{x}$ | 2220,00 | 1180,00 |
| $\sigma$ | 602,66 | 602,66 |
| $V_{zk}$ | 27,15% | 51,07% |
| $e_3$ | $-2{,}48 \cdot 10^7$ | $2{,}48 \cdot 10^7$ |
| $\gamma_3$ | −1,13 | 1,13 |
| $\gamma_4$ | 0,66 | 0,66 |
| $Q_1$ | 1925,00 | 766,67 |
| $M$ | 2395,24 | 1004,76 |
| $Q_3$ | 2633,33 | 1475,00 |
| $Q$ | 354,17 | 354,17 |
| $V_{zQ}$ | 14,79 | 35,25 |
| $D$ | 2508,00 | 892,00 |
| $x - D$ | −288,00 | 288,00 |
| $\frac{\bar{x} - D}{\sigma}$ | −0,478 | 0,478 |
| $\frac{Q_1 + Q_2 - 2M}{2Q}$ | −0,328 | 0,328 |
| $\frac{R}{\sigma}$ | 4,65 | 4,65 |

Źródło: obliczenia własne.

Rozkład *A* nie jest ani na jotę bardziej zróżnicowany od rozkładu *C*.

Trzeba więc pamiętać, że jeśli chcemy wykorzystać wiedzę o wartościach klasycznych współczynników zmienności do dokonania porównań siły zróżnicowania musimy wziąć pod uwagę obszary zmienności cechy (cech) oraz kierunki asymetrii rozkładów.

# Rozdział 10

# Miary związku między zmiennymi ilościowymi:

## analiza regresji i współczynnik korelacji liniowej

W dotychczasowych rozważaniach o metodach statystycznych uwzględniliśmy tylko takie z nich, które służyły do badania zbiorowości statystycznych przy uwzględnieniu jednej cechy. Przypomnijmy, że populacje analizowane ze względu na jedną cechę nazywa się jednowymiarowymi lub jednocechowymi. Dla każdej z jednostek dowolnej zbiorowości statystycznej liczba cech, czyli tego co orzeka się o danej rzeczy odpowiadając na pytanie „Jaka ona jest?", jest większa niż jeden. W wielu przypadkach np. dla człowieka jest ona nieskończona. Każda z cech może być poddana badaniu statystycznemu i stanowić analizowaną zmienną. Przykładowo populacja studentów może być badana ze względu na wyniki w nauce; pracownicy firmy ze względu na staż pracy, zarobki, wykształcenie czy wydajność; pacjenci ze względu na wiek, ciśnienie krwi, wagę ciała, poziom dobrego cholesterolu; przedsiębiorstwa z uwagi na rodzaj prowadzonej działalności, wielkość zysku operacyjnego, kapitał własny, przychody etc.

Dział statystyki zajmujący się analizą współzależności zjawisk stawia sobie za cel stwierdzenie, czy między zmiennymi występują zależności, a jeśli one zachodzą, to jaka jest ich siła i kierunek. W odróżnieniu od badań eksperymentalnych, w których manipulujemy zmiennymi niezależnymi aby poznać ich relacje przyczynowe ze zmiennymi zależnymi, w statystyce opisowej analizujemy współzmienność zmien-

nych[1]. Przedmiotem naszego zainteresowania jest najczęściej jedna zmienna, którą nazywamy zmienną wyjaśnianą lub zależną od jednej lub więcej zmiennych wyjaśniających nazywanych też predyktorami.

Dla zmiennych ilościowych[2], do których się tutaj ograniczymy, metodą analizy współzmienności jest analiza regresji z wyliczaniem współczynnika korelacji liniowej. W większości podręczników autorzy ograniczają się do przedstawienia narzędzi analizy związku dwóch zmiennych. Ponieważ wiele problemów jakie przychodzi nam rozwiązywać ma wielowymiarowy charakter, w pracy przedstawiono także regresję wieloraką (z więcej niż jednym predyktorem).

## 10.1. Związek liniowy między zmiennymi ilościowymi

Najbardziej popularną miarą opisującą związek dwóch zmiennych ilościowych, jest **współczynnik korelacji**. Mierzy on (po podniesieniu do kwadratu) **stopień dopasowania** rozkładu dwóch zmiennych do **linii prostej** (linii regresji).

Współczynnik korelacji liniowej pozwala na określenie wielkości i kierunku zależności między zmiennymi. Z sytuacją badania związku pomiędzy zmiennymi ilościowymi mamy do czynienia wtedy, gdy np. mierzymy motywację do pracy respondentów i ich zarobki i chcemy określić siłę tego związku — mówiąc inaczej — siłę wzajemnej współzmienności. Chcemy odpowiedzieć na pytanie o to, czy w ogóle istnieje istotny związek pomiędzy motywacją a zarobkami, oraz jaki ma charakter (dodatni/ujemny) i jaka jest jego siła. Nie będziemy jednak mogli rozstrzygnąć kierunku wpływu (tego czy motywacja wpływa na zarobki, czy zarobki wpływają na motywację).

Rozkład łączny dwóch zmiennych przedstawiamy w postaci wykresu korelacyjnego (nazywanego też wykresem rozrzutu), w którym na osi $OX$ są podane wartości zmiennej $X$, na osi $OY$ wartości zmiennej $Y$.

Zastanówmy się, czy istnieje związek między wysokością zarobków a wykształceniem (liczba lat nauki). Wykres korelacyjny (rozrzutu) przedstawia się następująco. Każda osoba biorąca udział w badaniu została oznaczona przez punkt o współrzędnych $X = edu$ $Y = zarobki$. Widzimy, że tym samym wartościom $X$ odpowiadają różne wartości $Y$. Związek między $X$ a $Y$, o ile istnieje, ma charakter **statystyczny**.

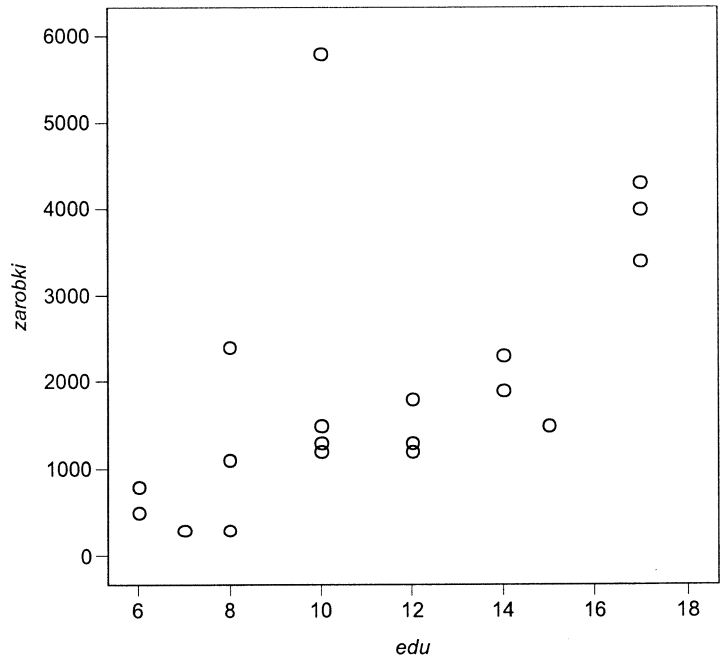
---

[1] Por. G. Wieczorkowska, J. Wierzbiński, *Badania sondażowe i eksperymentalne. Wybrane zagadnienia*, Wydawnictwa Naukowe Wydziału Zarządzania Uniwersytetu Warszawskiego, Warszawa 2005.

[2] Metody analizy związku między zmiennymi jakościowymi można znaleźć np. w: A. Aczel, *Statystyka w zarządzaniu*, PWN, Warszawa 2000.

**Tablica 10.1. Liczba lat nauki i zarobki — przykład dydaktyczny**

| | Wykształcenie (w latach nauki) | Zarobki (PLN) |
|---|---|---|
| 1 | 10 | 5800 |
| 2 | 6 | 500 |
| 3 | 8 | 300 |
| 4 | 12 | 1800 |
| 5 | 14 | 2300 |
| 6 | 15 | 1500 |
| 7 | 17 | 4000 |
| 8 | 17 | 3400 |
| 9 | 14 | 1900 |
| 10 | 10 | 1500 |
| 11 | 17 | 4300 |
| 12 | 8 | 1100 |
| 13 | 8 | 1100 |
| 14 | 8 | 2400 |
| 15 | 7 | 300 |
| 16 | 6 | 800 |
| 17 | 12 | 1200 |
| 18 | 12 | 1300 |
| 19 | 10 | 1300 |
| 20 | 10 | 1200 |




Wykres 10.1. Wykres korelacyjny (rozrzutu) wykształcenia i zarobków

Ze związkiem **funkcyjnym** (funkcją) mamy do czynienia, gdy jednej wartości $X$ odpowiada dokładnie jedna wartość $Y$ (np. wiek osoby jest jednoznacznie wyznaczony przez jej rok urodzenia).

