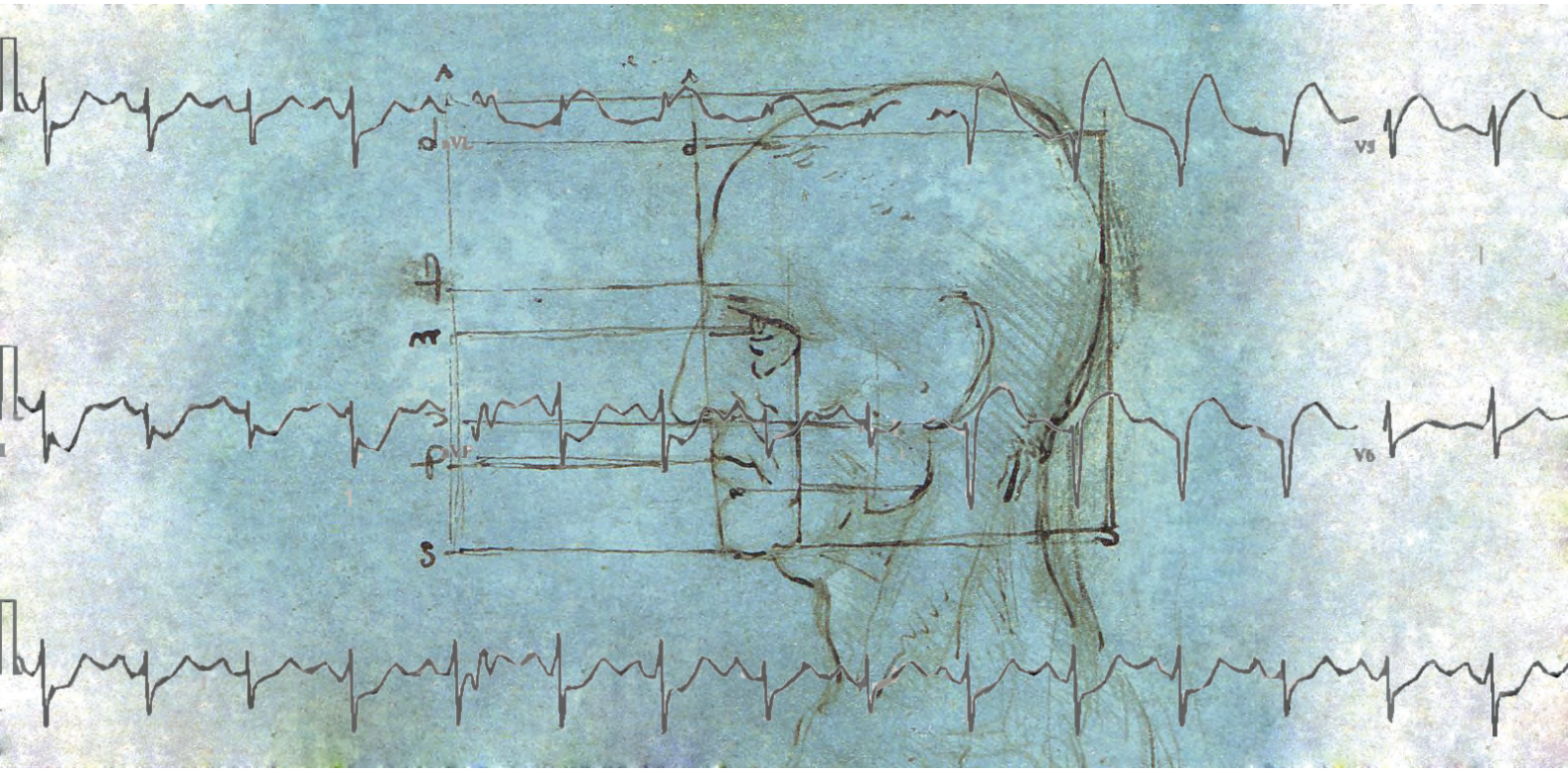


Elementy statystyki

laboratorium

Dorota Wejer

Gdańsk, 2013

Wydział Matematyki, Fizyki i Informatyki
Uniwersytetu Gdańskiego

Skrypt powstał w ramach projektu „Kształcimy profesjonalistów – kompleksowy program kształcenia skierowany do studentów, absolwentów oraz studentów niepełnosprawnych Uniwersytetu Gdańskiego”
współfinansowanego przez Unię Europejską w ramach środków Europejskiego Funduszu Społecznego.

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIWERSYTET GDAŃSKI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Dorota Wejer

Elementy statystyki

Gdańsk 2013

Spis treści

Rozdział 0. Podstawy rachunku prawdopodobieństwa . . . 3
0.1. Podstawowe pojęcia rachunku prawdopodobieństwa . . . 3
0.1.1. Definicja prawdopodobieństwa i jego własności . . . 5
0.1.2. Prawdpodobieństwo warunkowe. Zdarzenia niezależne . . . 6
0.1.3. Prawdpodobieństwo całkowite. Twierdzenie Bayesa . . . 6
Ćwiczenia nr 1 . . . 8
0.2. Zmienne losowe . . . 10
0.2.1. Definicja zmiennej losowej . . . 10
0.2.2. Niezależność zmiennych losowych . . . 11
0.2.3. Parametry zmiennej losowej . . . 12
0.3. Przykłady rozkładów zmiennych losowych dyskretnych . . . 13
0.3.1. Rozkład zero-jedynkowy . . . 13
0.3.2. Przykłady rozkładów zmiennych losowych ciągłych . . . 16
Ćwiczenia nr 2 . . . 19
Rozdział 1. Doświadczenia losowe, populacja i próba . . . 23
1.1. Statystyka jako narzędzie badawcze nauk przyrodniczych. Populacja generalna a próba . . . 23
1.2. Szeregi statystyczne . . . 24
1.3. Graficzne formy prezentacji szeregów statystycznych . . . 27
1.3.1. Histogram . . . 27
1.3.2. Łamana częstości. Wielobok częstości . . . 29
Ćwiczenia nr 3 . . . 30
Ćwiczenia nr 3 – Matlab . . . 31
Funkcje . . . 31
Przykładowe zadanie . . . 32
Rozdział 2. Statystyka opisowa . . . 35
2.1. Momenty statystyczne . . . 35
2.1.1. Momenty zwykłe i centralne . . . 35
2.1.2. Momenty absolutne zwykłe i centralne . . . 36
2.2. Miary położenia . . . 37
2.2.1. Miary tendencji centralnej . . . 37
2.2.2. Kwantyle . . . 44
2.3. Miary rozproszenia . . . 48
2.3.1. Rozstęp . . . 48
2.3.2. Odchylenie standardowe i wariancja . . . 49
2.3.3. Odchylenie przeciętne od średniej i odchylenie przeciętne od mediany . . . 50
2.3.4. Rozstęp ćwiartkowy i odchylenie ćwiartkowe . . . 51
2.3.5. Współczynnik zmienności względem średniej i współczynnik zmienności względem mediany . . . 52
Ćwiczenia nr 4 . . . 53

Ćwiczenia nr 4 – Matlab . . . 55
Funkcje . . . 55
Przykładowe zadanie . . . 58
2.4. Miary asymetrii rozkładu . . . 61
2.4.1. Współczynnik asymetrii . . . 62
2.4.2. Współczynnik skośności Pearsona . . . 63
2.4.3. Pozycyjny współczynnik skośności . . . 64
2.5. Miary koncentracji rozkładu . . . 65
2.5.1. Kurtoza i eksces . . . 65
2.6. Momenty a miary rozkładu . . . 66
Ćwiczenia nr 5 . . . 67
Ćwiczenia nr 5 – Matlab . . . 69
Funkcje . . . 69
Przykładowe zadanie . . . 69
Rozdział 3. Podstawowe testy statystyczne . . . 73
3.1. Estymacja . . . 73
3.1.1. Estymacja punktowa . . . 74
3.1.2. Estymacja przedziałowa . . . 75
3.1.3. Wyznaczenie liczebności próby . . . 79
3.1.4. Bootstrapowa estymacja parametrów . . . 80
Ćwiczenia nr 6 . . . 82
3.2. Testy nieparametryczne . . . 83
3.2.1. Testy zgodności rozkładu dla populacji jednowymiarowych . . . 85
3.2.2. Porównywanie rozkładów dwóch prób niezależnych . . . 87
3.2.3. Porównywanie rozkładów dwóch prób zależnych . . . 89
Ćwiczenia nr 7 . . . 91
3.3. Testy parametryczne . . . 92
3.3.1. Testy parametryczne służące do weryfikacji własności populacji jednowymiarowych . . . 92
3.3.2. Testy parametryczne służące do porównania własności dwóch populacji . . . 97
3.4. Testy bootstrapowe . . . 99
3.4.1. Testy bootstrapowe dla wartości średniej . . . 99
3.4.2. Testy bootstrapowe dla dwóch wartości średnich . . . 102
Ćwiczenia nr 8 . . . 107
Rozdział 4. Weryfikacja hipotez statystycznych w typowych doświadczeniach . . . 109
Ćwiczenia nr 9 . . . 112
Rozdział 5. Analiza wariancji . . . 115
5.1. ANOVA dla klasyfikacji pojedyńczej . . . 115
Ćwiczenia nr 10 . . . 117
Rozdział 6. Korelacja i regresja . . . 121
6.1. Diagram korelacyjny . . . 121
6.2. Tablica korelacyjna . . . 122
6.3. Współczynnik korelacji liniowej Pearsona . . . 123
6.4. Regresja liniowa . . . 124
Ćwiczenia nr 11 . . . 126
Bibliografia . . . 129

Rozdział 0

Podstawy rachunku prawdopodobieństwa

0.1. Podstawowe pojęcia rachunku prawdopodobieństwa

Podstawowym pojęciem rachunku prawdopodobieństwa jest *zdarzenie losowe*, czyli realizacja powtarzalnego doświadczenia, którego wyniku nie da się przewidzieć. Nie jest to ścisla definicja zdarzenia losowego, ponieważ jest ono pojęciem pierwotnym rachunku prawdopodobieństwa. Zdarzenia oznaczamy wielkimi literami alfabetu łacińskiego: A, B, C, ... Zbiór wszystkich możliwych wyników doświadczenia nazywamy *przestrzenią zdarzeń elementarnych*. Przestrzeń zrarzeń elementarnych oznaczamy symbolem Ω. *Zdarzeniem elementarnym* (ω_i) nazywamy zdarzenie losowe, które może być zrealizowane tylko w jeden sposób. Każde zdarzenie jest dowolnym podzbiorem przestrzeni zdarzeń elementarnych $\Omega = \{\omega_i : i \in \mathbb{N}\}$, czyli każde zdarzenie jest zbiorem sprzyjających mu zdarzeń elementarnych. Zdarzenie, które nigdy nie może być zrealizowane nazywamy *zdaraeniem niemożliwym*. Z kolei zdarzenie, które jest realizaowane zawsze to *zdarzenie pewne*. *Zdarzeniem przeciwnym* do zdarzenia A nazywamy zdarzenie realizowane przez wszystkie zdarzenia elementarne, ktore nie realizują zdarzenia A. Zdarzenie przeciwne oznaczamy symbolem A'.

Przykład 0.1:

Doświadczenie polega na rzucie kostką do gry.
W tym doświadczeniu można uzyskać nastepujące wyniki: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
Przykładowe zdarzenia dla tego doświadczenia to:

- A – liczba oczek jest równa 3,
- B – liczba oczek jest parzysta,
- C – liczba oczek jest większa od 6,
- D – liczba oczek jest dodatnia,
- E – liczba oczek jest nieparzysta.

Zdarzenie A jest zdarzeniem elementarnym.
Zdarzenie B jest zdarzeniem losowym realizowanym na 3 sposoby.
Zdarzenie C jest zdarzeniem nie możliwym.
Zdarzenie D jest zdarzeniem pewnym.
Zdarzenie E jest zdarzeniem przeciwnym do zdarzenia B ($E = B'$).

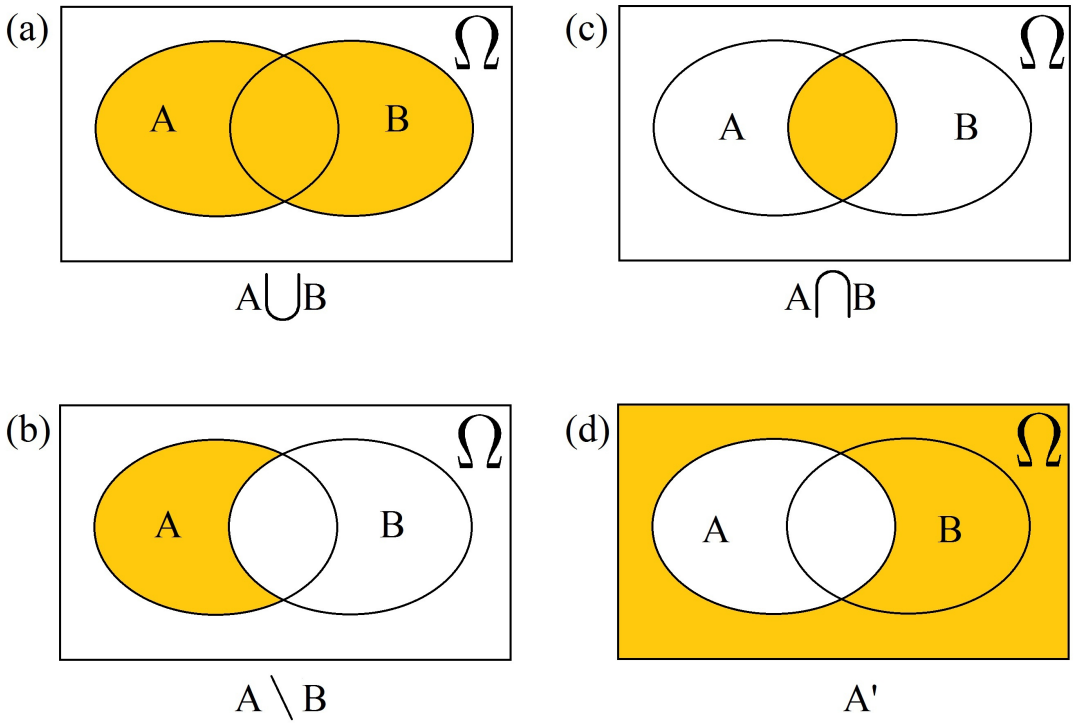
Jeżeli z faktu, że zachodzi zdarzenie A wynika, iż zachodzi zdażenie B, to zdarzenie A *zawiera się* zdarzeniu B ($A \subset B$). Z kolei, jeśli $A \subset B$ oraz $A \supset B$, to zdarzenia A i B są *rownoważne*. Zdarzenia wykluczają się, jeśli żadno zdarzenie elementarne jednocześnie nie realizuje obu zdarzeń.

Przykład 0.2:

Doświadczenie polega na rzucie kostką do gry.
$\Omega = \{1, 2, 3, 4, 5, 6\}$
A – liczba oczek jest równa 3

A = {3}
B – liczba oczek jest mniejsza od 5.
$B = \{\omega : \omega < 5 \wedge \omega \in \{1,2,3,4,5,6\}\} = \{1,2,3,4\}$
C – liczba oczek jest większa od 2 i mniejsza od 4.
$C = \{\omega : \omega < 4 \wedge \omega > 2 \wedge \omega \in \{1,2,3,4,5,6\}\} = \{3\}$
D – liczba oczek jest większa do 4.
$D = \{\omega : \omega > 4 \wedge \omega \in \{1,2,3,4,5,6\}\} = \{5,6\}$
Zdarzenie A zawiera się w zdarzeniu B ($A \subset B$).
Zdarzenia A i C są równoważne (A=B).
Zdarzenia B i D wykluczają się.

Skoro zdarzenia są zbiorami, to można na nich wykonywać między innymi następujące działania: *sumę*, *różnicę*, *iloczyn*, *dopełnienie*. Operacje na zdarzeniach tworzą nowe zdarzenia. Niech A i B są zdarzeniami losowymi. Sumą zdarzeń $A \cup B$ nazywamy zdarzenie takie, że zajdzie conajmiej jedno ze zdarzeń (A lub B). Różnicą zdarzeń $A \setminus B$ nazywamy zdarzenie polegające na tym, że zajdzie zdarzenie A i jednocześnie nie zajdzie zdarzenie B. Iloczynem zdarzeń $A \cap B$ jest zdarzenie zdarzenie takie, że zachodzą oba zdarzenia (A i B). W tym miejscu warto zauważyć, że jeśli iloczyn zdarzeń jest zbiorem pustym, to zdarzenia wykluczają się. Dopełnieniem zdarzenia A jest jego zdarzenie przeciwne (A'). Na rysunku 1 zostały przedstawione diagramy Venna reprezentujące działania na zdarzeniach.



Rysunek 1. Działania na zdarzeniach: (a) suma, (b) różnica, (c) iloczyn, (d) dopełnienie

0.1.1. Definicja prawdopodobieństwa i jego własności

Definicja prawdopodobienstwa ewoluowała w matematyce od definicji klasycznej (częstotliwościowej), przez definicję geometryczną, do obecnie przyjmowanej definicji aksjomatycznej. Aksjomatyczna definicja prawdopodobienstwa zostala podana w 1933 roku przez rosyjskiego matematyka A. Kołmogorowa.

Aby można było podać definicję Kołmogorowa należy najpierw zdefiniować pojęcie σ-ciała.

Definicja 0.1: (*σ-ciało*) Niech X jest ustalonym zbiorem, wówczas zbiór $\mathcal{A}$ złożony z podzbiorów zbioru X nazywmy σ-ciałem podzbiorów zbioru X, jeżeli spełnia następujące warunki:

— $\phi \in \mathcal{A}$
— $\mathrm{A} \in \mathcal{A} \;\Rightarrow\; X \setminus \mathrm{A} \in \mathcal{A}$
— $\forall_n\ \mathrm{A}_n \in \mathcal{A} \;\Rightarrow\; \bigcup_{n\in\mathbb{N}} \mathrm{A}_n \in \mathcal{A}$

Przykład 0.3:

W doświadczeniu polegającym na rzucie kostką do gry przestreń zdarzeń elementarnych jest następująca: $\Omega = \{1,2,3,4,5,6\}$.
Niech:
$\mathrm{A} = \{\phi, \Omega\}$
$\mathrm{B} = \{\{1\}, \{2,3,4,5,6\}, \Omega\}$
$\mathrm{C} = \{\phi, \{1,2\}, \{1,2,3\}, \{3,4,5,6\}, \Omega\}$
$\mathrm{D} = \{\phi, \{1,2\}, \{3,4,5,6\}, \{1,3\}, \{2,4,5,6\}, \Omega\}$
$\mathrm{E} = \{\phi, \{1,2\}, \{3,4,5,6\}, \Omega\}$
Zbiór A jest najmniejszym σ-ciałem zbioru Ω.
Zbiór B nie jest σ-ciałem zbioru Ω, ponieważ $\phi \notin \mathrm{B}$.
Zbiór C nie jest σ-ciałem zbioru Ω, ponieważ $\Omega \setminus \{1,2,3\} = \{4,5,6\} \notin \mathrm{C}$.
Zbiór D nie jest σ-ciałem zbioru Ω, ponieważ $\{1,2\} \cup \{1,3\} = \{1,2,3\} \notin \mathrm{D}$.
Zbiór E jest σ-ciałem zbioru Ω.

Znając już definicję σ-ciała możemy podać aksjomatyczną definicję prawdopodobieństwa.

Definicja 0.3: (*prawdopodobieństwo*) Niech Ω jest zbiorem zdarzeń elementarnych ω, zaś $\mathcal{A}$ jest σ-ciałem na zbiorze Ω. Prawdopodobieństwem nazywamy funkcję $P : \mathcal{A} \to \mathbb{R}$, spełniającą następujące warunki:

— $\forall_{\mathrm{A}\in\mathcal{A}}\ \ P(\mathrm{A}) \geqslant 0$,
— $P(\Omega) = 1$,
— jeśli (A_n) jest dowolnym ciągiem podzbiorów $\mathcal{A}$ takim, że $\forall_{i\neq j}\ \ A_i \cap A_j = \phi$, to:

$$P\left(\bigcup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} P(A_i)$$

Przykład 0.4:

Doświadczenie polega na rzucie symetryczną monetą.
Przestreń zdarzeń elementarnych: $\Omega = \{O, R\}$, gdzie:
— O – zdarzenie polegające na wurzuceniu orła,
— R – zdarzenie polegające na wurzuceniu reszki.

$\mathcal{A}$ – rodzina wszystkich podzbiorów Ω, tzn $\mathcal{A} = \{\phi, \{O\}, \{R\}, \{O, R\}\}$.
Niech: $P(\{\phi\}) = 0$, $P(\{\Omega\}) = 1$, $P(\{O\}) = \frac{1}{2}$ oraz $P(\{R\}) = \frac{1}{2}$
Tak zdefiniowana funkcja P jest poprawnie określonym prawdopodobienstwem.

Prawdopodobienstwo $P : \mathcal{A} \to \mathbb{R}$ jest funkcją posiadającą następujące własności:
— $P(\phi) = 0$,
— jeśli $\mathrm{A}, \mathrm{B} \in \mathcal{A}$ oraz $\mathrm{A} \subset \mathrm{B}$, to $P(\mathrm{A}) \leqslant P(\mathrm{B})$,
— jeśli $\mathrm{A} \in \mathcal{A}$, to $P(\mathrm{A}) \leqslant 1$,
— jeśli $\mathrm{A}, \mathrm{B} \in \mathcal{A}$ oraz $\mathrm{A} \subset \mathrm{B}$, to $P(\mathrm{B} \setminus \mathrm{A}) = P(\mathrm{B}) - P(\mathrm{A})$,
— jeśli $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_n \in \mathcal{A}$ oraz $\forall_{i \neq j} \mathrm{A}_i \cap \mathrm{A}_j = \phi$, to $P\left(\cup_{i=1}^{n} \mathrm{A}_i\right) = \sum_{i=1}^{n} P(\mathrm{A}_i)$
— jeśli $\mathrm{A}, \mathrm{B} \in \mathcal{A}$, to $P(\mathrm{A} \cup \mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B})$ - $P(\mathrm{A} \cap \mathrm{B})$,
— jeśli $\mathrm{A} \in \mathcal{A}$, to $P(\mathrm{A}') = 1 - P(\mathrm{A})$,
— jeśli $\Omega = \{\omega_i : i = 1, 2, \ldots, n\}$, $\forall_{i=1,2,\ldots,n} P(\{\omega_i\}) = \frac{1}{n}$ oraz $\mathrm{A} = \{\omega_{i_1}, \omega_{i_2}, \ldots, \omega_{i_k}\}$, to $P(\mathrm{A}) = \frac{\overline{\mathrm{A}}}{\overline{\Omega}} = \frac{k}{n}$.

0.1.2. Prawdpodobieństwo warunkowe. Zdarzenia niezależne

W wielu przypadkach interesują nas zdarzenia losowe, które są powiązane z innymi zdarzeniami losowymi. Nierzadko wystąpienie zdarzebia B ma wpływ na częstość występowania zdarzenia A. Prawdopodobienstwo zajścia zdarzenia A, przy założeniu, że zaszło zdarzenie B ($P(\mathrm{B}) > 0$) nazywamy *prawdopodobieństwem warunkowym* i wyznaczamy je za pomocą wzoru:

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{A} \cap \mathrm{B})}{P(\mathrm{B})}$$

Jeśli zajście zdarzenia A nie ma wpływu na zajście zdarzenia B i odwrotnie, to zdarzenia A oraz B są *niezależne*. Formalnie zdarzenia A i B są niezależne gdy:

$$P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A}) \cdot P(\mathrm{B})$$

Przykład 0.5:

Doświadczenie polega na rzucie kostką do gry.
$\Omega = \{1, 2, 3, 4, 5, 6\}$
$\forall_{i \in \Omega} P(i) = \frac{1}{6}$
A – liczba oczek jest mniejsza od 4.
$\mathrm{A} = \{1, 2, 3\} \quad P(\mathrm{A}) = \frac{1}{2}$
B – liczba oczek jest parzysta
$\mathrm{B} = \{2, 4, 6\} \quad P(\mathrm{B}) = \frac{1}{2}$
C – liczba oczek jest nieparzysta
$\mathrm{B} = \{1, 3, 5\} \quad P(\mathrm{C}) = \frac{1}{2}$
$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{A} \cap \mathrm{B})}{P(\mathrm{B})} = \frac{P(\{2\})}{P(\mathrm{B})} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$
Zatem zdarzenia A i B są zależne.
$P(\mathrm{C}|\mathrm{B}) = \frac{P(\mathrm{C} \cap \mathrm{B})}{P(\mathrm{B})} = \frac{P(\{\phi\})}{P(\mathrm{B})} = \frac{0}{\frac{1}{2}} = 0$
Natomiast zdarzenia B i C są niezależne.

0.1.3. Prawdpodobieństwo całkowite. Twierdzenie Bayesa

Często spotykamy się z sytuacjami, w których na dane zdarzenie B ma wpływ n zdarzeń losowych $(\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_n)$. Wówczas zdarzenie B można nazwać skutkiem zdarzeń $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_n$, które są nazywane przyczynami.

Jeśli zdarzenia A_i są parami rozłączne i któreś z A_i napewno następuje, to prawdopodobieństwo zdarzenia B można obliczyć *a priori*, to znaczy przed wykonaniem doświadczenia. Mówi o tym twierdzenie o przwdopodobieństwie całkowitym:

Twierdzenie 0.1: (*o przwdopodobieństwie całkowitym*) Niech $\mathrm{B} \subset \Omega$ będzie dowolnym zdarzeniem, a zdarzenia $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_n \subset \Omega$ spełniają następujące warunki:

1. $\forall_{\substack{i \neq j \\ i,j=1,2,\ldots,n}} \quad \mathrm{A}_i \cap \mathrm{A}_j = \phi$,
2. $\bigcup_{i=1}^{n} \mathrm{A}_i = \Omega$,
3. $\forall_{i=1,2,\ldots,n} \quad P(\mathrm{A}_i) > 0$.

Wówczas prawdopodobienstwo zdarzenia B dane jest wzorem:

$$P(\mathrm{B}) = \sum_{i=1}^{n} P(\mathrm{A}_i) \cdot P(\mathrm{B}|\mathrm{A}_i)$$

Przykład 0.6:

Apteka kupiła 50 termometrów w fabryce A i 75 w fabryce B. Wiadomo, że 5% termometrów produkowanych w fabryce A jest wadliwych. Natomiast w fabryce B produkuje się 3% wadliwych termometrów.

Jakie jest prawdopodobieństwo, że klient kupując losowo wybrany termometr dostanie niezepsuty?

Rozpatrywane zdarzenia losowe:

A – termometr pochodzi z fabryki A,

B – termometr pochodzi z fabryki B,

C – klient dostał dobry termometr.

Wiemy, że: $P(\mathrm{A}) = \frac{50}{50+75} = 0,4$ i $P(\mathrm{B}) = \frac{75}{50+75} = 0,6$

Ponadto: $P(\mathrm{C}|\mathrm{A}) = 1 - 0,05 = 0,95$ i $P(\mathrm{C}|\mathrm{B}) = 1 - 0,03 = 0,97$

Zatem: $P(\mathrm{C}) = P(\mathrm{A})P(\mathrm{C}|\mathrm{A}) + P(\mathrm{B})P(\mathrm{C}|\mathrm{B}) = 0,4 \cdot 0,95 + 0,6 \cdot 0,97 = 0,962$

Załóżmy teraz, że eksperyment został już wykonany i nastąpiło zdarzenie B, które zostało wywołane tylko jedną z rozłączych przyczyn $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_n$. Możemy wówczas wyznaczyć prawdopodobieństwo *a posteriori* („po zdarzeniu") przyczyny, która wywołała skutek B. Mówi o tym twierdzenie Bayes'a.

Twierdzenie 0.1: (*Bayes'a*) Niech $\mathrm{B} \subset \Omega$ będzie dowolnym zdarzeniem o $P(\mathrm{B}) > 0$ oraz zdarzenia $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_n \subset \Omega$ spełniają następujące warunki:

1. $\forall_{\substack{i \neq j \\ i,j=1,2,\ldots,n}} \quad \mathrm{A}_i \cap \mathrm{A}_j = \phi$,
2. $\bigcup_{i=1}^{n} \mathrm{A}_i = \Omega$,
3. $\forall_{i=1,2,\ldots,n} \quad P(\mathrm{A}_i) > 0$.

Wówczas dla każdego $i = 1, 2, \ldots, n$ prawdopodobieństwo warunkowe zdarzenia A_i przy warunku B jest dane wzorem:

$$P(\mathrm{A}_i|\mathrm{B}) = \frac{P(\mathrm{A}_i) \cdot P(\mathrm{B}|\mathrm{A}_i)}{\sum_{i=1}^{n} P(\mathrm{A}_i) \cdot P(\mathrm{B}|\mathrm{A}_i)}$$

Przykład 0.7:

Pewna choroba występuje u 5% ludzi. Test przeprowadzany w celu jej wykrycia daje wynki pozytwywny u 90% chorych i 3% osob zdrowych. Jakie jest prawdopodobieństwo, że osoba jest chora, jeśli wynki jej testu jest pozytywny? Rozpatrywane zdarzenia losowe:
A – osoba jest chora,
B – osoba jest zdrowa,
C – wynik testu jest pozytywny
Wiemy, że: $P(\mathrm{A}) = 0,05$ i $P(\mathrm{B}) = 0,95$
Ponadto: $P(\mathrm{C}|\mathrm{A}) = 0,9$ i $P(\mathrm{C}|\mathrm{B}) = 0,03$
Zatem: $P(\mathrm{A}|\mathrm{C}) = \frac{P(\mathrm{A})P(\mathrm{C}|\mathrm{A})}{P(\mathrm{A})P(\mathrm{C}|\mathrm{A})+P(\mathrm{B})P(\mathrm{C}|\mathrm{B})} = \frac{0,05 \cdot 0,9}{0,05 \cdot 0,9+0,95 \cdot 0,03} = \frac{0,045}{0,735} \approx 0,61$

Ćwiczenia nr 1

Zadanie 0.1.

W doświadczeniu rzucano dwiema sześciennymi kostkami do gry. Niech A jest zdarzeniem takim, że suma oczek jest parzysta, a B – suma oczek jest większa niż 5.
— Określ przestrzeń zdarzeń elementarnych.
— Określ rozkład prawdopodobienstwa.
— Opisz zdarzenia A∪B, A∩B, A', B', A\B.
— Wyznacz prawdopodobieństwa zdarzeń A∪B, A∩B, A', B', A\B.
— Wyznacz prawdopodobieństwo zdarzenia A pod warunkiem, że zaszło zdarzenie B.

Zadanie 0.2.

W doświadczeniu rzucano trzema symetrycznymi monetami. Niech A oznacza zdarzenie polegające na wyrzuceniu conajmniej dwóch orłów, a B – najwyżej trzech reszek.
— Określ przestrzeń zdarzeń elementarnych.
— Określ rozkład prawdopodobienstwa.
— Opisz zdarzenia A∪B, A∩B, A', B', A\B.
— Wyznacz prawdopodobieństwa zdarzeń A∪B, A∩B, A', B', A\B.
— Wyznacz prawdopodobieństwo zdarzenia A pod warunkiem, że zaszło zdarzenie B.

Zadanie 0.3.

W urnie znjdują się 3 kule białe i 8 czarnych. Losujemy dwie kule bez zwracania. Znajdź prawdopodobienstwo tego, że:
— obie są białe,
— obie są czarne,
— pierwsza jest czarna, a druga jest biała.

Zadanie 0.4.

W urnie znjdują się 4 kule białe i 11 czarnych. Losujemy dwie kule ze zwracaniem. Znajdź prawdopodobienstwo tego, że:
— obie są białe,
— obie są czarne,
— pierwsza jest czarna, a druga jest biała.

Zadanie 0.5.

Z urny, w ktorej znajdują się 2 kule białe, 5 kul czarnych i 3 kule zielone, losujemy kolejno dwie kule. Czy bardziej prawdopodobne jest wylosowanie dwóch kul zielonych w przypadku losowania ze zwracaniem, czy bez zwracania?

Zadanie 0.6.

Z wszystkich liczb trzycyfrowych o cyfrach: 0, 2, 5 losujemy jedną liczbę. Jakie jest prawdopodobiestwo otrzymanian liczby, która:
— jest nieparzysta,
— podzielna przez 3,
— większa niż 300.

Zadanie 0.7.

Niech A i B bedą zdarzeniami. Oblicz:
— $P(\mathrm{A} \cup \mathrm{B})$, jesli $P(\mathrm{A}) = 0,3$, $P(\mathrm{B}) = 0,6$ oraz $P(\mathrm{A} \cap \mathrm{B}) = 0,2$,
— $P(\mathrm{A} \setminus \mathrm{B})$, jesli $P(\mathrm{A}) = 0,9$ oraz $P(\mathrm{A} \cap \mathrm{B}) = 0,3$,
— $P(\mathrm{A} \cup \mathrm{B})$, $P(\mathrm{A} \cap \mathrm{B})$, $P(\mathrm{A} \setminus \mathrm{B})$, jesli A i B są zdarzeniami wykluczającymi się oraz $P(\mathrm{A}) = \frac{3}{8}$ oraz $P(\mathrm{B}) = \frac{1}{4}$.

Zadanie 0.8.

Niech A i B bedą zdarzeniami. Udowodnij, że jeśli $P(\mathrm{A}) = P(\mathrm{B}) = \frac{2}{3}$, to $P(\mathrm{A}|\mathrm{B}) \geqslant \frac{1}{2}$

Zadanie 0.9.

Doświadczenie polega na rzucie kostką do gry.
A – liczba oczek jest mniejsza od 4.
B – liczba oczek jest parzysta
C – liczba oczek jest nieparzysta
Które ze zdarzeń są zależne, a które niezależne?

Zadanie 0.10.

Apteka kupiła 50 termometrów w fabryce A i 75 w fabryce B. Wiadomo, że 5% termometrów produkowanych w fabryce A jest wadliwych. Natomiast w fabryce B produkuje się 3% wadliwych termometrów.
Jakie jest prawdopodobieństwo, że klient kupując losowo wybrany termometr dostanie niezepsuty?

Zadanie 0.11.

Pewna choroba występuje u 5% ludzi. Test przeprowadzany w celu jej wykrycia daje wynki pozytwywny u 90% chorych i 3% osob zdrowych. Jakie jest prawdopodobieństwo, że osoba jest chora, jeśli wynki jej testu jest pozytywny?

Zadanie 0.12.

W pewnej fabryce dwie maszyny wytwarzają części zamienne do samochodów. Wiadomo, że pierwsza maszyna wytwarza 96% części zamiennych dobrej jakości, a druga maszyna tylko 85%. W ciagu doby pierwsza maszyna produjkuje 10500 takich części, a druga 8160. Jakie jest prawdopodobieństwo, że klient kupując losowo wybraną część zamienną dostanie część dobrej jakości?

Zadanie 0.13.

Psychologowie w celu wykrycia pewnych cech osobowości stosują test, który daje wynk pozytwywny u 90% posiadających tą cechę i 3% osob jej nie posiadających. Wiadomo, że średnio co druga osoba ma badaną cechę osobowości. Jakie jest prawdopodobieństwo, że osoba posiada cechę, jeśli wynki jej testu jest pozytywny?

0.2. Zmienne losowe

0.2.1. Definicja zmiennej losowej

Zmienna losowa jest fuknkcją, która odwzorowuje przestrzeń zdarzeń elementarnych Ω w prostą rzeczywistą $\mathbb{R}$. Jeżeli badamy k cech populacji, wówczas zmienna losowa przekształca przestrzeń zdarzeń elementarnych Ω w iloczyn kartezjanski k prostych rzeczywistych $\mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R}$.

Definicja 0.3: (*zmienna losowa*) Niech Ω to przestrzeń zdarzeń elementarnych, natomiast $\mathcal{A}$ – σ-ciało zbioru Ω. Zmienną losową nazywamy funkcję $X : \Omega \to \mathbb{R}$ taką, że $X^{-1}([-\infty, x)) \in \mathcal{A}$, gdzie $X^{-1}([-\infty, x))$ – przeciwobraz przedziału $[-\infty, x)$.

Zmienne losowe, których zbiór wartości jest przeliczalny (skończony lub nieskończony) nazywamy *zmiennymi losowymi dyskretnymi* lub *skokowymi*. Dla zmiennej losowej dyskretnej X o wartościach $x_1, x_2, \ldots, x_n, \ldots$ prawdopodobieństwo z jakim zmienna losowa X przyjmuje wartość x_i dane jest wzorem:

$$p_i = P(X = x_i) \quad dla \quad i = 1, 2, \ldots, n$$

Należy zauważyć, że prawdopodobieństwa sumują się do jedności:

$$\sum_{i=1}^{n} p_i = 1$$

Zmienna losowa nazywana jest *zmienną losową ciągłą*, gdy istnieje taka nieujemna, całkowalna na prostej $\mathbb{R}$ funkcja $f(x)$, że dla dowolnej wartości x mamy równość:

$$P(X < x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$$

oraz

$$\int_{-\infty}^{\infty} f(t)\mathrm{d}t = 1$$

Każdą zmienną losową X można jednoznacznie określić za pomocą dystrybuanty $F(x)$, czyli funkcji $F : \mathbb{R} \to \mathbb{R}$ określonej wzorem:

$$F(x) = P(X < x)$$

W szczególności dla zmiennej losowej dyskretnej:

$$F(x) = P(X < x) = \sum_{x_i < x} p_i$$

A dla zmiennej losowej ciągłej:

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$$

Podstawowe własności dystrybuanty to:
- $0 \leqslant F(x) \leqslant 1$,
- $F(x)$ jest niemalejąca i lewostronnie ciągła,
- $\lim_{x\to-\infty} F(x) = 0$,
- $\lim_{x\to\infty} F(x) = 1$,
- $P(a \leqslant X \leqslant b) = F(b) - F(a)$.

Przykład 0.8:

Rzucamy kostką do gry na której ściankach znajduje się jedna jedynka, dwie dwójki i trzy trójki. Zatem mamy 3 zdarzenia elementarne: wyrzucenie 1 oczka, 2 oczek lub 3 oczek.
Na zbiorze zdarzeń elementarnych określamy zmienną losową przypisującą każdemu zdarzeniu liczbę rzeczywistą będącą ilością wurzuconych oczek.
Poniższa tabela przedstawia rozkład zmiennej losowej.

wartość zmiennej losowej x_i	1	2	3
prawdopodobieństwo p_i	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$

Rysunek 2 przedstawia rozkład zmiennej losowej dyskretnej, rysunek 3 – dystrybuantę zmiennej losowej dyskretnej.

0.2.2. Niezależność zmiennych losowych

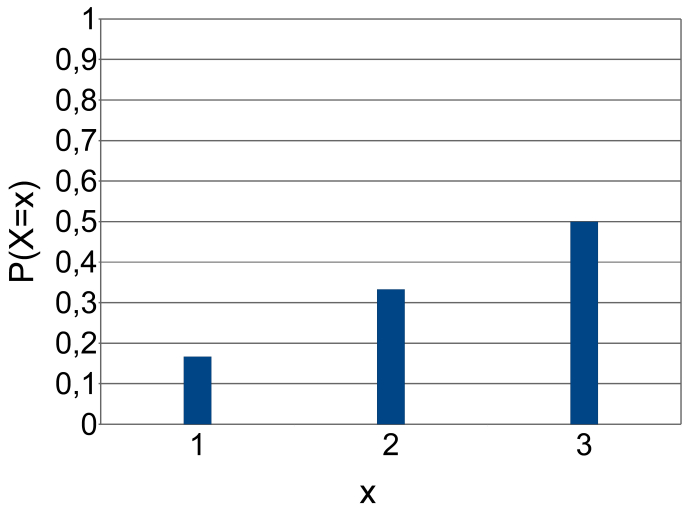
Dwie zmienne losowe X i Y na Ω są *niezależne* gdy:

$$P(X \in I, Y \in J) = P(X \in I) \cdot P(Y \in J)$$

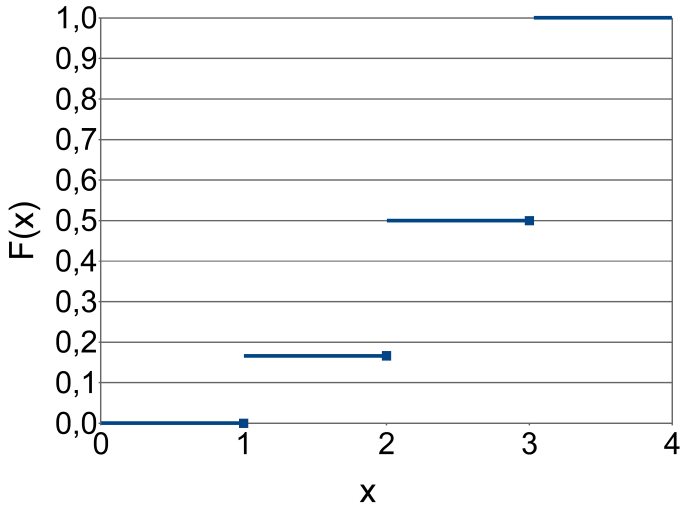
dla wszystkich przedziałów $I, J \subset \mathbb{R}$. Gdy Ω jest skończona to wyrażenie się upraszcza i zmienne losowe X i Y są *niezależne* gdy:

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$$

dla wszystkich $x \in \mathbb{R}$.



Rysunek 2. Rozkład zmiennej losowej dyskretnej



Rysunek 3. Dystrybuanta zmiennej losowej dyskretnej

Przykład 0.9:

Zmienna losowa X przyjmuje wartość 1 gdy wyrzucimy orła i 0 w przeciwnym wypadku. Natomiast zmienna losowa Y przyjmuje wartość liczby oczek wyrzuconych na czworościennej kostce. Możliwe zdarzenia elementarne:

$$(O,1),(O,2),(O,3),(O,4),(R,1),(R,2),(R,3),(R,4)$$

Zatem: $P(X=0,Y=2)=P(\{(R,2)\})=\frac{1}{8}$, $P(X=0)=\frac{1}{2}$, oraz $P(Y=2)=\frac{1}{4}$ Wówczas: $P(X=0,Y=2)=\frac{1}{8}=P(X=0)\cdot P(Y=2)$

0.2.3. Parametry zmiennej losowej

Najważniejszymi parametrami zmiennej losowej są:
— wartość oczekiwana,
— wariancja,

— odchylenie standardowe,
— moment zwykły rzędu k,
— moment centralny rzędu k.

W tabeli 0.2.3 przedstawione zostały zdfiniowane wzory parametrów dla zmiennych losowych dyskretnych i ciągłych.

parametr	zmienne losowe dyskretne	zmienne losowe ciągłe
wartość oczekiwana	$E[X] = \sum_i x_i P(X = x_i)$	$E[X] = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x$
moment zwykły rzędu k	$m_k = E[X^k] = = \sum_i x_i^k P(X = x_i)$	$m_k = E[X^k] = = \int_{-\infty}^{\infty} x^k f(x) \mathrm{d}x$
moment centralny rzędu k	$\mu_k = E[(X - m_1)^k] = = \sum_i (x_i - m_1)^k P(X = x_i)$	$\mu_k = E[X^k] = = \int_{-\infty}^{\infty} (x - m_1)^k f(x) \mathrm{d}x$
wariancja	$D^2[X] = E[(X - m_1)^2] = = \sum_i (x_i - m_1)^2 P(X = x_i)$	$D^2[X] = E[(X - m_1)^k] = = \int_{-\infty}^{\infty} (x - m_1)^k f(x) \mathrm{d}x$
odchylenie standardowe	$D[X] = \sqrt{D^2[X]}$	$D[X] = \sqrt{D^2[X]}$

0.3. Przykłady rozkładów zmiennych losowych dyskretnych

0.3.1. Rozkład zero-jedynkowy

Niech doświadczenie daje dwa różne rezultaty. Wówczas zmienna losowa X będąca wynikiem takiego doświadczenia może przyjmować tylko jedną z dwóch wartości: x_1 i x_2. Rozkład takiej zmiennej losowej ma następującą postać:

$$P(X = x_1) = p_1$$

$$P(X = x_2) = p_2$$

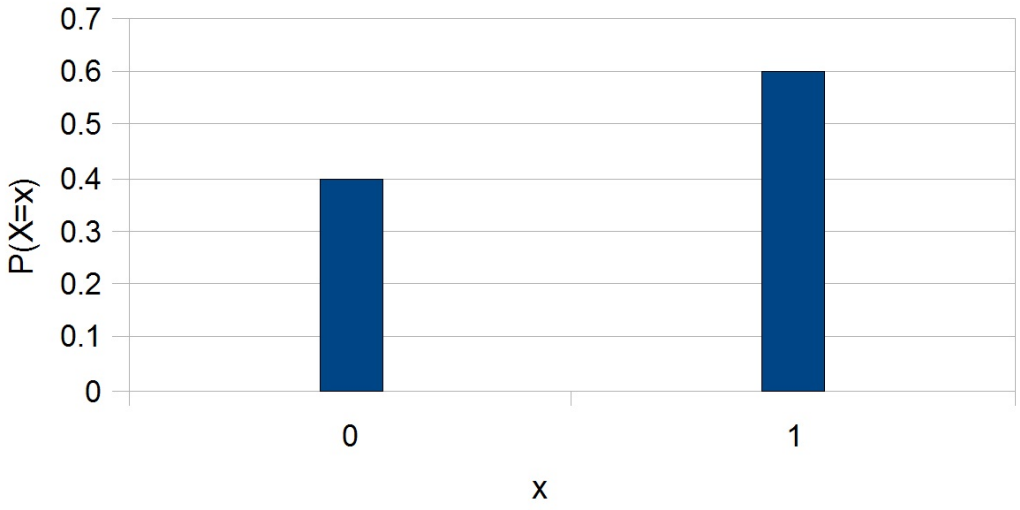
gdzie: $p_1 + p_2 = 1$. Taki rozkład nazywamy *dwupunktowym*.

Szczególnym przypadkiem rozkładu dwupunktowego jest rozklad *zero-jedynkowy*, dla którego $x_1 = 0$, a $x_2 = 1$. Funkcję rozkładu zero-jedynkowego zapisujemy w nastepujący sposób:

$$P(X = 1) = p$$

$$P(X = 0) = 1 - p$$

Podstawowe parametry rozkładu zero-jedynkowego	
wartość oczekiwana	p
wariancja	$p(1-p)$
odchylenie standardowe	$\sqrt{p(1-p)}$

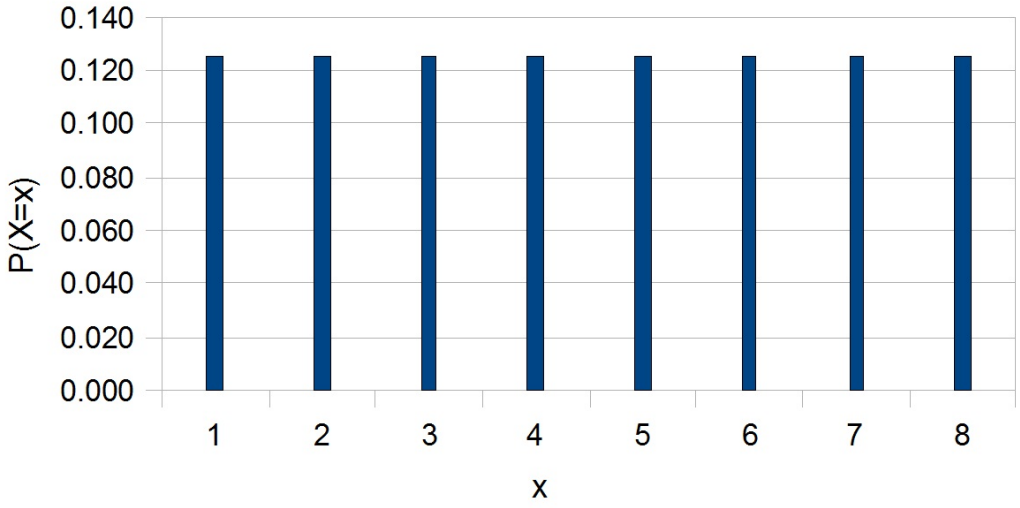


Rysunek 4. Rozkład dwumianowy z parametrem p=0,6

Rozkład jednostajny zmiennej losowej dyskretnej

Jeśli doświadczenie daje n różnych wyników $(x_1, x_2, \ldots, x_n)$ z jednakowym prawdopodobienstwem, to mówimy, że zmienna losowa X bedąca wynikiem tego doświadczenia ma rozkład jednostajny. Funkcja rozkładu prawdopodobieństwa dana jest wzorem:

$$\forall_{i=1,2,\ldots,n} P(X = x_i) = \frac{1}{n}$$



Rysunek 5. Rozkład jednostajny z n=8

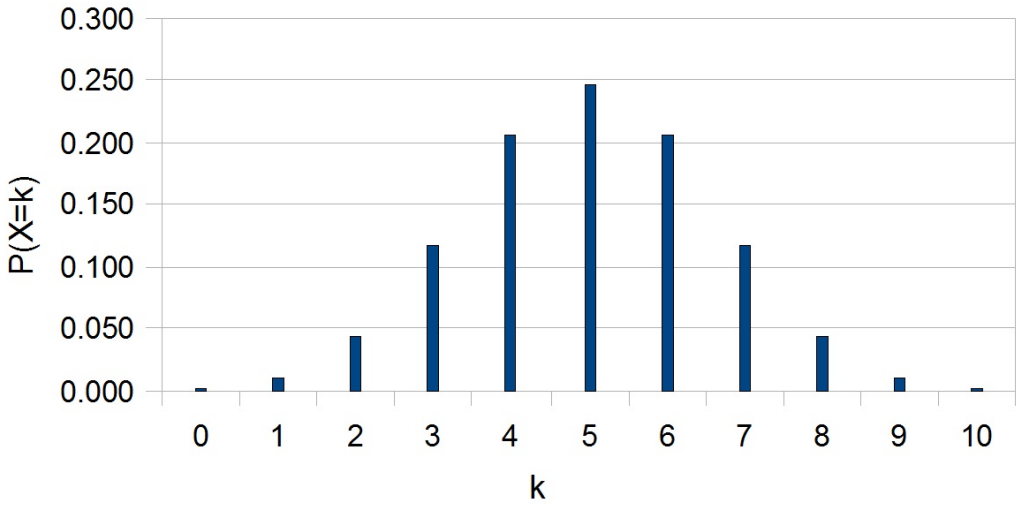
Podstawowe parametry rozkładu jednostajnego	
wartość oczekiwana	$\mathrm{E(X)}=\frac{1}{n}\sum_{i=1}^{n} x_i$
wariancja	$\frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))^2$
odchylenie standardowe	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))^2}$

Rozkład dwumianowy

Niech n doświadczeń spełnia *schemat Bernoulliego*, czyli nastepujące warunki:
— n jednakowych doswiadczeń przeprowadzanych w stałych warunkach,
— pojedyncze doświadczenia są niezależne.
— w wyniku doświadczenia może być zrealizowany tylko jeden ze stanów: „sukces" lub „porażka",
— prawdopodobieństwo sukcesu w każdym pojedyńczym doświadczeniu jest stałe i równe p $(0 < p < 1)$, a prawdopodobieństwo porażki jest równe $q = 1 - p$,
Wówczas zmienna losowa X okreslająca liczbę sukcesów w n doświadczeniach spełniających schemat Bernoulliego przyjmuje wartości $k = 1, 2, \ldots, n$ z prawdopodobienstwem:

$$P(X = k) = \binom{n}{k} p^k q^{n-k}$$

X jest zmienną losową o rozkładzie *dwumianowym*.



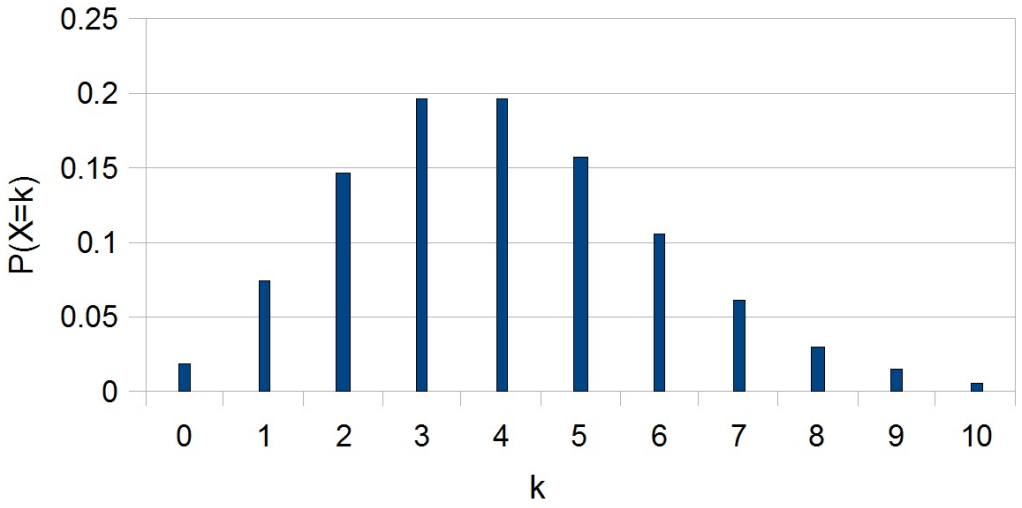
Rysunek 6. Rozkład dwumianowy z parametrami n=10 i p=0,5

Podstawowe parametry rozkładu dwumianowego	
wartość oczekiwana	np
wariancja	npq
odchylenie standardowe	$\sqrt{npq}$

Rozkład Poissona

Rozkład Poissona jest matematycznym modelem dla wielu zjawisk w przeyrodzie, między innymi oddzialywanie promieniowania na komórkę lub też rozmieszczenie zwierząt na danym terenie. Zmienna losowa X przyjmująca wartości $k = 0, 1, 2, \ldots$ ma rozkład Poissona o parametrze λ jeśli funkcja rozkładu prawdopodobienstwa ma rozkład:

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



Rysunek 7. Rozkład Poissona z parametrem λ=4

Podstawowe parametry rozkładu Poissona	
wartość oczekiwana	λ
wariancja	λ
odchylenie standardowe	$\sqrt{\lambda}$

0.3.2. Przykłady rozkładów zmiennych losowych ciągłych

Rozkład jednostajny zmiennej losowej ciągłej

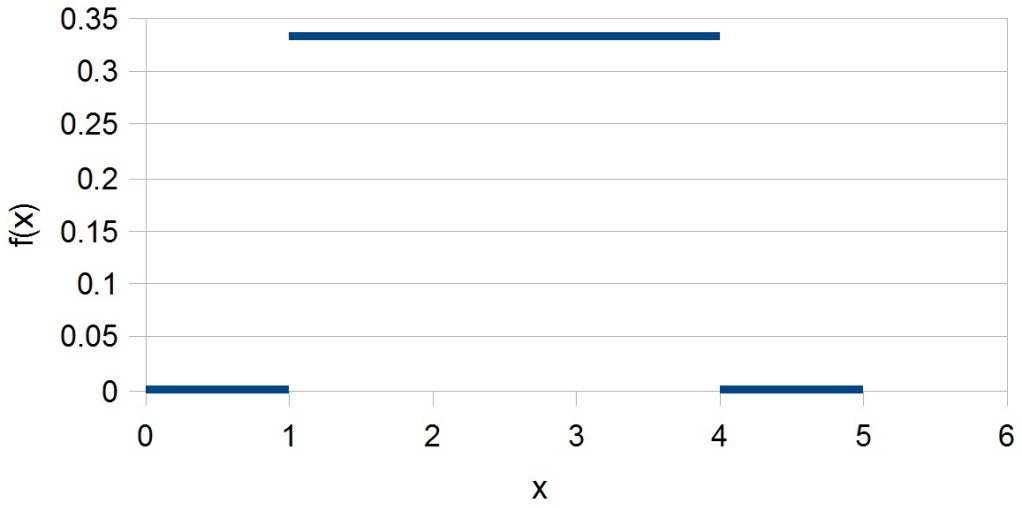
Zmienna losowa X ciągła ma rozkład jednostajny na przedziale $\langle a, b\rangle$, jeśli funkcja gęstości prawdopodobieństwa ma postać:

$$f(x) = \begin{cases} 0 & x < a \vee x > b \\ \frac{1}{b-a} & a \leqslant x \leqslant b \end{cases}$$

Natomiast dystrybuanta tego rozładu dana jest wzorem:

$$F(x) = \begin{cases} 0 & x \leqslant a \\ \frac{x-a}{b-a} & a < x \leqslant b \\ 1 & x > b \end{cases}$$

Podstawowe parametry rozkładu jednostajnego	
wartość oczekiwana	$\frac{b-a}{2}$
wariancja	$\frac{(b-a)^2}{12}$
odchylenie standardowe	$\frac{b-a}{2\sqrt{3}}$



Rysunek 8. Rozkład jednostajny na przedziale $\langle 1, 4\rangle$

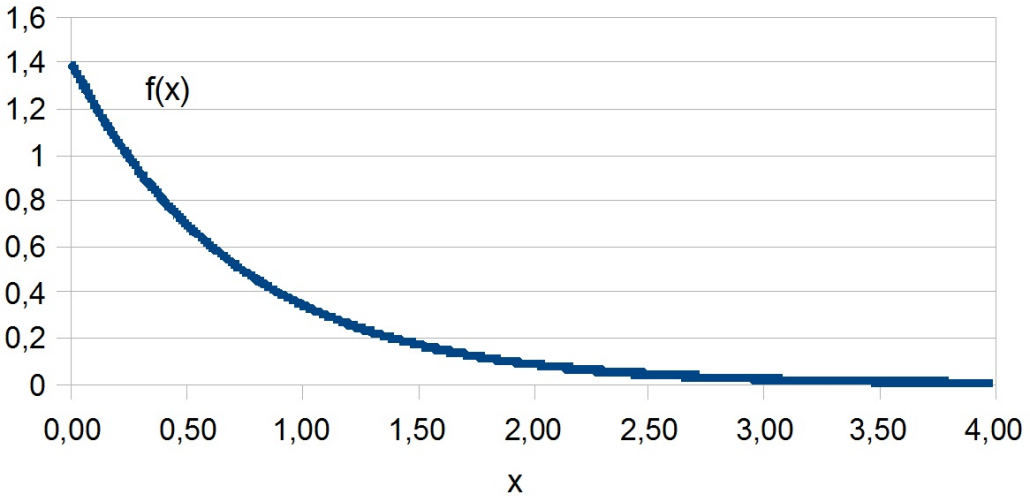
Rozkład wykładniczy

Ciągła zmienna losowa X ma rozkład wykładniczy, gdy jej funkcja gęstości prawdopodobieństwa jest dana wzorem:

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geqslant 0 \end{cases}$$

Z kolei dystrybuanta tego rozkładuma postać:

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geqslant 0 \end{cases}$$



Rysunek 9. Rozkład wykładniczy o parametrzy $\lambda = 1, 4$

Podstawowe parametry rozkładu jednostajnego	
wartość oczekiwana	$\frac{1}{\lambda}$
wariancja	$\frac{1}{\lambda^2}$
odchylenie standardowe	$\frac{1}{\lambda}$

Rozkład normalny

Rozkład normalny, nazywany również rozkładem Gaussa, jest najistotniejszym rozkładem w statystyce, ponieważ bardzo dużo zjawisk przyrodniczych podlega temy rozkladowi.

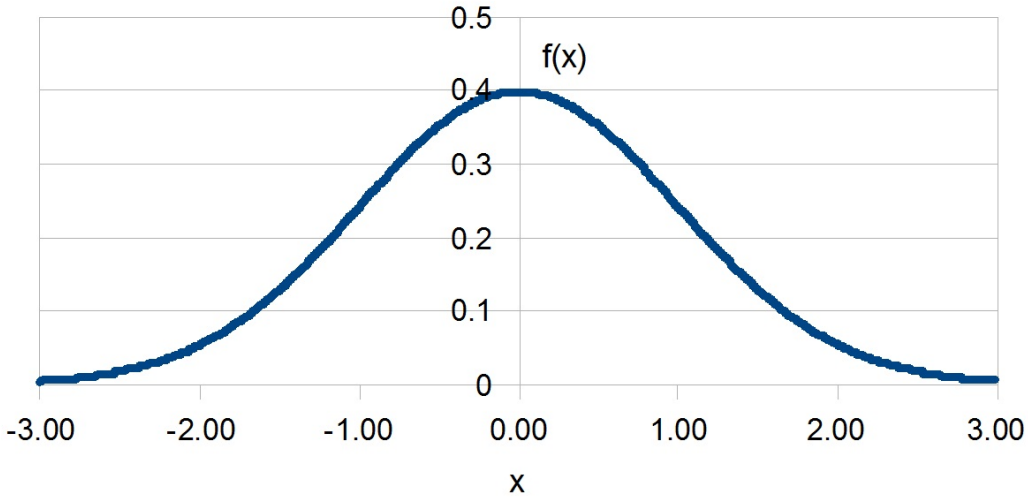
Zmienna losowa X ma rozkład normalny, jeśli funkcja gęstości ma postać:

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

gdzie m i $\sigma > 0$ są parametrami rozkładu narmalnego. Natomiast dystrybuanta rozkładu normalnego wyrażona jest wzorem:

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

Rozkład normalny o parametrach m i σ oznaczamy symbolem $N(m, \sigma)$. Rozkład normalny $N(0, 1)$ nazywamy rozkładem standardowym.



Rysunek 10. Rozkład normalny z parametrami m=0 i σ=1

Podstawowe parametry rozkładu normalnego	
wartość oczekiwana	m
wariancja	σ^2
odchylenie standardowe	σ

Ćwiczenia nr 2

Zadanie 0.14.

W taleli przedstawiono pewien rozkład prawdopodobieństwa:

x_i	1	2	3	4	5	6	7	8
p_i	0.1	0.2	0.1	0.05	0.15	0.3	0	0.1

Wyznacz dystrybuantę tego rozkładu.
Naszkicuj rozkład prawdopodobieństwa i dystrybuantę.
Oblicz:
— wartość oczekiwaną,
— wariancję,
— odchylenie standartowe.

Zadanie 0.15.

Niech $F(x)$ jest dystrybuantą pewnego rozkładu prawdopodobieństwa:

$$F(x) = \begin{cases} 0 & \text{dla } x \leqslant 1 \\ 0,2 & \text{dla } 1 < x \leqslant 2 \\ 0,3 & \text{dla } 2 < x \leqslant 4 \\ 0,5 & \text{dla } 4 < x \leqslant 5 \\ 0,55 & \text{dla } 5 < x \leqslant 6 \\ 0,7 & \text{dla } 6 < x \leqslant 7 \\ 0,9 & \text{dla } 7 < x \leqslant 8 \\ 1 & \text{dla } x > 81 \end{cases}$$

Wyznacz rozkład prawdopodobieństwa.
Naszkicuj rozkład prawdopodobieństwa i dystrybuantę.
Oblicz:
— wartość oczekiwaną,
— wariancję,
— odchylenie standartowe.

Zadanie 0.16.

Wśród 300 osób 16 choruje na pewną chorobę.
Oblicz:
— średnią liczbę osób chorujących na tą chorobę,
— odchylenie standartowe liczby osób chorujących.
Określ rozkład prawdopodobieństwa zmiennej losowej.
Jaki to typ rozkładu?
Wyznacz ogólne wzory na:
— wartość oczekiwaną,
— wariancję,
— odchylenie standartowe.

Zadanie 0.17.

Doświadczenie polega na rzucie sześcienną kostką do gry.
Oblicz:
— średnią liczbę oczek,
— odchylenie standartowe liczby oczek.
Określ rozkład prawdopodobieństwa zmiennej losowej.
Jaki to typ rozkładu?
Wyznacz ogólne wzory na:
— wartość oczekiwaną,
— wariancję,
— odchylenie standartowe.

Zadanie 0.18.

W sklepie medycznym znajdują sie opatrumki firmy A***. Wiadomo, że 20% tych opatrunków jest wadliwych. Pewna kobiet kupiła 5 opatrunków firmy A***.
Oblicz:
— średnią liczbę wadliwych opatrunków kupionych przez kobietę,
— odchylenie standartowe liczby wadliwych opatrunków.
Określ rozkład prawdopodobieństwa zmiennej losowej.
Jaki to typ rozkładu?
Wyznacz ogólne wzory na:
— wartość oczekiwaną,
— wariancję,
— odchylenie standartowe.

Zadanie 0.19.

W taleli przedstawiono liczbę zachorowań wzgledem czasu, który minął od momentu zetknięcia się osoby zdrowej z chorym.

czas (w dniach)	1	2	3	4	5	6	7	8
liczba chorych	18	37	49	49	39	26	15	7

Oblicz:
— średni czas rozwijania się obiawów choroby,
— odchylenie standartowe czasu rozwijania się obiawów choroby.
Określ rozkład prawdopodobieństwa zmiennej losowej.
Jaki to typ rozkładu?
Wyznacz ogólne wzory na:
— wartość oczekiwaną,
— wariancję,
— odchylenie standartowe.

Rozkład Poissona:

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

Zadanie 0.20.

Wiemy, że autobus pewnej lini kursuje co 10 minut. Nie mamy żadnych informacji informacji co do rozkładu jazdy. wychodzimy więc w losowym momęcie na przystanek.
Należy obliczyć prawdopodobieństwo, że będziemy czekać na autobus krócej niż 3 minuty oraz:
— średni czas oczekiwania na autobus,
— odchylenie standartowe czasu oczekiwania na autobus.
Określ rozkład prawdopodobieństwa zmiennej losowej. Jaki to typ rozkładu?
Wyznacz ogólne wzory na:
— wartość oczekiwaną,
— wariancję,
— odchylenie standartowe.

Zadanie 0.21.

Rozkład czasu leczenia pewnej choroby jest rozkładem wykładniczym z parametrem $\lambda = 0,2$.
Rozklad prawdopodobieństwa:

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geqslant 0 \end{cases}$$

Oblicz:
— średni czas leczenia,
— odchylenie standartowe.
Wyznacz ogólne wzory na:
— wartość oczekiwaną,
— wariancję,
— odchylenie standartowe.

Zadanie 0.22.

Wartość hematokrytu we krwi młodych kobiet podlega rozkładowi normalnemu o parametrach $\mu = 41\%$ oraz $\sigma = 2,4\%$. Obliczyć prawdopodobienstwo tego, że poziom hematokrytu we krwi młodych kobiet zawiera się w granicaach od 40% do $43,5\%$ oraz prawdopodobieństwo tego, że jest on większy niż $39,5\%$.
Rozklad prawdopodobieństwa:

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geqslant 0 \end{cases}$$

Oblicz:
— średni czas leczenia,
— odchylenie standartowe.
Wyznacz ogólne wzory na:
— wartość oczekiwaną,
— wariancję,
— odchylenie standartowe.

Rozdział 1

Doświadczenia losowe, populacja i próba

1.1. Statystyka jako narzędzie badawcze nauk przyrodniczych. Populacja generalna a próba

Statystyka zajmuje się badaniem prawidłowości występujących w zjawiskach masowych. Jej podstawowym zadaniem jest dostarczanie wiarygodnych informacji o cechach badanych zjawisk. Statystyka znajduje zastosowanie we wszystkich dziedzinach wiedzy, w których analizujemy duże ilości danych. Dzisiaj statystyka stosowana jest w wielu rozmaitych dziedzinach, między innymi w: biometrii, ekonomii, fizyce, biologii, medycynie, socjologii oraz psychologii.

Przykładowo wyniki pomiarów biometrycznych po opracowaniu metodami statystycznymi wykorzystywane są, między innymi w antropologii, fizjologii, genetyce oraz medycynie. W ekonomii statystyka jest wykorzystywana do badania ilościowych związków zachodzących między zjawiskami ekonomicznymi. Podstawową gałęzią fizyki w której wykorzystuje się statystykę jest fizyka statystyczna, która zajmuje się opisem własności układów złożonych z wielu oddziałujących ciał. Natomiast w biologii i medycynie statystykę stosuje się między innymi do ilościowego opisu wpływu działania leku na grupę pacjentów chorujących na tę samą chorobę. W socjologii wykorzystuje się statystykę do określania cech charakterystycznych dla danej zbiorowości. Wykorzystując metody statystyczne socjolodzy mogą badać preferencje polityczne studentów Uniwersytetu Gdańskiego lub inne temu podobne tematy. Natomiast psycholodzy wykorzystując metody statystyczne mogą badać przykładowo podatność na stres w grupie studentów gdańskich uczelni.

Aby uzyskać wiarygodne informacje na temat badanych zbiorowości wykonuje się badania statystyczne. Takie badania przeprowadza się na *zbiorowości statystycznej*, czyli na zbiorze elementów podobych pod względem określonych cech. Badania statystyczne można przeprowadzać na wszystkich elementach zbioru, czyli na *zbiorowości generalej*, nazywanej *populacją*, lub na zbiorze elementów wybranych z populacji, czyli na *próbie*. Badanie przeprowadzone na całej populacji nzwamy *badaniem kompletnym* (całkowitym, stuprocentowym), dostarcza ono pełnej informacji o badanej cesze, lecz prawie nigdy go nie stosujemy. Jest tak, ponieważ badania kompletne są trudne lub wręcz nie możliwe do wykonania. Badania kompletnego nie możmy wykonać naprzykład gdy zbiór badanych elementów jest nieskończony, lub też badanie ma charakter niszczący (np. badamy zawartość nieoznakowanych konserw otwierając je, po badaniu mamy pełną informację o ich zawartości, lecz nie mamy już konserw). Z badań kompletnych rezygnujemy również przy dużej populacji, ponieważ jest ono zbyt kosztowne i czasochłonne. Wówczas wykonujemy *badania niekompletne* przeprowadzane na próbie, a uzyskane wyniki uogólniamy na populację. Próba powinna być reprezentacyjną częścią populacji, to znaczy częstości występowania cech w próbie nie powinny różnić się od częstości występowania cech w

populacji. Aby to osiągnąć, elementy próby losujemy z populacji uzyskując w ten sposób *probę losową*. Próbę losową n-elementową nazywamy *prostą*, gdy każdy n-elementowy podzbiór populacji ma to samo prawdopodobienstwo wylosowania. Próby można podzielić na:

— duże – więcej niż 30 badanych elementów,
— małe – 30 lub mniej badanych elementów.

W badaniu statystycznym możemy analizować jedną lub więcej cech populacji. Badaniu mogą podlegać cechy mierzalne (ilościowe), ktore są wyrażane w jednostkach takich, jak np. kg, m, Ω, itd, a także cechy jakościowe (niemierzalne), np. kolor, płeć, wyznanie, itp. Badania ststystyczne mają cztery zasadnicze etapy:

1. Przygotowanie badania
2. Obserwacja statystyczna
3. Opracowanie i prezentacja materiału statystycznego
4. Opis zbiorowości lub/i wnioskowanie statystyczne

Statystyka składa się z dwóch głównych działów: statystyki opisowej i statystyki matematycznej. Opisem zbiorowości bez posługiwania się rachunkiem prawdopodobieństwa zajmuje się statystyka opisowa, której poświęcony jest rozdział 2. Natomiast statystyka matematyczna zajmuje się metodami wnioskowania statystycznego, polegającymi na formułowaniu wniosków o populacji na podstawie wyników uzyskanych z próby, przy czym używamy metod rachunku prawdopodobieństwa. Statystyce matematycznej poświęcone są kolejne rozdziały.

1.2. Szeregi statystyczne

Materiał statystyczny zebrany w badaniach zawiera dużą liczbę wyników. Wyniki te powinny zostać usystematyzowane i odpowiednio zestawione, co pozwoli na przejście od danych indywidualnych do danych zbiorowych. Opracowanie materiału statystycznego polega na porządkowaniu, zliczaniu i grupowaniu danych statystycznych. W wyniku czego otrzymujemy tak zwany szereg statystyczny. *Szeregiem statystycznym* nazywamy uporządkowany (malejąco lub rosnąco) ciąg wielkości statystycznych pogrupowanych według określonych kryteriów. Najczęściej szereg statystyczny składa się z dwóch kolumn:

— pierwsza – podaje wielkość cechy lub czas,
— druga – określa liczbę jednostek przypadających na daną cechę lub jednostkę czasu.

Na rysunku 1.1 została przedstawiona klasyfikacja szeregów statystycznych.

Szeregiem szczegółowym nazywamy uporządkowany (rosnąco lub malejąco) ciąg wszystkich wartości badanej cechy statystycznej. Jest on stosowany jedynie w przypadku, gdy badana próba jest niewielka.

Przykład 1.1:

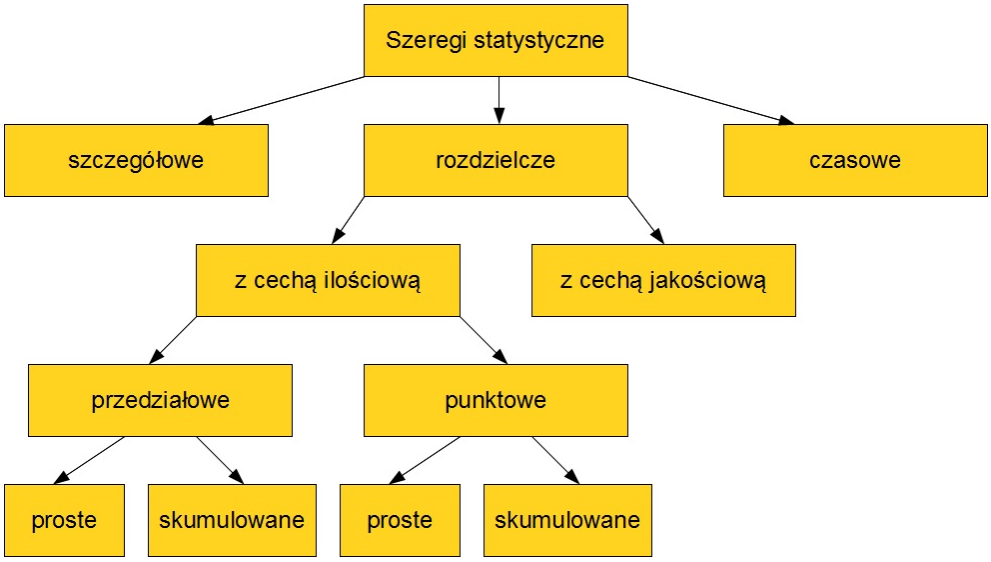
Dokonano pomiaru wzrostu (w cm) 10 studentów z jednej grupy ćwiczeniowej. Otrzymano następujące wyniki: 178, 192, 169, 185, 176, 178, 165, 171, 168, 170

Szereg szczegółowy uporządkowany rosnąco:

165 168 169 170 171 176 178 178 185 192

Szereg szczegółowy uporządkowany malejąco:

192 185 178 178 176 171 170 169 168 165



Rysunek 1.1. Klasyfikacja szeregów ststystycznych

Szereg czasowy otrzymujemy z pomiarów cechy wykonywanych w odstępach czasowych, jest on ciągiem informacji uporządkowanych w czasie.