Badania społeczne pokazują, że osoby, które mają takie same wykształcenie różnią się wysokością zarobków, co oznacza, że związek ma charakter statystyczny. Nie oznacza to, że rezygnujemy z próby opisania tej zależności przez związek funkcyjny i przewidywania jednej (takiej samej) wysokości zarobków dla wszystkich, którzy mają ten sam poziom wykształcenia.

### *Przewidywanie zmiennej objaśnianej ($Y'$) na podstawie jednego predyktora. Błąd predykcji*

Co powinniśmy zrobić, jeżeli chcemy przewidywać zarobki danej osoby nie mając żadnych dodatkowych informacji, a do tego jesteśmy karani finansowo za nietrafne predykcje? Jeżeli nie chcemy stracić, to najbezpieczniej byłoby dla każdej osoby przewidzieć... wynik średni w populacji, ponieważ, jak pamiętamy, z własności średniej wynika że

$$\sum\left(Y-\overline{Y}\right)^2 = \text{minimum}$$

Najlepszym predyktorem wysokości zarobków w sytuacji braku informacji jest średnia, czyli $\overline{Y} = 1900$.

Jeżeli jednak coś wiemy o osobach, których zarobki przewidujemy, np. znamy ich wykształcenie to możemy spróbować dopasować do tych wyników linię prostą, która pozwoli nam wyliczyć równanie regresji, na podstawie którego, będziemy mogli później przewidywać wysokość zarobków na podstawie wykształcenia.

Możemy sprawdzać różne predykcje dotyczące zarobków na podstawie informacji o wykształceniu np. zgadując, że zarobki mogą zależeć od wykształcenia w następujący sposób:

predykcja 1: $Y_1' = 1900$ (przewidujemy taką samą wartość dla wszystkich osób)

predykcja 2: $Y_2' = 100X - 300$

predykcja 3: $Y_3' = 200X - 400$

Równanie liniowe można zapisać w postaci ogólnej $Y = bX + a$.

Równanie prostej $\boldsymbol{Y = bX + a}$ zawiera dwa parametry, gdzie $\boldsymbol{a}$ wskazuje na miejsce przecięcia z osią $OY$, natomiast $\boldsymbol{b}$ informuje nas o kącie nachylenia prostej względem osi $OX$.

- Jeżeli $\boldsymbol{b = 0}$, to prosta jest równoległa do $OX$ (oznacza to, że dla każdej osoby przewidujemy tę samą wartość równą $a$ niezależnie od jej wykształcenia).

- Jeżeli $b < 0$, to wiemy, że wraz ze wzrostem $X$, maleje $Y$ (związek jest ujemny).
- Jeżeli $b > 0$, to wiemy, że wraz ze wzrostem $X$, rośnie $Y$ (związek jest dodatni).

Znając $b$ możemy mówić o kierunku zależności między zmiennymi.
Prostą $Y' = bX + a$ nazywamy prostą regresji, zaś $a$ i $b$ — współczynnikami regresji

predykcja 1: $Y' = 1900$ $(b = 0,\ a = 1900)$
predykcja 2: $Y' = 100X - 300$ $(b = 100,\ a = -300)$
predykcja 3: $Y' = 200X - 400$ $(b = 200,\ a = -400)$

Jak mówiliśmy wcześniej, możemy rozważać różne proste, różne predykcje dla $Y$. Jak wybrać najlepszą? Dobrą miarą dobroci predykcji może być *reszta*, mówiąca nam ile się pomyliliśmy, czyli jak bardzo rzeczywiste wartości $Y$ różnią się od wartości przewidywanych

$$reszta = Y - Y'$$

Dla pewnych osób reszta będzie dodatnia (gdy ich rzeczywiste zarobki są wyższe od przewidywanych na podstawie wykształcenia), dla innych — ujemna (gdy ich zarobki są niższe od przewidywanych).

Przykład: Jeżeli Jan zarabia $Y = 2400$, to *reszta* wyniesie:

$$reszta1 = [(\text{zarobki Jana}) - (\text{średnia } Y)] = 2400 - 1900 = Y - \overline{Y} = 500$$

Jeżeli wiemy, że Jan ma wykształcenie średnie (12 lat nauki) to zgodnie z równaniami predykcje wielkości jego zarobków wyniosą:

predykcja 2: zarobki Jana $= 100 \cdot 12 - 300 = 900$ $reszta2 = 2400 - 900 = 1500$
predykcja 3: zarobki Jana $= 200 \cdot 12 - 400 = 2000$ $reszta3 = 2400 - 2000 = 400$

Ponieważ nie interesuje nas znak reszty, dobrym pomysłem jest podniesienie różnic do kwadratu i zsumowanie ich po osobach.

W tablicy 10.2 przedstawione są wartości zmiennych (zarobki, wykształcenie) dla 20 osób i trzy przewidywane dla każdej z nich wysokości zarobków. Dla każdej z predykcji dla każdej z osób została wyliczona *reszta*. Różnice między wartościami przewidywanymi i rzeczywistymi zostały podniesione do kwadratu, a następnie zsumowane po wszystkich osobach, dając błąd predykcji.

$$błąd1 = \sum (Y - \overline{Y})^2 = \sum (Y - 1900)^2 = 39640000$$

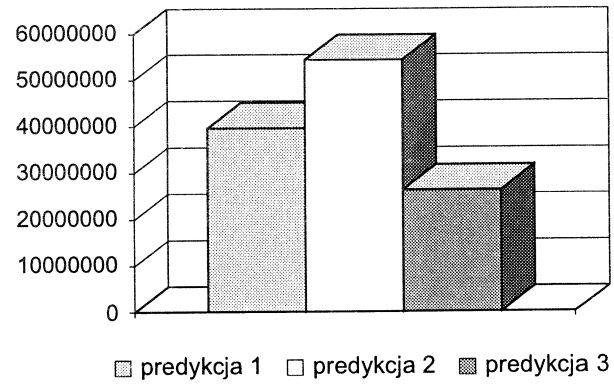
$$błąd2 = \sum (Y - Y')^2 = \sum (Y - 100X + 300)^2 = 54330000$$

$$błąd3 = \sum (Y - Y')^2 = \sum (Y - 200X + 400)^2 = 26240000$$

**Tablica 10.2. Reszty regresji dla trzech predykcji**

| Numer osoby | Liczba lat nauki | zarobki | $(Y-\overline{Y})$ | $(Y-\overline{Y})^2$ | predykcja 2 | $Y-Y'$ | $(Y-Y')^2$ | predykcja 3 | $Y-Y'$ | $(Y-Y')^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 10 | 5800 | 3900 | 15210000 | 700 | 5100 | 26010000 | 1600 | 4200 | 17640000 |
| 2 | 6 | 500 | –1400 | 1960000 | 300 | 200 | 40000 | 800 | –300 | 90000 |
| 3 | 8 | 300 | –1600 | 2560000 | 500 | –200 | 40000 | 1200 | –900 | 810000 |
| 4 | 12 | 1800 | –100 | 10000 | 900 | 900 | 810000 | 2000 | –200 | 40000 |
| 5 | 14 | 2300 | 400 | 160000 | 1100 | 1200 | 1440000 | 2400 | –100 | 10000 |
| 6 | 15 | 1500 | –400 | 160000 | 1200 | 300 | 90000 | 2600 | –1100 | 1210000 |
| 7 | 17 | 4000 | 2100 | 4410000 | 1400 | 2600 | 6760000 | 3000 | 1000 | 1000000 |
| 8 | 17 | 3400 | 1500 | 2250000 | 1400 | 2000 | 4000000 | 3000 | 400 | 160000 |
| 9 | 14 | 1900 | 0 | 0 | 1100 | 800 | 640000 | 2400 | –500 | 250000 |
| 10 | 10 | 1500 | –400 | 160000 | 700 | 800 | 640000 | 1600 | –100 | 10000 |
| 11 | 17 | 4300 | 2400 | 5760000 | 1400 | 2900 | 8410000 | 3000 | 1300 | 1690000 |
| 12 | 8 | 1100 | –800 | 640000 | 500 | 600 | 360000 | 1200 | –100 | 10000 |
| 13 | 8 | 1100 | –800 | 640000 | 500 | 600 | 360000 | 1200 | –100 | 10000 |
| 14 | 8 | 2400 | 500 | 250000 | 500 | 1900 | 3610000 | 1200 | 1200 | 1440000 |
| 15 | 7 | 300 | –1600 | 2560000 | 400 | –100 | 10000 | 1000 | –700 | 490000 |
| 16 | 6 | 800 | –1100 | 1210000 | 300 | 500 | 250000 | 800 | 0 | 0 |
| 17 | 12 | 1200 | –700 | 490000 | 900 | 300 | 90000 | 2000 | –800 | 640000 |
| 18 | 12 | 1300 | –600 | 360000 | 900 | 400 | 160000 | 2000 | –700 | 490000 |
| 19 | 10 | 1300 | –600 | 360000 | 700 | 600 | 360000 | 1600 | –300 | 90000 |
| 20 | 10 | 1200 | –700 | 490000 | 700 | 500 | 250000 | 1600 | –400 | 160000 |
| Razem | × | × | 0 | 39640000 | × | 21900 | 54330000 | × | 1800 | 26240000 |