Przykład 1.2:

Poniższa tabela obrazuje wzrost zachorowań na nowotwory złośliwe w pewnym regionie Polski w latach 1986-1995.

Lata	Zachorowania w tys.
1986	78,5
1987	80,9
1988	82,5
1989	82,8
1990	83,4
1992	94,8
1993	95,6
1994	96,1
1995	97,4

Szereg rozdzielczy otrzymamy, gdy próbę podzielimy na klasy według określonej cechy i podamy liczebność tych klas. *Szeregi rozdzielcze punktowe* budujemy, gdy liczba wariantów cechy jest niewielka.

Przykład 1.3:

Na pewnej uczelni przeprowadzono badanie grupy krwi studentów. Wylosowano 30 osób. Wyniki badania reprezentuje poniższy szereg rozdzielczy punktowy.

Grupa krwi	Liczba osób
A	11
B	5
AB	3
0	11

Szeregi rozdzielcze z przedziałami klasowymi budujemy, gdy jest dużo wariantów cech. Wartości próbki grupujemy w kalsach, czyli przedziałach (najczęściej tej samej długości) i przyjmujemy, że wszystkie wartości znajdujące się w daniej klasie są identyczne ze środkiem klasy.

W przypadku szeregów rozdzielczych z cechami jakościowymi nie powinno być trudności. Każda wartość cechy jest klasą. Natomiast w konstrukcji szeregu rozdzielczego z cechami ilościowymi dla prób n-elementowych wyróżnia się następujące etapy:

1. Ustalenie liczby klas – istnieje kilka reguł wyznaczanie liczby klas (k), na przykład:

$$k = 1 + 3,322 \cdot \ln n \qquad \text{czy} \qquad k = \sqrt{n}.$$

2. Wyznaczenie wielkości przedziałów klasowych i granic klas – ten etap składa się z 3 kroków:

— krok 1: wyznaczenie rozstępu R badanej cechy:

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

gdzie: x_{max} – największa wartość badanej cechy, x_{min} – najmniejsza wartość badanej cechy

— krok 2: wyznaczenie długości przedziału (b):

$$b \approx \frac{R}{k} \text{ takie, że } bk \geqslant R$$

— krok 3: ustalenie granic klas:
granice klas ustala się zwykle z dokładnością $\frac{1}{2}\alpha$, gdzie α – dokładność z jaką wyznaczono wartości w próbie

3. Zakwalifikowanie danych do przyjętych przedziałów klasowych.

4. Zliczenie liczby jednostek w każdej klasie. – W zależności czy wyznaczamy szereg rozdzielczy prosty, czy skumulowany inaczej zliczamy liczebności klas:

— szereg rozdzielczy prosty:
każdej klasie przyporządkowujemy liczbę danych do niej zakwalifikowanych (n_i), w ten sposób i-tej klasie przyporządkowujemy liczbę n_i, dla $i = 1, 2, \ldots, k$

— szereg rozdzielczy skumulowany:
wykorzystując dane z szeregu rozdzielczego prostego tworzymy szereg skumulowany w ten sposób, że i-tej klasie przyporządkowujemy liczbę $\sum_{j=1}^{i} n_j$, dla $i = 1, 2, \ldots, k$, gdzie n_j jest liczebnością j-tej klasy szeregu rozdzielczego prostego

Przykład 1.4:

W grupie 40 osób przeprowadzono badanie ze względu na minutową pojemność płuc (w litrach) i otrzymano wyniki: 7,6 6,6 8,9 7,6 8,0 6,7 8,5 7,8 7,4 6,8 8,0 8,1 7,7 8,2 7,9 7,1 6,3 6,7 8,7 6,8 7,5 7,9 7,0 6,1 8,0 7,5 7,2 8,4 7,4 9,0 7,8 6,1 7,7 7,2 7,6 7,6 7,7 6,4 7,8 7,3.
Konstrukcja szeregu rozdzielczego:
Liczba elementów szeregu: $n = 40$

1. Ustalenie liczby klas – $k = \sqrt{n} = \sqrt{40} \approx 6$
2. Wyznaczenie wielkości przedziałów klasowych i granic klas:

— krok 1: wyznaczenie rozstępu R badanej cechy:

$$x_{max} = 9 \qquad x_{min} = 6,1$$

$$R = x_{max} - x_{min} = 2,9$$

— krok 2: wyznaczenie długości przedziału (b):

$$b \approx \frac{R}{k} \approx 0,48 \approx 0,5$$

— krok 3: ustalenie granic klas:

$$\alpha = 0,1 \qquad \frac{1}{2}\alpha = 0,05$$

Wyznaczone klasy: (6,05 , 6,55), (6,55 , 7,05), (7,05 , 7,55), (7,55 , 8,05), (8,05 , 8,55), (8,55 , 9,05)

3. zakwalifikowanie danych do przyjętych przedziałów klasowych

Klasa	zakwalifikowane wartości
(6,05 , 6,55)	6,3 6,1 6,1 6,4
(6,55 , 7,05)	6,6 6,7 6,8 6,7 6,8
(7,05 , 7,55)	7,4 7,1 7,5 7,0 7,5 7,2 7,4 7,2 7,3
(7,55 , 8,05)	7,6 7,6 8,0 7,8 8,0 7,7 7,9 7,9 8,0 7,8 7,7 7,6 7,6 7,7 7,8
(8,05 , 8,55)	8,5 8,1 8,2 8,4
(8,55 , 9,05)	8,9 8,7 9,0

4. Zliczenie liczby jednostek w każdej klasie

— szereg rozdzielczy prosty:

Klasa	Liczebność klasy
(6,05 , 6,55)	4
(6,55 , 7,05)	5
(7,05 , 7,55)	9
(7,55 , 8,05)	15
(8,05 , 8,55)	4
(8,55 , 9,05)	3

— szereg rozdzielczy skumulowany:

Klasa	Liczebność klasy
(6,05 , 6,55)	4
(6,55 , 7,05)	4+5=9
(7,05 , 7,55)	4+5+9=18
(7,55 , 8,05)	4+5+9+15=33
(8,05 , 8,55)	4+5+9+15+4=37
(8,55 , 9,05)	4+5+9+15+4+3=40

1.3. Graficzne formy prezentacji szeregów statystycznych

Szeregi rozdzielcze można przedstawiać graficznie w postaci histogramu, łamanej częstości lub wieloboku częstości.

1.3.1. Histogram

Histogram jest wykresem kolumnowym, na którego osi poziomej odłożone są środki lub granice klas, a na osi pionowej liczności klas n_i, częstości klas $w_i = \frac{n_i}{\sum_{i=1}^{k} n_i}$ bądź

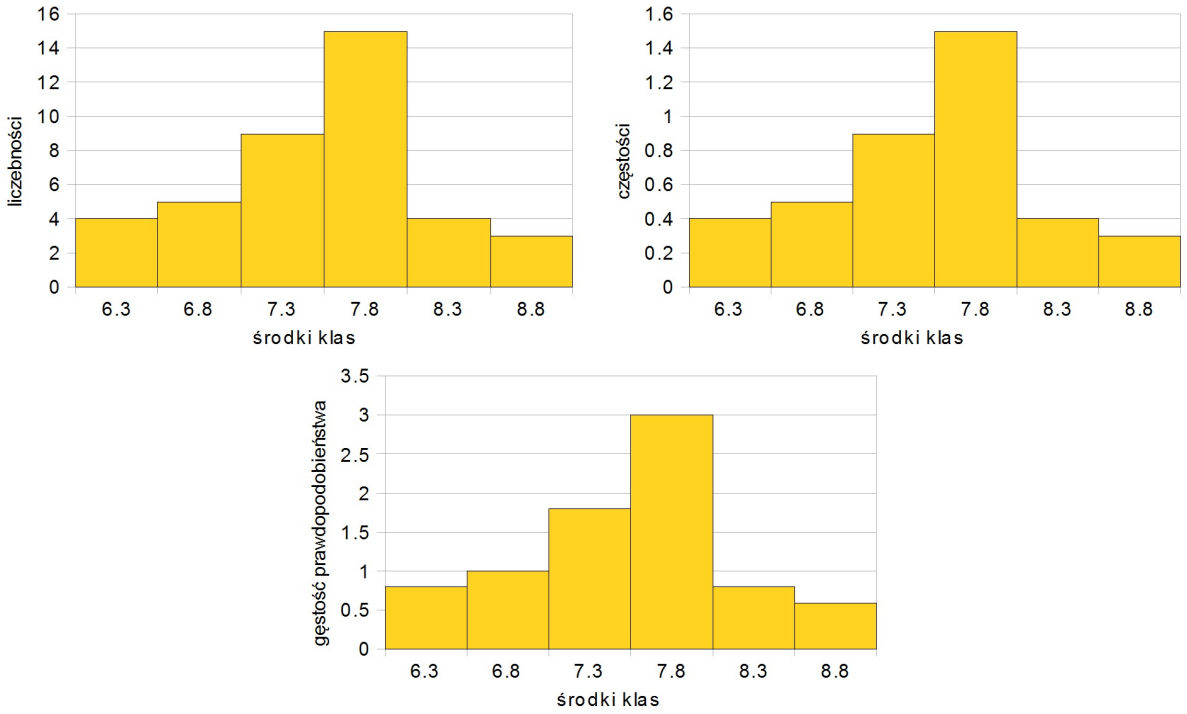
gęstość prawdopodobieństwa $v_i = \frac{w_i}{b}$. W ostatnim wypadku pole histogramu równe jest jedności, ponieważ:

$$\sum_{i=1}^{k} b \cdot v_i = \sum_{i=1}^{k} b \cdot \frac{w_i}{b} = \sum_{i=1}^{k} w_i = \sum_{i=1}^{k} \frac{n_i}{\sum_{i=1}^{k} n_i} = \frac{1}{\sum_{i=1}^{k} n_i} \cdot \sum_{i=1}^{k} n_i = 1$$

Przykład 1.5:

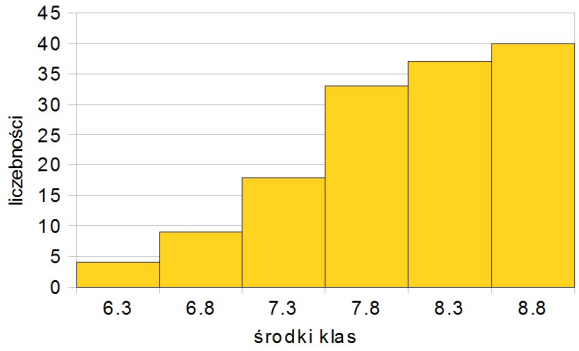
Histogramy dla danych z przykładu 1.4:

— szereg rozdzielczy prosty:



Rysunek 1.2.

— szereg rozdzielczy skumulowany:



Rysunek 1.3.

1.3.2. Łamana częstości. Wielobok częstości

Łamana częstości jest wykresem liniowym. Na osi poziomej tego wykresu są odłożone środki klas, natomiast na osi pionowej gęstości prawdopodobieństwa v_i. Łamaną częstości tworzą linie łączące punkty o współrzędnych $(\overline{x}_i, v_i)$ i $(\overline{x}_{i+1}, v_{i+1})$ dla $i = 1, 2, \ldots, k-1$, gdzie $\overline{x}_i$ - środek i-tej klasy.

Przykład 1.6:

Łamana częstości dla danych z przykładu 1.4:

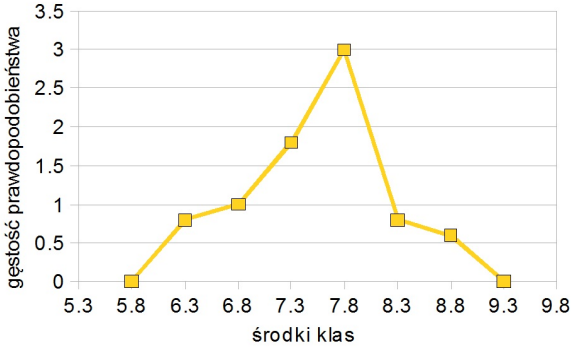


Rysunek 1.4. Łamana częstości

Wielobok częstości jest wykresem liniowym zawierającym łamaną częstości i dodatkowe dwa odcinki łączące punkty o współrzędnych:
— $(x_1 - b, 0)$ i (x_1, v_1),
— (x_k, v_k) i $(x_k + b, 0)$.
Pole wieloboku częstości, co nie trudno udowodnić, jest równe jedności.

Przykład 1.7:

Wielobok częstości dla danych z przykładu 1.4:



Rysunek 1.5. Wielobok częstości

Ćwiczenia nr 3

Zadanie 1.1. [6]

Grupę 50 mężczyzn w wieku 40 lat zbadano ze względu na liczbę brakujących zębów w uzębieniu i otrzymano następujące wyniki:

1	3	0	1	2	5	0	2	1	2
2	3	2	3	6	3	2	4	0	1
4	1	1	2	0	5	1	1	3	2
2	3	4	3	6	3	4	5	3	0
4	0	1	2	4	0	2	0	1	2

— Zbudować szereg rozdzielczy punktowy.
— Zbudować histogram.
— Zbudować wielobok częstości.

Zadanie 1.2. [2]

Z partii bawełny pobrano próbkę złożoną z 64 włókien, a następnie zmierzono długość tych włókien (w mm). Otrzymano następujące wyniki:

23	8	15	35	21	20	10	4	28	12	9	7	24	25	31	26
23	17	13	33	29	27	24	22	32	16	9	29	22	20	8	16
21	25	31	29	23	15	32	22	23	19	24	15	21	20	29	27
23	19	16	18	24	31	28	21	8	17	24	13	12	18	23	25

— Zbudować szereg rozdzielczy.
— Narysować histogram, tak dobierając skalę na osi pionowej, aby pole histogramu było równe 1.
— Narysować łamaną częstości.

Zadanie 1.3.

Wygeneruj próbkę (n=1000) liczb losowych o rozkładzie normalnym $N(0,1)$. Dla otrzymanego ciągu liczb:
— zbudować szereg rozdzielczy,
— narysować histogram (na osi pionowej odłożyć liczebności klas),
— narysować łamaną częstości.

Zadanie 1.4.

Wygeneruj próbkę (n=5000) liczb losowych z przedziału $[0,1]$ o rozkładzie normalnym. Dla otrzymanego ciągu liczb:
— zbudować szereg rozdzielczy,
— narysować histogram (na osi pionowej odłożyć częstości dla poszczególnych klas),
— narysować wielobok częstości.

Zadanie 1.5.

Wygeneruj próbkę (n=2000) liczb losowych o rozkładzie normalnym $N(3, \frac{1}{2})$. Dla otrzymanego ciągu liczb:

— zbudować szereg rozdzielczy,

— narysować histogram (na osi pionowej odłożyć gęstości prawdopodobieństwa poszczególnych klas),

— narysować wielobok częstości.

Ćwiczenia nr 3 – Matlab

Funkcje

hist

`hist(A)` – wyświetla histogram wartości tablicy A w 10 klasach.
`L=hist(A)` – tworzy tablicę L liczebności klas histogramu wartości tablicy A podzielonych na 10 klas.
`hist(A, k)` – wyświetla histogram wartości tablicy A w k klasach.
`L=hist(A, k)` – tworzy tablicę L liczebności klas histogramu wartości tablicy A podzielonych na k klas.
`[L, S]=hist(A, k)` – tworzy tablice L – liczebności klas histogramu oraz S – środków klas dla wartości tablicy A podzielonych na k klas.

plot

`plot(A)` – wyświetla łamaną łączącą punkty $(i, A(i))$, gdzie i jest indeksem kolejnych wartości tablicy A.
`plot(X, A)` – wyświetla łamaną łączącą punkty $(X(i), A(i))$, gdzie i jest indeksem kolejnych wartości tablic X i A.
`plot(X, A,'specyfikacje')` – elementy wyświetlane są jak w powyższym przykładzie, ponadto specyfikacje pozwalają na określenie typu lini, koloru wyświetlenia oraz typu punktora. W poniższej tabeli znajdują się możliwe specyfikacje.

typ lini		kolor		typ punktora	
-	linia ciągła	r	czerwony	+	znak plus
- -	linia przerywana	g	zielony	o	okrąg
:	linia kropkowana	b	niebieski	*	gwiazdka
-.	linia kreskowano-kropkowana	c	cyjan	.	punkt
		m	magenta	x	krzyżyk
		y	żółty	s	kwadrat
		k	czarny	d	romb

Na przykład ':b+' da nam wykres z linią kropkowaną w kolorze niebieskim i punktorami +.

randn

`r=randn` – generuje jedną liczbę losową z rozkładu normalnego standardowego ($N(0,1)$).
`r=randn(n)` – generuje macierz $n \times n$ liczb losowych z rozkładu normalnego standardowego ($N(0,1)$).
`r=randn(m,n)` – generuje macierz $m \times n$ liczb losowych z rozkładu normalnego standardowego ($N(0,1)$).

round

`round(A)` – wszystkie elementy macierzy zaokrąglane są do najbiższej liczby całkowitej.

set

`set(gca,'XTick',A)` – instrukcja pozwala ustawić na aktualnym wykresie podziałkę na osi X po przez zaznaczenie na osi X elementów zbioru A.

sum

`sum(A)` – jeśli A jest wektorem, to funkcja zwraca sumę wartości wektora A, natomiast jeśli A jest macierzą, to funkcja zwraca wektor zawierający sumy wartości w kolumnach macierzy A.

xlabel

`xlabel('Tytuł X');` – instrukcja pozwalająca podpisać poziomą oś na aktualnym wykresie tekstem: `Tytuł X`.

ylabel

`ylabel('Tytuł Y');` – instrukcja pozwalająca podpisać pionową oś na aktualnym wykresie tekstem: `Tytuł Y`.

Przykładowe zadanie

Zadanie I.

Wygeneruj próbkę (n=20000) liczb losowych o rozkładzie normalnym $N(100,10)$. Dla otrzymanego ciągu liczb:
— zbudować szereg rozdzielczy,
— narysować histogram (na osi pionowej odłożyć liczebność poszczególnych klas),
— narysować łamaną częstości.

krok 1: Generowanie próby o rozkładzie $N(100,10)$
Aby wygenerować próbę o rozkładzie normalnym $N(0,1)$ wykorzystujemy funkcję `randn`. Próbę o rozkładzie $N(m,1)$ otrzymujemy dodając do każdej wylosowanej wartości m. Natomiast próbę o rozkładzie $N(0,\sigma)$ – mnożąc każdą wylosowaną wartość przez σ. Gdy chcemy mieć próbę o rozkładzie $N(m,\sigma)$ mnożymy każdą wylosowaną wartość przez σ, a następnie dodajemy m.

```
A=randn(20000,1)*10+100;
```

krok 2: Budowanie szeregu rozdzielczego

Aby zbudować szereg rozdzielczy należy ustalić liczbę klas k Dla próby n-elementowej $k \approx \sqrt{n}$. Wykorzystamy funkcje: `sqrt` oraz `round`.

```
k=round(sqrt(20000));
```

Mając liczbę klas można przejść do ustalenia ich liczebności. Wykorzystamy funkcję `hist`.

```
[L,S]=hist(A,k)
```

Tablicę R zawierającą szereg rozdzielczy tworzymy w następujący sposób:

```
R(1,:)=S;
R(2,:)=L;
  R=R'
```

krok 3: Histogram

Histogram wyświetlamy instrukcją `hist`. Przy dużych liczbach klas warto wcześniej zadbać o estetyczne wyświetlenie wartości na osi poziomej. Aby podziałka na osi X zawierała tylko niektóre wartości środków klas należy utworzyć nową tablicę w następujący sposób:

```
for i=1:k/10
S1(i,1)=S(i*10-9);
end
```

Wowczas histogram wyświetlamy korzystając z nastepujących instrukcji:

```
hist(A,k)
set(gca,'XTick',S1)
xlabel('Środki klas');
ylabel('Liczebność');
```

krok 4: Łamana częstości

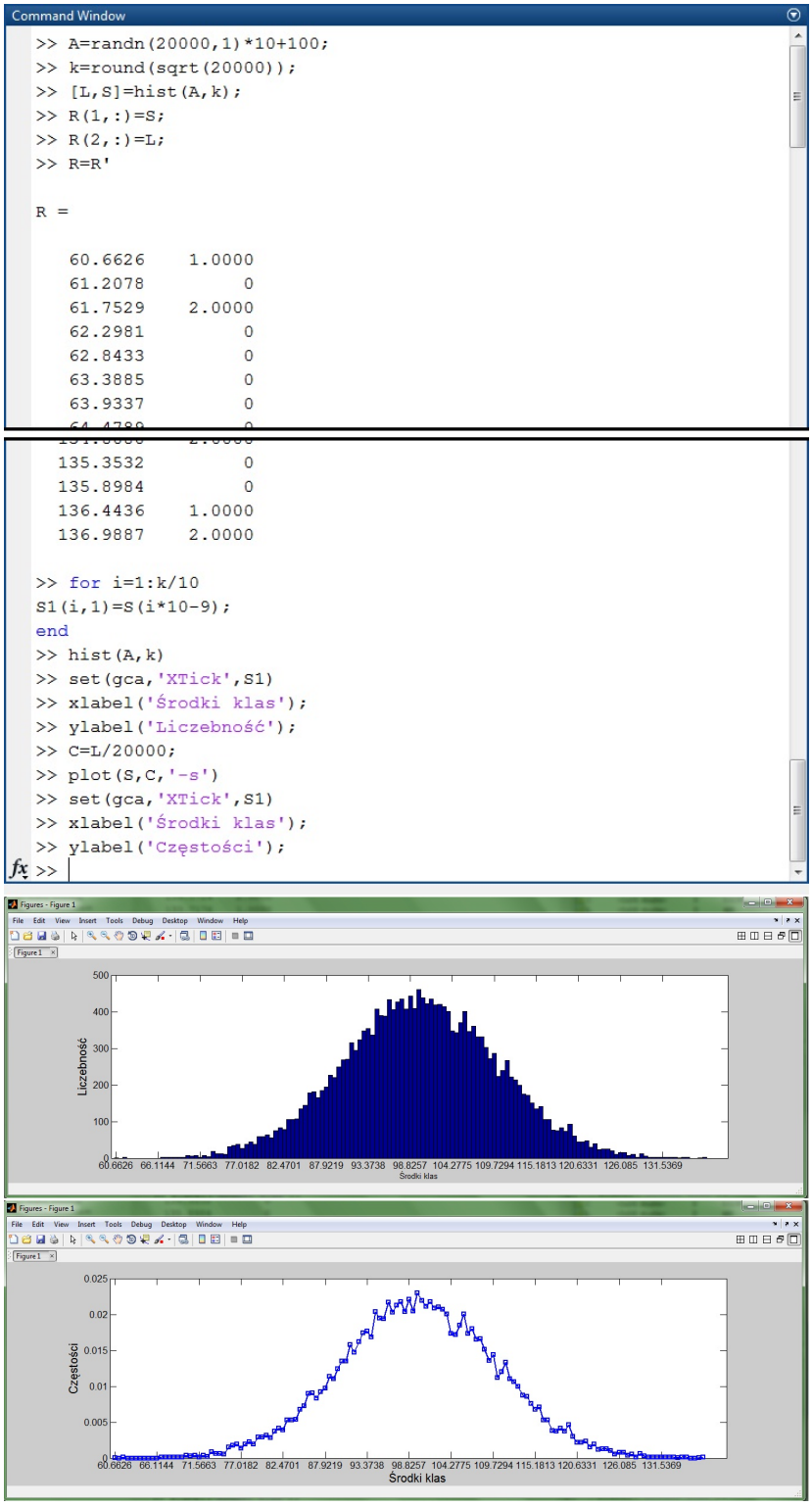
W tablicy L znajdują się liczebności poszczególnych klas, aby uzyskać tablicę z wartościami częstości należy każdy wyrza z tablicy L podzielić przez liczebność próby.

```
C=L/20000;
```

Łamaną częstości wyświetlamy instrukcją `plot`. Wówczas łamaną częstości wyświetlamy poleceniem:

```
plot(S,C,'-s')
set(gca,'XTick',S1)
xlabel('Środki klas');
ylabel('Częstości');
```

Na rysunku 1.6 znajduje się okno poleceń z poprawnie wykonanym zadaniem oraz uzyskane wykresy.



Rysunek 1.6. Rozwiązanie zadania I

Rozdział 2

Statystyka opisowa

Statystyka opisowa zajmuje się wstępną analizą próby bez użycia rachunku prawdopodobieństwa. Jedną z podstawowych technik statystyki opisowej jest wyznaczanie *miar rozkładu*, które są ściśle zdefiniowane ze pomocą wzorów. Miary rozkładu stanowią różnego rodzaju wielkości, które opisują zbiorowość w taki sposób, że określają jej podstawowe własności oraz pozwalają porównywać różne zbiorowości. Miary rozkładu dzielimy na cztery podstawowe grupy:

— miary położenia,
— miary rozproszenia,
— miary asymetrii,
— miary koncentracji.

2.1. Momenty statystyczne

Momenty statytstyczne są wilelkościami odzwierciedlającymi rozmaite cechy rozkładu wartości w próbie. Wyróżniamy:

— momenty zwykłe,
— momenty centralne,
— momenty absolutne zwykłe,
— momenty absolutne centralne.

2.1.1. Momenty zwykłe i centralne

Moment zwykły rzędu k z próby n-elementowej wyznaczamy korzystając z wzorów:

— dla szeregu szczegółowego o wartościach x_i:

$$m_k = \frac{1}{n} \sum_{i=1}^{n} {x_i}^k$$

— dla szeregu rozdzielczego punktowego o l wartościach x_i, które występują n_i-krotnie:

$$m_k = \frac{1}{n} \sum_{i=1}^{l} {x_i}^k n_i$$

— dla szeregu rozdzielczego przedzialowego o l klasach o środkach $\overline{x_i}$ i liczebności n_i:

$$m_k = \frac{1}{n} \sum_{i=1}^{l} {\overline{x_i}}^k n_i$$

Z łatwością można zauważyć, że pierwszy moment zwykły jest poprostu średnią arytmetyczną z wartości próby.

Z kolei moment centralny rzędu k z próby n-elementowej wyznaczamy za pomocą wzorów:

— dla szeregu szczegółowego o wartościach x_i:

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

— dla szeregu rozdzielczego punktowego o l wartościach x_i, które występują n_i-krotnie:

$$\mu_k = \frac{1}{n}\sum_{i=1}^{l}(x_i - \bar{x})^k n_i$$

— dla szeregu rozdzielczego przedzialowego o l klasach o środkach $\overline{x_i}$ i liczebności n_i:

$$\mu_k = \frac{1}{n}\sum_{i=1}^{l}(\overline{x_i} - \bar{x})^k n_i$$

Pierwszy moment centralny posiada swoistą cechę – zawsze jset rowny zeru ($\mu_1 = 0$), a dowód tego fakktu jest trywialny. Pnoadto łatwo zauważyć, że drugi moment centralny jest wariancją wartości próby.

2.1.2. Momenty absolutne zwykłe i centralne

Moment absolutne zwykłe i centralne rzędu k z próby wyznaczamy podobnie jak momenty zwykłe i centralne, z tą różnicą, że do potęgi k-tej podnosimy wartości bezwzględne z wartości próby (momenty absolutne zwykłe) lub wartości bezwzględne różnic wartości próby i średniej z próby (momenty absolutne centralne). Zatem moment absolutne zwykłe z próby n-elementowej wyznaczamy ze wzorów:

— dla szeregu szczegółowego o wartościach x_i:

$$a_k = \frac{1}{n}\sum_{i=1}^{n}|x_i|^k$$

— dla szeregu rozdzielczego punktowego o k wartościach x_i, które występują n_i-krotnie:

$$a_k = \frac{1}{n}\sum_{i=1}^{m}|x_i|^k n_i$$

— dla szeregu rozdzielczego przedzialowego o k klasach o środkach $\overline{x_i}$ i liczebności n_i:

$$a_k = \frac{1}{n}\sum_{i=1}^{m}|\overline{x_i}|^k n_i$$

Natomiast moment absolutny centralny rzędu k z próby n-elementowej wyznaczamy korzystając ze wzorów:

— dla szeregu szczegółowego o wartościach x_i:

$$b_k = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|^k$$

— dla szeregu rozdzielczego punktowego o k wartościach x_i, które występują n_i-krotnie:

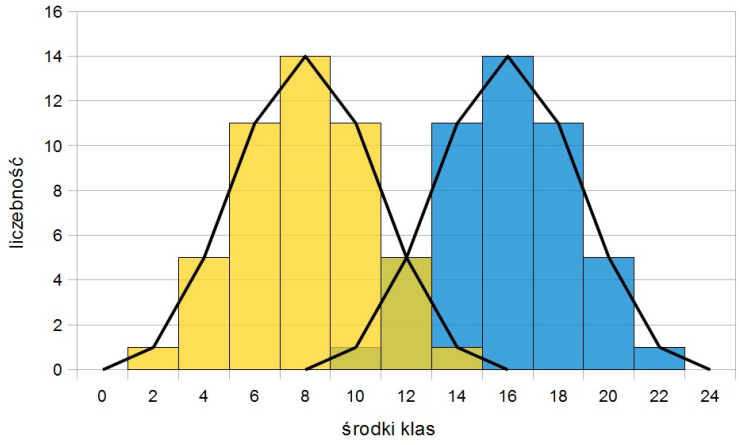
$$b_k = \frac{1}{n}\sum_{i=1}^{m}|x_i - \bar{x}|^k n_i$$

— dla szeregu rozdzielczego przedzialowego o k klasach o środkach $\overline{x_i}$ i liczebności n_i:

$$b_k = \frac{1}{n}\sum_{i=1}^{m}|\overline{x_i} - \bar{x}|^k n_i$$

2.2. Miary położenia

Miary położenia opisują poziom i rozmieszczenie wartości badanej cechy. Wskazują wartości cechy, które najlepiej reprezentują wszystkie wartości. Na rysunku 2.1 zostały przedstawione histogrmy dla dwóch prób losowych różniących się położeniem na prostej.



Rysunek 2.1. Histogramy szeregów statystycznych o różnym położeniu

Miary położenia można podzielić na:
— miary tendencji centralnej – opisują typowy poziom wartości badanej cechy,
— kwantyle – opisują rozmieszczenie wartości badanej cechy.

2.2.1. Miary tendencji centralnej

Do *miar tendencji centralnej* zaliczamy:
— średnie klasyczne: arytmetyczną, geometryczną, harmoniczną, kwadratową,
— medianę,
— modę.

Średnia arytmetyczna

Średnia arytmetyczna jest jedną z najczęściej używanych miar charakteryzujących rozkład. Jak nie trudno zauważyć, średnia arytmetyczna jest momentem zwykłym pierwszego rzędu.

Średnią arytmetyczną $\bar{x}$ dla szeregu szczegółowego o długosci n obliczamy według wzoru:

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

gdzie x_i – kolejne wartości w szeregu szczegółowym.

Przykład 2.1:

Zbadano czas snu 12 pacjentów chorujących na pewną chorobę i otrzymano następujące wyniki w minutach: 435 389 533 324 561 395 416 500 499 397 356 398
Średni czas snu obliczamy następująco:

$$\bar{x} = \frac{435+389+533+324+561+395+416+500+499+397+356+398}{12} =$$

$$\bar{x} = \frac{5203}{12} = 433\frac{7}{12}$$

Średnią arytmetyczną $\bar{x}$ dla szeregu rozdzielczego punktowego definiujemy wzorem:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \cdot x_i}{\sum_{i=1}^{k} n_i}$$

gdzie:
- x_i – i-ta wartość cechy,
- n_i – liczebność i-tej wartości cechy,
- k – liczba różnych wartości cechy.

Przykład 2.2:

Badano grupę 20 pacjentów podwzględem czasu reakcji na podanie leku. Otrzymane wyniki reprezentuje tabela:

czas reakcji (min)	liczba pacjentów
1	2
2	3
3	5
4	8
5	2

Średni czas reakcji organizmu pacjenta na podanie leku obliczamy następująco:

$$\bar{x} = \frac{2 \cdot 1 + 3 \cdot 2 + 5 \cdot 3 + 8 \cdot 4 + 2 \cdot 5}{2 + 3 + 5 + 8 + 2} = \frac{65}{20} = 3,25$$

Średnią arytmetyczną $\bar{x}$ dla szeregu rozdzielczego przedziałowego wyznaczamy za pomocą wzoru:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \cdot \overline{x_i}}{\sum_{i=1}^{k} n_i}$$

gdzie:
- $\overline{x_i}$ – środek i-tej klasy,
- n_i – liczebność i-tej klasy,
- k – liczba klas.

Przykład 2.3:

Poniższa tabela przedstawia liczbę pacjentow pewnego szpiltala, pogrupowanych według czasu działania leku przeciwbólowego. Pierwszym krokiem jest wyznaczenie środków poszczegolnych

czas działania leku	1 – 3	4 – 6	7 – 9	10 – 12	13 – 15	16 – 18
liczba pacjentów	2	5	10	11	8	4

kas:

$$\overline{x_1} = 2 \qquad \overline{x_2} = 5 \qquad \overline{x_3} = 8 \qquad \overline{x_4} = 11 \qquad \overline{x_5} = 14 \qquad \overline{x_6} = 17$$

Następnie średni czas działania leku przeciwbólowego na pacjentów szpitala wyznaczamy następująco:

$$\bar{x} = \frac{2 \cdot 2 + 5 \cdot 5 + 10 \cdot 8 + 11 \cdot 11 + 8 \cdot 14 + 4 \cdot 17}{2 + 5 + 10 + 11 + 8 + 4} = \frac{410}{40} = 10,25$$

Szczególną własnością średniej arytmetycznej jest, że suma odchyleń wartości od średniej arytmetycznej $(\bar{x} - x_i)$ jest równa zeru, co można łatwo udowodnić.

$$\sum_{i=1}^{n}(\bar{x} - x_i) = 0$$

Średnia geometryczna

Średnia geometryczna jest kolejną miarą tendencji centralnej.

Średnia geometryczna $\bar{x}_g$ dla szeregu szczegółowego o długości n dana jest wzorem:

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

gdzie x_i – kolejne wartości w szeregu szczegółowym.

Przykład 2.4:

Średnią geometryczną dla danych z przykładu 2.1 wyznaczamy w następujący sposób:

$$\bar{x}_g = \sqrt[12]{435 \cdot 389 \cdot 533 \cdot 324 \cdot 561 \cdot 395 \cdot 416 \cdot 500 \cdot 499 \cdot 397 \cdot 356 \cdot 398} \approx 428,02$$

Średnią geometryczną $\bar{x}_g$ dla szeregu rozdzielczego punktowego wyznaczamy ze wzoru:

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^{k} {x_i}^{n_i}}, \qquad n = \sum_{i=1}^{k} n_i$$

gdzie:

— x_i – i-ta wartość cechy,
— n_i – liczebność i-tej wartości cechy,
— k – liczba różnych wartości cechy.

Przykład 2.5:

Średnią geometryczną dla danych z przykładu 2.2 wyznaczamy w następujący sposób:

$$\bar{x}_g = \sqrt[20]{1^2 \cdot 2^3 \cdot 3^5 \cdot 4^8 \cdot 5^2} = \sqrt[20]{65813} \approx 1,74$$

Natomiast dla szeregu rozdzielczego przedziałowego średnią geometryczną $\bar{x}$ definiujemy wzorem:

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^{k} {\overline{x_i}}^{n_i}}, \qquad n = \sum_{i=1}^{k} n_i$$

gdzie:

— $\overline{x_i}$ – środek i-tej klasy,
— n_i – liczebność i-tej klasy,
— k – liczba klas.

Przykład 2.6:

Wyznaczmy średnią geometryczną dla danych z przykładu 2.3:

$$\bar{x}_g = \sqrt[40]{2^2 \cdot 5^5 \cdot 8^{10} \cdot 11^{11} \cdot 14^8 \cdot 17^4} \approx 9,27$$

Średnia harmoniczna

Następną miariarą tendencji centralnej jest średnia harmoniczna.
Dla szeregu szczegółowego o długosci n średnia harmoniczna $\bar{x}_h$ wyrażona jest wzorem:

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \qquad \sum_{i=1}^{n} \frac{1}{x_i} \neq 0$$

gdzie x_i – kolejne wartości w szeregu szczegółowym.

Przykład 2.7:

Wyznaczmy średnią harminiczną dla danych z przykładu 2.1:

$$\bar{x}_h = \frac{12}{\frac{1}{435} + \frac{1}{389} + \frac{1}{533} + \frac{1}{324} + \frac{1}{561} + \frac{1}{395} + \frac{1}{416} + \frac{1}{500} + \frac{1}{499} + \frac{1}{397} + \frac{1}{356} + \frac{1}{398}}$$

$$\bar{x}_h \approx 422,62$$

Średnią harmoniczną $\bar{x}_g$ dla szeregu rozdzielczego punktowego obliczamy korzyztając ze wzoru:

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}} \qquad \sum_{i=1}^{n} \frac{n_i}{x_i} \neq 0 \qquad \sum_{i=0}^{k} n_i = n$$

gdzie:
— x_i – i-ta wartość cechy,
— n_i – liczebność i-tej wartości cechy,
— k – liczba różnych wartości cechy.