Błędy dla trzech analizowanych predykcji są przedstawione na wykresie 10.2. Widzimy, że błąd wynikający z predykcji 2 jest większy niż wtedy gdy dla wszystkich osób przewidujemy tę samą wartość (średnią równą 1900). Predykcja 3 charakteryzuje się najmniejszym błędem.




Wykres 10.2. Błąd (suma kwadratów różnic między wartością przewidywaną na podstawie wykształcenia a rzeczywistą wysokością zarobków) dla trzech predykcji dla 20 osób

Dzięki pakietom statystycznym, to komputer szuka takich wartości *a* i *b*, aby błąd był najmniejszy. Nie musimy więc zgadywać, jakie wartości należy podstawić pod *a* i *b*, aby znaleźć najlepszą predykcję. Zostało udowodnione, że dla każdego wykresu korelacyjnego możemy jednoznacznie wyznaczyć linię prostą, która daje minimalny błąd. Trzeba jednak pamiętać, że różne metody wyznaczania *a* i *b* (np. najmniejszych kwadratów czy największej wiarygodności) mogą dać różne proste. W badaniu PGSS[3] dla 453 pracujących osób pochodzących z badań w 2005 reprezentatywnej próby dorosłych Polaków ta najlepsza prosta ma równanie:

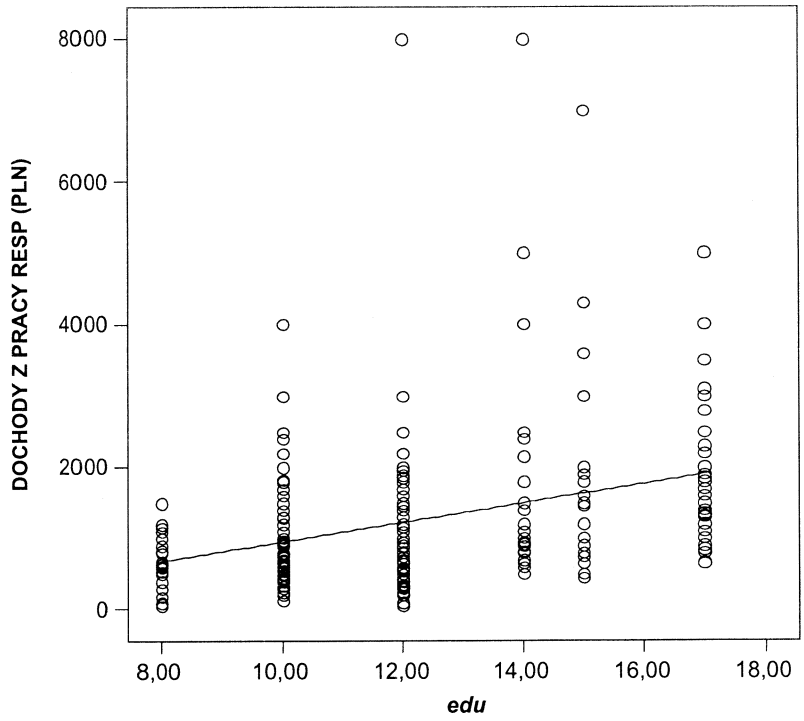
$$Y' = 135{,}82X - 406{,}66$$

gdzie:

$Y'$ — wysokość zarobków przewidywana na podstawie wykształcenia,
$X$ — oznaczone na wykresie jako *edu* — liczba lat nauki (wykształcenie),
$b = 135{,}82$ — współczynnik kierunkowy (nachylenia) linii regresji,
$a = -406{,}66$ — wyraz wolny (punkt przecięcia z osią zarobków) [stała].

[3] B. Cichomski, *Polskie Generalne Sondaże Społeczne: skumulowany komputerowy zbiór danych 1992–2005*, Instytut Studiów Społecznych, Uniwersytet Warszawski, Warszawa 2005.

Interpretując współczynnik kierunkowy równania regresji powiemy, że różnica wykształcenia o rok jest związana z różnicą 135 zł w miesięcznym dochodzie. Interpretacja wyrazu wolnego musi być przeprowadzona ostrożnie. Choć można powiedzieć, że wyraz wolny jest wartością zmiennej wyjaśnianej gdy $X = 0$, to jest to nieuzasadniona ekstrapolacja, ponieważ w analizowanym zbiorze najmniejsza wartość $X$ wynosiła 8. Nie ma sensu też przewidywanie ujemnej wartości $Y$, ponieważ zarobki z natury rzeczy nie przyjmują wartości ujemnych.




Wykres 10.3. Dopasowanie prostej dla danych z Polskiego Generalnego Sondażu Społecznego w roku 2005

## Standaryzowane współczynniki regresji

Zmienna $Y$ była mierzona w złotówkach, zmienna $X$ w latach nauki. Aby uciec od różnych jednostek, w których są wyrażone te zmienne, dobrym pomysłem jest wystandaryzowanie ich. Gdy zmienne są wystandaryzowane linia prosta przechodzi przez początek układu współrzędnych, czyli punkt (0,0), gdzie $a = 0$.

Równanie regresji dla zmiennych wystandaryzowanych ma postać

$$Z_{Y'} = \beta \cdot Z_X$$

$\beta$ nazywana jest standaryzowanym współczynnikiem regresji.

Standaryzowane współczynniki (wagi) regresji $\beta = b\frac{s_X}{s_Y}$

$\beta = 0{,}2$ — oznacza, że wzrost wartości $X$ o jedno odchylenie standardowe[4] jest równoznaczny ze wzrostem wartości $Y$ o 0,2 odchylenia standardowego.

$\beta = -0{,}5$ — oznacza, że wzrost wartości $X$ o jedno odchylenie standardowe jest równoznaczny ze spadkiem wartości $Y$ o 0,5 odchylenia standardowego.

W badaniu PGSS $\beta = 0{,}397$ co oznacza, że równanie $\boldsymbol{Y' = 135{,}82X - 406{,}66}$ możemy zapisać w postaci $\boldsymbol{Z_{Y'} = 0{,}397 \cdot Z_X}$.

Oznacza to, że różnica jednego odchylenia standardowego w wykształceniu odpowiada różnicy wynoszącej 0,397 odchylenia standardowego zarobków.

### *Współczynnik korelacji liniowej*

Znajomość równania regresji pozwala na przewidywanie wyników. Czy jednak „najlepsza prosta" jest dobra? Potrzebna jest miara, która określi jak dobrze linia prosta opisuje nasz zbiór danych. Tą miarą jest współczynnik korelacji.

Jeżeli do prawej strony równania $Y - \overline{Y} = Y - \overline{Y}$ dodamy i odejmiemy wartość przewidywaną $Y'$

$$(Y - \overline{Y}) = (Y - Y') + (Y' - \overline{Y})$$

to otrzymana równość oznacza, że:

odchylenie całkowite = odchylenie regresyjne + błąd

[4] Odchylenie standardowe wyliczane na podstawie próby oznaczamy przez $s$.