Przykład 2.8:

Średnią harmoniczną dla danych z przykładu 2.2 obliczamy w następujący sposób:

$$\bar{x}_h = \frac{20}{\frac{2}{1} + \frac{3}{2} + \frac{5}{3} + \frac{8}{4} + \frac{2}{5}} \approx 2,64$$

Z kolei średnią harmoniczną $\bar{x}_g$ dla szeregu rozdzielczego przedzialowego wyznaczamy ze wzoru:

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{\overline{x_i}}} \qquad \sum_{i=1}^{n} \frac{n_i}{\overline{x_i}} \neq 0 \qquad \sum_{i=0}^{k} n_i = n$$

gdzie:
— $\overline{x_i}$ – środek i-tej klasy,
— n_i – liczebność i-tej klasy,
— k – liczba klas.

Przykład 2.9:

Średnia harmoniczna dla danych z przykładu 2.3:

$$\bar{x}_h = \frac{40}{\frac{2}{2} + \frac{5}{5} + \frac{10}{8} + \frac{11}{11} + \frac{8}{14} + \frac{4}{17}} \approx 7,91$$

Średnia kwadratowa

Średnia kwadratowa $\bar{x}_k$ jest kolejną miariarą tendencji centralnej.
Dla szeregu szczegółowego o długosci n wyznaczamy ją węglug wzoru:

$$\bar{x}_k = \sqrt{\frac{\sum_{i=1}^{n} {x_i}^2}{n}}$$

gdzie x_i – kolejne wartości w szeregu szczegółowym.

Przykład 2.10:

Wyznaczmy średnią kwadratową dla danych z przykładu 2.1:

$$\bar{x}_k = \sqrt{\frac{435^2 + 389^2 + 533^2 + 324^2 + 561^2 + 395^2 + 416^2 + 500^2 + 499^2 + 397^2 + 356^2 + 398^2}{12}}$$

$$\bar{x}_k = \sqrt{\frac{2315163}{12}} \approx 439.24$$

Natomiast dla szeregu rozdzielczego punktowego średnią kwadratową $\bar{x}_k$ definiujemy za pomocą wzoru:

$$\bar{x}_k = \sqrt{\frac{\sum_{i=1}^{k} n_i \cdot {x_i}^2}{n}} \qquad n = \sum_{i=1}^{k} n_i$$

gdzie:
— x_i – i-ta wartość cechy,
— n_i – liczebność i-tej wartości cechy,
— k – liczba różnych wartości cechy.

Przykład 2.11:

Średnią harmoniczną dla danych z przykładu 2.2 obliczamy w następujący sposób:

$$\bar{x}_k = \sqrt{\frac{2 \cdot 1^2 + 3 \cdot 2^2 + 5 \cdot 3^2 + 8 \cdot 4^2 + 2 \cdot 5^2}{20}} \approx 3,44$$

Średnią kwadratową $\bar{x}_g$ dla szeregu rozdzielczego przedzialowego wyznaczamy korzystając ze wzoru:

$$\bar{x}_k = \sqrt{\frac{\sum_{i=1}^{k} n_i \cdot \overline{x_i}^2}{n}} \qquad n = \sum_{i=1}^{k} n_i$$

gdzie:
— $\overline{x_i}$ – środek i-tej klasy,
— n_i – liczebność i-tej klasy,
— k – liczba klas.

Przykład 2.12:

Wyznaczymy średnią kwadratową dla danych z przykładu 2.3:

$$\bar{x}_k = \sqrt{\frac{2 \cdot 2^2 + 5 \cdot 5^2 + 10 \cdot 8^2 + 11 \cdot 11^2 + 8 \cdot 14^2 + 4 \cdot 17^2}{40}} = \sqrt{\frac{4828}{40}} \approx 10,99$$

Mediana

Mediana (M_e), nazywana rownież *wartością środkową*, jest miarą położenia. Dla szeregu szczegółowego jest określona wzorem:

$$M_e = \begin{cases} x_{\frac{n+1}{2}} & \text{jeżeli } n \text{ nieparzyste} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & \text{jeżeli } n \text{ parzyste} \end{cases}$$

gdzie x_i – kolejne wartości w uporządkowanym szeregu szczegółowym.

Przykład 2.13:

Obliczymy medianę dla szeregu szczegółowego z przykładu 2.1:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
324	356	389	395	397	398	416	435	499	500	533	561

$$M_e = \frac{1}{2}(x_{\frac{12}{2}} + x_{\frac{12}{2}+1}) = \frac{1}{2}(x_6 + x_7) = \frac{1}{2}(398 + 416) = 407$$

Aby wyznaczyć medianę szeregu rozdzielczego punktowego należy zamienić go na szereg szczegółowy w taki sposób, że:

$$x_1 = x_2 = \ldots = x_{n_1} = \hat{x}_1 \qquad x_{n_1+1} = x_{n_1+2} = \ldots = x_{n_1+n_2} = \hat{x}_2$$

$$\ldots \qquad x_{\sum_{i=1}^{k-1} n_i+1} = x_{\sum_{i=1}^{k-1} n_i+2} = \ldots = x_{\sum_{i=1}^{k} n_i} = \hat{x}_k$$

gdzie:
— x_i – wartości utworzonego szeregu szczegółowego,
— k – liczba różnych wartości w szeregu punktowym,
— $\hat{x}_i$ – wartości szeregu rozdzielczego punktowego,
— n_i – liczebność wartości $\hat{x}_i$.
Następnie, aby wyznaczyć medianę, stosujemy wzór dla szeregu szczegółowego.

Przykład 2.14:

Wyznaczymy medianę dla szeregu rozdzielczego punktowego z przykładu 2.2:

$$x_1 = x_2 = 1 \qquad x_3 = \ldots = x_5 = 2 \qquad x_6 = \ldots = x_{10} = 3$$

$$x_{11} = \ldots = x_{18} = 4 \qquad x_{19} = x_{20} = 5$$

$$M_e = \frac{1}{2}(x_{\frac{20}{2}} + x_{\frac{20}{2}+1}) = \frac{1}{2}(x_{10} + x_{11}) = \frac{1}{2}(3 + 4) = 3,5$$

Medianę dla szeregu rozdzielczego przedziałowego wyznaczamy korzystając ze wzoru:

$$M_e = \tilde{x}_m + \frac{b}{n_m}\left(\frac{n}{2} - \sum_{i=1}^{m-1} n_i\right) \qquad \text{dla} \qquad n = \sum_{i=1}^{k} n_i$$

gdzie:
— k – liczba klas w szeregu rozdzielczym przedziałowym,

— m – numer klasy zawierającej medianę, czyli spelniającej dwa warunki:

$$\sum_{i=1}^{m-1} n_i < \frac{n}{2} \qquad \text{oraz} \qquad \sum_{i=1}^{m} n_i \geqslant \frac{n}{2}$$

— $\tilde{x}_m$ – dolna granica m-tej klasy,
— n_i – liczebności i-tej klasy,
— b – długość przedziału klasowego,

Przykład 2.15:

Wyznaczymy medianę dla szeregu rozdzielczego przedziałowego z przykładu 2.3:

$$n = 2 + 5 + 10 + 11 + 8 + 4 = 40$$

$$2 + 5 + 10 < 20 \qquad \text{i} \qquad 2 + 5 + 10 + 11 \geqslant 20 \qquad \text{zatem} \qquad m = 4$$

$$b = 3$$

$$\tilde{x}_m = 7$$

$$\sum_{i=1}^{4-1} n_i = 2 + 5 + 10 = 17$$

Zatem:

$$M_e = 10 + \frac{3}{11} \cdot \left(\frac{40}{2} - 17 \right) = 10 + \frac{9}{11} \approx 10,82$$

Moda

Moda (M_o) jest wartością najczęściej występującą w szeregu statystycznym. Inne stosowane nazywy dla tej miary to:
— modalna,
— dominanata,
— wartość najczęstsza.

Wyznaczenie mody w szeregu szczegółowym sprowadza się do określenia, która wartość występuje najczęściej. W szeregu rozdzielczym punktowym jest to wartość, która ma największą liczebność. Natomiast w szeregu rozdzielczym przedziałowym modę wyznaczamy ze wzoru:

$$M_o = \tilde{x}_m + \frac{n_m - n_{m-1}}{(n_m - n_{m-1}) + (n_m - n_{m+1})} b$$

gdzie:
— m – numer klasy o największej liczebności,
— $\tilde{x}_m$ – dolna granica klasy o numerze m,
— n_m, n_{m-1}, n_{m+1} – liczebności klas o numerach m, $m-1$ oraz $m+1$,
— b – długość przedziału klasowego.

Przykład 2.16:

W szeregu szczegółowym z przykładu 2.1 wszystkie wartości występują tylko jeden raz, zatem w tym szeregu nie ma mody.

Przykład 2.17:

Dla szeregu rozdzielczego punktowego z przykładu 2.2 najczęściej występującą wartością jest 4, dlatego też:

$$M_o = 4$$

Przykład 2.18:

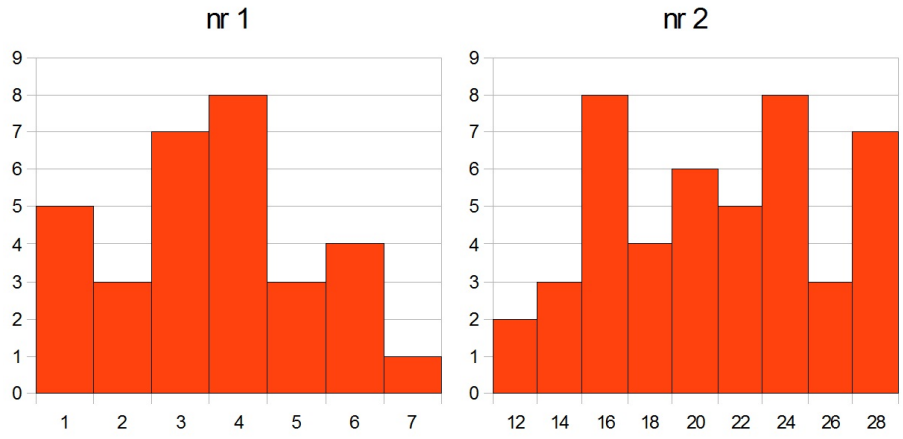
Wyznaczymy modę dla szeregu rozdzielczego przedziałowego z przykładu 2.3. Najwięcej wartości znajduje się w przedziale 10 - 12. Jest to czwarty przedział ($m = 4$), którego liczebność to: $n_4 = 11$. Liczebności sąsiednich przedziałów to: $n_3 = 10$ i $n_5 = 8$. Długość przedziału klasowego to: $b = 3$. Dolna granica przedziału to: $\tilde{x}_4 = 10$. Zatem:

$$M_o = 11 + \frac{11 - 10}{(11 - 10) + (11 - 8)} \cdot 3 = 11,75$$

W przypadku, gdy w szeregu statystycznym występują wartości o takiej samej liczebności i jednocześnie jest to największa liczebność, wówczas mamy większą liczbę modalnych.

Przykład 2.19:

Na rysunku 2.2 histogram nr 1 przedstawia szereg jednomodowy ($M_o = 4$), natomiast histogram nr 2 szereg dwumodowy ($M_{o_1} = 16, \quad M_{o_2} = 24$).



Rysunek 2.2. Histogramy szeregów statystycznych: jednomodowego (nr 1) i dwumodowego (nr2)

2.2.2. Kwantyle

Do miar opisujących rozmieszczenie wartości badanej cechy należą kwantyle. *Kwantylem rzędu* p ($\tilde{x}_p$), dla $0 \leqslant p \leqslant 1$, jest taka wartość badanej cechy w n-elementowym szeregu statystycznym, że $p \cdot n$ wartości jest mniejszych lub równych od $\tilde{x}_p$, a $(1 - p) \cdot n$ wartości jest większych lub równych od $\tilde{x}_p$. Szczególnymi, a zarazem często stosowanymi kwantylami są:

— *kwartyle*,
— *kwintyle*,
— *decyle*,

— *percentyle.*

Kwartyle dzielą szereg statystyczny na 4 równe części, kwintyle na 5, decyle na 10, natomiast percentyle dzielą szereg na 100 równych części. W praktyce najczęściej stosowanymi kwantylami są kwartyle.

Kwartyle

Kwartyle, nazywane także *wartościami ćwiartkowymi*, są trzema wartościami badanej cechy (Q_1, Q_2, Q_3) dzielącymi szereg statystyczny na 4 równe części. *Kwartyl pierwszy* Q_1 jest kwantylem rzędu $0,25$, czyli 25% elementów szeregu statystycznego jest mniejszych lub równych Q_1. *Kwartyl drugi* Q_2 jest kwantylem rzędu $0,5$. Kwartyl drugi dzieli szereg statystyczny na dwie równe części, zatem jest medianą:

$$Q_2 = M_e$$

Kwartyl trzeci Q_3 jest kwantylem rzędu $0,75$, czyli 75% elementów szeregu statystycznego jest mniejszych lub równych Q_3. Kwartyl pierwszy jest również nazywany *kwartylem dolnym*, a kwartyl trzeci – *kwartylem górnym.*

Wyznaczenie kwartyli dla uporządkowanego szeregu szczegółowego i rozdzielczego punktowego:

1. wyznaczamy położenie pierwszego kwartyla:

$$\tilde{n}_1 = \frac{n+3}{4}$$

gdzie n – długość szeregu statystycznego.

2. wyznaczamy część całkowitą wartości $\tilde{n}_1$:

$$\tilde{n}_{1c} = \lfloor \tilde{n}_1 \rfloor$$

3. wyznaczamy część ułamkową wartości $\tilde{n}_1$:

$$\tilde{n}_{1u} = \tilde{n}_1 - \lfloor \tilde{n}_1 \rfloor$$

4. wyznaczamy wartość pierwszego kwartyla:

$$Q_1 = x_{\tilde{n}_{1c}} + \tilde{n}_{1u} \cdot (x_{\tilde{n}_{1c}+1} - x_{\tilde{n}_{1c}})$$

gdzie $x_{\tilde{n}_{1c}},\ x_{\tilde{n}_{1c}+1}$ – wartości szeregu statystycznego.

5. kwartyl drugi jest medianą, zatem korzystamy ze wzoru na medianę:

$$Q_2 = M_e = \begin{cases} x_{\frac{n+1}{2}} & \text{jeżeli } n \text{ nieparzyste} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & \text{jeżeli } n \text{ parzyste} \end{cases}$$

gdzie: n – długość szeregu statystycznego, $x_{\frac{n+1}{2}}\ x_{\frac{n}{2}}\ x_{\frac{n}{2}+1}$ – wartości szeregu statystycznego.

6. wyznaczamy położenie trzeciego kwartyla:

$$\tilde{n}_3 = \frac{3n+1}{4}$$

gdzie n – długość szeregu statystycznego.

7. wyznaczamy część całkowitą wartości $\tilde{n}_3$:

$$\tilde{n}_{3c} = \lfloor \tilde{n}_3 \rfloor$$

8. wyznaczamy część ułamkową wartości $\tilde{n}_1$:

$$\tilde{n}_{3u} = \tilde{n}_3 - \lfloor \tilde{n}_3 \rfloor$$

9. wyznaczamy wartość trzeciego kwartyla:

$$Q_3 = x_{\tilde{n}_{3c}} + \tilde{n}_{3u} \cdot (x_{\tilde{n}_{3c}+1} - x_{\tilde{n}_{3c}})$$

gdzie $x_{\tilde{n}_{3c}},\ x_{\tilde{n}_{3c}+1}$ – wartości szeregu statystycznego.

Przykład 2.20:

Kwartyle dla szeregu szczegółowego z przykładu 2.1 wyznaczamy w następujący sposób:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
324	356	389	395	397	398	416	435	499	500	533	561

$$\tilde{n}_1 = \frac{12+3}{4} = 4,75 \qquad \tilde{n}_{1c} = 4 \qquad \tilde{n}_{1u} = 4,75 - 4 = 0,75$$

$$Q_1 = x_4 + 0,75 \cdot (x_5 - x_4) = 395 + 0,75 \cdot (397 - 395) = 396,5$$

$$Q_2 = M_e = 407$$

$$\tilde{n}_3 = \frac{3 \cdot 12 + 1}{4} = 9,25 \qquad \tilde{n}_{3c} = 9 \qquad \tilde{n}_{3u} = 9,25 - 9 = 0,25$$

$$Q_3 = x_9 + 0,25 \cdot (x_{10} - x_9) = 499 + 0,25 \cdot (500 - 499) = 499,25$$

Przykład 2.21:

Dla szeregu rozdzielczego punktowego z przykładu 2.2 kwartyle obliczamy w następujący sposób:

$$x_1 = x_2 = 1 \qquad x_3 = \ldots = x_5 = 2 \qquad x_6 = \ldots = x_{10} = 3$$

$$x_{11} = \ldots = x_{18} = 4 \qquad x_{19} = x_{20} = 5$$

$$\tilde{n}_1 = \frac{20+3}{4} = 5,75 \qquad \tilde{n}_{1c} = 5 \qquad \tilde{n}_{1u} = 5,75 - 5 = 0,75$$

$$Q_1 = x_5 + 0,75 \cdot (x_6 - x_5) = 2 + 0,75 \cdot (3 - 2) = 2,75$$

$$Q_2 = M_e = 3,5$$

$$\tilde{n}_3 = \frac{3 \cdot 20 + 1}{4} = 15,25 \qquad \tilde{n}_{3c} = 15 \qquad \tilde{n}_{3u} = 15,25 - 15 = 0,25$$

$$Q_3 = x_{15} + 0,25 \cdot (x_{16} - x_{15}) = 4 + 0,25 \cdot (4 - 4) = 4$$

Aby wyznaczyć kwartyle dla szeregu rozdzielczego przedzialowego wykorzystujemy wzory:

$$Q_1 = \tilde{x}_{m_1} + \frac{b}{n_{m_1}} \left(\frac{n}{4} - \sum_{i=1}^{m_1-1} n_i \right)$$

$$Q_2 = M_e = \tilde{x}_{m_2} + \frac{b}{n_{m_2}} \left(\frac{n}{2} - \sum_{i=1}^{m_2-1} n_i \right)$$

$$Q_3 = \tilde{x}_{m_3} + \frac{b}{n_{m_3}} \left(\frac{3n}{4} - \sum_{i=1}^{m_3-1} n_i \right)$$

gdzie:
— k – liczba klas w szeregu rozdzielczym przedziałowym,
— m_k – numer klasy zawierającej k-ty kwartyl, czyli spelniającej dwa warunki:

$$\sum_{i=1}^{m_k-1} n_i < \frac{kn}{4} \qquad \text{oraz} \qquad \sum_{i=1}^{m_k} n_i \geqslant \frac{kn}{4} \qquad \text{dla} \qquad k = 1, 2, 3$$

— $n = \sum_{i=1}^{k} n_i$ – liczba badanych jednostek zbiorowości
— $\tilde{x}_{m_k}$ – dolna granica m_k-tej klasy,
— n_i – liczebności i-tej klasy,
— b – długość przedziału klasowego.

Przykład 2.22:

Wyznaczenie kwartyli dla szeregu rozdzielczego przedziałowego z przykładu 2.3:

$$2 + 5 < 10 \qquad \text{i} \qquad 2 + 5 + 10 \geqslant 10 \qquad \text{zatem} \qquad m_1 = 3$$

$$Q_1 = \tilde{x}_{m_1} + \frac{b}{n_{m_1}} \left(\frac{n}{4} - \sum_{i=1}^{m_1-1} n_i \right) = 7 + \frac{3}{10} \cdot (10 - 7) = 7,90$$

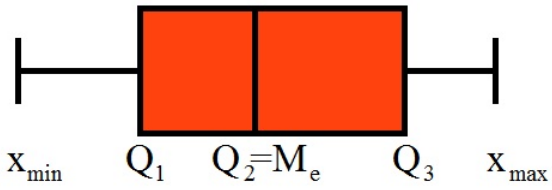
$$2 + 5 + 10 < 20 \qquad \text{i} \qquad 2 + 5 + 10 + 11 \geqslant 10 \qquad \text{zatem} \qquad m_2 = 4$$

$$Q_2 = M_e = \tilde{x}_{m_2} + \frac{b}{n_{m_2}} \left(\frac{n}{2} - \sum_{i=1}^{m_2-1} n_i \right) = 10 + \frac{3}{11} \cdot (20 - 17) \approx 10,82$$

$$2 + 5 + 10 + 11 < 30 \qquad \text{i} \qquad 2 + 5 + 10 + 11 + 8 \geqslant 30 \qquad \text{zatem} \qquad m_3 = 5$$

$$Q_3 = \tilde{x}_{m_3} + \frac{b}{n_{m_3}} \left(\frac{3n}{4} - \sum_{i=1}^{m_3-1} n_i \right) = 13 + \frac{3}{8} \cdot (30 - 28) = 13\frac{3}{4} = 13,75$$

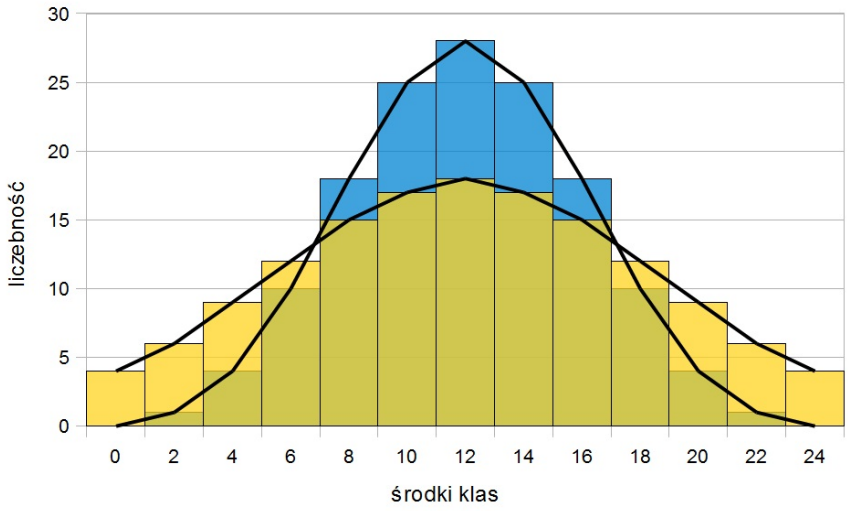
Gdy mamy wyznaczone kwartyle oraz warości minimalną i maksymalną w szeregu statystycznym, możemy przedstawić szereg statystyczny w formie diagramu pudełkowego. Na poniższym rysunku został ptzedstawiony wzór diagramu pudelkowego.



Rysunek 2.3. Diagram pudełkowy

2.3. Miary rozproszenia

Miary rozproszenia, nazywane rownież *miarami zmienności*, określają granice obszaru, w którym znajdują się wartości badanej cechy, a także opisują zróżnicowanie wartości cechy wokół wartości centralnych. Na rysunku 2.4 zostały przedstawione histogrmy dla dwóch prób losowych o różnym rozproszeniu.



Rysunek 2.4. Histogramy szeregów statystycznych o różnej zmienności

Podstawowymi miarami rozproszenia są:
— rozstęp,
— odchylenie standardowe,
— wariancja,
— odchylenie przeciętne od średniej,
— odchylenie przeciętne od mediany,
— rozstęp ćwiartkowy,
— odchylenie ćwiartkowe,
— współczynnik zmienności względem średniej,
— współczynnik zmienności względem mediany.

2.3.1. Rozstęp

Rozstęp (R) jest to najprostsza miara charakteryzująca empiryczny obszar zmienności badanej cechy. Jest on zdefiniowany wzorem:

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

gdzie:
— x_{min} – najmniejsza wartość cechy w szeregu statystycznym,
— x_{max} – największa wartość cechy w szeregu statystycznym.

Przykład 2.23:

Dla szeregu szczegółowego z przykładu 2.1 rozstęp jest równy:

$$R = 561 - 324 = 237$$

Natomiast dla szeregu rozdzielczego punktowego z przykładu 2.2 rozstęp jest równy:

$$R = 5 - 1 = 4$$

Z kolei dla szeregu rozdzielczego przedzialowego z przykładu 2.3 rozstęp jest równy:

$$R = 18 - 1 = 17$$

2.3.2. Odchylenie standardowe i wariancja

Najczęściej stosowanymi w praktyce miarami rozproszenia są *odchylenie standardowe* i *wariancja*. Inną często stosowaną nazwą wariancji jest *dyspersja*.

Odchylenie standardowe dla szeregów szczegółowych definiowane jest wzorem:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

gdzie:
- n – liczebność próby,
- $\bar{x}$ – wartość średnia z próby,
- x_i – wartości próby.

Z kolei dla szeregów rozdzielczych punktowych odchylenie standardowe wyznacza się ze wzoru:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i \cdot (x_i - \bar{x})^2}$$

gdzie:
- k – liczba różnych wartości cechy,
- n_i – liczebność i-tej wartości cechy,
- $n = \sum_{i=1}^{k} n_i$ – liczebność próby,
- $\bar{x}$ – wartość średnia z próby,
- x_i – wartości próby.

Natomiast odchylenie standardowe dla szeregów rozdzielczych przedziałowych wyrażone jest wzorem:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i \cdot (\overline{x_i} - \bar{x})^2}$$

- k – liczba klas wartości cechy,
- n_i – liczebność i-tej klasy,
- $n = \sum_{i=1}^{k} n_i$ – liczebność próby,
- $\bar{x}$ – wartość średnia z próby,
- $\overline{x_i}$ – wartość środkowa i-tej klasy.

Przykład 2.24:

Odchylenie standardowe dla szeregu szczegółowego z przykładu 2.1 jest równe:

$$s \approx 70,25$$

Dla szeregu rozdzielczego punktowego z przykładu 2.2 odchylenie standardowe jest równe:

$$s \approx 1,13$$

Z kolei odchylenie standardowe dla szeregu rozdzielczego przedziałowego z przykładu 2.3 równa się:

$$s \approx 3,95$$

Wariancja jest kwadratem odchylenia standardowego. Zatem dla szeregów szczegółowych definiowana jest wzorem:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

gdzie tak, jak poprzednio:
— n – liczebność próby,
— $\bar{x}$ – wartość średnia z próby,
— x_i – wartości próby.
Natomiast dla szeregów rozdzielczych punktowych wariancję opisuje wzór:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot (x_i - \bar{x})^2$$

w którym wielkości: k, n_i, $n = \sum_{i=1}^{k} n_i$, $\bar{x}$ oraz x_i są zdefiniowane tak jak w przypadku odchylenia standardowego. Z kolei wariancję dla szeregów rozdzielczych przedziałowych definuje wzór:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot (\overline{x_i} - \bar{x})^2$$

Oznaczenia we wzorze są identyczne jak dla odchylenia standardowego.

Przykład 2.25:

Wariancja dla szeregu szczegółowego z przykładu 2.1 równa się:

$$s^2 \approx 4935,74$$

Dla szeregu rozdzielczego punktowego z przykładu 2.2 wariancja wynosi:

$$s^2 \approx 1,29$$

Z kolei wariancja dla szeregu rozdzielczego przedziałowego z przykładu 2.3 równa się:

$$s^2 \approx 15,64$$

Istotną cechą wariancji i odchylena standardowego jest wzrost ich wartości wraz ze zwiększeniem zróżnicowania w próbie.

2.3.3. Odchylenie przeciętne od średniej i odchylenie przeciętne od mediany

Odchylenie przeciętne od średniej ($d_{\bar{x}}$) definujemy jako średnią arytmetyczną wartości bezwzględnych różnic $(x_i - \bar{x})$, gdzie $\bar{x}$ – średnia z próby. Natomiast odchylenie przeciętne od mediany (d_{M_e})definujemy jako średnią arytmetyczną wartości bezwzględnych różnic $(x_i - M_e)$, gdzie M_e – mediana próby. Szczegółowe wzory dla poszczególnych typów szeregu zawiera poniższa tabela:

typ szeregu statystycznego	odchylenie przeciętne od średniej	odchylenie przeciętne od mediany
szczegółowy	$d_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} \lvert x_i - \bar{x}\rvert$	$d_{M_e} = \frac{1}{n}\sum_{i=1}^{n} \lvert x_i - M_e\rvert$
rozdzielczy punktowy	$d_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot \lvert x_i - \bar{x}\rvert$	$d_{M_e} = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot \lvert x_i - M_e\rvert$
rozdzielczy przedziałowy	$d_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} \lvert \overline{x_i} - \bar{x}\rvert$	$d_{M_e} = \frac{1}{n}\sum_{i=1}^{n} \lvert \overline{x_i} - M_e\rvert$

W tabeli zastosowano oznaczenia takie same jak przy poprzednich definicjach.

Przykład 2.26:

Wyznaczenie odchyleń przeciętnych dla:

1. Dla szeregu szczegółowego z przykładu 2.1:

— odchylenie przeciętne od średniej:

$$n = 12 \qquad \bar{x} \approx 433,58$$

$$d_{\bar{x}} \approx 60,01$$

— odchylenie przeciętne od mediany:

$$n = 12 \qquad M_e = 407$$

$$d_{M_e} \approx 57,08$$

2. Dla szeregu rozdzielczego punktowego z przykładu 2.2:

— odchylenie przeciętne od średniej:

$$n = 20 \qquad \bar{x} = 3,25$$

$$d_{\bar{x}} \approx 0,95$$

— odchylenie przeciętne od mediany:

$$n = 20 \qquad M_e = 3,5$$

$$d_{M_e} \approx 0,95$$

3. Dla szeregu rozdzielczego przedziałowego z przykładu 2.3:

— odchylenie przeciętne od średniej:

$$n = 40 \qquad \bar{x} = 10,25$$

$$d_{\bar{x}} \approx 3,26$$

— odchylenie przeciętne od mediany:

$$n = 40 \qquad M_e = 10,9$$

$$d_{M_e} \approx 3,17$$

2.3.4. Rozstęp ćwiartkowy i odchylenie ćwiartkowe

Rozstęp ćwiartkowy (*rozstęp międzykwartylowy*) jest miarą zmienności cechy charakteryzującą jej zmienność względem miar pozycyjnych (mediana, moda). Rozstęp kwartylowy zdefiniowany jest wzorem:

$$R_Q = Q_3 - Q_1$$

gdzie:

— Q_1 – kwartyl dolny,
— Q_3 – kwartyl górny.

Odchylenie ćwiartkowe jest miarą rozproszenia proporcjonalną do rozstępu między kwartylowego. Jest ono zdefiniowane naatępującym wzorem:

$$Q = \frac{R_Q}{2}$$

.

Przykład 2.27:

W poniższej tabeli zostały zestawione kwartyle dolne i górne oraz wartości rozstępu międzykwartylowego oraz odchylenia ćwiartkowego dla danych z przykladów 2.1, 2.1 oraz 2.3.

nr przykładu oraz typ szeregu statystycznego	kwartyl dolny (Q_1)	kwartyl górny (Q_3)	rozstęp międzykwartylowy	odchylenie ćwiartkowe
2.1 szczegółowy	396,5	499,25	102,75	51,375
2.2 rozdzielczy punktowy	2,75	4	1,25	0,625
2.3 rozdzielczy przedziałowy	7,9	13,75	5,85	2,925

2.3.5. Współczynnik zmienności względem średniej i współczynnik zmienności względem mediany

Współczynnik zmienności względem średniej znajduje zastosowanie w porównywaniu względnej zmienności różnych cech względem ich wartości średnich. Jest on zdefiniowany jako:

$$V_{\bar{x}} = \frac{s}{\bar{x}} \cdot 100\%$$

gdzie:

— s – odchylennie standardowe;
— $\bar{x}$ – średnia arytmetyczna z próby.

Współczynnik zmienności względem mediany jest stosowany w porównywaniu względnej zmienności różnych cech względem ich median. Współczynnik ten jest zdefiniowany wzorem:

$$V_{M_e} = \frac{Q}{M_e} \cdot 100\%$$

gdzie:

— Q – odchylennie ćwiartkowe;
— M_e – mediana.

Przykład 2.28:

W poniższej tabeli zostały zestawione wartości średnie ($\bar{x}$), odchylenia standardowe (s), współczynniki zmienności względem średniej ($V_{\bar{x}}$), odchylenia ćwiartkowe (Q), mediany (M_e) oraz współczynniki zmienności względem mediany (V_{M_e}) dla danych z przykładów 2.1, 2.1 oraz 2.3.

nr przykładu oraz typ szeregu statystycznego	$\bar{x}$	s	$V_{\bar{x}}$	Q	M_e	V_{M_e}
2.1 szczegółowy	433,58	70,25	16,2%	407	51,38	12,6%
2.2 rozdzielczy punktowy	3,25	1,13	34,9%	3,5	0,63	17,9%
2.3 rozdzielczy przedziałowy	10,25	3,95	38,6%	10,9	2,37	21,8%

Ćwiczenia nr 4

Zadanie 2.1. [2]

Pojazd przebył drogę złożoną z trzech odcinków, każdy o długości s. Pierwszy odcinek ze stałą prędkością $v_1 = 50\frac{\text{km}}{\text{h}}$, drugi ze stałą prędkością $v_2 = 60\frac{\text{km}}{\text{h}}$ i trzeci ze stałą prędkością $v_3 = 70\frac{\text{km}}{\text{h}}$. Z jaką stałą średnią prędkością powinien poruszać się pojazd, aby całą drogę $3s$ przebyć w tym samym czasie?

Zadanie 2.2.

Liczba pacjentów przyjmowanych przez pewien szpital w ciągu roku kształtowała się następująco:

w roku 2005 – 182000 pacjentów
w roku 2006 – 202410 pacjentów
w roku 2007 – 223130 pacjentów
w roku 2008 – 243980 pacjentów
w roku 2009 – 264890 pacjentów
w roku 2010 – 285860 pacjentów
w roku 2011 – 306980 pacjentów
w roku 2012 – 328440 pacjentów

Wyznaczyć wskaźniki wzrostu liczby przyjmowanych pacjentów w ciągu roku v_i, $i = 1, 2, \ldots, 7$ dla lat 2005-2012, a następnie wyznaczyć średni wskaźnik wzrostu liczby przyjmowanych pacjentów v.

Uwaga: Wskaźnik wzrostu liczby przyjmowanych pacjentów jest równy stosunkowi liczby pacjentów przyjmowanych w danym roku do liczby pacjentów przyjmowanych w roku wcześniejszym.

Zadanie 2.3.

W pewnym doświadczeniu medycznym grupę pacjentów scharakteryzowano pod względem masy ciała. Uzyskano następujące wyniki (w kg):

53	46	77	74	91	57	78	74	59	73
57	52	42	68	68	62	65	78	50	76
76	56	68	92	60	73	63	68	62	56
48	69	73	60	60	50	72	66	69	59
56	52	63	77	61	61	73	66	73	52

Wyznacz:
- średnią masę ciała pacjenta,
- odchylenie standardowe od średniej masy ciała,
- modalną,
- medianę,
- kwartyle,
- rozstęp,
- odchylenie przeciętne od średniej,
- odchylenie przeciętne od mediany,
- rozstęp ćwiartkowy i odchylenie ćwiartkowe,
- współczynnik zmienności względem średniej,
- współczynnik zmienności względem mediany.

Zadanie 2.4.

Grupę 100 mężczyzn w wieku 35 lat zbadano ze względu na liczbę brakujących zębów w uzębieniu i otrzymano następujące wyniki:

liczba brakujących zębów	liczba mężczyzn
0	10
1	18
2	26
3	29
4	11
5	3
6	0
7	1
8	1

Wyznacz:
- średnią liczbę brakujących zębów,
- odchylenie standardowe od średniej liczby brakujących zębów,
- modalną,
- medianę,
- kwartyle,
- rozstęp,
- odchylenie przeciętne od średniej,
- odchylenie przeciętne od mediany,
- rozstęp ćwiartkowy i odchylenie ćwiartkowe,
- współczynnik zmienności względem średniej,
- współczynnik zmienności względem mediany.

Zadanie 2.5.

W pewnym szpitalu przebadano 100 pacjentów pod względem czasu działania leku przeciwbólowego. Poniższa tabela zawiera zestawienie wyników. Wyznacz:

czas działania	1 – 2	3 – 4	5 – 6	7 – 8	9 – 10	11 – 12	13 – 14	15 –16
liczba pacjentów	2	5	11	23	23	21	10	5

— średni czas działania leku przeciwbólowego,
— odchylenie standardowe od średniego czasu działania leku przeciwbólowego,
— modalną,
— medianę,
— kwartyle,
— rozstęp,
— odchylenie przeciętne od średniej,
— odchylenie przeciętne od mediany,
— rozstęp ćwiartkowy i odchylenie ćwiartkowe,
— współczynnik zmienności względem średniej,
— współczynnik zmienności względem mediany.

Zadanie 2.6. [7]

Wśród 500 pracowników pewnego zakładu palących tytoń przeprowadzono ankietę pod kątem wypalanych dziennie papierosów. Uzyskano następujące dane:

Liczba wypalanych papierosów	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35	35 –40
liczba pacjentów	33	57	162	185	43	11	7	2

Ćwiczenia nr 4 – Matlab

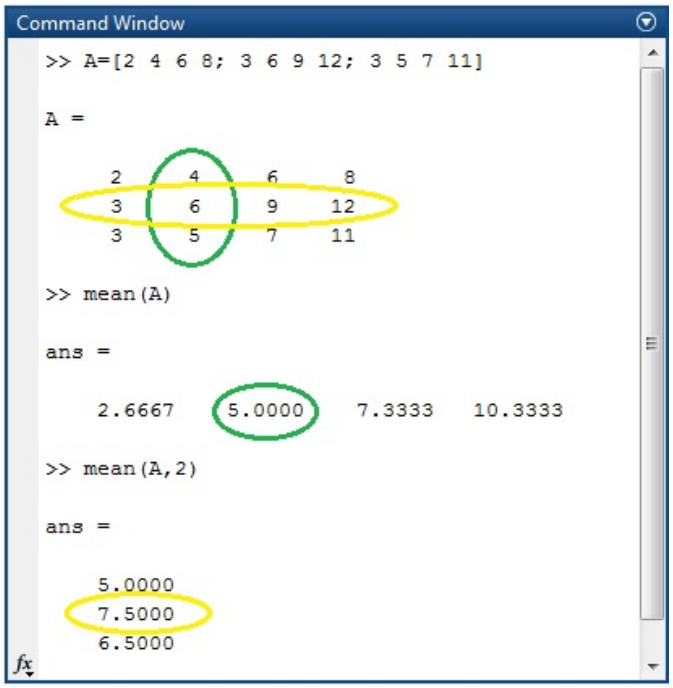
Funkcje

mean

`mean(A)` – jeśli A jest wektorem, to funkcja zwraca średnią arytmetyczną wartości wektora A, natomiast jeśli A jest macierzą, to funkcja zwraca wektor zawierający średnie arytmetyczne wartości w kolumnach macierzy A.
`mean(A,2)` – dla A będącego macierzą 2 wymiarową funkcja zwraca wektor zawierający średnie arytmetyczne wartości we wierszach macierzy A.

rms

`rms(A)` – jeśli A jest wektorem, to funkcja zwraca średnią kwadratową wartości wektora A, natomiast jeśli A jest macierzą, to funkcja zwraca wektor zawierający średnie kwadratowe wartości w kolumnach macierzy A.
`rms(A,2)` – dla A będącego macierzą 2 wymiarową funkcja zwraca wektor zawierający średnie kwadratowe wartości we wierszach macierzy A.



Rysunek 2.5. Zastosowanie funkcji `mean`

mode

`mode(A)` – jeśli A jest wektorem, to funkcja zwraca modę wartości wektora A, natomiast jeśli A jest macierzą, to funkcja zwraca wektor zawierający mody wartości w kolumnach macierzy A.
`mode(A,2)` – dla A będącego macierzą 2 wymiarową funkcja zwraca wektor zawierający mody wartości we wierszach macierzy A.
Uwaga: Jeżeli istnieje więcej wartości mających tę samą częstość występowania to funkcja `mode` zwróci najmniejszą z nich.

median

`median(A)` – jeśli A jest wektorem, to funkcja zwraca medianę wartości wektora A, natomiast jeśli A jest macierzą, to funkcja zwraca wektor zawierający mediany wartości w kolumnach macierzy A.
`median(A,2)` – dla A będącego macierzą 2 wymiarową funkcja zwraca wektor zawierający mediany wartości we wierszach macierzy A.

std

`std(A)` – jeśli A jest wektorem, to funkcja zwraca odchylenie standardowe wartości wektora A, natomiast jeśli A jest macierzą, to funkcja zwraca wektor zawierający odchylenia standardowe wartości w kolumnach macierzy A, gdzie odchylenie standardowe liczone jest według wzoru: $s = \frac{1}{n-1}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.
`std(A,1)` – jeśli A jest wektorem, to funkcja zwraca odchylenie standardowe wartości wektora A, natomiast jeśli A jest macierzą, to funkcja zwraca wektor zawierający odchylenia standardowe wartości w kolumnach macierzy A, gdzie odchylenie standardowe liczone jest według wzoru: $s = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.
`std(A,0,2)` – dla A będącego macierzą 2 wymiarową funkcja zwraca wektor zawie-

rający odchylenia standardowe wartości we wierszach macierzy A, gdzie odchylenie standardowe liczone jest według wzoru: $s = \frac{1}{n-1}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.
`std(A,1,2)` – dla A będącego macierzą 2 wymiarową funkcja zwraca wektor zawierający odchylenia standardowe wartości we wierszach macierzy A, gdzie odchylenie standardowe liczone jest według wzoru: $s = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

var

`var(A)` – jeśli A jest wektorem, to funkcja zwraca wariancję wartości wektora A, natomiast jeśli A jest macierzą, to funkcja zwraca wektor zawierający wariancje wartości w kolumnach macierzy A, gdzie odchylenie standardowe liczone jest według wzoru: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
`var(A,1)` – jeśli A jest wektorem, to funkcja zwraca wariancję wartości wektora A, natomiast jeśli A jest macierzą, to funkcja zwraca wektor zawierający wariancje wartości w kolumnach macierzy A, gdzie odchylenie standardowe liczone jest według wzoru: $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
`var(A,0,2)` – dla A będącego macierzą 2 wymiarową funkcja zwraca wektor zawierający odchylenia standardowe wartości we wierszach macierzy A, gdzie odchylenie standardowe liczone jest według wzoru: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
`std(A,1,2)` – dla A będącego macierzą 2 wymiarową funkcja zwraca wektor zawierający odchylenia standardowe wartości we wierszach macierzy A, gdzie odchylenie standardowe liczone jest według wzoru: $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

abs

`abs(A)` – funkcja zwraca macierz o takim samym rozmiarze jak macierz A zawierającą wartości bezwzględne elementów macierzy A.

floor

`floor(A)` – funkcja zwraca macierz o takim samym rozmiarze jak macierz A zawierającą najbliższe całkowite wartości mniejsze lub równe wartościom macierzy A.

min

`min(A)` – jeśli A jest wektorem, to funkcja zwraca najmniejszą wartości wektora A, natomiast jeśli A jest macierzą, to funkcja zwraca wektor zawierający najmniejsze wartości w kolumnach macierzy A.
`min(A,[],1)` – ma identyczne działanie jak `min(A)`.
`min(A,[],2)` – dla A będącego macierzą 2 wymiarową funkcja zwraca wektor zawierający najmniejsze wartości we wierszach macierzy A.
`[w,i]=min(A)` – funkcja `min` zwraca dwie wartości: `w` – wartość minimalną oraz `i` – indeks określający położenie wartości minimalnej w wektorze A.

max

`max` – jeśli A jest wektorem, to funkcja zwraca największą wartości wektora A, natomiast jeśli A jest macierzą, to funkcja zwraca wektor zawierający największe wartości w kolumnach macierzy A.
`max(A,[],1)` – ma identyczne działanie jak `max(A)`.

`max(A,[],2)` – dla A będącego macierzą 2 wymiarową funkcja zwraca wektor zawierający największe wartości we wierszach macierzy A.
`[w,i]=min(A)` – funkcja `max` zwraca dwie wartości: `w` – wartość maksymalną oraz `i` – indeks określający położenie wartości maksumalnej w wektorze A.

Przykładowe zadanie

Zadanie II.

Badano grupę 1000 osób pod względem czasu reakcji na podanie środka znieczulającego. Wyniki zostaly zebrane w poniższej tabeli:

czas reakcji (min)	liczba osób
3	13
4	70
5	168
6	288
7	253
8	149
9	46
10	13

Wyznacz:
- średni czas reakcji organiazmu pacjęta na znieczulenie,
- odchylenie standartowe czasu reakcji organiazmu na znieczulenie,
- modalną,
- medianę,
- kwartyle,
- rozstęp,
- odchylenie przeciętne od średniej,
- odchylenie przeciętne od mediany,
- rozstęp ćwiartkowy i odchylenie ćwiartkowe,
- współczynnik zmienności wzglendem średniej,
- współczynnik zmienności wzglendem mediany.

krok 1: Utworzenie macierzy zawierającej szereg rozdzielczy punktowy
Aby wyznaczyć miary rozkładu należy utworzyć macierz zawierającą wszysykie dane.

```
A=[3 13; 4 70; 5 168; 6 288; 7 253; 8 149; 9 46; 10 13];
```

krok 2: Wyznaczenie średniego czasu reakcji
Aby wyznaczyć średnią arytmetyczną dla szeregu rozdzialczego punktowego stosujemy wzór:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$$

gdzie:
- n – liczebność próby ($n = 1000$),
- k – liczba różnych wartości cechy ($k = 8$),
- x_i – i-ta wartość cechy,
- n_i – krotność i-tej wartości cechy.

Najpierw tworzymy macierz zawierającą iloczyny $n_i x_i$:

```
B=A(:,1).*A(:,2);
```

Następnie sumujemy macierz B i sumę dzielimy przez n, w wyniku czego otrzymujemy średnią arytmetyczną.

```
srednia=sum(B)/1000
```

krok 3: Wyznaczenie odchylenia standardowego od średniej

Aby wyznaczyć odchylenie standardowe od średniaj dla szeregu rozdzialczego punktowego stosujemy wzór:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2}$$

W pierwszej kolejności tworzymy macierz różnic $x_i - \bar{x}$.

```
C=A(:,1)-srednia;
```

Następnie tworzymy macierz zawierającą iloczyny $n_i(x_i - \bar{x})^2$:

```
D=A(:,2).*(C.^2);
```

Następnie sumujemy macierz D i sumę dzielimy przez n, w wyniku czego otrzymujemy kwadrat odchylenia standardowego, czyli wariancję.

```
var=sum(D)/1000
```

Zatem odchylenie standardowe obliczamy jako pierwiastek z wariancji.

```
s=sqrt(var)
```

krok 4: Wyznaczenie mody

Po pierwsze należy ustalić, które cechy mają największą liczebność, tzn. znaleźć największą wartość w drugiej kolumnie macierzy A oraz określić jej indeks.

```
[w,i]=max(A(:,2));
```

Wowczas wartością modalną jest wartość o indeksie `i` z pierwszej kolumny macierzy A.

```
moda=A(i,1)
```

krok 5: Wyznaczenie mediany

Aby wyznaczyć medianę warto najpierw stworzyć szereg skumulowany.

```
for i=1:8
E(i,1)=A(i,1);
E(i,2)=sum(A(1:i,2));
end
E
```

Skono $n = 1000$ medianę wyznaczamy ze wzoru: $M_e = \frac{x_{500}+x_{501}}{2}$. Z szmacierzy E łatwo można odczytać, że: $x_{500} = x_{501} = 6$. Zatem:

```
mediana=(6+6)/2
```

krok 6: Wyznaczenie kwartyli

Najpierw wyznaczamy położenie pierwszego kwartylu, tzn. obliczmy wartości wyrażeń: $\tilde{n}_1 = \frac{n+3}{4}$, $\tilde{n}_{1c} = [n_1]$ oraz $\tilde{n}_{1u} = n_1 - n_{1c}$.

```
n1=(1000+3)/4;
n1c=floor(n1);
n1u=n1-n1c;
```

Wówczas pierwszy kwartyl wyznaczamy korzystając ze wzoru:

$$Q_1 = x_{\tilde{n}_{1c}} + \tilde{n}_{1u} \cdot (x_{\tilde{n}_{1c}+1} - x_{\tilde{n}_{1c}})$$

Mamy, że x_i są wartościami z pierwszej kolumny macierzy E odpowiadającymi skumulowanej liczebności `n1c` i `n1c1`.

```
Xn1c=5;
Xn1c1=5;
```

Co daje:

```
Q1=Xn1c+n1u*(Xn1c1-Xn1c)
```

Drugi kwartyl to mediana, zatem;

```
Q2=mediana
```

Trzeci kwartyl wyznaczamy podobnie jak pierwszy, tzn. obliczmy wartości wyrażeń: $\tilde{n}_3 = \frac{3n+1}{4}$, $\tilde{n}_{3c} = [n_3]$ oraz $\tilde{n}_{3u} = n_3 - n_{3c}$.

```
n3=(3*1000+1)/4;
n3c=floor(n3);
n3u=n3-n3c;
```

Wówczas trzeci kwartyl wyznaczamy korzystając ze wzoru:

$$Q_3 = x_{\tilde{n}_{3c}} + \tilde{n}_{3u} \cdot (x_{\tilde{n}_{3c}+1} - x_{\tilde{n}_{3c}})$$

gdzie x_i są wartościami z pierwszej kolumny macierzy E odpowiadającymi skumulowanej liczebności `n3c` i `n3c1`.

```
Xn3c=7;
Xn3c1=7;
```

Co daje:

```
Q3=Xn3c+n3u*(Xn3c1-Xn3c)
```

krok 7: Wyznaczenie rozstępu

Wyznaczenie rozstępu jest stosunkowo proste, ponieważ od największej wartości w pierwszej kolumnie macierz A odejmujemy najmniejszą wartość w tej kolumnie:

```
rozstep=max(A(:,1))-min(A(:,1))
```

krok 8: Wyznaczenie odchylenia przecientnego od śreniniej

W pierwszej kolejności tworzymy wektor zawierający wielkości $n_i \cdot |x_i - \bar{x}|$. Następnie sumujemy wartości wektora i sumę dzielimy przez n.

```
for i=1:8
F(i,1)=A(i,2)*abs(A(i,1)-srednia);
end
d_srednia=sum(F)/1000
```

krok 9: Wyznaczenie odchylenia przecientnego od mediany

Po pierwsze tworzymy wektor zawierający wielkości $n_i \cdot |x_i - M_e|$. Następnie sumujemy wartości wektora i sumę dzielimy przez n.

```
for i=1:8
G(i,1)=A(i,2)*abs(A(i,1)-mediana);
end
d_mediana=sum(G)/1000
```

krok 10: Wyznaczenie rozstępu ćwiartkowego i odchylenia ćwiartkowego
Rozstęp ćwiartkowy to różnica trzeciego i pierwszego kwartylu, zatem:

```
rozstep_cwiartkowy=Q3-Q1
```

Odchylenie ćwiartkowe to połowa rozstąpu ćwiartkowego, więc:

```
od_ćwiartkowe=rozstep_cwiartkowy/2
```

krok 11: Wyznaczenie współczynnika zmienności względem średniej
Współczynnik zmienności wzgledem średniej jest stosunkiem odchylenia standardowego i średniej arytmetycznej wyrażonym w procentach, zatem;

```
wsp_zmiennosci_srednia=100*s/srednia
```

krok 12: Wyznaczenie współczynnika zmienności względem mediany
Współczynnik zmienności względem mediany jest stosunkiem odchylenia ćwiartkowego i mediany, więc

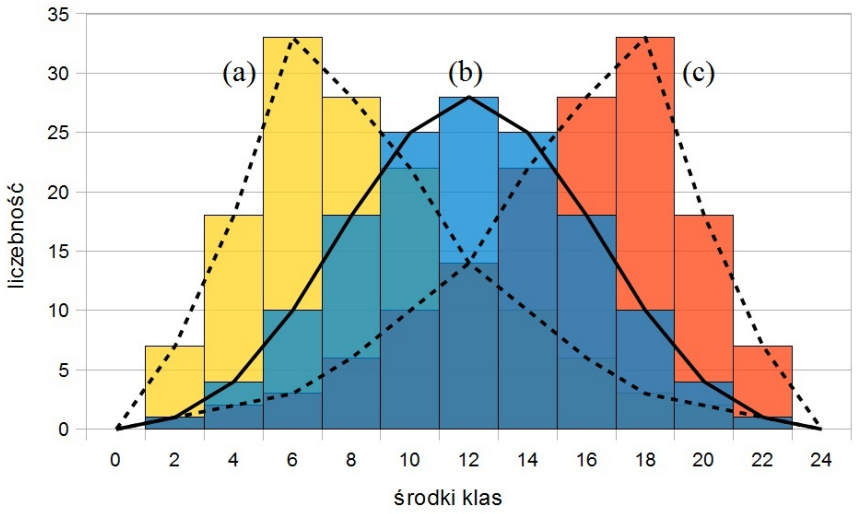
```
wsp_zmiennosci_mediana=od_ćwiartkowe=100*rozstep/mediana
```

Zestawienie wyników:
— średni czas reakcji organizmu pacjenta na znieczulenie – 6,3940 min
— odchylenie standardowe czasu reakcji organizmu pacjenta na znieczulenie – 1,3714 min,
— modalną – 6 min,
— medianę – 6 min,
— kwartyle – $Q_1 = 5$ min, $Q_2 = 6$ min, $Q_3 = 7$ min,
— rozstęp – 7 min,
— odchylenie przeciętne od średniej – 1,1187 min,
— odchylenie przeciętne od mediany – 1,0880 min,
— rozstęp ćwiartkowy – 2 min,
— odchylenie ćwiartkowe – 1 min,
— współczynnik zmienności względem średniej – 21.4484%,
— współczynnik zmienności względem mediany – 16,6667%.

2.4. Miary asymetrii rozkładu

Miary asymetrii opisują rozkład wartości wokół wartości centralnych. Pozwalają one określić czy rozkład wartości jest symetryczny, czy nie. Miary asymetrii są rownież nazywane miarami skośności.

Wyróżniamy dwie asymetrie:
— prawostronną, gdy $\bar{x} > M_e > M_o$,
— lewostronną, gdy $\bar{x} < M_e < M_o$,
gdzie: $\bar{x}$ – wartość średnia z próby, M_e – mediana próby oraz M_o – wartość modalna próby. Natomiast dla rozkladu symetrycznego zachodzą następujące równosci: $\bar{x} = M_e = M_o$. Na rysunku 2.6 zostały przedstawione histogrmy dla jednej próby losowej o rozkładzie symetrycznym i dwóch prób losowych o różnych asymetriach.



Rysunek 2.6. Histogramy szeregów statystycznych o różnej asymetrii: (a) asymetria prawostronna, (b) rozkład symetryczny, (c) asymetria lewostronna

Podstawowymi klasycznymi miarami opisującymi asymetrię są:
— współczynnik asymetrii,
— współczynnik skośności Pearsona.
Oba wymienione współczynniki przyjmują wartości z przedziału $\langle -1, 1\rangle$. Dla rozkładów o asymetrii prawostronnej przyjmują wartości dodatnie, natomiast dla rozkładów o asymetrii lewostronnej są ujemne. Z kolei dla rozkladów symetrycznych przyjmują wartość 0. Oba współczynniki, jeśli są różne od zera określają jednoznacznie asymertię prawo- i lewostronną, lecz są niejednoznaczne w zerze. Znaczy to, że istnieją rozkłady niesymetryczne dla których wyżej wymienione współczynniki są równe zeru, jednak takie rozkłady są rzadko spotykane w praktyce. Zatem z faktu, że wspólczynniki przyjmują wartość zero nie wynika, iż rozkład musi być symetryczny. Istnieje również pozycyjny współczynnik skośności, który ma podobne własności jak miary klasyczne.

Jeśli asymetria rozkładu jest niewielka to między średnią ($\bar{x}$), modą (M_o) i medianą (M_e) zachodzi przybliżony związek:

$$\bar{x} - M_o \approx 3\,(\bar{x} - M_e)$$

2.4.1. Współczynnik asymetrii

Podstwową miarą asymetrii jest współczynnik asymetrii. Dla szeregu szczegółowego o długosci n zdefinowany jest wzorem:

$$A_s = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

gdzie:
— x_i – kolejne wartości w szeregu szczegółowym,
— $\bar{x}$ – średnia arytmetyczna wartości w szeregu szczegółowym dana wzorem: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
— s – odchylenie standardowe wartości w szeregu szczegółowym dane wzorem: $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Natomiast dla szeregu rozdzielczego punktowego współczynnik asymetrii wyznaczamy ze wzoru:

$$A_s = \frac{\frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^3 n_i}{s^3}$$

gdzie:

— x_i – i-ta wartość cechy,

— n_i – liczebność i-tej wartości cechy,

— k – liczba różnych wartości cechy,

— n – wielkość próby: $n = \sum_{i=1}^{k} n_i$,

— $\bar{x}$ – średnia arytmetyczna dla szeregu rozdzielczego punktowego dana wzorem: $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot x_i$

— s – odchylenie standardowe dla szeregu rozdzielczego punktowego dane wzorem: $s = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i \cdot (x_i - \bar{x})^2}$

Z kolei dla szeregu rozdzielczego przedziałowego współczynnik asymetrii wyznaczamy ze wzoru:

$$A_s = \frac{\frac{1}{n}\sum_{i=1}^{k}(\overline{x_i} - \bar{x})^3 n_i}{s^3}$$

gdzie:

— $\overline{x_i}$ – środek i-tej klasy,

— n_i – liczebność i-tej klasy,

— k – liczba klas,

— n – wielkość próby: $n = \sum_{i=1}^{k} n_i$,

— $\bar{x}$ – średnia arytmetyczna dla szeregu rozdzielczego przedziałowego dana wzorem: $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot \overline{x_i}$,

— s – odchylenie standardowe dla szeregu rozdzielczego punktowego dane wzorem: $s = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i \cdot (\overline{x_i} - \bar{x})^2}$.

Przykład 2.26:

W poniższej tabeli zostały zestawione wartości współczynnika asymetrii dla danych z przykładów 2.1, 2.1 oraz 2.3.

nr przykładu oraz typ szeregu statystycznego	A_s	typ asymetrii
2.1: szczegółowy	$0,38$	prawostornna
2.2: rozdzielczy punktowy	$-0,50$	lewostronna
2.3: rozdzielczy przedziałowy	$-0,12$	lewostronna

2.4.2. Współczynnik skośności Pearsona

Karl Paerson zaproponował miarę asymetrii łączącą parametry klasyczne (średnią arytmetyczną) i parametry pozycyjne (medianę), korzystając z faktu, że w rozkładzie symetrycznym wartość średnia, mediana oraz moda pokrywają się.

Współczynnik skośności Pearsona dla wszystkich typów szeregów statystycznych wyznaczamy korzystając ze wzoru:

$$W_s = \frac{\bar{x} - M_e}{s}$$

gdzie:

— $\bar{x}$ – średnia arytmetyczna z próby,

— M_e – mediana,
— s – odchylenie standardowe.

Należy pamiętać, że wartości $\bar{x}$, M_e oraz s są wyznaczane zgodnie ze wzorami dla odpowiednich typow szeregów statystycznych.

Przykład 2.27:

Poniższa tabeli zawiera zestawione wartości średnie ($\bar{x}$), mediany (M_e), odchylenia standardowe (s) oraz współczynniki skośności Pearsona (W_s) dla danych z przykładów 2.1, 2.1 oraz 2.3.

nr przykładu oraz typ szeregu statystycznego	$\bar{x}$	M_e	s	W_s	typ asymetrii
2.1: szczegółowy	433,58	407	70,25	0,38	prwostronna
2.2: rozdzielczy punktowy	3,25	3,5	1,13	$-0,22$	lewostronna
2.3: rozdzielczy przedziałowy	10,25	10,82	3,95	$-0,14$	lewostronna

2.4.3. Pozycyjny współczynnik skośności

Inną miarą skośności jest pozycyjny współczynnik asymetrii, który jest wyliczany ze wzoru:

$$A_Q = \frac{Q_1 + Q_3 - 2M_e}{Q_3 - Q_1}$$

gdzie:
— Q_1 – pierwszy kwartyl,
— Q_3 – trzeci kwartyl,
— M_e – mediana.

Dla wszystkich typów szeregów statystycznych wykorzystywany jest ten sam wzór, lecz wartości Q_1, Q_3 i M_e wyznaczane są ze wzorów dla konkretnego typu.

Pozycyjny współczynnik skośności przyjmuje wartosci z przedziału $\langle -1, 1\rangle$. Jeśli $A_Q < 0$ rzkład cechy jest prawostronnie asymetryczny, natomiast jeśli $A_Q > 0$ to lewostronnie asymetryczny. Dla rozkladów symetrycznyh $A_Q = 0$.

Przykład 2.28:

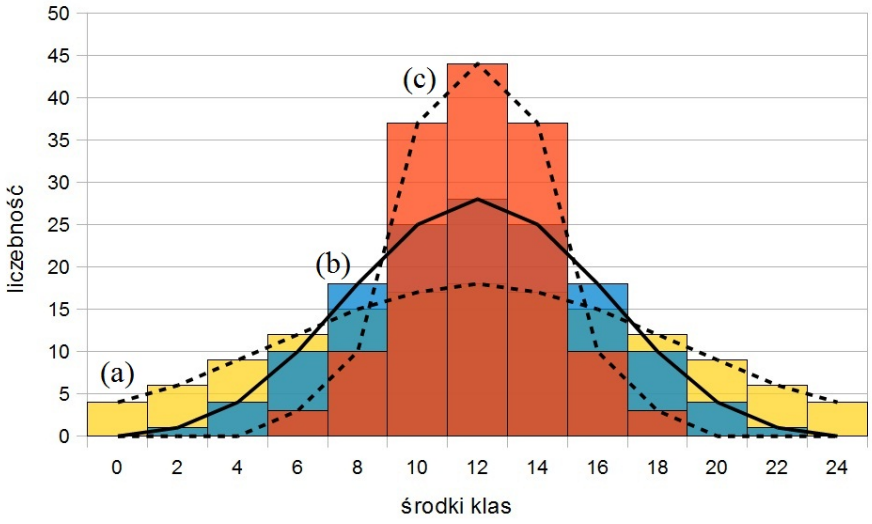
Poniższa tabela zawiera zestawione wartości średnie ($\bar{x}$), mediany (M_e), odchylenia standardowego (s) oraz współczynniki skośności Pearsona (W_s) dla danych z przykładów 2.1, 2.1 oraz 2.3.

nr przykładu oraz typ szeregu statystycznego	Q_1	Q_3	M_e	A_Q	typ asymetrii
2.1: szczegółowy	396,5	499,25	407	0,796	prawostronna
2.2: rozdzielczy punktowy	2,75	4	3,5	$-0,200$	lewostronna
2.3: rozdzielczy przedziałowy	7,9	13,75	10,82	0,002	prawostronna

2.5. Miary koncentracji rozkładu

Kolejną istotną cechą rozkładu wartości badanej cechy jest koncentracja. Miary koncentracji są nazywane również miarami spłaszczenia, ponieważ im bardziej rozkład cechy jest skupiony wokół średniej tym mniej jest spłaszczony i odwrotnie. Taki opis wymaga jednak punktu odniesienia względem, którego będziemy określać czy rozkład cechy ma dużą, czy małą koncentrację. Takim punktem odniesienia jest rozkład normalny, dla którego kurtoza jest zawsze równa 3.

Na rysunku 2.7 zostały przedstawione histogrmy dla trzech prób losowych o różnej koncentracji. Podstawowymi miarami koncentracji rozkładu są: kurtoza i eksces.



Rysunek 2.7. Histogramy szeregów statystycznych o różnej mierze koncentracji: (a) rozkład spłaszczony, (b) rozkład normalny, (c) rozkład wysmukły

2.5.1. Kurtoza i eksces

Podstawową miarą koncentaracji jest kurtoza. Dla szeregów szczegółowych definiuje się ją wzorem:

$$Ku = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4}$$

gdzie:

— n – liczebność próby,

— x_i – wartości próby,

— $\bar{x}$ – średnia arytmetyczna z próby dana wzorem:
$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

— s – odchylenie standardowe z próby dane wzorem:
$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Kurtozę dla szeregu rozdzielczego punktowego wyznaczamy korzystając ze wzoru:

$$Ku = \frac{\frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^4 n_i}{s^4}$$

gdzie:

— x_i – i-ta wartość cechy,

— n_i – liczebność i-tej wartości cechy,
— k – liczba różnych wartości cechy,
— n – liczebność próby: $n = \sum_{i=1}^{k} n_i$,
— $\bar{x}$ – średnia arytmetyczna z próby dana wzorem: $\bar{x} = \frac{1}{n} \sum_{i=1}^{k} n_i \cdot x_i$
— s – odchylenie standardowe z próby dane wzorem:
$s = \sqrt{\frac{1}{n} \sum_{i=1}^{k} n_i \cdot (x_i - \bar{x})^2}$

Natomiast dla szeregu rozdzielczego przedziałowego kurtozę obliczamy za pomocą następującego wzoru:

$$Ku = \frac{\frac{1}{n} \sum_{i=1}^{k} (\overline{x_i} - \bar{x})^4 n_i}{s^4}$$

gdzie:
— $\overline{x_i}$ – środek i-tej klasy,
— n_i – liczebność i-tej klasy,
— k – liczba klas,
— n – liczebność próby: $n = \sum_{i=1}^{k} n_i$,
— $\bar{x}$ – średnia arytmetyczna z wartości cechy dana wzorem:
$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} n_i \cdot \overline{x_i}$,
— s – odchylenie standardowe z wartości próby dane wzorem:
$s = \sqrt{\frac{1}{n} \sum_{i=1}^{k} n_i \cdot (\overline{x_i} - \bar{x})^2}$.

Jeżeli $Ku > 3$ to mówimy, że rozkład jest bardziej skoncentrowany („mniej płaski") niż rozkład normalny, a gdy $Ku < 3$ to, że rozkład jest mniej skoncentrowany („bardziej płaski") niż rozkład normalny. Jeśli $Ku = 3$ mówimy, że koncentracja (spłaszczenie) rozkładu jest taka sama jak rozkładu normalnego.

W przypadku kurtozy, aby określić koncentrację rozkładu cechy, należy wartość kurtozy porownywać z 3. DLatego wygodniejszą miarą wygodniejszą do porównań jest eksces definiowany wzorem:

$$Ex = Ku - 3$$

W tym przypadku wartości ekscesu porownywane są z zerem, ponieważ dla rozkładu normalnego zawsze mamy następującą równość: $Ex = 0$.

Przykład 2.29:

W poniższej tabeli zostały zestawione wartości kurtozy i ekscesu dla danych z przykładów 2.1, 2.1 oraz 2.3.

nr przykładu oraz typ szeregu statystycznego	Ku	Ex
2.1: szczegółowy	$2,02$	$-0,98$
2.2: rozdzielczy punktowy	$2,41$	$-0,59$
2.3: rozdzielczy przedziałowy	$2,37$	$-0,63$

W przykladach 2.1, 2.2 oraz 2.3 rozkłady są mniej skoncentrowane niż rozkład normalny.

2.6. Momenty a miary rozkładu

Momenty statystyczne są wilekościami powiązanymi z wileloma miarami rozkładu. Poniższa tabela przedstawia powiązania momentów z miarami rozkładu.

miara rozkładu	moment	wzór z wykorzystaniem momentów
średnia arytmetyczna	m_1	$\bar{x} = m_1$
średnia kwadratowa	m_2	$\bar{x}_k = \sqrt{m_2}$
odchylenie standartdowe	μ_2	$s = \sqrt{\mu_2}$
wariancja	μ_2	$s^2 = \mu_2$
odchylenie przecietne od średniej	b_1	$d_{\bar{x}} = b_1$
współczynnik zmienności	m_1, μ_2	$V_{\bar{x}} = \frac{\sqrt{\mu_2}}{m_1} \cdot 100\%$
współczynnik asymetrii	μ_2, μ_3	$A_s = \frac{\mu_3}{\sqrt{\mu_2}^3}$
kurtoza	μ_2, μ_4	$Ku = \frac{\mu_4}{{\mu_2}^2}$
eksces	μ_2, μ_4	$Ex = \frac{\mu_4}{{\mu_2}^2} - 3$

Ćwiczenia nr 5

Zadanie 2.7.

W pewnym doświadczeniu medycznym grupę pacjentów scharakteryzowano pod wzglendem masy ciała. Uzyskano następujące wyniki (w kg):

53	46	77	74	91	57	78	74	59	73
57	52	42	68	68	62	65	78	50	76
76	56	68	92	60	73	63	68	62	56
48	69	73	60	60	50	72	66	69	59
56	52	63	77	61	61	73	66	73	52

Wyznacz:
- współczynnik asymetrii rozkładu mas pacjentów,
- współczynnik skośności Pearsona,
- pozycyjnu współczynnik skośności,
- kurtozę i eksces.

Zadanie 2.8.

Grupę 100 mężczyzn w wieku 35 lat zbadano ze względu na liczbę brakujących zębów w uzębieniu i otrzymano następujące wyniki: Wyznacz:

liczba brakujących zębów	liczba mężczyzn
0	10
1	18
2	26
3	29
4	11
5	3
6	0
7	1
8	1

- współczynnik asymetrii rozkładu liczby brakujących zębów,
- współczynnik skośności Pearsona,

— pozycyjnu współczynnik skośności,
— kurtozę i eksces.

Zadanie 2.9.

W pewnym szpitalu przebadano 100 pacjentów pod wzglendem czasu działania leku przeciwbólowego. Poniższa tabela zawiera zestawienie wyników.

czas działania	1 – 2	3 – 4	5 – 6	7 – 8	9 – 10	11 – 12	13 – 14	15 –16
liczba pacjentów	2	5	11	23	23	21	10	5

Wyznacz:
— współczynnik asymetrii rozkładu czasu działania leku przeciw bólowego,
— współczynnik skośności Pearsona,
— pozycyjnu współczynnik skośności,
— kurtozę i eksces.

Zadanie 2.10.

W grupie 100 mężczyzn w wieku 40 lat zbadano liczbę erytrocytów w $1mm^3$ krwi. Wyniki zebrano w tabeli.

liczba erytrocytów (w mln)	liczba mężczyzn
4,5 – 4,6	6
4,7 – 4,8	8
4,9 – 5,0	23
5,1 – 5,2	27
5,3 – 5,4	14
5,5 – 5,6	18
5,8 – 5,7	4

Wyznacz wszystkie momenty zwykłe i centralne do 4 rzędu, a następnie wszystkie miary rozkładu jakie można z ich pomocą obliczyć.

Zadanie 2.11. [2]

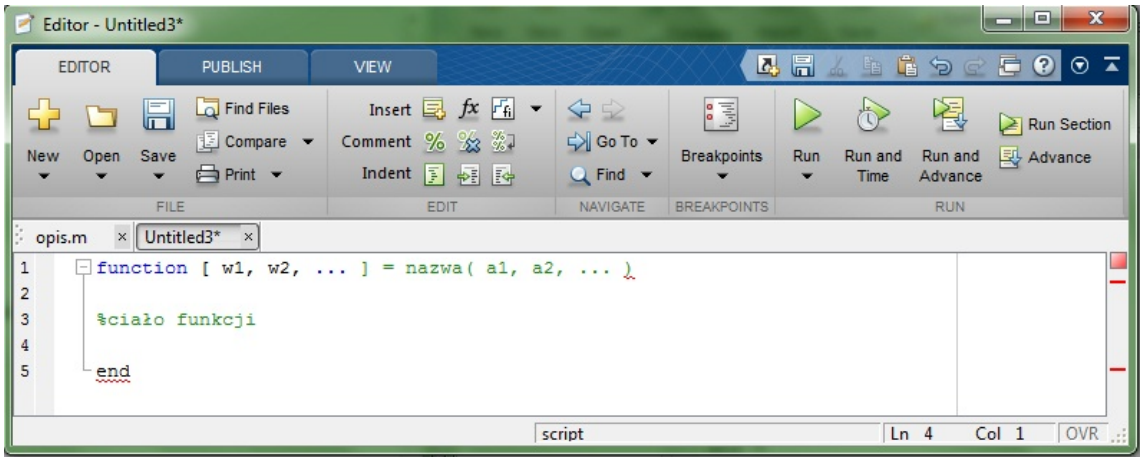
Wyznacz wszystkie momenty zwykłe i centralne do 4 rzędu, a następnie wszystkie miary rozkładu jakie można z ich pomocą obliczyć. Dla czterech szeregów rozdzielczych:

środki klas	liczebności			
$\overline{x}_i$	szereg I	szereg II	szereg III	szereg IV
1	0	2	0	2
2	6	2	2	4
3	12	10	20	10
4	14	22	12	12
5	12	10	10	20
6	6	2	4	2
7	0	2	2	0

Ćwiczenia nr 5 – Matlab

Funkcje

Nie zawsze funkcje dostępne w programie Matlab są wystarczające w codziennej pracy z tym programem. Dlatego też istnieje możliwość tworzenia własnych funkcji. Na rysunku 2.8 został przedstawiony schemat stryktury nowej funkcji stworzonej przez użytkownika.



Rysunek 2.8. Schemat stryktury funkcji stworzonej przez użytkownika

Pierwsze słowo `function` jest obowiązkowe, ponieważ definiuje typ nowego elementu. Następnie w nawiasach `[]` znajdują się nazwy zmiennych zwracanych przez fukcję. W miejscu słowa `nazwa` wstawiamy nazwę nowej funkcji. Ważną rzeczą jest nadanie nazwy nie występującej wcześniej w programie Matlab. W nawiasach `()` znajdują się argumenty funkcji, które muszą być podane podczas jej wywoływania. Następnie mamy cialo funkcji. Tworzenie nowej funkcji kończy słówko `end`.

Praktyczne tworzenie nowej funkcji zostanie przedstawione w zadaniu III.

Przykładowe zadanie

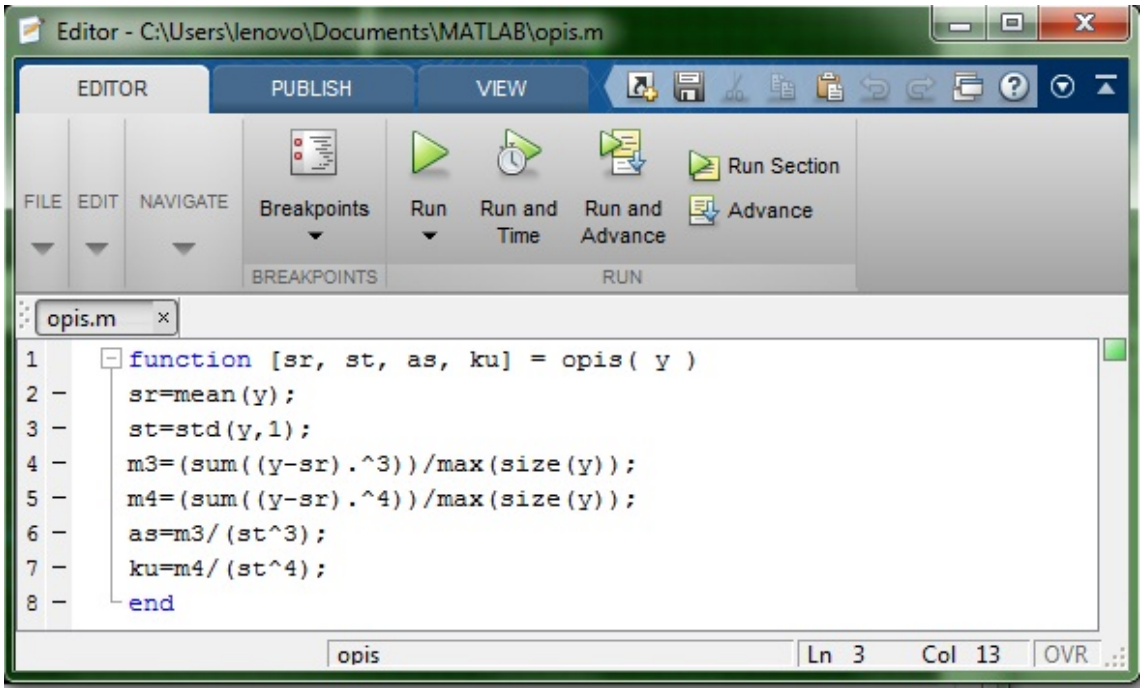
Zadanie III.