Jeżeli wszystkie trzy odchylenia podniesiemy do kwadratu i zsumujemy[5] po wszystkich danych punktach to otrzymamy następującą równość

$$\sum(Y-\bar{Y})^2 = \sum(Y-Y')^2 + \sum(Y'-\bar{Y})^2$$

gdzie:

$\sum(Y'-\bar{Y})^2$ — oznacza zmienność całkowitą zmiennej $Y$,

$\sum(Y'-\bar{Y})^2$ — różnica między przewidywaniem zakładającym związek między zmiennymi ($Y'$), a przewidywaniem zakładającym brak związku między zmiennymi określana jako wariancja wyjaśniona przez zmienną $X$,

$\sum(Y-Y')^2$ — to zsumowane reszty regresji określane jako wariancja niewyjaśniona przez zmienną $X$ — błąd regresji.

Jako miarę siły związku miedzy $X$ i $Y$ możemy wykorzystać stosunek zmienności wyjaśnionej do zmienności całkowitej nazywany **współczynnikiem determinacji**, informującym nas o proporcji zmienności $Y$ wyjaśnianej przez zmienność $X$:

$$r^2 = \frac{\sum(Y'-\bar{Y})^2}{\sum(Y-\bar{Y})^2}$$

Np. jeżeli zmienna $X$ koreluje z $Y$ na poziomie 0,3, to znaczy, że $r^2=(0,3)^2=0,09$, a to oznacza, że 9% zmienności zmiennej $X$ możemy wyjaśnić poprzez zmienność zmiennej $Y$. Pierwiastek z powyższego wyrażenia z odpowiednim znakiem, to współczynnik korelacji liniowej obu zmiennych.

$$r = \mp\sqrt{\frac{\sum(Y'-\bar{Y})^2}{\sum(Y-\bar{Y})^2}}$$

**Współczynnik determinacji $r^2$** jest interpretowany jako stopień zmienności $Y$ wyjaśnianej przez zmienność $X$.

[5] Dla uproszczenia wzorów pominięto indeksy.

Zależność między współczynnikiem korelacji $r$ a współczynnikiem nachylenia linii regresji $b$ przedstawia się następująco:

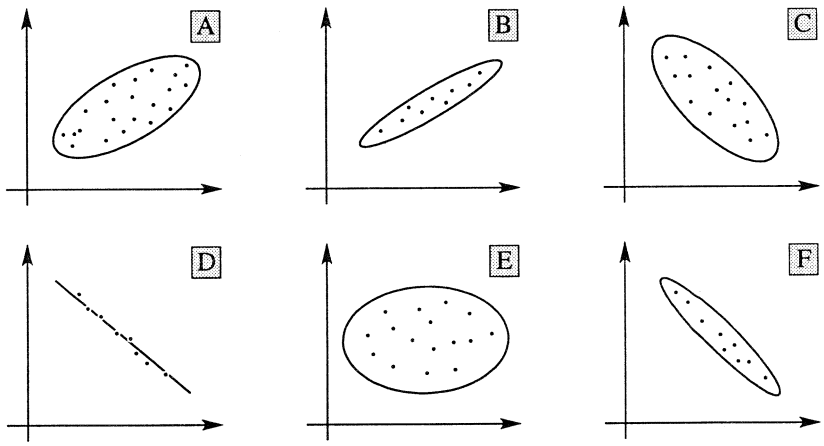
$$b_Y = r\frac{s_Y}{s_X}$$

Wyraz wolny linii regresji wynosi:

$$A_Y = \overline{Y} - b_Y\overline{X}$$

Warto zauważyć, że $\beta = r$, ale tylko w przypadku zależności między dwiema zmiennymi. Współczynnik korelacji można zdefiniować także odwołując się do miary współzmienności dwóch zmiennych.

Oglądając wykresy rozrzutu (patrz wykres 10.4) możemy określić na ile dopasowanie linii prostej do danych ma sens i czy mamy do czynienia ze związkiem dodatnim ($b > 0$) czy ujemnym.




Wykres 10.4. Przykładowe wykresy rozrzutu dla różnych wartości współczynnika korelacji

Najwyższą korelację dodatnią możemy przewidywać dla diagramu B, najsilniejszą korelację ujemną dla diagramu D, diagram E ilustruje brak związku liniowego.

### Kowariancja

Innym sposobem przedstawienia związku pomiędzy zmiennymi ilościowymi jest określenie ich współzmienności wyrażonej wzorem:

$$SS_{XY} = \sum(X - \overline{X})(Y - \overline{Y})$$

gdzie $X$, $Y$ są wartościami zmiennych, a $\overline{X}$ i $\overline{Y}$ odpowiednio ich średnimi.

Kiedy podzielimy współzmienność przez odpowiadającą im liczbę stopni swobody, co w naszym przypadku oznacza liczebność próby minus jeden (oczywiste jest, że oba pomiary $X$ i $Y$ dotyczą tej samej próby), otrzymamy wzór na kowariancję:

$$\mathrm{cov}_{XY} = \frac{SS_{XY}}{n-1}$$

Powyższa miara nie nadaje się do porównywania siły dwóch różnych związków, ponieważ uzależniona jest od jednostek skali, na których opisane są zmienne. Użycie tej miary np. do porównywania siły związku motywacji i zarobków, oraz zarobków i samooceny pozycji społecznej jest nieuprawnione ze względu na różne jednostki skali tych zmiennych. Możemy uniknąć uzależnienia od jednostek skali, jeżeli będziemy stosować zmienne w postaci standaryzowanej, przekształcone w $z$. W ten sposób wprowadzamy wzór na współczynnik **korelacji liniowej *r* Pearsona**. Jest on miarą współzmienności dwóch zmiennych $X$ i $Y$ wyrażonych w postaci standaryzowanej. Gdy:

$$z_X = \frac{(X - \overline{X})}{s_X} \quad \text{i} \quad z_Y = \frac{(X - \overline{Y})}{s_Y}$$

wtedy współczynnik korelacji ma postać:

$$r_{XY} = \frac{\sum z_X z_Y}{n-1}$$

Wartość współczynnika korelacji $r$ zawiera się w granicach $\langle 0; 1 \rangle$, gdzie **1** lub **−1** oznacza maksymalną siłę liniowego związku pomiędzy zmiennymi, **0** oznacza brak liniowego związku pomiędzy badanymi zmiennymi.

***Interpretacja współczynnika korelacji***

W podręcznikach można spotkać następujące określenia słowne siły związku liniowego:

| | |
|---|---|
| $r = 0$ | — zmienne nie są skorelowane, |
| $0 < r < 0{,}1$ | — korelacja nikła, |
| $0{,}1 \leq r < 0{,}3$ | — korelacja słaba, |
| $0{,}3 \leq r < 0{,}5$ | — korelacja przeciętna, |
| $0{,}5 \leq r < 0{,}7$ | — korelacja wysoka, |
| $0{,}7 \leq r < 0{,}9$ | — korelacja bardzo wysoka, |
| $0{,}9 \leq r < 1$ | — korelacja prawie pełna. |

Trzeba jednak podkreślić, że można porównywać wielkości współczynników korelacji tylko z prób o zbliżonej liczebności. Dla próby 12 elementowej współczynniki korelacji, których wartość bezwzględna jest mniejsza od 0,576 nie różnią się istotnie od zera, dla próby 102 elementowej analogiczna wartość krytyczna wynosi 0,1637. W tablicy 10.3 pokazano progowe wartości, jakie musi osiągnąć wartość bezwzględna współczynnika korelacji aby mógł być on uznany za istotnie większy od zera, w zależności od liczebności próby.

**Tablica. 10.3. Wartości bezwzględne współczynnika korelacji istotnie różnego od zera w zależności od liczebności próby**

| $n$ — liczebność próby | $df$ (stopnie swobody) $= n-2$ | Wartość progowa (krytyczna) $R$ |
|---|---|---|
| 12 | 10 | 0,5760 |
| 22 | 20 | 0,4227 |
| 32 | 30 | 0,3494 |
| 42 | 40 | 0,2573 |
| 52 | 50 | 0,2306 |
| 62 | 60 | 0,2108 |
| 72 | 70 | 0,1954 |
| 82 | 80 | 0,1829 |
| 92 | 90 | 0, 1726 |
| 102 | 100 | 0,1638 |

Należy też zawsze pamiętać, że współczynnik korelacji mówi nam jedynie o możliwości dopasowania zbioru danych do linii prostej. Na wykresie 10.5 przedstawiono wykresy korelacyjne czterech różnych grup wyników[6], dla których współczynnik korelacji wyników jest taki sam i wynosi $r = 0{,}816$. We wszystkich przypadkach zmienne mają takie same średnie $\overline{X} = 9, \overline{Y} = 7{,}5$, równanie regresji jest dokładnie takie samo: $Y' = 3 + 0{,}5X$, choć zbiory wyników różnią się dramatycznie.