Wyznacz średnią arytmetyczną, odchylenie standardowe, współczynnik asymetrii oraz kurtozę dla trzech szeregów statystycznych:

szereg I	15	26	36	11	20	30	44	32	24	20	17	16
szereg II	31	18	29	30	26	22	30	22	22	15	15	17
szereg III	33	21	26	16	31	29	8	46	12	27	34	10

krok 1: Stworzenie funkcji wysnaczającej wszystkie badane parametry

Na rysunku 2.9 zostala przedstawiona funkcja pozwalająca wyznaczyć średnią arytmetyczną, odchylenie standardowe, współczynnik asymetrii oraz kurtozę.



Rysunek 2.9. Schemat stryktury funkcji stworzonej przez użytkownika

W lini `1` w nawiasach `[]` zostały określone 4 parametry wyjściowe:

- `sr` – średnia arytmetyczna,
- `st` – odchylenie standardowe,
- `as` – współczynnik asymetrii,
- `ku` – kurtoza.

Z kolei w liniach `2` i `3` znajdują się instrukcje podstawiające wartości parametrów pod zmienne `sr` i `st`. Wiadomo, że współczynnik asymetrii powiązany jest z trzecim momentem centralnym, a kurtoza z czwartym momentem centralnym. Dlategoteż w liniach `4` i `5` zostały wyznaczone wspomniane momenty centralne, które posłożyły w liniach `6` i `7` wuznaczeniu wartości parametrów `as` i `ku`.

krok 2: Wprowadzenie danych

Należy utworzyć trzy wektory z danymi:

```
I=[15 26 36 11 20 30 44 32 24 20 17 16];

II=[1 18 29 30 26 22 30 22 22 15 15 17];

III=[33 21 26 16 31 29 8  46 12 27 34 10];
```

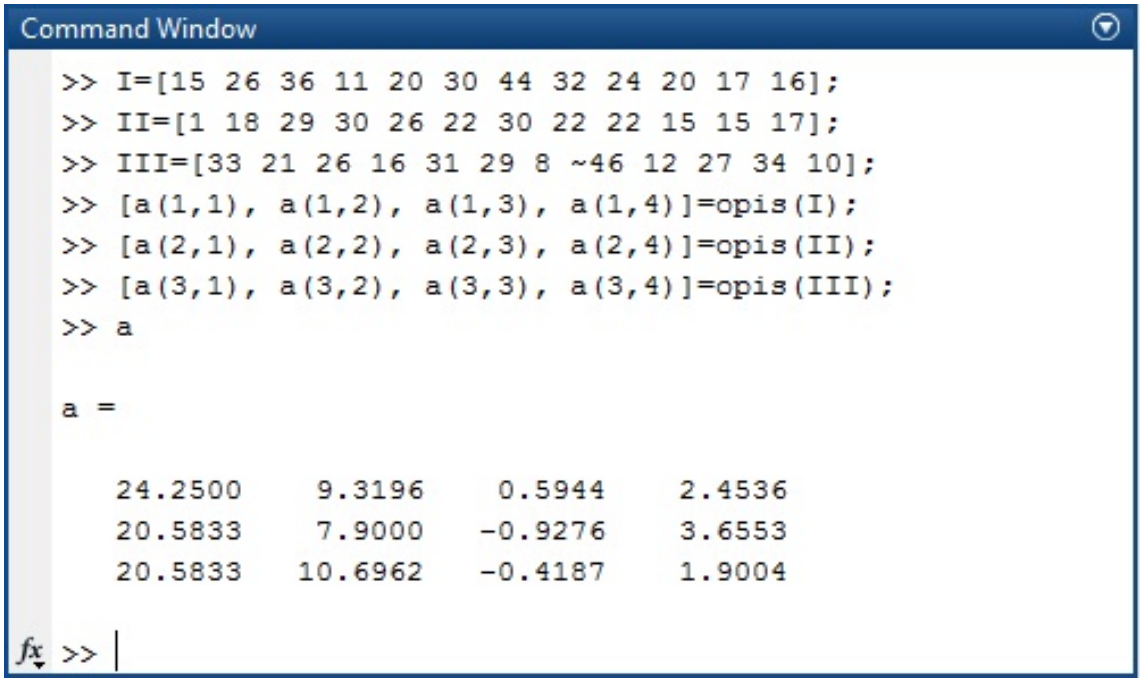
krok 3: Zastosowanie funkcji `opis`

```
[a(1,1), a(1,2), a(1,3), a(1,4)]=opis(I);

[a(2,1), a(2,2), a(2,3), a(2,4)]=opis(II);
```

```
[a(3,1), a(3,2), a(3,3), a(3,4)]=opis(III);
```

Na rysunku 2.10 zostało przedstawione rozwiązanie zadania III.



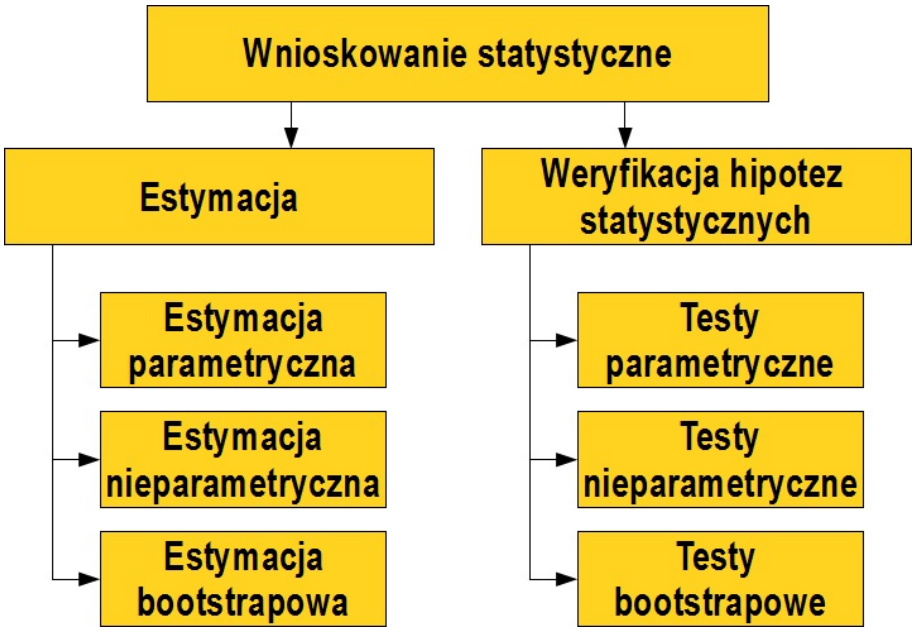
Rysunek 2.10. Rozwiązanie zadania III.

Rozdział 3

Podstawowe testy statystyczne

Statystyka matematyczna polega na wnioskowaniu o cechach populacji na podstawie wyników uzyskanych z próby. Przy wnioskowaniu statystycznym używamy metod rachunku prawdopodobieństwa. To znaczy: losujemy n elementów populacji generalnej i badamy je ze względu na pewną cechę (zmienną losową) X. W ten sposób uzyskujemy ciąg wartości dla n-elenentowej próby: $x_1\ x_2\ \ldots\ x_n$. Jest to jeden z możliwych ciągów bedących realizacją zniennej losowej X.

W statytstyce matematycznej wyróżnia się dwie techniki wnioskowania statystycznego: *estymację* i *weryfikację hipotez statystycznych*. W obu technikach wyróżnia się wnioskowanie parametryczne (pozwalające oszacować wartości parametrów rozkładu), nieparametrczne (pozwalające oszacować postać rozkładu) oraz bootstrapowe (polegające na szacowaniu parametrów, bądź postaci rozkładu na podstawie małych prób). Na rysunku 3.1 został przedstawiony podział metod wnioskowania statystycznego.



Rysunek 3.1. Podział metod wnioskowania statystycznego

Wnioskowanie statystyczne dostarcza zaledwie wniosków wiarygodnych a nie całkowicie pewnych, ponieważ jest oparte jedynie na częściowej informacji.

3.1. Estymacja

Jedną z dwóch technik wnioskowania statystycznego jest *estymacja*. Stanowi ona zbiór metod, które dają możliwość uogólniania wyników uzyskanych z próby losowej na całą po-

pulację oraz pozwala szacować błędy wynikające z tego uogólnienia. Estymacja pozwala szacować parametry i postaci nieznanych rozkładów, w przeciwieństwie do weryfikacji hipotez statystycznych, gdzie najpierw stawiamy hipotezę o rozkładzie, a następnie dowodzimy jej poprawność.

Wyróżniamy dwa podstawowe typy estymacji:

— *parametryczną*, polegającą na znajdowaniu nieznanych wartości parametrów populacji generalnej,

— *nieparametryczną*, pozwalającą znaleść postać rozkładu cechy populacji.

Omówimy wyłącznie estymację parametryczną i nieklasyczną: estymację bootstrapową.

Estymację parametryczną można podzielić na dwie grupy:

— *estymację punktową* – parametr opisujący populację jest estymowany (oceniany) konkretną wartością uzyskaną z próby,

— *estymację przedziałową* – parametr opisujący populację z pewnym prawdopodobienstwem należy do przedziału, nazywanego przedzialem ufności.

3.1.1. Estymacja punktowa

Funkcje $\hat{\theta}_n$ zmiennych losowych $X_1\ X_2\ \ldots\ X_n$ nazywamy *statystyką*. Statystyki jako funkcje zmiennych losowych są także zmiennymi losowymi o własnych rozkładach. Każdą statystykę $\hat{\theta}_n(X_1, X_2, \ldots, X_n)$ wykorzystywaną do oceny nieznanego porametru populacji θ nazywamy *estymatorem parametru* θ. Natomiast wartość estymatora $\hat{\theta}_n$ wyznaczoną dla konkretnej próby $x_1, x_2, \ldots, x_n$ nazywamy *oceną* t porametru θ:

$$t = \hat{\theta}_n(x_1, x_2, \ldots, x_n)$$

Dla danego parametru θ istnieje wiele estymatorów, lecz urzyteczne są jedynie te, które mają specyficzne własności. Możliwie „optymalnyęstymator jest:

— nieobciążony,

— najefektywniejszy,

— zgodny.

Jednak nawet najlepsze estymatory nie dają oszacowania precyzji i wiarygodności uzyskanych wyników.

Własności estymatorów

Nieobciążoność

Mówimy, że estymator jest *nieobciążony*, jeśli wartość oczekiwana tego estymatora jest równa szacowanemu parametrowi θ, tzn.

$$E[\hat{\theta}_n(X_1, X_2, \ldots, X_n)] = \theta$$

Natomiast jeśli $E[\hat{\theta}_n(X_1, X_2, \ldots, X_n)] \neq \theta$, to estymator $\hat{\theta}_n$ nazywamy *obciążonym*, a wielkość $B_n = E[\hat{\theta}_n(X_1, X_2, \ldots, X_n)] - \theta$ – *obciążeniem*. Estymator obciążony nazywamy *asymptotycznie nieobciążonym*, jeśli $\lim_{n\to\infty} B_n = 0$.

Efektywność

Efektywność estymatora $\hat{\theta}_n$ jest uzależniona od jego wariancji $D^2[\hat{\theta}_n]$, tzn. im mniejsza jest wariancja $D^2[\hat{\theta}_n]$, tym estymator $\hat{\theta}_n$ jest efektywniejszy. Z tego powodu szukamy estymatora o możliwie najminiejszej wariancji (*estymatora najefektywniejszego*).

Zgodność

Estymator nazywamy *zgodnym*, gdy jest stochastycznie zgodny z szacowanym parametrem. To znaczy, że im mamy większą próbę, tym większe jest prawdopodobieństwo, że błąd w ocenie parametru θ przez estymator $\hat{\theta}_n$ jest dowolnie mały:

$$\forall \epsilon > 0 \quad \lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1$$

Podstawowe estymatory

Nieznany parametr	Estymator	Własności
Wartość oczekiwana μ	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	zgodny, nieobciążony
	mediana z próby	zgodny, asymptotycznie nieobciążony
Wariancja σ^2, gdy μ – znane	$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$	zgodny, nieobciążony
Wariancja σ^2, gdy μ – nieznane	$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$	zgodny, nieobciążony
	$S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$	zgodny, nieobciążony
Odchylenie standardowe σ	S_1	zgodny
	S	zgodny
	S^*	zgodny
Wskaźnik struktury	$\hat{\theta} = \frac{k}{n}$	zgodny, nieobciążony, efektywny (tylko dla rozkładu Bernouliego i $n > 100$)
Współczynnik zmienności względem średniej V_μ	$V = \frac{S}{\bar{X}}$	zgodny

3.1.2. Estymacja przedziałowa

Innym sposobem estymacji jest *estymacja przedziałowa*, która daje możliwość oceny dokładności. Estymacja przedzialowa polega na podaniu *przedziału ufności* dla nieznanego parametru danej populacji.

Przedział (t_1, t_2) nazywamy przedziałem ufności dla parametru θ na *poziomie ufności* $1-\alpha$ ($\alpha \in (0,1)$), gdy spełnia dwa następujące warunki:

— t_1 jest wartością statystyki $\hat{\theta}_1(X_1, X_2, \ldots, X_n)$ dla konkretnej próby losowej, a t_2 wartością statystyki $\hat{\theta}_1(X_1, X_2, \ldots, X_n)$ dla tej samej próby losowej

— prawdopodobieństwo, że przedział pokrywa nieznany parametr θ jest równe $1-\alpha$:

$$P(t_1 < \theta < t_2) = 1 - \alpha$$

Liczbę α nazywamy *poziomem istotności*, a liczbę $1-\alpha$ – *współczynnikiem ufności* Przedział ufności może, lecz nie musi, pokrywać nieznanej wartość parametru θ. Z powyższej definicji wynika, że dla wielu różnych prób losowych znajdziemy wiele realizacji przedziału ufności, a częstość występowania przedziałów ufności zawierających rzeczywistą wartość parametru θ będzie w przybliżeniu równa $1-\alpha$.

Długość przedziału ufności jest zależna od współczynnika ufności $1-\alpha$. Najczęściej przyjmowane wartości współczynnika ufności to: $0,90$, $0,95$, $0,99$.

Wraz ze wzrostem współczynnika ufności następuje wzrost długości tego przedzialu oraz pogorszenie precyzji dokonywanej oceny, a z drugiej strony zwiększa się dokladność estymacji. Przy doborze przedziału ufności warto ustalić *względną precyzję oszacowania danego parametru* $B(\theta)$. Jeżeli:

— $B(\theta) \leqslant 5\%$ – duża precyzja oszacowania,

— $5\% < B(\theta) \leqslant 10\%$ – średnia precyzja oszacowania, należy zachować ostrożność przy uogólnianiu wyników,

— $B(\theta) < 5\%$ – słaba precyzja oszacowania, nie powinno się przeprowadzać uogólnień wyników.

Wyznaczanie przedziału ufności dla średniej

Model 1

Założenia:

— populacja generalna ma rozkład $N(\mu, \sigma)$,

— μ – nieznane,

— σ – znane.

Końce przedziału ufności (t_1, t_2) dla wartości średniej określają statystyki:

$$\hat{\theta}_1(X_1, X_2, \ldots, X_n) = \bar{X} - u_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

$$\hat{\theta}_2(X_1, X_2, \ldots, X_n) = \bar{X} + u_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

gdzie:

— n - liczebność próby

— $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X$,

— u_α – wartość taka, że dla zmiennej losowej U o rozkładzie $N(0,1)$ zachodzi równość:

$$P\{-u_\alpha < U < u_\alpha\} = 1 - \alpha$$

W tabeli poniżej przedstawione są wartości u_α dla najczęściej przyjmowanych współczynników ufności $1 - \alpha$.

$1-\alpha$	u_α
$0,90$	$1,64$
$0,95$	$1,96$
$0,99$	$2,58$

Względna precyzja oszacowania wartości średniej dana jest wzorem:

$$B(\mu) = \frac{u_\alpha \cdot \sigma}{\bar{X} \cdot \sqrt{n}} \cdot 100\%$$

Model 2

Założenia:

— populacja generalna ma rozkład $N(\mu, \sigma)$,

— μ – nieznane,

— σ – nieznane,

— mała próba ($n \leqslant 30$)

Końce przedziału ufności (t_1, t_2) dla wartości średniej określają statystyki:

$$\hat{\theta}_1(X_1, X_2, \ldots, X_n) = \bar{X} - t_\alpha \cdot \frac{S^*}{\sqrt{n}} = \bar{X} - t_\alpha \cdot \frac{S}{\sqrt{n-1}}$$

$$\hat{\theta}_2(X_1, X_2, \ldots, X_n) = \bar{X} + t_\alpha \cdot \frac{S^*}{\sqrt{n}} = \bar{X} + t_\alpha \cdot \frac{S}{\sqrt{n-1}}$$

gdzie:

— n - liczebność próby
— $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X$,
— $S^* = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$, a $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$,
— t_α – wartość taka, że dla zmiennej losowej T o rozkładzie t-Studenta o $n-1$ stopniach swobody zachodzi równość:

$$P\{-t_\alpha < T < t_\alpha\} = 1 - \alpha$$

Wartość t_α dla współczynnika ufności $1-\alpha$ można odczytać z tabeli na stronie **??**.

Względna precyzja oszacowania wartości średniej dana jest wzorem:

$$B(\mu) = \frac{t_\alpha \cdot S^*}{\bar{X} \cdot \sqrt{n}} \cdot 100\% = \frac{t_\alpha \cdot S}{\bar{X} \cdot \sqrt{n-1}} \cdot 100\%$$

Model 3

Założenia:
— populacja generalna ma rozkład $N(\mu, \sigma)$,
— μ – nieznane,
— σ – nieznane,
— duża próba ($n > 30$)

Końce przedziału ufności (t_1, t_2) dla wartości średniej określają statystyki:

$$\hat{\theta}_1(X_1, X_2, \ldots, X_n) = \bar{X} - u_\alpha \cdot \frac{S}{\sqrt{n}}$$

$$\hat{\theta}_2(X_1, X_2, \ldots, X_n) = \bar{X} + u_\alpha \cdot \frac{S}{\sqrt{n}}$$

gdzie:
— n - liczebność próby
— $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X$,
— $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$,
— u_α – wartość taka, że dla zmiennej losowej U o rozkładzie $N(0,1)$ zachodzi równość:

$$P\{-u_\alpha < U < u_\alpha\} = 1 - \alpha$$

Wartości u_α dla najczęściej przyjmowanych współczynników ufności $1-\alpha$ zestwione są w tabeli na stronie 76.

Względna precyzja oszacowania wartości średniej dana jest wzorem:

$$B(\mu) = \frac{u_\alpha \cdot S}{\bar{X} \cdot \sqrt{n-1}} \cdot 100\%$$

Wyznaczanie przedziału ufności dla wariancji i odchylenia standardowego

Model 1

Założenia:
— populacja generalna ma rozkład $N(\mu, \sigma)$,
— μ – nieznane,
— σ – znane,
— mala próba $n \leqslant 30$.

Końce przedziału ufności (t_1, t_2) dla wariancji określają statystyki:

$$\hat{\theta}_1(X_1, X_2, \ldots, X_n) = \frac{(n-1)S^{*2}}{c_1} = \frac{nS^2}{c_1}$$

$$\hat{\theta}_2(X_1, X_2, \ldots, X_n) = \frac{(n-1)S^{*2}}{c_2} = \frac{nS^2}{c_2}$$

gdzie:

- n - liczebność próby
- $S^* = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$, a $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ dla $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X$,
- c_1, c_2 – wartości takie, że dla zmiennej losowej C o rozkładzie χ^2 zachodzą relacje:

$$P\{C \geqslant c_1\} = \frac{1}{2}\alpha$$

$$P\{C < c_2\} = \frac{1}{2}\alpha$$

Wartości c_1 i c_2 dla współczynników ufności $1 - \alpha$ można odczytać z tabeli na stronie **??**.

W celu otrzymania przedziału ufności dla odchylenia standardowego pierwiastkujemy końce przedziału ufności dla wariancji. Zatem przedziału ufności dla odchylenia standardowego to: $(\sqrt{t_1}, \sqrt{t_2})$.

Model 2

Założenia:

- populacja generalna ma rozkład $N(\mu, \sigma)$,
- μ – nieznane,
- σ – nieznane,
- duża próba $(n > 30)$

Końce przedziału ufności (t_1, t_2) dla odchylenia standardowego określają statystyki:

$$\hat{\theta}_1(X_1, X_2, \ldots, X_n) = S - u_\alpha \cdot \frac{S}{\sqrt{2n}}$$

$$\hat{\theta}_2(X_1, X_2, \ldots, X_n) = S + u_\alpha \cdot \frac{S}{\sqrt{2n}}$$

gdzie:

- n - liczebność próby
- $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ dla $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X$,
- u_α – wartość taka, że dla zmiennej losowej U o rozkładzie $N(0, 1)$ zachodzi równość:

$$P\{-u_\alpha < U < u_\alpha\} = 1 - \alpha$$

Wartości u_α dla najczęściej przyjmowanych współczynników ufności $1 - \alpha$ zestwione są w tabeli na stronie 76.

Względna precyzja oszacowania odchylenia standardowego dana jest wzorem:

$$B(\sigma) = \frac{u_\alpha}{\sqrt{2n}} \cdot 100\%$$

3.1.3. Wyznaczenie liczebności próby

Wyznaczone przedziały ufności nie zawsze dostarczają ocenę szacowanego parametru o wymaganym stopniu precyzji. Na względną precyzję oszacowania parametru θ ($B(\theta)$) mają wpływ:

— wariancja cechy populacji generalnej – nie mamy na nią wplywu,

— współczynnik ufności – pozwala jedynie w ograniczony sposób zmieniać watrosci $B(\theta)$,

— liczebność próby – jest zależna od wyboru badacza.

Zatem, aby uzyskać wymagny stopień precyzji modyfikujemy liczebność próby.

Omówimy jedynie ustalanie liczebności próby dla estymacji wartości średniej, pomijając ustalanie liczebności próby dla oszacowania wartości wariancji i odchylenia standardowego, ponieważ wyznaczane przdziały ufności nie są symetryczne dla tych dwóch parametrów.

Dla modelu 1 estymacji wartości średniej oszacowanie liczebności n o precyzji nie przekraczającej wartości d daje nierówność:

$$\frac{u_\alpha \cdot \sigma}{\sqrt{n}} < d$$

Zatem:

$$n > \left\lfloor \left(\frac{u_\alpha \cdot \sigma}{d}\right)^2 \right\rfloor$$

gdzie: $\lfloor x \rfloor$ to część całkowita liczby x.

Natomiast dla modelu 2 estymacji wartości średniej oszacowanie liczebności n o precyzji nie przekraczającej wartości d daje nierówność:

$$\frac{t_\alpha \cdot S}{\sqrt{n-1}} < d$$

Wówczas:

$$n > \left\lfloor \left(\frac{t_\alpha \cdot S}{d}\right)^2 + 1 \right\rfloor$$

Wyznaczanie przedziału ufności dla wskaźnika struktury

Model

Założenia:

— populacja generalna ma rozkład dwupunktowy z parametrem p,

— duża próba ($n > 100$)

Końce przedziału ufności (t_1, t_2) dla wskaźnika struktury określają statystyki:

$$\hat{\theta}_1(X_1, X_2, \ldots, X_n) = \frac{m}{n} - u_\alpha \cdot \sqrt{\frac{\frac{m}{n} \cdot \left(1 - \frac{m}{n}\right)}{n}}$$

$$\hat{\theta}_2(X_1, X_2, \ldots, X_n) = \frac{m}{n} + u_\alpha \cdot \sqrt{\frac{\frac{m}{n} \cdot \left(1 - \frac{m}{n}\right)}{n}}$$

gdzie:

— n – liczebność próby,

— $\frac{m}{n}$ – wskaźnik struktury próby,

— u_α – wartość taka, że dla zmiennej losowej U o rozkładzie $N(0,1)$ zachodzi równość:

$$P\{-u_\alpha < U < u_\alpha\} = 1 - \alpha$$

Wartości u_α dla najczęściej przyjmowanych współczynników ufności $1-\alpha$ zestwione są w tabeli na stronie 76.

Względna precyzja oszacowania wskaźnika struktury dana jest wzorem:

$$B\Big(\frac{m}{n}\Big) = \frac{u_\alpha}{\frac{m}{n}} \cdot \sqrt{\frac{\frac{m}{n} \cdot \Big(1 - \frac{m}{n}\Big)}{n}} \cdot 100\%$$

3.1.4. Bootstrapowa estymacja parametrów

Metody bootstrapowe są metodami symulacyjnymi zaproponowanymi przez Eforda w 1979r. Te metody są stosowane miedzy innymi do aproksymacji rozkładów estymatorów i statystyk testowych. Wnioskowanie statystyczne w tym przypadku bazuje na wartościach wygenerowanych wielokrotnie z tej samej próby. Załużmy, że badamy populację generalną ze względu na zmienną losową X o nieznanym nam rozkładzie F oraz, że $X_1, X_2, \ldots, X_n$ jest prostą próbą losową z badanej populacji, a $x_1, x_2, \ldots, x_n$ jest pewną realizacją zmiennych losowych $X_1, X_2, \ldots, X_n$. Niech $R(x_1, x_2, \ldots, x_n)$ jest statystyką określoną na przestrzeni prób. *Rozkład bootstrapowy z próby* $\hat{F}$ dany jest wzorem:

$$P(X_B = x_k) = \frac{1}{n} \quad \text{dla} \quad k = 1, 2, \ldots, n$$

Próbą bootstrapową nazywamy losowy ciąg $X_1^*, X_2^*, \ldots, X_m^*$ niezależnych zmiennych losowych X_k^* dla $k = 1, 2, \ldots, m$ o rozkładzie bootstrapowym. Konkretną realizację próby bootstrapowej będziemy oznaczać $x_1^*, x_2^*, \ldots, x_m^*$. Natomiast *rozkładem bootstrapowym statystyki* R nazywamy rozkład zmiennej losowej $R^* = R(x_1^*, x_2^*, \ldots, x_m^*)$. Zatem rozkład statystyki R przybliżamy rozkładem zmiennej losowej R^*, co stanowi podstawową zasadę metody bootstrapowej.

Rozkład zmiennej losowej R^* można wyznaczyć trzema metodami:
— poprzez rozważania teoretyczne,
— wykorzystując rozwinięcie w szereg Taylora momentów zmiennej losowej R^*,
— wykorzystując aproksymacje Monte Carlo zmiennej losowej R^*.

W praktyce wykorzystuje się metodę Monte Carlo. Aby uzyskać zadowalające oszacowanie rozkładu zmiennej losowej R^* należy wylosować co najmniej 1000 ($N \geqslant 1000$) prób bootstrapowych wygenerowanych z rozkładu bootstrapowego.

Bootstrapowa estymacja punktowa

Niech θ będzie szacowanym parametrem rozkładu zmiennej losowej X, a R estymatorem tego parametru. Wówczas losujemy N prób bootstrabowych $x_1^*, x_2^*, \ldots, x_n^*$ z losowej próby prostej $x_1, x_2, \ldots, x_n$ pochodzącej z populacji. Jest to oczywiście losowanie ze zwracaniem wylosowanych wartości. Następnie dla każdej próby bootstrapowej wyznaczamy wartość statystyki R^*. W ten sposób otrzymujemy ciąg realizacji zmiennej losowej R^*: $r_1^*, r_2^*, \ldots, r_N^*$. Oceną parametru θ jest wartość:

$$\bar{r}^* = \frac{1}{N} \sum_{i=1}^{N} r_i^*$$

Zatem bootstrapowyn estymatorem zmiennej losowej X jest nastepująca statystyka:

$$\bar{R}^* = \frac{1}{N}\sum_{i=1}^{N} R_i^*$$

gdzie:
— N – licba prób bootstrapowych (z góry ustalona i $N \geqslant 1000$);
— R_i^* – wartość statystyki R^* dla i-tej próby bootstrapowej.

Bootstrapowa estymacja przedziałowa

Metody bootstrapowe dają również możlowość estymacji przedziałowej. Istnieje kilka metod bootstrapowych wyznaczania przedziałów ufności dla parametru θ rozkładu. pomniżej zostaną omówione dwie z nich:
— metoda percentyli,
— metoda t-bootstrapowa.

Metoda percentyli

Niech $\hat{\theta} = R(X_1, X_2, \ldots, X_n)$ będzie estymatorem parametru θ. Wówczas aproksymujemy rozkład estymatora $\hat{\theta}$ rozkładem zmiennej losowej

$$\hat{\theta}^* = R^*(X_1, X_2, \ldots, X_n) = R(X_1^*, X_2^*, \ldots, X_m^*)$$

Wówczas przedziałem ufności parametru θ przy poziomie istotności α obliczonym metodą percentyli, jest przedział postaci:

$$(\hat{\theta}^*_{\frac{\alpha}{2}}, \hat{\theta}^*_{1-\frac{\alpha}{2}})$$

gdzie:
— $\hat{\theta}^*_{\frac{\alpha}{2}}$ – kwantyl rzędu $\frac{\alpha}{2}$ rozkładu bootstrapowego statystyki θ,
— $\hat{\theta}^*_{1-\frac{\alpha}{2}}$ – kwantyl rzędu $1-\frac{\alpha}{2}$ rozkładu bootstrapowego statystyki θ.

Metoda t-bootstrapowa

Niech $\hat{\theta} = R(X_1, X_2, \ldots, X_n)$ będzie estymatorem parametru θ, a rozkład estymatora $\hat{\theta}$ niech jest aproksymowany rozkładem zmiennej losowej

$$\hat{\theta}^* = R^*(X_1, X_2, \ldots, X_n) = R(X_1^*, X_2^*, \ldots, X_m^*)$$

W mdetodzie t-bootstrapowe wykorzystujemy statystykę postaci:

$$t^* = \frac{\hat{\theta}^* - \hat{\theta}}{S_{\hat{\theta}^*}} \qquad \text{gdzie:} \qquad S_{\hat{\theta}^*} = \frac{1}{n}\sum_{i=1}^{N}(\hat{\theta}^* - \bar{\hat{\theta}}^*)^2 \quad \text{i} \quad \bar{\hat{\theta}}^* = \frac{1}{n}\sum_{i=1}^{N}\hat{\theta}^*$$

Wówczas przedziałem ufności parametru θ przy poziomie istotności α obliczonym metodą t-bootstrapową, jest przedział postaci:

$$(\hat{\theta} - t^*_{1-\frac{\alpha}{2}}, \hat{\theta} - t^*_{\frac{\alpha}{2}})$$

gdzie:
— $\hat{\theta} - t^*_{\frac{\alpha}{2}}$ – kwantyl rzędu $\frac{\alpha}{2}$ rozkładu bootstrapowego statystyki t^*,
— $\hat{\theta} - t^*_{1-\frac{\alpha}{2}}$ – kwantyl rzędu $1-\frac{\alpha}{2}$ rozkładu bootstrapowego statystyki t^*.

Ćwiczenia nr 6

Zadanie 3.1. [7]

W związku z zaopatrzeniem Banku Krwi chcemy oszacować procent krwiodawców posiadających zerową grupę krwi. W tym celu pobrano losową próbe 144 krwiodawców, wśród których 52 osoby miały zerową grupę krwi. Zbudować przedział, który z prawdopodobieństwem 0,95 pokrywa oszacowany procent krwiodawców posiadających zerową grupę krwi.

Zadanie 3.2. [6]

W doświadczeniu medycznym bada się czas snu pacjentow leczonych na pewną chorobę. Zmierzono u 16 wylosowanych niezaleznie pacjentów czas snu i otrzymano następujące wyniki w minutach: 435 533 393 458 525 481 324 437 348 503 383 395 416 553 500 488. Przyjmując, że czas snu ma rozkład $N(m, 70)$ oszacować metodą przedziałową średni czas snu pacjentów, przyjmując współczynnik ufności 0,99.

Zadanie 3.4. [6]

W doswiadczeniu biochemicznym badano czas życia komórek. Dokonano 5 niezależnych pomiarów i otrzymano wyniki w godzinach: 4,7 5,3 4,0 3,8 4,7. Przy założeniu normalności rozkładu czasu życia komórek, oszacować średni czas życia komórek z prawdopodobieństwem 0,90.

Zadanie 3.5. [6]

W doswiadczeniu bada się utlenowanie tkankowe wątroby królików (ilość mililitrów tlenu zużytego w ciągu 1 godziny przez 100 mg wilgotnej tkanki). Dokonano 40 pomiarów i otrzymano wyniki (w postaci szeregu rozdzielczego):

Ilość zurzytego tlenu	Liczba pomiarow
15-25	4
25-35	6
35-45	13
45-55	8
55-65	6
65-75	3

Oszacować z prawdopodobieństwem 0,95 średnią ilość zużytego tlenu przy utlenianiu tkankowym wątroby krolika.

Zadanie 3.6. [6]

W doswiadczeniu farmakologicznym bada się poziom kreatyniny we krwi królików. Dokonano 9 oznaczen na wylosowanych zwierzętach i otrzymano wyniki (w mg%): 1,7 2,0 1,8 1,2 1,5 1,4 2,0 1,2 1,6. Przyjmując współczynnik ufności 0,98 zbudować przedział ufności dla wariancji oznaczeń poziomu kreatyniny we krwi królików.

Zadanie 3.7. [7]

Wyznaczyć minimalną liczebność próby do oszacowania średniego wzrostu uczniów u klasach piątych szkół podstawowych, jeżeli w probie wstępnej liczącej 10 uczniów otrzymano następujace wyniki (w cm): 125 150 145 130 155 140 160 125 155 135. Zakładamy dopuszczalny błąd szacunku 5 cm, przy współczynniku ufności 0,95.

Zadanie 3.8.

W doświadczeniu medycznym bada się czas snu pacjentow leczonych na pewną chorobę. Zmierzono u 16 wylosowanych niezaleznie pacjentów czas snu i otrzymano następujące wyniki w minutach: 435 533 393 458 525 481 324 437 348 503 383 395 416 553 500 488. Oszacować metodami bootstrapowymi przedział ufności dla średniego czasu snu pacjentów, przyjmując współczynnik ufności 0,99. Porównać otrzymane wyniki z klasycznym wynikiem z zadania 3.2

Zadanie 3.9. [6]

W doswiadczeniu biochemicznym badano czas życia komórek. Dokonano 5 niezależnych pomiarów i otrzymano wyniki w godzinach: 4,7 5,3 4,0 3,8 4,7. Oszacować metodami bootstrapowymi średni czas życia komórek z prawdopodobieństwem 0,90. Porównać otrzymane wyniki z klasycznym wynikiem z zadania 3.4

Zadanie 3.10.

W związku z zaopatrzeniem Banku Krwi chcemy oszacować procent krwiodawców posiadających zerową grupę krwi. W tym celu pobrano losową próbe 144 krwiodawców, wśród których 52 osoby miały zerową grupę krwi. Wykorzystując metody bootstrapowe zbudować przedział, który z prawdopodobieństwem 0,95 pokrywa oszacowany procent krwiodawców posiadających zerową grupę krwi. Porównać otrzymane wyniki z klasycznym wynikiem z zadania 3.1

3.2. Testy nieparametryczne

Jednym z rodzajów wnioskowania statystycznego jest testowanie (weryfikacja) hipotez statystycznych. Polega ono na badaniu przypuszczeń o wartościach parametrów charakteryzujących daną populację lub o rozkładzie badanych cech populacji. Wnioskowanie odbywa się na podstawie wyników uzyskanych dla wybranej części populacji (próby). Każde przypuszczenie na temat populacji nazywane jest *hipotezą statystyczną*. Wyróżnia się dwa rodzaje hipotez statystycznych:

— *hipotezy parametryczne*, które dotyczą wartości parametrów rozkładu,

— *hipotezy nieparametryczne*, które dotyczą funkcji rozkładu populacji.

Hipotezy statystyczne weryfikujemy wykonując pewne ustalone reguły postępowania, które nazywamy *testami statystycznymi*. Testy statystyczne przyporządkowują każdej losowo wybranej części populacji decyzję o przyjęciu bądź odrzuceniu postawionej hipotezy.

Weryfikacja hipotez statystycznych składa się z następujących etapów:

1. Formułowanie hipotezy zerowej (H_0), czyli hipotezy, którą testujemy.

2. Formułowanie hipotezy alternatywnej (H_1), czyli hipotezy, którą przyjmiemy po odrzuceniu hipotezy zerowej.

3. Wyboru testu statystycznego pozwalającego na weryfikację postawionej hipotezy.

4. Obliczenie wartości parametru charakterystycznego dla danego testu, nazywanego *statystyką testową*.

5. Określenie poziomu istotności (α), czyli prawdopodobieństwa popełnienia błędu I rodzaju,

6. Podjęcie decyzji o odrzuceniu lub przyjęciu hipotezy zerowej.

Podczas testowania hipotez można popełnić dwa rodzaje błędów:

— *błąd I rodzaju*, czyli odrzucenie hipotezy zerowej pomimo tego, że jest prawdziwa,

— *błąd II rodzaju*, przyjęcie hipotezy zerowej pomimo tego, że jest fałszywa.

Prawdopodobieństwo popełnienia błędu I rodzaju nazywamy *poziomem istotności* i oznaczamy grecką literą α, prawdopodobieństwo popełnienia błędu II rodzaju oznaczamy grecką literą β.

Testy statystyczne, w których uwzględniamy tylko popełnienie błędu I rodzaju, pomijając kwestię popełnienia błędu II rodzaju nazywamy testami istotności. Testy istotności pozwalają jedynie na podjęcie jednej z dwóch decyzji:

— odrzucenie hipotezy zerowej na rzecz hipotezy alternatywnej na poziomie istotności α,

— stwierdzenie, że nie ma podstaw do odrzucenia hipotezy zerowej.

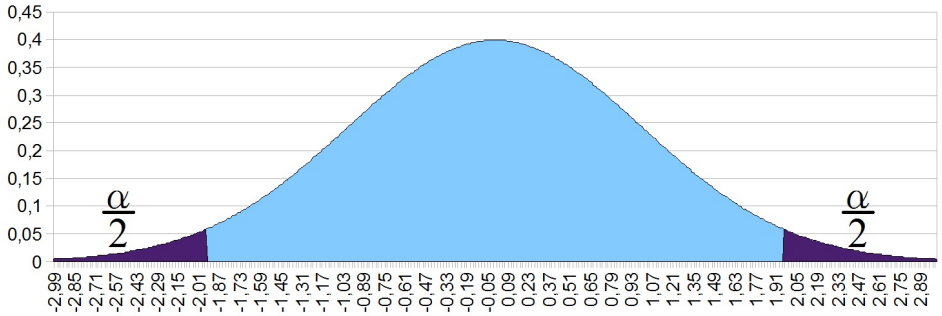
W testach statystycznych przyjmuje się zwykle, że poziom istotności wynosi 0,1 lub 0,05. Im mniejszy przyjmiemy poziom istotności tym trudniej jest odrzucić hipotezę zerową i tym bardziej wiarygodna jest hipoteza alternatywna.

Wybierając poziom istotności wyznaczamy jednocześnie *obszar krytyczny testu*. Obszarem krytycznym testu nazywamy zbior tych wartości statystyki testowej, które przemawiają za odrzuceniem hipotezy zerowej. Położenie obszaru krytycznego jest uzależnione od postaci hipotezy alternatywnej H_1. Załużmy następującą postać hipotezy zerowej: $H_0 : m = A$.

Wówczas obszar krytyczny może być:

1. dwustronny:

Hipoteza alternatywna: $H_1 : m \neq A$

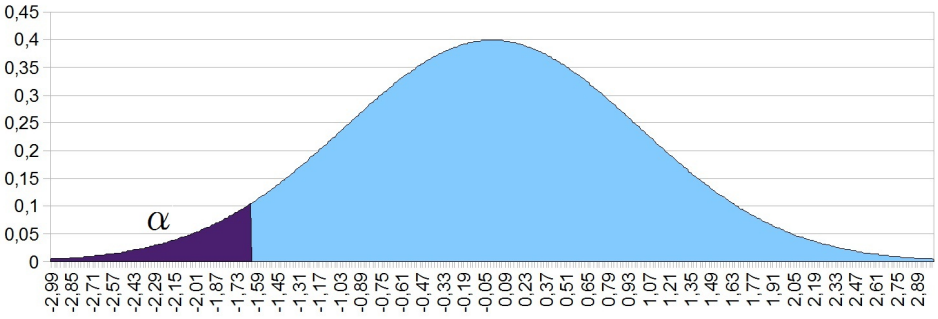


Rysunek 3.2. Obszar krytyczny dwustronny dla $\alpha = 5\%$

2. jednostronny:

— lewostronny:

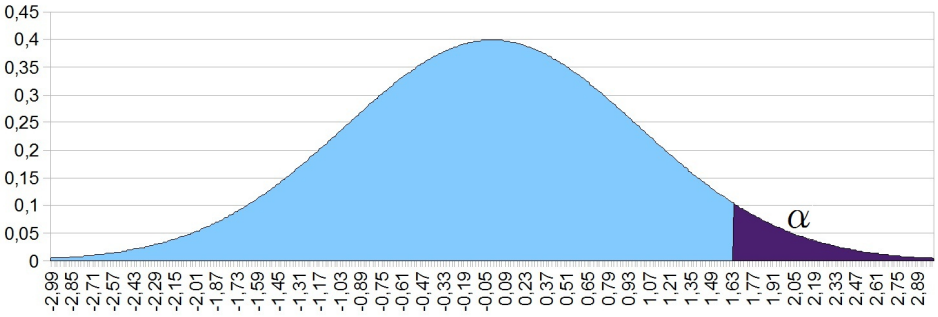
Hipoteza alternatywna: $H_1 : m < A$



Rysunek 3.3. Obszar krytyczny lewostronny dla $\alpha = 5\%$

— prawostronny:

Hipoteza alternatywna: $H_1 : m > A$



Rysunek 3.4. Obszar krytyczny prawostronny dla $\alpha = 5\%$

3.2.1. Testy zgodności rozkładu dla populacji jednowymiarowych

Test zgodności rozkładu χ^2

Test zgodności rozkładu χ^2 (test χ^2 Pearsona) jest testem pozwalającym zweryfikować hipotezę o tym, że rozkład badanej cechy należy do pewnej określonej rodziny rozkładów. Test ten jest przeznaczony do badania dużych prób.

Niech $F(x)$ oznacza dystrybuantę badanej cechy, a $\mathcal{F}_N$ rodzinę dystrybuant o określonej postaci dystrybuanty. Wówczas testem χ^2 Pearsona weryfikuje się następujące hipotezy:

$$H_0 : F(x) \in \mathcal{F}_N$$

$$H_1 : F(x) \notin \mathcal{F}_N.$$

Badaną próbę przedstawiamy w postaci szeregu rozdzielczego o k klasach. Statystykę testową obliczamy według wzoru:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

We wzorze zastosowano następujące oznaczenia:

— n – liczebność próby,

— n_i – liczebność i-tej klasy,

— $\hat{p}_i = \mathbf{F}(b_i) - \mathbf{F}(a_i)$ – teoretyczne prawdopodobieństwo odpowoadające wartościom badanej cechy w i-tej klasie.

Wartość krytyczna $\chi^2_{\alpha,k-1}$ statystyki χ^2 jest zależna od ustalonego poziomu istotności α i liczebystopni swobody n. Wartość $\chi^2_{\alpha,k-1}$ odczytujemy z tablic wartości krytycznych rozkładu χ^2. Obszar krytyczny w teście χ^2 jest lewostronny. Zatem gdy $\chi^2 \geqslant \chi^2_{\alpha,k-1}$ to odrzucamy hipotezę H_0 na rzecz hipotezy H_1, to znaczy, że rozkład badanej cechy nie należy do rodziny rozkładów $\mathcal{F}_N$ Jeśli $\chi^2 < \chi^2_{\alpha,k-1}$ to przyjmujemy, że nie ma podstaw do odrzucenia hipotezy H_0.

Test λ Kołmogorowa

Test λ Kołmogorowa jest testem, który pozwala zweryfikować hipotezę o tym, że rozkład badanej cechy jest rozkładem ciągłym o dystrybuancie $F(x)$ równej hipotetycznej dystrybuancie rozkładu cięgłego $F_0(x)$. Wówczas testem λ Kołmogorowa weryfikuje się następujące hipotezy:

$$H_0 : F(x) = F_0(x)$$

$$H_1 : F(x) \neq F_0(x).$$

Najpierw wyznaczamy dystrybuantę empiryczną $F_n(x)$ dla próby n-elementowej według wzoru:

— dla szeregu szczegółowego:

$$F_n(x) = \begin{cases} 0 & \text{dla } x \leqslant x_1 \\ \frac{i}{n} & \text{dla } x_i < x \leqslant x_{i+1} \\ 1 & \text{dla } x > x_k \end{cases}$$

— dla szeregu rozdzielczego:

$$F_n(x_j) = \frac{n_{sj}}{n}, \quad \text{gdzie:} n_{sj} = \sum_{i=1}^{j} n_i$$

Potem wyznaczamy dla każdego x_i wartości dystrybuanty teoretycznej $F(x)$ oraz wyznaczamy wartość statystyki $D = \sup_x |F_n(x) - F(x)|$.

Wartość statystyki testowej wyznaczamy za pomocą wzoru:

$$\lambda = D \cdot \sqrt{n}$$

Wartość krytyczna λ_α statystyki λ jest zależna jedynie od ustalonego poziomu istotności α. Wartość λ_α odczytujemy z tablic wartości granicznych testu λ Kołmogorowa. Obszar krytyczny w teście λ Kołmogorowa jest lewostronny. Zatem gdy $\lambda \leqslant \lambda_\alpha$ to nastepuje odrzucenie hipotezy H_0 na rzecz hipotezy H_1, w przeciwnym wypadku, gdy $\lambda > \lambda_\alpha$ to przyjmujemy, że nie ma podstaw do odrzucenia hipotezy H_0.

Test Shapiro-Wilka

Test Shapiro-Wilka pozwala zweryfikować hipotezę o tym, że rozkład badanej cechy jest rozkładem normalnym. Zatem jest testem normalności rozkładu. Jest on przeznaczony do badania prób o liczebności $n \leqslant 50$.

Niech $F(x)$ oznacza dystrybuantę badanej cechy, a $\mathcal{F}_N$ rodzinę dystrybuant rozkładu normalnego. Testem Shapiro-Wilka weryfikuje się następujące hipotezy:

$$H_0 : F(x) \in \mathcal{F}_N$$

$$H_1 : F(x) \notin \mathcal{F}_N.$$

Statystykę testową obliczamy za pomocą wzoru:

$$W = \frac{(\sum_{i=1}^{[\frac{n}{2}]} a_i(n)(x_{n-i+1} - x_i))^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

W powyższym wzorze zastosowano następujące oznaczenia:
— n – liczebność próby,
— $[\frac{n}{2}]$ – część całkowita z liczby $\frac{n}{2}$,
— x_i – wartości próby uporządkowane niemalejąco,
— $\bar{x}$ – średnia arytmetyczna z próby,
— $a_i(n)$ – stablicowane współczynniki testu odczytane dla liczebności próby n.

Wartość krytyczna $W_{\alpha,n}$ statystyki W jest zależna od ustalonego poziomu istotności α i liczebności próby n. Wartość $W_{\alpha,n}$ odczytujemy z tablic wartości krytycznych dla testu Shapiro-Wilka. Obszar krytyczny w teście Shapiro-Wilka jest lewostronny, zatem jeśli $W \leqslant W_{\alpha,n}$ to odrzucamy hipotezę H_0 na rzecz hipotezy H_1, to znaczy, że rozkład badanej cechy nie jest rozkładem normalnym. Natomiast jeżeli $W > W_{\alpha,n}$ to przyjmujemy, że nie ma podstaw do odrzucenia hipotezy H_0.

3.2.2. Porównywanie rozkładów dwóch prób niezależnych

Test Kołmogorowa-Smirnowa

Test Kołmogorowa-Smirnowa pozwala zweryfikować hipotezę o tym, że rozkłady dwóch baradnych cech populacji o ciągłych dystrybuantach $F_1(x)$ i $F_2(x)$ są takie same, czyli, że ich dystrybuanty są równe.

Testem Kołmogorowa-Smirnowa weryfikuje się następujące hipotezy:

$$H_0 : F_1(x) = F_2(x)$$

$$H_1 : F_1(x) \neq F_2(x)$$

Aby przeprowadzić test Kołmogorowa-Smirnowa losuje się dwie niezależne próby o liczebnościach n_1 i n_2. Niech x_{ij} oznacza j-tą wartość i-tej próby. Dla obu prób losowych tworzymy szeregi rozdzialcze o tych samych końcach x_k. Niech n_{ij} oznacza liczebność j-tej klasy dla i-tej próby. Następnie wyznaczamy dystrybuanty empiryczne według wzorów:

$$F_{n_1}(x_k) = \frac{n_{1sk}}{n_1}, \quad \text{gdzie: } n_{1sk} = \sum_{j=1}^{k} n_{1j}$$

$$F_{n_2}(x_k) = \frac{n_{2sk}}{n_2}, \quad \text{gdzie: } n_{2sk} = \sum_{j=1}^{k} n_{2j}$$

Potem wyznaczamy wartość statystyki $D*$ za pomocą wzoru:

$$D^* = \sup_x |F_{n_1}(x) - F_{n_2}(x)|$$

Statystykę testową obliczamy według wzoru:

$$\lambda = D^* \cdot \sqrt{n}$$

gdzie: $n = \frac{n_1 \cdot n_2}{n_1 + n_2}$.

Wartość krytyczna λ_α statystyki λ jest zależna od ustalonego poziomu istotności α. Wartość λ_α odczytujemy z tablic wartości granicznych testu λ Kołmogorowa. Obszar krytyczny w teście Kołmogorowa-Smirnowa jest lewostronny, zatem jeśli $\lambda \leqslant \lambda_\alpha$ to następuje odrzucenie hipotezy H_0 na rzecz hipotezy H_1, to znaczy, że rozkłady badanych cech są różne. Natomiast jeżeli $\lambda > \lambda_\alpha$ to przyjmujemy, że nie ma podstaw do odrzucenia hipotezy H_0, czyli nie ma podstaw do stwierdzenia, iż równość $F_1(x) = F_2(x)$ nie jest prawdziwa.

Test Walda-Wolfowitza

Test Walda-Wolfowitza daje możliwość badania dwóch populacji generalnych o rozkładach z ciągłymi dystrybuantami $F_1(x)$ i $F_2(x)$. Ten test pozwala badać hipotezę o równości tych dystrybuant.

Testem Walda-Wolfowitza weryfikuje się następujące hipotezy:

$$H_0 : F_1(x) = F_2(x)$$

$$H_1 : F_1(x) \neq F_2(x)$$

W celu przeprowadzenia testu Walda-Wolfowitza losuje się dwie niezależne próby o liczebnościach n_1 i n_2. Niech x_{ij} oznacza j-tą wartość i-tej próby. Wyniki obu prób ustawiamy w jeden niemalejący ciąg y_n. Następnie odczytujemy z ciagu y_n liczbę serii k, czyli podciągów złożonych z wartości z jednej próby.

Wartość krytyczna k_{α,n_1,n_2} statystyki k jest zależna od ustalonego poziomu istotności α oraz liczebności dwóch prób niezaleznych n_1 i n_2. Wartość k_{α,n_1,n_2} odczytujemy z tablic dla rozkładu serii. Obszar krytyczny w teście Walda-Wolfowitza jest lewostronny, zatem jeżeli $k > k_{\alpha,n_1,n_2}$ to przyjmujemy, że nie ma podstaw do odrzucenia hipotezy H_0. To znaczy, że na poziomie istotność α nie ma podstaw do stwirdzenia, iż rozkłady tych cech są różne. A jeśli $k \leqslant k_{\alpha,n_1,n_2}$ to następuje odrzucenie hipotezy H_0 na rzecz hipotezy H_1. Wówczas stwierdzamy, że rozkłady badanych cech są różne.

Test U Manna-Whitneya

Test U Manna-Whitneya weryfikuje hipotezę, że dwie próby pochodzą populacji generalnych o takich samych rozkładach. Zastosowanie testu U Manna-Whitneya nie wymaga spełnienia zbyt wielu założeń, co sprawia, że może być on szeroko stosowany. Test U Manna-Whitneya jest alternatywą dla testu t-Studenta dla prób niezależnych. Można go wykorzystywać, gdy dane nie spełniają założeń testu t-Studenta.

Testem U Manna-Whitneya weryfikuje się następujące hipotezy:

$$H_0 : F_1(x) = F_2(x)$$

$$H_1 : F_1(x) \neq F_2(x)$$

W celu przeprowadzenia testu U Manna-Whitneya losuje się dwie niezależne próby o liczebnościach n_1 i n_2. Niech x_{ij} oznacza j-tą wartość i-tej próby. Wyniki obu prób ustawiamy w jeden niemalejący ciąg y_n. W ten sposób każdej wartości x_{ij} przyporządkowany

zostaje indeks w ciągu y_n. Ten indeks nazywamy rangą. Nastepnie obliczamy wartości dwóch statystyk:

$$u_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1+1)}{2} - R_1 \qquad u_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2+1)}{2} - R_2$$

gdzie:

— $R_1 = \sum_{\{n:y_n=x_{1j} \wedge n=1,2,\ldots,n_1+n_2\}} n$ – jest sumą rang w pierwszej próbie,

— $R_2 = \sum_{\{n:y_n=x_{2j} \wedge n=1,2,\ldots,n_1+n_2\}} n$ – jest sumą rang w drugiej próbie,

Jeśli $n_1 \leqslant 20 \vee n_2 \leqslant 20$ statystka testowa jest równa mniejszej z wartości u_1 i u_2:

$$U = \min\{u_1, u_2\}$$

Wartość krytyczna U_{α,n_1,n_2} statystyki U jest zależna od ustalonego poziomu istotności α oraz liczebności dwóch prób niezaleznych n_1 i n_2. Wartość U_{α,n_1,n_2} odczytujemy z tablic wartości krytycznych testu U Manna-Whitneya. Obszar krytyczny w teście U Manna-Whitneya jest lewostronny, zatem jeżeli $U > U_{\alpha,n_1,n_2}$ to uznajemy, że nie ma podstaw do odrzucenia hipotezy H_0. A jeśli $U \leqslant U_{\alpha,n_1,n_2}$ to odrzucamy H_0 na rzecz hipotezy H_1.

A jeżeli $n_1, n_2 > 20$ obliczamy wartości statystki testowej węgług wzoru:

$$U = \frac{\min\{u_1, u_2\} - \frac{n_1 \cdot n_2}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1+n_2+1)}{12}}}$$

Przy dużych próbach ($n_1, n_2 > 20$) statystyka U, przy założeniu prawdziwosci hipotezy H_0 ma rozkład $N(0,1)$. Zatem wartość krytyczną u_α statystyki U odczytujemy z tablic wartości krytycznych rozkładu normalnego standardowego dla ustalonego poziomu istotności α, tak aby spełniona była równość:

$$P\{U \leqslant u_\alpha\} = \alpha$$

Obszar krytyczny w teście Walda-Wolfowitza jest lewostronny, zatem jeżeli $U > u_\alpha$ to na poziomie istotność α przyjmujemy hipotezę zerową. A jeśli $U \leqslant u_\alpha$ to odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej.

3.2.3. Porównywanie rozkładów dwóch prób zależnych

Test znaków

Test znaków pozwala badać pary obserwacji należących do dwóch n-elementowych prób zależnych pochodzących z populacji generalnych o ciągłych dystrybuantach $F_1(x)$ i$F_2(x)$. Niech x_{ij} oznacza j-tą wartość i-tej próby, gdzie $i = 1, 2$, a $j = 1, 2, \ldots, n$. Ten test jest najczęściej wykorzystywany:

— przy badaniu zmian cechy populacji generalnej w czasie,

— do porownywania wartości cechy populacji generalnej przed i po wykonaniu eksperymenu.

Test znaków nie jest testem precyzyjnym, ponieważ jedyną wykorzystywaną informacją pochodzącą z prób jest znak różnicy pomiedzy odpowieadajacymi sobie wynikami w tych dwóch próbach.

Hipotezy weryfikowane testem znaków są następujące:

$$H_0 : F_1(x) = F_2(x)$$

$$H_1 : F_1(x) \neq F_2(x)$$

Aby zweryfikować hipotezę badamy znak różnicy $x_{1j} - x_{1j}$ dla każdej pary wyników. Następnie zliczamy znaki ujemne r_- oraz znaki dodatnie r_+.

$$r_- = \sum_{\{j:x_{1j}-x_{1j}<0\}} 1 \qquad\qquad r_+ = \sum_{\{j:x_{1j}-x_{1j}>0\}} 1$$

Jeśeli para wyników ma tą samą wartość ($x_{1j} = x_{1j}$), to nie bierzemy jej pod uwagę.

Statystka testowa jest równa mniejszej z wartości r_- i r_+:

$$R = \min\{r_-, r_+\}$$

Wartość krytyczna $R_{\alpha,n}$ statystyki R jest zależna od ustalonego poziomu istotności α oraz liczebności prób zaleznych n. Wartość $R_{\alpha,n}$ odczytujemy z tablicy rozkładu liczby znaków. Obszar krytyczny w teście znaków jest lewostronny, więc jeżeli $R \leqslant R_{\alpha,n}$ to odrzucamy hipotezę H_0. A jeśli $R > R_{\alpha,n}$ to na poziomie ufności α przyjmujemy hipotezę zerową.

Test Wilcoxoma

Test Wilcoxoma jest także naztwany testem rangowych znaków. Umożliwia on badanie obserwacji należących do dwóch n-elementowych prób zależnych pochodzących z populacji generalnych o ciągłych dystrybuantach $F_1(x)$ i$F_2(x)$. Niech x_{ij} oznacza j-tą wartość i-tej próby, gdzie $i = 1, 2$, a $j = 1, 2, \ldots, n$. Wyniki prób odpowiadają sobie parami $(x_{1j}\tilde{x}_{2j})$. Ten test stosujemy gdy mamy male próby, a badana cecha nie ma rozkladu normalnego. Test Wilcoxoma jest testem precyzyjniejszym niż test znaków, ponieważ wykorzystuje nie tylko informację o znaku różnicy, ale także informację o jej bezwzględnej wartości.

Hipotezy weryfikowane testem znaków są następujące:

$$H_0 : F_1(x) = F_2(x)$$

$$H_1 : F_1(x) \neq F_2(x)$$

Aby zweryfikować hipotezę badamy znak różnicy $x_{1j} - x_{1j}$ dla każdej pary wyników. Następnie wartości bezwzgledne różnic $|x_{1j} - x_{1j}|$ ustawiamy w jeden niemalejący ciąg y_n. W ten sposób dla każdego j przyporządkowana zostaje ranga r_j, która jest równa wartości indeksu w ciągu y_n. Następnie sumujemy rangi dla znaków ujemnych t_- oraz rangi dla znaków dodatnich t_+.

$$t_- = \sum_{\{j:x_{1j}-x_{1j}<0\}} r_j \qquad\qquad t_+ = \sum_{\{j:x_{1j}-x_{1j}>0\}} r_j$$

Wartość statystyki testowej jest równa mniejszej z wartości t_- i t_+:

$$T = \min\{t_-, t_+\}$$

Wartość krytyczną $T_{\alpha,n}$ statystyki T odczytujemy z tablicy wartości krytycznych dla testu Wilcoxoma dla ustalonego poziomu istotności α oraz liczebności prób zaleznych n. Obszar krytyczny w teście Wilcoxoma jest lewostronny, zatem gdy $T \leqslant T_{\alpha,n}$ to podejmujemy decyzję o odrzuceniu hipotezy H_0. A gdy $T > T_{\alpha,n}$ to przyjmujemy hipotezę zerową na poziomie ufności α.

Ćwiczenia nr 7

Zadanie 3.11. [6]

W pewnym teście psychologicznym przeprowadzonym na 50 wylosowanych niezależnie dzieciach otrzymano następujący rozkład wyników liczby zapamiętanych przez dzieci elementów:

Liczba zapamietanych elementów	Liczba dzieci
10 – 15	2
15 – 20	6
20 – 25	8
25 – 30	12
30 – 35	7
35 – 40	6
40 – 45	4
45 – 50	3
50 – 55	2

Na poziomie istotności 0,05 testem λ Kołmogorowa zweryfikować hipotezę, że ilości zapamiętanych przez dzieci elementów w teście mają rozkład normalny.

Zadanie 3.12. [6]

W pewnym eksperymencie zbadano w wylosowanej grupie 42 chorych na pewną chorobę psychiatryczną procenty czasu snu w pewnej fazie. Otrzymano następujące wyniki (w %):
34,8 33,9 32,6 49,4 55,2 48,5 40,3 34,0 42,1 17,9 36,0 21,2 35,9 41,2 40,9 16,9 42,9 28,7 51,9 24,1 29,1 44,6 41,2 17,0 29,8 35,0 51,7 42,9 54,2 25,9 30,3 36,9 19,2 59,1 31,3 50,0 19,8 30,6 31,7 28,8 30,0

Zadanie 3.13. [2]

Pobrano próbę dotyczącą cechy mierzalnej X o liczebności $n = 19$; wyniki uporządkowano według wielkości: 12,4 14,2 14,9 15,6 16,1 16,8 17,3 17,9 18,2 18,6 19,3 19,7 20,4 21,9 22,8 23,7 25,2 25,9 27,4. Na poziomie istotności $\alpha = 0,10$ zweryfikować testem Shapiro-Wilka hipotezę o normalności rozkładu badanej cechy X w populacji generalnej.

Zadanie 3.14. [2]

Wynikami pięcioelementowej próby są: 0,18 0,56 0,87 1,37 2,46. Na poziomie istotności $\alpha = 0,05$ testem λ Kołmogorowa zweryfikować hipotezę, że próbka została pobrana z populacji, w ktorej dystrybuantą badanej cechy X jest $F(x) = 1 - e^{-x}$ dla $x > 0$, tzn. że rozkładem jest rozkład wykładniczy z parametrem $\lambda = 1$.

Zadanie 3.15. [2]

Z dwóch różnych partii wlukien bawełny wylosowano próbki i zmierzono długości wylosowanych włukien, otrzymując w milimetrach:
próbka 1: 18 22 7 12 25 19 17 10 18 23 9 7 24 30 23 22 17 8 19

próbka 2: 8 14 217 19 23 4 19 15 13 28 18 24 33 28 28 17 19 20 18
Testem Smirnowa-Kołmogorowa na poziomie istotności $\alpha = 0,05$ zweryfikować hipotezę, że rozkłady długości tych partiiwłokien bawełny są identyczne, wobec hipotezy alternatywnej, że rozkłady te różnią się między sobą.

Zadanie 3.16.

Napisz funkcję w Matlabie, która będzie wyznaczala wartość statystyki dla wybranego testu nieparametrycznego.

3.3. Testy parametryczne

Testy parametryczne są testami pozwalającymi zweryfikować hipotezy mówiące o wartościach badanych parametrów populacji. Najczęściej wyznaczanymi parametrami populacji generalnej są: wartość średnia, wariancja i wskaźnik struktury.

3.3.1. Testy parametryczne służące do weryfikacji własności populacji jednowymiarowych

Test istotności dla wartości średniej

W tych testach weryfikujemy hipotezy o wartości średniej populacji. Wyróżniamy trzy podstawowe modele, dla których buduje się testy istotności dla średniej.

Model 1

Założenia:

— populacja generalna ma rozkład $N(\mu, \sigma)$,

— σ – znane.

	test dwustronny	test lewostronny	test prawostronny
hipotezy	$H_0: \ m = \mu_0$ $H_1: \ m \neq \mu_0$	$H_0: \ m = \mu_0$ $H_1: \ m < \mu_0$	$H_0: \ m = \mu_0$ $H_1: \ m > \mu_0$
statystyka testowa	$u = \frac{\bar{x}-\mu_0}{\sigma}\sqrt{n}$, gdzie: $\bar{x}$ –średnia arytmetyczna z próby Statystyka u przy założeniu prawdziwości H_0 ma rozkład normalny $N(0,1)$.		
wartość krytyczna	Z tablic wartości krytycznych rozkładu normalnego odczytujemy wartości krytyczne u_α.		
	$P\{\|U\| \geqslant u_\alpha\} = \alpha$	$P\{U \leqslant u_\alpha\} = \alpha$	$P\{U \geqslant u_\alpha\} = \alpha$
decyzja	jeżeli:		
	$\|u\| \geqslant u_\alpha$	$u \leqslant u_\alpha$	$u \geqslant u_\alpha$
	odrzucamy hipotezę zerową H_0 na korzyść hipotezy alternatywnej H_1		
	jeżeli:		
	$\|u\| < u_\alpha$	$u > u_\alpha$	$u < u_\alpha$
	nie ma podstaw do odrzucenia hipotezy zerowej H_0 na poziomie istotności α		

Model 2

Założenia:

— populacja generalna ma rozkład $N(\mu, \sigma)$,

— σ – nieznane,
— mała próba o liczebności $n \leqslant 30$.

<table>
<tr><td></td><td>test dwustronny</td><td>test lewostronny</td><td>test prawostronny</td></tr>
<tr><td>hipotezy</td><td>$H_0: \quad m = \mu_0$
$H_1: \quad m \neq \mu_0$</td><td>$H_0: \quad m = \mu_0$
$H_1: \quad m < \mu_0$</td><td>$H_0: \quad m = \mu_0$
$H_1: \quad m > \mu_0$</td></tr>
<tr><td>statystyka

testowa</td><td colspan="3">$t = \frac{\bar{x}-\mu_0}{S}\sqrt{n-1}$, gdzie:
$\bar{x}$ – średnia arytmetyczna z próby, $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
Statystyka t przy założeniu prawdziwości H_0
ma rozkład t Studenta o $n-1$ stopniach swobody.</td></tr>
<tr><td>wartość
krytyczna</td><td colspan="3">Z tablicy wartości krytycznych rozkładu t-Studenta odczytujemy
wartości krytyczne $t_{\alpha,n-1}$.</td></tr>
<tr><td></td><td>$P\{|t| \geqslant t_{\alpha,n-1}\} = \alpha$</td><td>$P\{t \leqslant t_{\alpha,n-1}\} = \alpha$</td><td>$P\{t \geqslant t_{\alpha,n-1}\} = \alpha$</td></tr>
<tr><td>decyzja</td><td colspan="3">jeżeli:</td></tr>
<tr><td></td><td>$|t| \geqslant t_{\alpha,n-1}$</td><td>$t \leqslant t_{\alpha,n-1}$</td><td>$t \geqslant t_{\alpha,n-1}$</td></tr>
<tr><td></td><td colspan="3">odrzucamy hipotezę zerową H_0
na korzyść hipotezy alternatywnej H_1</td></tr>
<tr><td></td><td colspan="3">jeżeli:</td></tr>
<tr><td></td><td>$|t| < t_{\alpha,n-1}$</td><td>$t > t_{\alpha,n-1}$</td><td>$t < t_{\alpha,n-1}$</td></tr>
<tr><td></td><td colspan="3">nie ma podstaw do odrzucenia hipotezy zerowej H_0
na poziomie istotności α</td></tr>
</table>

Model 3

Założenia:

— populacja generalna ma rozkład $N(\mu, \sigma)$,

— σ – nieznane,

— duża próba o liczebności $n > 30$.

	test dwustronny	test lewostronny	test prawostronny
hipotezy	$H_0: \quad m = \mu_0$ $H_1: \quad m \neq \mu_0$	$H_0: \quad m = \mu_0$ $H_1: \quad m < \mu_0$	$H_0: \quad m = \mu_0$ $H_1: \quad m > \mu_0$
statystyka testowa	$U = \frac{\bar{x}-\mu_0}{S}\sqrt{n}$, gdzie: $\bar{x}$ – średnia arytmetyczna z próby, $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ Statystyka u przy założeniu prawdziwości H_0 ma rozkład normalny $N(0,1)$.		
wartość krytyczna	Z tablic wartości krytycznych rozkładu normalnego odczytujemy wartości krytyczne u_α.		
	$P\{\lvert U\rvert \geqslant u_\alpha\} = \alpha$	$P\{U \leqslant u_\alpha\} = \alpha$	$P\{U \geqslant u_\alpha\} = \alpha$
decyzja	jeżeli:		
	$\lvert u\rvert \geqslant u_\alpha$	$u \leqslant u_\alpha$	$u \geqslant u_\alpha$
	odrzucamy hipotezę zerową H_0 na korzyść hipotezy alternatywnej H_1		
	jeżeli:		
	$\lvert u\rvert < u_\alpha$	$u > u_\alpha$	$u < u_\alpha$
	nie ma podstaw do odrzucenia hipotezy zerowej H_0 na poziomie istotności α		

Test istotności dla wariancji

W tych testach weryfikujemy hipotezy o wartości wariancji. Wyróżniamy dwa podstawowe modele, dla których buduje się testy istotności dla wariancji.

Model 1

Założenia:

— populacja generalna ma rozkład $N(\mu, \sigma)$,

— μ, σ – nieznane,

— mała próba o liczebności $n \leqslant 30$.

	test dwustronny	test lewostronny	test prawostronny
hipotezy	$H_0: \quad \sigma^2 = \sigma_0^2$ $H_1: \quad \sigma^2 \neq \sigma_0^2$	$H_0: \quad \sigma^2 = \sigma_0^2$ $H_1: \quad \sigma^2 < \sigma_0^2$	$H_0: \quad \sigma^2 = \sigma_0^2$ $H_1: \quad \sigma^2 > \sigma_0^2$
statystyka testowa	$\chi^2 = \frac{n \cdot S^2}{\sigma_0^2}$, gdzie: $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ Statystyka χ^2 przy założeniu prawdziwości H_0 ma rozkład χ^2 o $n-1$ stopniach swobody.		
wartość krytyczna	Z tablic wartości krytycznych rozkładu χ^2 odczytujemy wartości krytyczne $\chi^2_{\alpha,n-1}$.		
	$P\{\chi^2 \geqslant \chi^2_{\alpha,n-1,1}\} = \frac{\alpha}{2}$ $P\{\chi^2 \leqslant \chi^2_{\alpha,n-1,2}\} = \frac{\alpha}{2}$	$P\{\chi^2 \leqslant \chi^2_{\alpha,n-1}\} = \alpha$	$P\{\chi^2 \geqslant \chi^2_{\alpha,n-1}\} = \alpha$
decyzja	jeżeli:		
	$\chi^2 \leqslant \chi^2_{\alpha,n-1,1}$ $\vee\ \chi^2 \geqslant \chi^2_{\alpha,n-1,2}$	$\chi^2 \leqslant \chi^2_{\alpha,n-1}$	$\chi^2 \geqslant \chi^2_{\alpha,n-1}$
	odrzucamy hipotezę zerową H_0 na korzyść hipotezy alternatywnej H_1		
	jeżeli:		
	$\chi^2 > \chi^2_{\alpha,n-1,1}$ $\wedge\ \chi^2 < \chi^2_{\alpha,n-1,2}$	$\chi^2 > \chi^2_{\alpha,n-1}$	$\chi^2 < \chi^2_{\alpha,n-1}$
	nie ma podstaw do odrzucenia hipotezy zerowej H_0 na poziomie istotności α		

Model 2

Założenia:

— populacja generalna ma rozkład $N(\mu, \sigma)$,

— μ, σ – nieznane,

— duża próba o liczebności $n > 30$.

	test dwustronny	test lewostronny	test prawostronny
hipotezy	$H_0: \ \sigma^2 = \sigma_0^2$ $H_1: \ \sigma^2 \neq \sigma_0^2$	$H_0: \ \sigma^2 = \sigma_0^2$ $H_1: \ \sigma^2 < \sigma_0^2$	$H_0: \ \sigma^2 = \sigma_0^2$ $H_1: \ \sigma^2 > \sigma_0^2$
statystyka testowa	$U = \sqrt{\frac{2nS^2}{\sigma_0^2}} - \sqrt{2n-3}$, gdzie: $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ Statystyka U przy założeniu prawdziwości H_0 ma rozkład normalny $N(0,1)$.		
wartość krytyczna	Z tablic wartości krytycznych rozkładu normalnego odczytujemy wartości krytyczne u_α.		
	$P\{\lvert U\rvert \geqslant u_\alpha\} = \alpha$	$P\{U \leqslant u_\alpha\} = \alpha$	$P\{U \geqslant u_\alpha\} = \alpha$
decyzja	jeżeli:		
	$\lvert u\rvert \geqslant u_\alpha$	$u \leqslant u_\alpha$	$u \geqslant u_\alpha$
	odrzucamy hipotezę zerową H_0 na korzyść hipotezy alternatywnej H_1		
	jeżeli:		
	$\lvert u\rvert < u_\alpha$	$u > u_\alpha$	$u < u_\alpha$
	nie ma podstaw do odrzucenia hipotezy zerowej H_0 na poziomie istotności α		

3.3.2. Testy parametryczne służące do porównania własności dwóch populacji

Test istotności dla dwóch wartości średnich

Model 1 (Test U)

Założenia:

— populacje generalne mają rozkłady $N(\mu_1, \sigma_1)$, $N(\mu_2, \sigma_2)$,

— σ_1, σ_2 – znane,

	test dwustronny	test lewostronny	test prawostronny
hipotezy	$H_0: \ \mu_1 = \mu_2$ $H_1: \ \mu_1 \neq \mu_2$	$H_0: \ \mu_1 = \mu_2$ $H_1: \ \mu_1 < \mu_2$	$H_0: \ \mu_1 = \mu_2$ $H_1: \ \mu_1 > \mu_2$
statystyka testowa	$U = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ gdzie: $\bar{x}_1$, $\bar{x}_2$ – średnie arytmetyczne z prób Statystyka U przy założeniu prawdziwości H_0 ma rozkład normalny $N(0,1)$.		
wartość krytyczna	Z tablic wartości krytycznych rozkładu normalnego odczytujemy wartości krytyczne u_α.		
	$P\{\lvert U\rvert \geqslant u_\alpha\} = \alpha$	$P\{U \leqslant u_\alpha\} = \alpha$	$P\{U \geqslant u_\alpha\} = \alpha$
decyzja	jeżeli:		
	$\lvert u\rvert \geqslant u_\alpha$	$u \leqslant u_\alpha$	$u \geqslant u_\alpha$
	odrzucamy hipotezę zerową H_0 na korzyść hipotezy alternatywnej H_1		
	jeżeli:		
	$\lvert u\rvert < u_\alpha$	$u > u_\alpha$	$u < u_\alpha$
	nie ma podstaw do odrzucenia hipotezy zerowej H_0 na poziomie istotności α		

Model 2 (Test t Stusenta)

Założenia:

— populacje generalne mają rozkłady $N(\mu_1, \sigma_1)$, $N(\mu_2, \sigma_2)$,

— $\sigma_1 = \sigma_2$ – nieznane,

— małe próby o liczebnościach $n_1, n_2 \leqslant 30$.