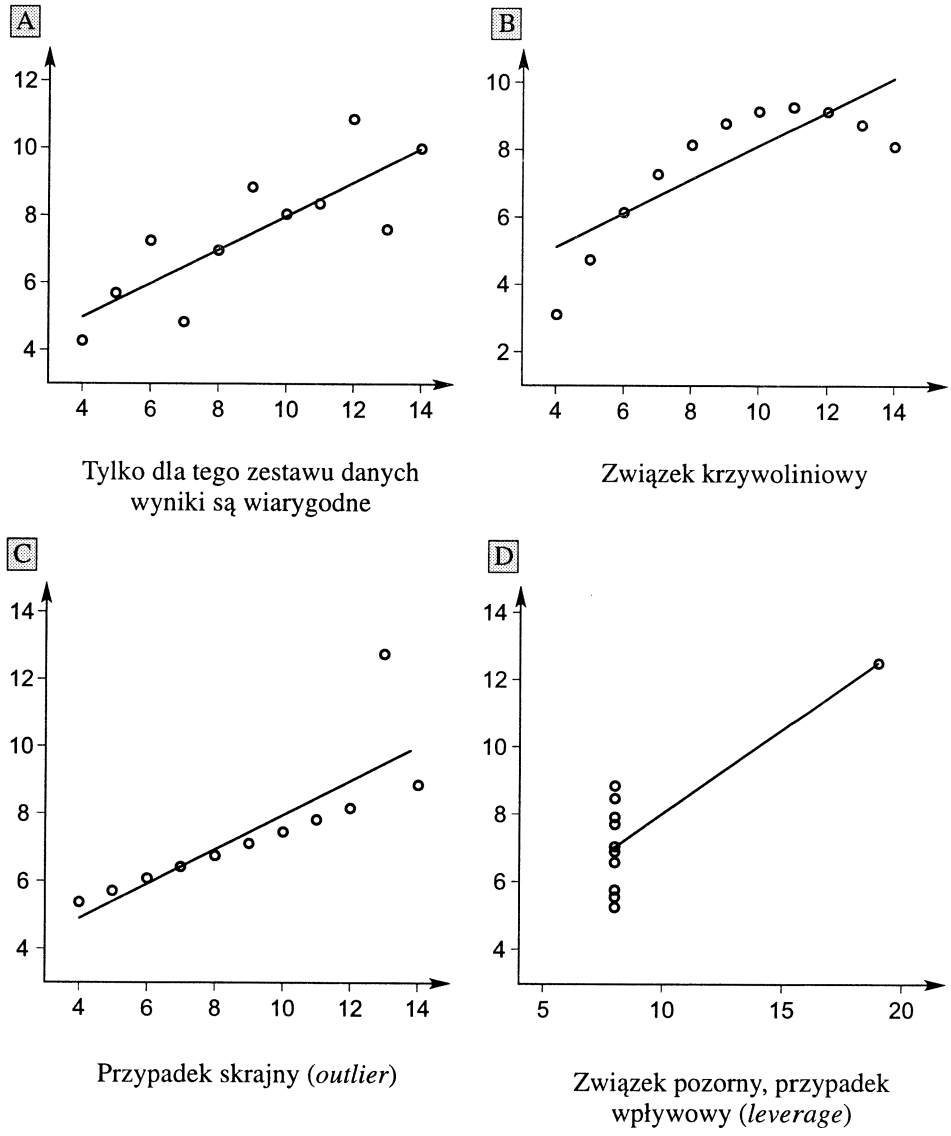
Pierwszy i podstawowy wniosek brzmi: **należy oglądać dane**[7]. Współczynnik korelacji liniowej służy do badania siły i kierunku związku liniowego między pomiarami, reprezentowanego przez linię prostą. Zatem adekwatny jest tylko do danych

[6] Przedstawione w: F.J. Anscombe, Graphs in Statistical Analysis. The American Statistician, 27, 17–21, 1973.

[7] O metodach wizualizacji danych można przeczytać w: G. Wieczorkowska, P. Kochański, M. Eljaszuk, *Statystyka: wprowadzenie do analizy badań sondażowych i eksperymentalnych*, Wydawnictwo Naukowe Scholar, Warszawa 2005.

typu (A). Gdy związek nie jest liniowy, możemy zastosować regresję krzywoliniową, czyli zamiast linii prostej dopasowywać krzywą.

Powszechnym błędem popełnianym przez badaczy jest stwierdzanie braku związku miedzy zmiennymi, gdy współczynnik korelacji jest nie istotnie statystycznie różny od zera. W takim wypadku możemy stwierdzić jedynie **brak związku linowego**.




Tylko dla tego zestawu danych wyniki są wiarygodne

Związek krzywoliniowy

Przypadek skrajny (*outlier*)

Związek pozorny, przypadek wpływowy (*leverage*)

Wykres 10.5. Przykład danych Anscombe'a. Ta sama linia regresji może być „najlepszym" dopasowaniem do bardzo różnych zestawów danych

Innym problemem jest omawiana dokładnie w literaturze wrażliwość współczynnika korelacji na przypadki skrajne i ograniczenie zakresu zmienności zmiennej niezależnej[8].

## 10.2. Regresja wielokrotna. Określanie związku zmiennej objaśnianej z więcej niż jednym predyktorem

Jednokrotna analiza regresji (jedna zmienna zależna, jeden predyktor) daje się łatwo uogólnić na przypadek **regresji wielokrotnej** (wielorakiej). Dopasowujemy wtedy nie prostą, ale hiperpłaszczyznę regresji opisaną w przypadku dwuwymiarowym równaniem:

$$Y = a + b_1 X_1 + b_2 X_2$$

lub gdy zmienne są w postaci standaryzowanej równaniem:

$$Z_Y = \beta_1 Z_{X_1} + \beta_2 Z_{X_1}$$

Dla ułatwienia zapisu w dalszych wzorach:

$r_{Y_1}$ — oznacza korelację zmiennej objaśnianej $Y$ z $X_1$,
$r_{Y_2}$ — oznacza korelację zmiennej objaśnianej $Y$ z $X_2$,
$r_{12}$ — oznacza korelację predyktorów $X_1$ i $X_2$,
$\beta_1 = \beta_{1.2}$ — oznacza współczynnik regresji $X_1$ przy kontroli $X_2$,
$\beta_2 = \beta_{2.1}$ — oznacza współczynnik regresji $X_2$ przy kontroli $X_1$.

$$\beta_1 = \beta_{1.2} = \frac{r_{Y_1} - r_{Y_2} r_{12}}{1 - r_{12}^2}, \qquad \beta_2 = \beta_{2.1} = \frac{r_{Y_2} - r_{Y_1} r_{12}}{1 - r_{12}^2}$$

$$a = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2$$

$$b_1 = \beta_1 \frac{s_Y}{s_1}, \quad b_2 = \beta_2 \frac{s_Y}{s_2}$$

[8] Por. J. Cohen, P. Cohen, *Applied multiple regression/correlation analysis for the behavioral sciences*, Lawrence Erlbaum, Hillsdale 1983.

Regresja wielokrotna ma olbrzymie znaczenie, ponieważ pozwala ocenić **ważność każdego z predyktorów** przy **kontrolowaniu wpływu pozostałych predyktorów**. Wagi β (standaryzowane współczynniki regresji) pokazują o ile zmienia się zmienna zależna wskutek standaryzowanej zmiany predyktora przy zachowanej kontroli pozostałych zmiennych, i dzięki testom istotności pozwalają ocenić, który z predyktorów jest istotny.

Możemy także policzyć **współczynnik korelacji wielokrotnej *R*** między wartościami zmiennej objaśnianej a wartościami przewidywanymi na podstawie równania regresji (kombinacją liniową predyktorów). Współczynnik ten podniesiony do kwadratu, pozwala ocenić procent zmienności zmiennej objaśnianej, wyjaśniony łącznie przez dany zestaw predyktorów.

$$R^2_{Y(12)} = \frac{r^2_{Y_1} + r^2_{Y_2} - 2r_{Y_1}r_{Y_2}r_{12}}{1 - r^2_{12}}$$

***Przykład 10.1.***

Zarobki na uczelni w zależności od stażu pracy i liczby publikacji.

Korzyści z wprowadzeniem do równania regresji więcej niż jednego predyktora możemy prześledzić na przykładzie[9] zawartym w tablicy 10.4. Predyktorami były: $X_1$ i $X_2$ oznaczające odpowiednio liczbę lat na uczelni po doktoracie (staż) i liczbę publikacji. Zmienną wyjaśnianą była wysokość zarobków.

***Przykład 10.2.***

Analizy regresji wysokości zarobków w zależności od płci, wieku, wykształcenia i liczby dzieci.

Analizę przeprowadzono na 495 osobach pracujących, biorących udział w badaniach w 2005 roku reprezentatywnej próby dorosłych Polaków Polskiego Generalnego Sondażu Społecznego[10], którzy podali swoje średnie dochody w miesiącu w ostatnim roku (zmienna: *zarobki*), wiek (zmienna: *age*), liczbę dzieci (zmienna: *dzieci*), płeć (zmienna: *płeć*) i wykształcenie w latach nauki (zmienna: *edu*).

Celem przeprowadzonej analizy regresji było wyjaśnienie zmienności wysokości zarobków przez 4 predyktory. Wielkość współczynników regresji i $R^2$ zależy m.in. od

---

[9] Przykład pochodzi z książki: J. Cohen, P. Cohen, *op. cit.*

[10] B. Cichomski, *op. cit.*

**Tablica 10.4. Porównanie współczynników regresji z jednym i dwoma predyktorami**

| Liczba predyktorów | Predyktory | Reszty regresji podniesione do kwadratu | Wyliczone równanie regresji | Interpretacja współczynników regresji |
|---|---|---|---|---|
| 1 | $X_1$ — staż | $(Y' - Y)^2 = (b_1X_1 + a_1 - Y)^2$ | $Y = \$520X_1 + \$20411$ | Każdy dodatkowy rok pracy po doktoracie związany jest z przyrostem wynagrodzenia średnio o **$520** |
| 1 | $X_2$ — liczba publikacji | $(Y' - Y)^2 = (b_2X_2 + a_2 - Y)^2$ | $Y = \$566X_2 + \$21106$ | Każda dodatkowa publikacja związana jest z przyrostem średnio wynagrodzenia o **$566** |
| 2 | $X_1$ — staż<br>$X_2$ — liczba publikacji | $(Y' - Y)^2 = (b_1X_1 + b_2X_2 + a - Y)^2$ | $Y = \$479X_1 + \$88X_2 + \$20138$ | Jeżeli przewidując zarobki uwzględnimy dwa predyktory równocześnie: każda dodatkowa publikacja związana jest z przyrostem wynagrodzenia średnio jedynie o **$88**, każdy dodatkowy rok pracy po doktoracie związany jest z przyrostem wynagrodzenia średnio o **$479** |

wielkości próby, więc musimy brać pod uwagę wyniki testów istotności. Tablica 10.5 zawiera estymatory współczynników regresji (standaryzowane i nie), współczynnik korelacji wielokrotnej i odpowiednie błędy standardowe statystyk potrzebne do testowania ich istotności. Współczynnik korelacji wielokrotnej wyliczony na podstawie próby jest zbyt „optymistycznym" estymatorem współczynnika korelacji wielokrotnej w populacji.

Skorygowane $R^2$ bierze pod uwagę relację między liczbą predyktorów i liczbą obserwacji. Na podstawie $N$ obserwacji można bowiem bardzo prosto wyjaśnić zmienność $Y$ za pomocą $p = N - 1$ predyktorów, ale tego typu równanie nie wnosi nic istotnego.

Wzór na skorygowane $R^2$ przedstawia się następująco:

$$\text{skorygowane } R^2 = R^2 - \frac{p(1-R^2)}{N-p-1}$$

Jeżeli liczba obserwacji ($N$) wynosi 100 a liczba predyktorów ($p$) równa jest 10 to korekta $R^2$ spowoduje zmianę $R^2 = 0{,}7$ na 0,67, zaś $R^2 = 0{,}4$ na 0,33. Jeżeli jednak liczba obserwacji przewyższa liczbę predyktorów tylko dwukrotnie ($N = 20\ p = 10$) to po korekcie $R^2 = 0{,}7$ będzie wynosić 0,37 zaś skorygowane $R^2 = 0{,}4$ przyjmie wartość spoza dopuszczalnego zakresu (będzie ujemne).

Wyniki analizy regresji zawiera tablica 10.5.

**Tablica 10.5. Wyniki omawianego przykładu analizy regresji zarobków z czterema predyktorami**

| $R$ | $R^2$ | Skorygowane $R^2$ | Błąd standardowy oszacowania | | |
|---|---|---|---|---|---|
| ,436 | ,190 | ,185 | 823,639 | | |
| predyktory | Współczynniki niestandaryzowane | | Współczynniki standaryzowane | $t$ | $p$ |
| | $B$ | Błąd standardowy | β | | |
| (Stała) | −1092,798 | 227,804 | | −4,797 | ,000 |
| *dzieci* | 44,133 | 30,756 | ,064 | 1,435 | ,152 |
| *age* | 12,305 | 3,712 | ,145 | 3,314 | ,001 |
| *edu* | 167,508 | 13,786 | ,489 | 12,150 | ,000 |
| *płeć* | −585,497 | 71,764 | −,320 | −8,159 | ,000 |

Analiza regresji (tablica 10.5) wysokości zarobków wykazała istotny wpływ wieku, wykształcenia i płci respondenta. Przeprowadzona analiza osobno dla kobiet i mężczyzn wykazała interesujące różnice, które są przedstawione w tablicy 10.6.

**Tablica 10.6. Wyniki analizy regresji zarobków z czterema predyktorami przeprowadzonej osobno dla grupy mężczyzn i kobiet**

| płeć | $R$ | $R^2$ | Skorygowane $R^2$ | Błąd standardowy oszacowania | | |
|---|---|---|---|---|---|---|
| mężczyźni | ,513 | ,263 | ,255 | 940,919 | | |
| kobiety | ,556 | ,309 | ,300 | 459,123 | | |
| | | Współczynniki niestandaryzowane | | Współczynniki standaryzowane | | |
| | | $B$ | Błąd standardowy | β | $t$ | Istotność |
| mężczyźni | (Stała) | −1595,793 | 358,624 | | −4,450 | ,000 |
| | *dzieci* | 111,348 | 54,171 | ,131 | 2,055 | ,041 |
| | *age* | 7,619 | 6,170 | ,078 | 1,235 | ,218 |
| | *edu* | 217,759 | 22,856 | ,510 | 9,527 | ,000 |
| kobiety | (Stała) | −856,925 | 219,670 | | −3,901 | ,000 |
| | *dzieci* | −37,806 | 25,560 | −,095 | −1,479 | ,141 |
| | *age* | 14,958 | 3,302 | ,278 | 4,530 | ,000 |
| | *edu* | 105,578 | 12,232 | ,524 | 8,632 | ,000 |

Istotnym predyktorem zarobków dla mężczyzn okazało się być wykształcenie i liczba dzieci (im więcej tym wyższe zarobki). Dla kobiet β dla liczby dzieci jest ujemna (co sugerowałoby, że im mniej dzieci, tym wyższe zarobki) ale nie jest istotna statystycznie.

W przypadku kobiet zarówno wiek jak i wykształcenie są istotnymi predyktorami wysokości zarobków. Więcej zarabiają kobiety starsze i lepiej wykształcone. Wiek nie jest istotnym predyktorem zarobków dla mężczyzn (jednak inerakcja wieku i płci nie była istotna statystycznie).

Te różnice sugerują wprowadzenie do równania regresji interakcji płci z wykształceniem i liczbą dzieci. Zanim to zrobimy warto sprawdzić różnice w analizowanych zmiennych w grupie mężczyzn i kobiet. W tablicy 10.7 znajdują się statystyki opisowe (średnie, odchylenia standardowe i przedziały ufności), w tablicy 10.8 współczynniki korelacji.

**Tablica 10.7. Statystyki opisowe dla analizowanych zmiennych**

| Nazwa zmiennej | Grupa | *N* | Średnia | Odchylenie standardowe | Błąd standardowy | 95% przedział ufności dla średniej | | Minimum | Maksimum |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Dolna granica | Górna granica | | |
| *zarobki* | mężczyźni | 270 | 1452,11 | 1090,212 | 66,348 | 1321,48 | 1582,74 | 50 | 8000 |
| | kobiety | 225 | 1010,08 | 548,587 | 36,572 | 938,01 | 1082,15 | 50 | 4000 |
| | ogółem | 495 | 1251,19 | 912,260 | 41,003 | 1170,62 | 1331,75 | 50 | 8000 |
| *dzieci* | mężczyźni | 270 | 1,50 | 1,281 | ,078 | 1,35 | 1,65 | 0 | 7 |
| | kobiety | 225 | 1,72 | 1,385 | ,092 | 1,53 | 1,90 | 0 | 8 |
| | ogółem | 495 | 1,60 | 1,332 | ,060 | 1,48 | 1,72 | 0 | 8 |
| *edu* | mężczyźni | 270 | 11,8037 | 2,55104 | ,15525 | 11,4980 | 12,1094 | 8,00 | 17,00 |
| | kobiety | 225 | 12,6889 | 2,72244 | ,18150 | 12,3312 | 13,0465 | 8,00 | 17,00 |
| | ogółem | 495 | 12,2061 | 2,66442 | ,11976 | 11,9708 | 12,4414 | 8,00 | 17,00 |
| *age* | mężczyźni | 270 | 40,7556 | 11,10776 | ,67600 | 39,4246 | 42,0865 | 20,00 | 76,00 |
| | kobiety | 225 | 39,5911 | 10,20994 | ,68066 | 38,2498 | 40,9324 | 20,00 | 60,00 |
| | ogółem | 495 | 40,2263 | 10,71405 | ,48156 | 39,2801 | 41,1724 | 20,00 | 76,00 |

**Tabela 10.8. Współczynniki korelacji miedzy analizowanymi zmiennymi liczone w całej grupie i osobno dla mężczyzn i kobiet**

| | | liczba dzieci | | | *edu* | | | *age* | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | cała próba | tylko mężczyźni | tylko kobiety | cała próba | tylko mężczyźni | tylko kobiety | cała próba | tylko mężczyźni | tylko kobiety |
| *zarobki* | *r* | –,019 | ,083 | –,181 | ,397 | ,480 | ,495 | ,115 | ,106 | ,117 |
| | *p* | ,671 | ,176 | ,006 | ,000 | ,000 | ,000 | ,011 | ,082 | ,081 |
| *dzieci* | *r* | | | | –,258 | –,178 | –,379 | ,475 | ,547 | ,405 |
| | *p* | | | | ,000 | ,003 | ,000 | ,000 | ,000 | ,000 |
| *edu* | *r* | | | | | | | –,159 | –,085 | –,235 |
| | *p* | | | | | | | ,000 | ,164 | ,000 |

Zarobki mężczyzn w analizowanej grupie są istotnie wyższe od zarobków kobiet ($F = 30{,}5$ $p < 0{,}001$). Pracujące kobiety są istotnie lepiej (dłużej) wykształcone niż badani mężczyźni ($F = 13{,}9$ $p < 0{,}001$). Obie grupy nie różnią się istotnie wiekiem i liczbą posiadanych dzieci.

Analiza prostych korelacji (tablica 10.8) wskazuje, że w przypadku mężczyzn związek zarobków z liczbą dzieci jest dodatni ale nieistotny, w przypadku zaś kobiet istotnie ujemny.

Wykształcenie koreluje w obu grupach ujemnie z liczbą dzieci (ale silniej dla kobiet) i dodatnio z wysokością zarobków. Wykształcenie koreluje z wiekiem istotnie ujemnie tylko w grupie kobiet (starsze są gorzej wykształcone niż młodsze).

Analiza regresji pozwala na wprowadzenie do modelu zmiennych jakościowych jeżeli przyjmują one dwie wartości[11]. Tak właśnie zrobiliśmy wprowadzając zmienną *płeć*.

Można też wprowadzać interakcję zmiennych będącą iloczynem wycentrowanych zmiennych. Centrowanie polega na odejmowaniu od każdego wyniku miary tendencji centralnej (średniej, mediany lub środkowego punktu skali).

W tablicy 10.9 przedstawiono wyniki analizy regresji z i bez interakcji predyktorów. W kroku 1. wprowadzono wszystkie predyktory (płeć, liczba dzieci, wykształcenie, wiek). W kroku 2. wprowadzono dwie interakcje: płci z liczbą dzieci oraz płci z wykształceniem.

Wprowadzenie interakcji płci z liczbą dzieci (zmienna: *płeć_dzieci*) i wykształceniem (*płeć_edu*) spowodowało istotny wzrost procentu wyjaśnionej wariancji (zmiana $R^2$). Ujemne współczynniki regresji dla interakcji możemy interpretować następująco:

- kobiety mające większą liczbę dzieci i mężczyźni z mniejszą liczbą dzieci zarabiają mniej niż mężczyźni z większą liczbą dzieci i kobiety z mniejszą liczbą dzieci,
- lepiej wykształceni mężczyźni i słabiej wykształcone kobiety zarabiają więcej niż słabiej wykształceni mężczyźni i lepiej wykształcone kobiety.

[11] Wprowadzenie zmiennych jakościowych przyjmujących większą liczbę wartości wymaga zakodowania ich w postaci zmiennych instrumentalnych, o czym można przeczytać np. w: A. Aczel, *op. cit.*

**Tablica 10.9. Wyniki analizy regresji zarobków z wprowadzonymi interakcjami**

| Krok | $R$ | $R^2$ | Skorygowane $R^2$ | Błąd standardowy oszacowania | Statystyki zmiany | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Zmiana $R^2$ | Zmiana $F$ | $df1$ | $df2$ | Istotność zmiany $F$ |
| 1 | ,535(a) | ,287 | ,281 | 773,607 | ,287 | 49,237 | 4 | 490 | ,000 |
| 2 | ,559(b) | ,312 | ,304 | 761,054 | ,026 | 9,148 | 2 | 488 | ,000 |

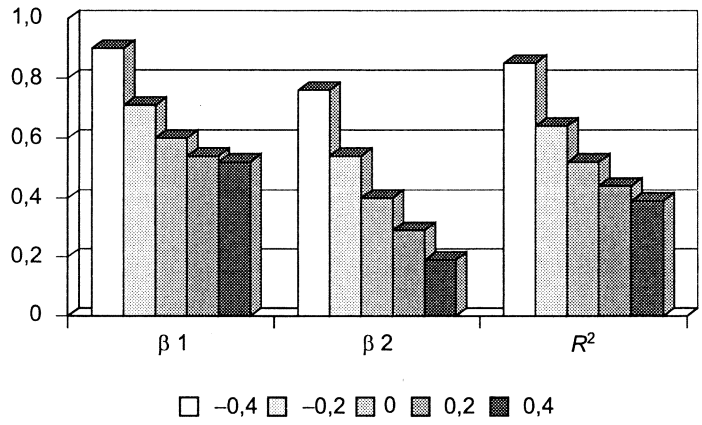
| Krok | | Współczynniki niestandaryzowane | | Współczynniki standaryzowane | $t$ | Istotność |
|---|---|---|---|---|---|---|
| | | $B$ | Błąd standardowy | β | | |
| 1 | (Stała) | −1092,798 | 227,804 | | −4,797 | ,000 |
| | *płeć* | −585,497 | 71,764 | −,320 | −8,159 | ,000 |
| | *dzieci* | 44,133 | 30,756 | ,064 | 1,435 | ,152 |
| | *age* | 12,305 | 3,712 | ,145 | 3,314 | ,001 |
| | *edu* | 167,508 | 13,786 | ,489 | 12,150 | ,000 |
| 2 | (Stała) | −978,776 | 245,623 | | −3,985 | ,000 |
| | *płeć* | −446,905 | 157,592 | −,244 | −2,836 | ,005 |
| | *dzieci* | 34,290 | 30,372 | ,050 | 1,129 | ,259 |
| | *age* | 10,951 | 3,688 | ,129 | 2,969 | ,003 |
| | *edu* | 160,861 | 13,695 | ,470 | 11,746 | ,000 |
| | *płeć_dzieci* | −122,380 | 54,633 | −,183 | −2,240 | ,026 |
| | *płeć_edu* | −113,425 | 27,410 | −,284 | −4,138 | ,000 |

### *Stopień skorelowania (współliniowości) predyktorów*

Waga danego predyktora (współczynnik regresji) w wyjaśnianiu zmienności $Y$ zależy nie tylko od jego korelacji z zmienną wyjaśnianą jak to było w przypadku jednowymiarowym, ale i od korelacji z pozostałymi predyktorami. Ilustracją są dane przedstawione w tablicy 10.10, pokazujące, jak zmieniają się współczynniki regresji przy takich samych korelacjach ze zmienną wyjaśnianą w zależności od wielkości i znaku współczynnika korelacji między $X_1$ i $X_2$.

**Tablica 10.10. Zmiana współczynników regresji w zależności od stopnia skorelowania predyktorów**

| | |
|---|---|
| Przykład 1. Predyktory nieskorelowane | |
| $R_{12}=R(X_1,X_2)=\mathbf{0}$ | $R^2_{Y(12)}=0{,}54$ |
| $r_{Y_1}=r(Y,X_1)=0{,}6$ | $\beta_1=0{,}6$ |
| $r_{Y_2}=r(Y,X_2)=0{,}4$ | $\beta_2=0{,}4$ |
| Przykład 2. Predyktory umiarkowanie skorelowane dodatnio | |
| $r_{12}=r(X_1,X_2)=\mathbf{0{,}2}$ | $R^2_{Y(12)}=0{,}44$ |
| $r_{Y_1}=r(Y,X_1)=0{,}6$ | $\beta_1=0{,}54$ |
| $r_{Y_2}=r(Y,X_2)=0{,}4$ | $\beta_2=0{,}29$ |
| Przykład 3. Predyktory silniej skorelowane dodatnio | |
| $r_{12}=r(X_1,X_2)=\mathbf{0{,}4}$ | $R^2_{Y(12)}=0{,}39$ |
| $r_{Y_1}=r(Y,X_1)=0{,}6$ | $\beta_1=0{,}52$ |
| $r_{Y_2}=r(Y,X_2)=0{,}4$ | $\beta_2=0{,}19$ |
| Przykład 4. Predyktory skorelowane ujemnie | |
| $r_{12}=r(X_1,X_2)=\mathbf{-0{,}2}$ | $R^2_{Y(12)}=0{,}54$ |
| $r_{Y_1}=r(Y,X_1)=0{,}6$ | $\beta_1=0{,}71$ |
| $r_{Y_2}=r(Y,X_2)=0{,}4$ | $\beta_2=0{,}54$ |
| Przykład 5. Predyktory silniej skorelowane ujemnie | |
| $r_{12}=r(X_1,X_2)=\mathbf{-0{,}4}$ | $R^2_{Y(12)}=0{,}85$ |
| $r_{Y_1}=r(Y,X_1)=0{,}6$ | $\beta_1=0{,}90$ |
| $r_{Y_2}=r(Y,X_2)=0{,}4$ | $\beta_2=0{,}76$ |




Wykres 10.6. Wartość współczynników regresji i współczynnika determinacji w zależności od stopnia skorelowania predyktorów występujących w równaniu regresji

Trzeba koniecznie pamiętać, że skorelowanie predyktorów (ich współliniowość) niekorzystnie wpływa na jakość estymatorów współczynników regresji. W sytuacji optymalnej najlepiej wprowadzać predyktory nieskorelowane, bo wtedy każda zmienna wyjaśnia odrębną porcje zmienności *Y*. Wskaźnikiem występowania problemu skorelowanych predyktorów jest duży procent wyjaśnionej wariancji przy nieistotnych współczynnikach regresji. Takie równanie regresji nie nadaje się do wykorzystania i należy zastanowić się nad usunięciem niektórych predyktorów lub przekształceniu ich w nieskorelowane zmienne za pomocą analizy składowych głównych z rotacją ortogonalną.

## *Wykorzystanie reszt regresji w dalszych analizach. Korelacje cząstkowe*

Jednym z najważniejszych problemów procesu badawczego jest kontrola zmiennych ubocznych, które mogą wpływać na kształt związku między analizowanymi przez nas zmiennymi. Wpływ ten może mieć charakter **mediatora** lub **modyfikatora (moderatora)**. Zależność między liczbą dzieci a zarobkami może być dodatnia dla mężczyzn i ujemna dla kobiet. Powiemy wtedy, że płeć jest zmienną modyfikującą (moderatorem) związek między liczbą dzieci a zarobkami. Zależność między zarobkami a liczbą dzieci zależy od płci respondenta. O tym, czy zmienna *M* jest moderatorem dla związku *X* z *Y* przekonujemy się testując istotność interakcji *M* z *X* w wyjaśnianiu zmienności *Y*.

Wiek może być mediatorem relacji między tymi zmiennymi. Jeżeli tak by było to związek między zarobkami a wykształceniem powinien się zmienić po uwzględnieniu wieku w analizie regresji. Jeżeli zarobki i wykształcenie są skorelowane z trzecią

zmienną — np. wiekiem to należy policzyć **korelację cząstkową** uwzględniającą związek obu zmiennych z wiekiem.

Jeżeli przeprowadzilibyśmy analizę regresji **wysokości zarobków** z jednym predyktorem: wiekiem, to jak pokazywaliśmy to wcześniej, zróżnicowanie zarobków zostałoby podzielone na część wyjaśnioną przez wiek i **część nie wyjaśnioną** nazywaną błędem regresji lub **resztą regresji**.

Jeżeli przeprowadzilibyśmy analizę regresji **wykształcenia** z jednym predyktorem: wiekiem, to zróżnicowanie zarobków zostałoby podzielone na część wyjaśnioną przez wiek i część nie wyjaśnioną (resztę regresji).

Te dwie reszty regresji określają zróżnicowanie zmiennych z wyłączeniem zróżnicowania wyjaśnionego przez wiek.

**Korelacja tych dwóch reszt regresji, to korelacja cząstkowa między zarobkami a wykształceniem przy kontroli wieku.** Zostało to schematycznie przedstawione w tabeli:

| Zróżnicowanie zmiennej | Zróżnicowanie zmiennej wyjaśnione przez wiek | Błąd (reszta regresji) |
|---|---|---|
| Zarobki | *Zw* — zróżnicowanie zarobków przez wiek | *Zb* — zróżnicowanie zarobków NIEwyjaśnione przez wiek |
| Wykształcenie | *Ew* — zróżnicowanie wykształcenia wyjaśnione przez wiek | *Eb* — zróżnicowanie wykształcenia NIEwyjaśnione przez wiek |

Prosty współczynnik korelacji dotyczy związku *Z* i *E*. Współczynnik korelacji cząstkowej (przy kontroli wieku respondenta) to współczynnik korelacji między *Zb* i *Eb*.

Stosując regresję wielokrotną możemy liczyć reszty regresji przy większej liczbie predyktorów np. wiek, liczba dzieci, ... itd. co oznacza możliwość analizy współczynników korelacji cząstkowej przy większej liczbie zmiennych kontrolowanych. Nie podajemy wzorów, bo obliczenia te dokonywane są za pomocą programu komputerowego. Reszty regresji można zapisywać jako osobne zmienne i dokonywać na nich dalszych analiz.

# Bibliografia


1. Aczel A.D., *Statystyka w zarządzaniu*, PWN, Warszawa 2000.
2. Adamkiewicz H.G., *Statystyka. Zastosowania w ekonomii*, ODDK, Gdańsk 1996.
3. Freund J.E., *Modern elementary statistics*, Pearson Prentice Hall 2004, 11 ed.
4. Kordos J., *Jakość danych statystycznych*, PWE, Warszawa 1995.
5. Lange O., Banasiński A., *Teoria statystyki*, PWE, Warszawa 1970.
6. Luszniewicz A., *Statystyka ogólna*, PWE, Warszawa 1987.
7. Makać W., Urbanek-Krzysztofiak D., *Metody opisu statystycznego*, Wyd. UG, Gdańsk 1995.
8. Marszałkowicz T., *Metody statystyki opisowej w badaniach ekonomiczno-rolniczych*, SGGW-AR, Warszawa 1986.
9. Sobczyk M., *Statystyka. Podstawy teoretyczne, przykłady — zadania*, Wyd. UMCS, Lublin 1998.
10. Szulc B., *Statystyka dla ekonomistów*, t.1, PWE, Warszawa 1972.
11. Wieczorkowska G., Kochański P., Eljaszuk M., *Statystyka: wprowadzenie do analizy badań sondażowych i eksperymentalnych*, Wydawnictwo Naukowe Scholar. Warszawa 2005, wydanie 3.
12. Wieczorkowska G., Wierzbiński J., *Badania sondażowe i eksperymentalne. Wybrane zagadnienia*, Wydawnictwa Naukowe Wydziału Zarządzania, Warszawa 2005.
13. Yule G.U., Kendall M.G., *Wstęp do teorii statystyki*, PWN, Warszawa 1966.