Model 3 (Test c Cochrana-Coxa)

Założenia:

— populacje generalne mają rozkłady $N(\mu_1, \sigma_1)$, $N(\mu_2, \sigma_2)$,

— $\sigma_1 \neq \sigma_2$ – nieznane,

— małe próby o liczebnościach $n_1, n_2 \leqslant 30$.

	test dwustronny	test lewostronny	test prawostronny
hipotezy	$H_0: \quad \mu_1 = \mu_2$ $H_1: \quad \mu_1 \neq \mu_2$	$H_0: \quad \mu_1 = \mu_2$ $H_1: \quad \mu_1 < \mu_2$	$H_0: \quad \mu_1 = \mu_2$ $H_1: \quad \mu_1 > \mu_2$
statystyka testowa	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ gdzie: $\bar{x}_1$, $\bar{x}_2$ – średnie arytmetyczne z prób $S_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2$, $S_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(x_{2i}-\bar{x}_2)^2$ Statystyka t przy założeniu prawdziwości H_0 ma rozkład t Studenta o n_1+n_2-2 stopniach swobody.		
wartość krytyczna	Z tabeli wartości krytycznych rozkładu t Studenta odczytujemy wartości krytyczne t_{α,n_1+n_2-2}.		
	$P\{\lvert t\rvert \geqslant t_{\alpha,n_1+n_2-2}\} = \alpha$	$P\{t \leqslant t_{\alpha,n_1+n_2-2}\} = \alpha$	$P\{t \geqslant t_{\alpha,n_1+n_2-2}\} = \alpha$
decyzja	jeżeli:		
	$\lvert t\rvert \geqslant t_{\alpha,n_1+n_2-2}$	$t \leqslant t_{\alpha,n_1+n_2-2}$	$t \geqslant t_{\alpha,n_1+n_2-2}$
	odrzucamy hipotezę zerową H_0 na korzyść hipotezy alternatywnej H_1		
	jeżeli:		
	$\lvert t\rvert < t_{\alpha,n_1+n_2-2}$	$t > t_{\alpha,n_1+n_2-2}$	$t < t_{\alpha,n_1+n_2-2}$
	nie ma podstaw do odrzucenia hipotezy zerowej H_0 na poziomie istotności α		

	test dwustronny	test lewostronny	test prawostronny
hipotezy	$H_0: \quad \mu_1 = \mu_2$ $H_1: \quad \mu_1 \neq \mu_2$	$H_0: \quad \mu_1 = \mu_2$ $H_1: \quad \mu_1 < \mu_2$	$H_0: \quad \mu_1 = \mu_2$ $H_1: \quad \mu_1 > \mu_2$
statystyka testowa	$c = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{z_1+z_2}}$ gdzie: $\bar{x}_1$, $\bar{x}_2$ – średnie arytmetyczne z prób, $z_1 = \frac{S_1^2}{n_1-1}$, $z_2 = \frac{S_2^2}{n_2-1}$, $S_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2$, $S_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(x_{2i}-\bar{x}_2)^2$		
wartość krytyczna	Z tablicy wartości krytycznych rozkładu t-Studenta odczytujemy wartości $t_{1\alpha,n_1-1}$, $t_{2\alpha,n_2-1}$. Wartość krytyczną wyznaczamy ze wzoru: $c_\alpha = \frac{t_1 z_1 + t_2 z_2}{z_1+z_2}$		
decyzja	jeżeli:		
	$\lvert c\rvert \geqslant c_\alpha$	$c \leqslant c_\alpha$	$c \geqslant c_\alpha$
	odrzucamy hipotezę zerową H_0 na korzyść hipotezy alternatywnej H_1		
	jeżeli:		
	$\lvert c\rvert < c_\alpha$	$c > c_\alpha$	$c < c_\alpha$
	nie ma podstaw do odrzucenia hipotezy zerowej H_0 na poziomie istotności α		

Model 4 (Test U)

Założenia:

— populacje generalne mają rozkłady $N(\mu_1, \sigma_1)$, $N(\mu_2, \sigma_2)$,

— σ_1, σ_2 – nieznane,

— duże próby o liczebnościach $n_1, n_2 > 30$.

	test dwustronny	test lewostronny	test prawostronny
hipotezy	$H_0: \quad \mu_1 = \mu_2$ $H_1: \quad \mu_1 \neq \mu_2$	$H_0: \quad \mu_1 = \mu_2$ $H_1: \quad \mu_1 < \mu_2$	$H_0: \quad \mu_1 = \mu_2$ $H_1: \quad \mu_1 > \mu_2$
statystyka testowa	$u = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ gdzie: $\bar{x}_1$, $\bar{x}_2$ – średnie arytmetyczne z prób, $S_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2$, $S_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2$ Statystyka u przy założeniu prawdziwości H_0 ma rozkład normalny $N(0,1)$.		
wartość krytyczna	Z tablic wartości krytycznych rozkładu normalnego odczytujemy wartości krytyczne u_α.		
	$P\{\|U\| \geqslant u_\alpha\} = \alpha$	$P\{U \leqslant u_\alpha\} = \alpha$	$P\{U \geqslant u_\alpha\} = \alpha$
decyzja	jeżeli:		
	$\|u\| \geqslant u_\alpha$	$u \leqslant u_\alpha$	$u \geqslant u_\alpha$
	odrzucamy hipotezę zerową H_0 na korzyść hipotezy alternatywnej H_1		
	jeżeli:		
	$\|u\| < u_\alpha$	$u > u_\alpha$	$u < u_\alpha$
	nie ma podstaw do odrzucenia hipotezy zerowej H_0 na poziomie istotności α		

Test istotności dla dwóch wariancji

Model

Założenia:

— populacje generalne mają rozkłady $N(\mu_1, \sigma_1)$, $N(\mu_2, \sigma_2)$,

— μ_1, μ_2, σ_1, σ_2 – nieznane.

3.4. Testy bootstrapowe

Metody bootstrapowe, którym poświęcony jest punkt 3.1.4, mają również zastosowanie przy weryfikacji hipotez statystycznych. Wykorzystuje się je do aproksymacji statystyk testowych.

Poniżej zostaną przedstwione procedury pozwalające weryfikować:

— hipotezę o wartości średniej nieznanego rozkładu zmiennej losowej,

— hipotezę o równości dwóch średnich.

Te procedury nie są jedynymi możliwościami wykorzystania metod bootstrapowych przy weryfikacji hipotez statystycznych.

3.4.1. Testy bootstrapowe dla wartości średniej

W tych testach weryfikujemy następującą hipotezę o wartości średniej populacji:

$$H_0: \quad \mu = \mu_0$$

Wyróżniamy dwa modele, dla których bedziemy budować testy bootstrapowe dla wartości średniej:

— Model 1 – populacja generalna ma nieznany rozkład oraz σ jest znane,

— Model 2 – populacja generalna ma nieznany rozkład oraz σ jest nieznane.

<table>
<tr><td></td><td>test dwustronny</td><td>test lewostronny</td><td>test prawostronny</td></tr>
<tr><td>hipotezy</td><td>$H_0: \quad \sigma_1^2 = \sigma_2^2$
$H_1: \quad \sigma_1^2 \neq \sigma_2^2$</td><td>$H_0: \quad \sigma_1^2 = \sigma_2^2$
$H_1: \quad \sigma_1^2 < \sigma_2^2$</td><td>$H_0: \quad \sigma^2 = \sigma_2^2$
$H_1: \quad \sigma_1^2 > \sigma_2^2$</td></tr>
<tr><td>statystyka testowa</td><td colspan="3">$F = \frac{\frac{n_1}{n_1-1}S_1^2}{\frac{n_2}{n_2-1}S_2^2}$
gdzie: $S_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2$, $S_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2$.
Statystyka F przy założeniu prawdziwości H_0
ma rozkład F Snedecora o $n_1 - 1$ i $n_2 - 1$ stopniach swobody.</td></tr>
<tr><td>wartość krytyczna</td><td colspan="3">Z tablicy wartości krytycznych rozkładu F Snedecora odczytujemy wartości krytyczne F_α.</td></tr>
<tr><td></td><td>$P\{F \geqslant F_{\alpha,1}\} = \frac{\alpha}{2}$
$P\{F \geqslant F_{\alpha,2}\} = \frac{\alpha}{2}$</td><td>$P\{F \leqslant F_\alpha\} = \alpha$</td><td>$P\{F \geqslant F_\alpha\} = \alpha$</td></tr>
<tr><td>decyzja</td><td colspan="3">jeżeli:</td></tr>
<tr><td></td><td>$F \leqslant F_\alpha$
$\vee \; F \geqslant F_{\alpha,n-1,2}$</td><td>$F \leqslant F_\alpha$</td><td>$F \geqslant F_\alpha$</td></tr>
<tr><td></td><td colspan="3">odrzucamy hipotezę zerową H_0
na korzyść hipotezy alternatywnej H_1</td></tr>
<tr><td></td><td colspan="3">jeżeli:</td></tr>
<tr><td></td><td>$F > F_{\alpha,1}$
$\wedge \; F < F_{\alpha,2}$</td><td>$F > F_\alpha$</td><td>$F < F_\alpha$</td></tr>
<tr><td></td><td colspan="3">nie ma podstaw do odrzucenia hipotezy zerowej H_0
na poziomie istotności α</td></tr>
</table>

Model 1

Załużmy, że populacja badana jest ze względu na cehę X o nieznanej średniej μ i znanym odchyleniu standardowym σ. Z tej populacji wylosowano próbę niezależną $X_1, X_2, \ldots, X_n$, której realizacją jest ciąg $x_1, x_2, \ldots, x_n$. Naszym zadaniem jest weryfikacja na poziomie istotności α jednego z zestawów hipotez:

$$H_0: \quad \mu = \mu_0 \qquad H_0: \quad \mu = \mu_0 \qquad H_0: \quad \mu = \mu_0$$

$$H_1: \quad \mu \neq \mu_0 \qquad H_1: \quad \mu < \mu_0 \qquad H_1: \quad \mu > \mu_0$$

Statytstyką testową jest:

$$U = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n}$$

gdzie: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Aby zweryfikować hipotezę zerową metodą bootstrapową należy wyznaczyć rozkład bootstrapowy z próby o wartościach $x_1, x_2, \ldots, x_n$. Następnie generujemy N ($N > 1000$) prób bootstrapowych $x_{1k}^*, x_{2k}^*, \ldots, x_{nk}^*$, gdzie $k = 1, 2, \ldots, N$ oznacza numer próby bootstrapowej. Dla każdej próby bootstrapowej wyznaczamy:

$$\bar{x}_k^* = \frac{1}{n}\sum_{i=1}^{n} x_{ik}^* \qquad \text{oraz} \qquad u_k^* = \frac{\bar{x}_k^* - \mu_0}{\sigma}\sqrt{n}$$

Zatem uzyskujemy ciąg wartości $u_1^*, u_2^*, \ldots, u_N^*$. Wprowadźmy następujące oznaczenia:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{oraz} \qquad u_0 = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n}$$

Dla testu dwustronnego ($H_0:\ \mu = \mu_0$ i $H_1:\ \mu \neq \mu_0$) obliczamy wartość:

$$\alpha_u^* = \frac{\overline{\{u \in \{u_1^*, u_2^*, \ldots, u_N^*\} : |u| \geqslant |u_0|\}}}{N}$$

gdzie: $\overline{\{u \in \{u_1^*, u_2^*, \ldots, u_N^*\} : |u| \geqslant |u_0|\}}$ – moc zbioru $\{u \in \{u_1^*, u_2^*, \ldots, u_N^*\} : |u| \geqslant |u_0|\}$. Wartość α_u^* porównujemy z poziomem istotności α. Jeżeli $\alpha_u^* > \alpha$, to nie mamy podstaw do odrzucenia hipotezy zerowej. A jeżeli $\alpha_u^* \leqslant \alpha$, to odrzucamy hipotezę H_0 na rzecz hipotezy H_1.

Natomiast dla testu lewostronnego ($H_0:\ \mu = \mu_0$ i $H_1:\ \mu < \mu_0$) wyznaczamy wartość:

$$\alpha_{u^-}^* = \frac{\overline{\{u \in \{u_1^*, u_2^*, \ldots, u_N^*\} : u \leqslant u_0\}}}{N}$$

gdzie: $\overline{\{u \in \{u_1^*, u_2^*, \ldots, u_N^*\} : u \leqslant u_0\}}$ – moc zbioru $\{u \in \{u_1^*, u_2^*, \ldots, u_N^*\} : u \leqslant u_0\}$. Wartość $\alpha_{u^-}^*$ porównujemy z poziomem istotności α. Jeżeli $\alpha_{u^-}^* > \alpha$, to nie mamy podstaw do odrzucenia hipotezy H_0. A jeżeli $\alpha_{u^-}^* \leqslant \alpha$, to odrzucamy hipotezę i przyjmujemy hipotezę alternatywną.

Z kolei dla testu prawostronnego ($H_0:\ \mu = \mu_0$ i $H_1:\ \mu > \mu_0$) obliczamy wartość:

$$\alpha_{u^+}^* = \frac{\overline{\{u \in \{u_1^*, u_2^*, \ldots, u_N^*\} : u \geqslant u_0\}}}{N}$$

gdzie: $\overline{\{u \in \{u_1^*, u_2^*, \ldots, u_N^*\} : u \geqslant u_0\}}$ – moc zbioru $\{u \in \{u_1^*, u_2^*, \ldots, u_N^*\} : u \geqslant u_0\}$. Wartość $\alpha_{u^+}^*$ porównujemy z poziomem istotności α. Jeżeli $\alpha_{u^+}^* > \alpha$, to nie mamy podstaw do odrzucenia hipotezy zerowej. A jeżeli $\alpha_{u^+}^* \leqslant \alpha$, to odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej.

Model 2

W modelu 2 zakładamy, że populacja badana jest ze względu na cehę X o nieznanej średniej μ i nieznanym odchyleniu standardowym σ oraz, że z populacji generalnej wylosowano próbę niezależną $X_1, X_2, \ldots, X_n$, której realizacją jest ciąg $x_1, x_2, \ldots, x_n$. Podobnie jak w modelu 1 nasze zadanie polega na weryfikacji na poziomie istotności α jednego z zestawów hipotez:

$$\begin{array}{lll} H_0:\ \mu = \mu_0 & H_0:\ \mu = \mu_0 & H_0:\ \mu = \mu_0 \\ H_1:\ \mu \neq \mu_0 & H_1:\ \mu < \mu_0 & H_1:\ \mu > \mu_0 \end{array}$$

W tym wypadku statytstyką testową jest:

$$t = \frac{\bar{X} - \mu_0}{S}\sqrt{n}$$

gdzie: $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ oraz $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2}$.

Aby w tym modelu zweryfikować hipotezę zerową metodą bootstrapową należy wyznaczyć rozkład bootstrapowy z próby o wartościach $x_1, x_2, \ldots, x_n$, a nstępnie wygenerować N ($N > 1000$) prób bootstrapowych $x_{1k}^*, x_{2k}^*, \ldots, x_{nk}^*$, gdzie $k = 1, 2, \ldots, N$ oznacza numer próby bootstrapowej. Następnie dla każdej próby bootstrapowej wyznaczamy:

$$\bar{x}_k^* = \frac{1}{n}\sum_{i=1}^n x_{ik}^* \qquad s_k^* = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_{ik}^* - \bar{x}_k^*)^2} \qquad \text{oraz} \qquad t_k^* = \frac{\bar{x}_k^* - \mu_0}{s^*}\sqrt{n}$$

Zatem uzyskujemy ciąg wartości $t_1^*, t_2^*, \ldots, t_N^*$. Wprowadźmy następujące oznaczenia:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{oraz} \qquad t_0 = \frac{\bar{x} - \mu_0}{s}\sqrt{n}$$

Dla testu dwustronnego ($H_0: \ \mu = \mu_0$ i $H_1: \ \mu \neq \mu_0$) wyznaczamy wartość:

$$\alpha_t^* = \frac{\overline{\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : |t| \geqslant |t_0|\}}}{N}$$

gdzie: $\overline{\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : |t| \geqslant |t_0|\}}$ – moc zbioru $\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : |t| \geqslant |t_0|\}$. Wartość α_t^* porównujemy z poziomem istotności α. Jeżeli $\alpha_t^* > \alpha$, to nie mamy podstaw do odrzucenia hipotezy H_0. A jeżeli $\alpha_t^* \leqslant \alpha$, to odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej.

Z kolei testu lewostronnego ($H_0: \ \mu = \mu_0$ i $H_1: \ \mu < \mu_0$) obliczamy wartość:

$$\alpha_{t^-}^* = \frac{\overline{\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : t \leqslant t_0\}}}{N}$$

gdzie: $\overline{\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : t \leqslant t_0\}}$ – moc zbioru $\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : t \leqslant t_0\}$. Wartość $\alpha_{u^-}^*$ porównujemy z poziomem istotności α. Jeżeli $\alpha_{u^-}^* > \alpha$, to nie mamy podstaw do odrzucenia hipotezy zerowej. A jeżeli $\alpha_{u^-}^* \leqslant \alpha$, to odrzucamy hipotezę H_0 na rzecz hipotezy H_1.

Dla testu prawostronnego ($H_0: \ \mu = \mu_0$ i $H_1: \ \mu > \mu_0$) obliczamy wartość:

$$\alpha_{t^+}^* = \frac{\overline{\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : t \geqslant t_0\}}}{N}$$

gdzie: $\overline{\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : t \geqslant t_0\}}$ – moc zbioru $\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : t \geqslant t_0\}$. Wartość $\alpha_{t^+}^*$ porównujemy z poziomem istotności α. Jeżeli $\alpha_{t^+}^* > \alpha$, to nie mamy podstaw do odrzucenia hipotezy zerowej. A jeżeli $\alpha_{t^+}^* \leqslant \alpha$, to odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną.

3.4.2. Testy bootstrapowe dla dwóch wartości średnich

W tych testach weryfikujemy następującą hipotezę o dwóch wartościach średnich dwóch populacji:

$$H_0: \quad \mu_1 = \mu_2$$

Wyróżniamy trzy modele, dla których bedziemy budować testy bootstrapowe pozwalające porównać dwie wartości średnie:

— Model 1 – populacje generalne mają nieznane rozkłady oraz σ_1 i σ_2 są znane,

— Model 2 – populacje generalne mają nieznane rozkłady, σ_1 i σ_2 są nieznane oraz $\sigma_1 = \sigma_2$,

— Model 3 – populacje generalne mają nieznane rozkłady, σ_1 i σ_2 są nieznane oraz nie wiemy czy $\sigma_1 = \sigma_2$.

Model 1

Załużmy, że jedna populacja jest badana ze względu na cehę X o nieznanej średniej μ_1 i znanym odchyleniu standardowym σ_1, a druga populacja jest badana ze względu na cehę Y o nieznanej średniej μ_2 i znanym odchyleniu standardowym σ_2. Z pierwszej populacji

wylosowano próbę niezależną $X_1, X_2, \ldots, X_{n_1}$, której realizacją jest ciąg $x_1, x_2, \ldots, x_{n_1}$. Natomiast z drugiej populacji wylosowano próbę niezależną $Y_1, Y_2, \ldots, Y_{n_2}$, której realizacją jest ciąg $y_1, y_2, \ldots, y_{n_2}$. Naszym zadaniem jest weryfikacja na poziomie istotności α jednego z zestawów hipotez:

$$H_0: \ \mu_1 = \mu_2 \qquad H_0: \ \mu_1 = \mu_2 \qquad H_0: \ \mu_1 = \mu_2$$

$$H_1: \ \mu_1 \neq \mu_2 \qquad H_1: \ \mu_1 < \mu_2 \qquad H_1: \ \mu_1 > \mu_2$$

Statytstyką testową jest:

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma}\sqrt{n}$$

gdzie: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ i $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

Aby zweryfikować hipotezę $H_0: \ \mu_1 = \mu_2$ metodą bootstrapową musimy wyznaczyć dwa rozkłady bootstrapowe z z dwóch prób:

— próby o wartościach $x_1, x_2, \ldots, x_n$

$$P(X_B = x_k) = \frac{1}{n_1} \quad \text{dla} \quad k = 1, 2, \ldots, n_1$$

— próby o wartościach $y_1, y_2, \ldots, y_n$

$$P(Y_B = y_k) = \frac{1}{n_2} \quad \text{dla} \quad k = 1, 2, \ldots, n_2$$

Następnie generujemy N ($N > 1000$) prób bootstrapowych $x^*_{1k}, x^*_{2k}, \ldots, x^*_{n_1 k}$ według rozkładu $P(X_B = x_k)$ i tyle samo prób bootstrapowych $y^*_{1k}, y^*_{2k}, \ldots, y^*_{n_2 k}$ według rozkładu $P(Y_B = y_k)$, gdzie $k = 1, 2, \ldots, N$ oznacza numer próby bootstrapowej. Dla każdej próby bootstrapowej wygenerowanej według rozkładu $P(X_B = x_k)$ wyznaczamy:

$$\bar{x}^*_k = \frac{1}{n_1}\sum_{i=1}^{n} x^*_{ik}$$

Analogiczne parametry wyznaczamy dla każdej próby bootstrapowej wygenerowanej według rozkładu $P(Y_B = y_k)$:

$$\bar{y}^*_k = \frac{1}{n_2}\sum_{i=1}^{n} y^*_{ik}$$

Gdy mamy już wielkości $\bar{x}^*_k$ i $\bar{y}^*_k$, to dla każdego $k = 1, 2, \ldots, N$ wyznaczamy:

$$z^*_k = \frac{\bar{x}^*_k - \bar{y}^*_k}{\sigma}\sqrt{n}$$

W ten sposób uzyskujemy ciąg wartości: $z^*_1, z^*_2, \ldots, z^*_N$ Wprowadźmy następujące oznaczenia:

$$\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i \qquad \bar{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i \qquad \text{oraz} \qquad z_0 = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}}}$$

Dla testu dwustronnego ($H_0: \ \mu_1 = \mu_2$ i $H_1: \ \mu_1 \neq \mu_2$) obliczamy wartość:

$$\alpha^*_z = \frac{\overline{\{z \in \{z^*_1, z^*_2, \ldots, z^*_N\} : |z_1| \geqslant |z_1|\}}}{N}$$

gdzie: $\overline{\{z \in \{z_1^*, z_2^*, \ldots, z_N^*\} : |z| \geqslant |z_0|\}}$ – moc zbioru $\{z \in \{z_1^*, z_2^*, \ldots, z_N^*\} : |z| \geqslant |z_0|\}$. Wartość α_z^* porównujemy z poziomem istotności α. Jeżeli $\alpha_z^* > \alpha$, to nie mamy podstaw do odrzucenia hipotezy zerowej. A jeżeli $\alpha_z^* \leqslant \alpha$, to odrzucamy hipotezę H_0 na rzecz hipotezy H_1.

Natomiast dla testu lewostronnego ($H_0: \ \mu_1 = \mu_2$ i $H_1: \ \mu_1 < \mu_2$) wyznaczamy wartość:

$$\alpha_{z^-}^* = \frac{\overline{\{z \in \{z_1^*, z_2^*, \ldots, z_N^*\} : z \leqslant z_0\}}}{N}$$

gdzie: $\overline{\{z \in \{z_1^*, z_2^*, \ldots, z_N^*\} : z \leqslant z_0\}}$ – moc zbioru $\{z \in \{z_1^*, z_2^*, \ldots, z_N^*\} : z \leqslant z_0\}$. Wartość $\alpha_{z^-}^*$ porównujemy z poziomem istotności α. Jeżeli $\alpha_{z^-}^* > \alpha$, to nie mamy podstaw do odrzucenia hipotezy H_0. A jeżeli $\alpha_{z^-}^* \leqslant \alpha$, to odrzucamy hipotezę i przyjmujemy hipotezę alternatywną.

Z kolei dla testu prawostronnego ($H_0: \ \mu_1 = \mu_2$ i $H_1: \ \mu_1 > \mu_2$) obliczamy wartość:

$$\alpha_{z^+}^* = \frac{\overline{\{z \in \{z_1^*, z_2^*, \ldots, z_N^*\} : z \geqslant z_0\}}}{N}$$

gdzie: $\overline{\{z \in \{z_1^*, z_2^*, \ldots, z_N^*\} : z \geqslant z_0\}}$ – moc zbioru $\{z \in \{z_1^*, z_2^*, \ldots, z_N^*\} : z \geqslant z_0\}$. Wartość $\alpha_{z^+}^*$ porównujemy z poziomem istotności α. Jeżeli $\alpha_{z^+}^* > \alpha$, to nie mamy podstaw do odrzucenia hipotezy zerowej. A jeżeli $\alpha_{z^+}^* \leqslant \alpha$, to odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej.

Model 2

Załużmy, że jedna populacja jest badana ze względu na cehę X o nieznanej średniej μ_1 i nieznanym odchyleniu standardowym σ_1, a druga populacja jest badana ze względu na cehę Y o nieznanej średniej μ_2 i nieznanym odchyleniu standardowym σ_2. Ponadto wiemy, że $\sigma_1 = \sigma_2$. Z każdej populacji wylosowano po jednej próbie niezależnej ($X_1, X_2, \ldots, X_{n_1}$ i $Y_1, Y_2, \ldots, Y_{n_2}$). Realizacjami tych prób są ciągi: $x_1, x_2, \ldots, x_{n_1}$ oraz $y_1, y_2, \ldots, y_{n_2}$. Naszym zadaniem jest weryfikacja na poziomie istotności α jednego z zestawów hipotez:

$H_0: \ \mu_1 = \mu_2$	$H_0: \ \mu_1 = \mu_2$	$H_0: \ \mu_1 = \mu_2$
$H_1: \ \mu_1 \neq \mu_2$	$H_1: \ \mu_1 < \mu_2$	$H_1: \ \mu_1 > \mu_2$

W tym wypadku statytstyką testową jest:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-1}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

gdzie:
- $\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$,
- $\bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$,
- $S_1 = \sqrt{\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2}$,
- $S_2 = \sqrt{\frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2}$.

Aby w tym modelu zweryfikować hipotezę zerową metodą bootstrapową należy przeprowadzić identyczne czynności jak w przypadku modelu 1, tzn wyznaczamy rozkłady bootstrapowe z dwóch pób oraz obliczmy wartości:

$$\bar{x}_k^* = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{ik}^* \qquad s_{1k}^* = \sqrt{\frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{ik}^* - \bar{x}_k^*)^2}$$

$$\bar{y}_k^* = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{ik}^* \qquad s_{2k}^* = \sqrt{\frac{1}{n_2-1}\sum_{i=1}^{n_2}(y_{ik}^* - \bar{y}_k^*)^2}$$

oraz

$$t_k^* = \frac{\bar{x}_k^* - \bar{y}_k^*}{\sqrt{\frac{(n_1-1)s_1^{*2}+(n_2-1)s_2^{*2}}{n_1+n_2-1}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Zatem uzyskujemy ciąg wartości $t_1^*, t_2^*, \ldots, t_N^*$. Wprowadźmy następujące oznaczenia:

$$\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i \qquad s_1 = \sqrt{\frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2}$$

$$\bar{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i \qquad s_2 = \sqrt{\frac{1}{n_2-1}\sum_{i=1}^{n_2}(y_i - \bar{y})^2}$$

oraz

$$t_0 = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(n_1-1)s_1{}^2+(n_2-1)s_2{}^2}{n_1+n_2-1}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Dla testu dwustronnego ($H_0:\ \mu_1 = \mu_2$ i $H_1:\ \mu_1 \neq \mu_2$) wyznaczamy wartość:

$$\alpha_t^* = \frac{\overline{\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : |t| \geqslant |t_0|\}}}{N}$$

gdzie: $\overline{\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : |t| \geqslant |t_0|\}}$ – moc zbioru $\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : |t| \geqslant |t_0|\}$. Wartość α_t^* porównujemy z poziomem istotności α. Jeżeli $\alpha_t^* > \alpha$, to nie mamy podstaw do odrzucenia hipotezy H_0. A jeżeli $\alpha_t^* \leqslant \alpha$, to odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej.

Z kolei testu lewostronnego ($H_0:\ \mu_1 = \mu_2$ i $H_1:\ \mu_1 < \mu_2$) obliczamy wartość:

$$\alpha_{t^-}^* = \frac{\overline{\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : t \leqslant t_0\}}}{N}$$

gdzie: $\overline{\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : t \leqslant t_0\}}$ – moc zbioru $\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : t \leqslant t_0\}$. Wartość $\alpha_{u^-}^*$ porównujemy z poziomem istotności α. Jeżeli $\alpha_{u^-}^* > \alpha$, to nie mamy podstaw do odrzucenia hipotezy zerowej. A jeżeli $\alpha_{u^-}^* \leqslant \alpha$, to odrzucamy hipotezę H_0 na rzecz hipotezy H_1.

Dla testu prawostronnego ($H_0:\ \mu_1 = \mu_2$ i $H_1:\ \mu_1 > \mu_2$) obliczamy wartość:

$$\alpha_{t^+}^* = \frac{\overline{\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : t \geqslant t_0\}}}{N}$$

gdzie: $\overline{\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : t \geqslant t_0\}}$ – moc zbioru $\{t \in \{t_1^*, t_2^*, \ldots, t_N^*\} : t \geqslant t_0\}$. Wartość $\alpha_{t^+}^*$ porównujemy z poziomem istotności α. Jeżeli $\alpha_{t^+}^* > \alpha$, to nie mamy podstaw do odrzucenia hipotezy zerowej. A jeżeli $\alpha_{t^+}^* \leqslant \alpha$, to odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną.

Model 3

W modelu 3 zakładamy, że jedna populacja jest badana ze względu na cehę X o nieznanej średniej μ_1 i nieznanym odchyleniu standardowym σ_1, a druga populacja jest badana ze względu na cehę Y o nieznanej średniej μ_2 i nieznanym odchyleniu standardowym σ_2 i nie wiemy czy $\sigma_1 = \sigma_2$. Z każdej populacji losujemy po jednej próbie niezależnej ($X_1, X_2, \ldots, X_{n_1}$ i $Y_1, Y_2, \ldots, Y_{n_2}$). Realizacjami tych prób są ciągi: $x_1, x_2, \ldots, x_{n_1}$ oraz $y_1, y_2, \ldots, y_{n_2}$. Naszym zadaniem jest weryfikacja na poziomie istotności α jednego z zestawów hipotez:

$$H_0: \ \mu_1 = \mu_2 \qquad H_0: \ \mu_1 = \mu_2 \qquad H_0: \ \mu_1 = \mu_2$$

$$H_1: \ \mu_1 \neq \mu_2 \qquad H_1: \ \mu_1 < \mu_2 \qquad H_1: \ \mu_1 > \mu_2$$

W tym wypadku statytstyką testową jest:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

gdzie:
- $\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i$,
- $\bar{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i$,
- $S_1 = \sqrt{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2}$,
- $S_2 = \sqrt{\frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}$.

Aby w modelu 3 zweryfikować hipotezę zerową metodą bootstrapową należy przeprowadzić identyczne czynności jak w przypadku poprzednich dwóch modeli, tzn wyznaczamy rozkłady bootstrapowe z dwóch pób oraz obliczmy wartości:

$$\bar{x}_k^* = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{ik}^* \qquad s_{1k}^* = \sqrt{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_{ik}^* - \bar{x}_k^*)^2}$$

$$\bar{y}_k^* = \frac{1}{n_2} \sum_{i=1}^{n_2} y_{ik}^* \qquad s_{2k}^* = \sqrt{\frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_{ik}^* - \bar{y}_k^*)^2}$$

oraz

$$c_k^* = \frac{\bar{x}_k^* - \bar{y}_k^*}{\sqrt{\frac{s_1^{*2}}{n_1} + \frac{s_2^{*2}}{n_2}}}$$

Zatem uzyskujemy ciąg wartości $t_1^*, t_2^*, \ldots, t_N^*$. Wprowadźmy następujące oznaczenia:

$$\bar{x} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i \qquad s_1 = \sqrt{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2}$$

$$\bar{y} = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i \qquad s_2 = \sqrt{\frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_i - \bar{y})^2}$$

oraz

$$c_0 = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 1} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Dla testu dwustronnego ($H_0: \ \mu_1 = \mu_2$ i $H_1: \ \mu_1 \neq \mu_2$) wyznaczamy wartość:

$$\alpha_c^* = \frac{\overline{\{c \in \{c_1^*, c_2^*, \ldots, c_N^*\} : |c| \geqslant |c_0|\}}}{N}$$

gdzie: $\overline{\{c \in \{c_1^*, c_2^*, \ldots, c_N^*\} : |c| \geqslant |c_0|\}}$ – moc zbioru $\{c \in \{c_1^*, c_2^*, \ldots, c_N^*\} : |c| \geqslant |c_0|\}$. Wartość α_c^* porównujemy z poziomem istotności α. Jeżeli $\alpha_c^* > \alpha$, to nie mamy podstaw do odrzucenia hipotezy zerową. A jeżeli $\alpha_c^* \leqslant \alpha$, to odrzucamy hipotezę zerową na rzecz hipotezy H_1.

Z kolei testu lewostronnego ($H_0: \ \mu_1 = \mu_2$ i $H_1: \ \mu_1 < \mu_2$) obliczamy wartość:

$$\alpha_{c^-}^* = \frac{\overline{\{c \in \{c_1^*, c_2^*, \ldots, c_N^*\} : c \leqslant c_0\}}}{N}$$

gdzie: $\overline{\{c \in \{c_1^*, c_2^*, \ldots, c_N^*\} : c \leqslant c_0\}}$ – moc zbioru $\{c \in \{c_1^*, c_2^*, \ldots, c_N^*\} : c \leqslant c_0\}$. Wartość $\alpha_{c^-}^*$ porównujemy z poziomem istotności α. Jeżeli $\alpha_{c^-}^* > \alpha$, to nie mamy podstaw do odrzucenia hipotezy zerowej. A jeżeli $\alpha_{c^-}^* \leqslant \alpha$, to odrzucamy hipotezę H_0 na rzecz hipotezy H_1.

Dla testu prawostronnego ($H_0: \ \mu_1 = \mu_2$ i $H_1: \ \mu_1 > \mu_2$) obliczamy wartość:

$$\alpha_{c^+}^* = \frac{\overline{\{c \in \{c_1^*, c_2^*, \ldots, c_N^*\} : c \geqslant c_0\}}}{N}$$

gdzie: $\overline{\{c \in \{c_1^*, c_2^*, \ldots, c_N^*\} : c \geqslant c_0\}}$ – moc zbioru $\{c \in \{c_1^*, c_2^*, \ldots, c_N^*\} : c \geqslant c_0\}$. Wartość $\alpha_{c^+}^*$ porównujemy z poziomem istotności α. Jeżeli $\alpha_{c^+}^* > \alpha$, to nie mamy podstaw do odrzucenia hipotezy zerowej. A jeżeli $\alpha_{c^+}^* \leqslant \alpha$, to odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną.

Ćwiczenia nr 8

Zadanie 3.17. [6,7]

W pewnym biochemicznym doświadczeniu bada się czas życia pewnych żywych komórek w pewnym środowisku. Rozkład tego czasu można przyjąć za normalny. Dokonano 8 pomiarów i otrzymano następujące czasy życia tych komórek w badanym środowisku (w godz.): 4,7 5,3 4,0 3,8 6,2 5,5 4,5 6,0. Na poziomie istorności $\alpha = 0,05$ sprawdzić hipotezę, że średni czas życia tych komórek w tym środowisku wynosi: 4,0 godziny.

Zadanie 3.18.

W pewnym biochemicznym doświadczeniu bada się czas życia pewnych żywych komórek w pewnym środowisku. Rozkład tego czasu można przyjąć za normalny. Dokonano 8 pomiarów i otrzymano następujące czasy życia tych komórek w badanym środowisku (w godz.): 4,7 5,3 4,0 3,8 6,2 5,5 4,5 6,0. Sprawdzić hipotezę, że średni czas życia tych komórek w tym środowisku wynosi: 4,0 godziny. Przeprowadzić test bootstrapowy, a następnie porównać rezultat z wynikiem z zadania 3.21.

Zadanie 3.19. [7]

W pewnym doświadczeniu medycznym bada się czas snu pacjentów po zastosowaniu pewnego leku. Przyjmuje się, że czas snu (min) na rozkład $N(m, 67)$. Zmierzono czas snu u 16 losowo wybranych pacjentów i otrzymano: 435, 533, 393, 458, 525, 481, 324, 433, 503, 383, 395, 416, 555, 500, 488. Na poziomie istotności 0,05 sprawdzić hipotezę, że średni czas snu wynosi 7 godz. Przy jakim α podjeta decyzja weyfikacyjna ulegnie zmianie?

Zadanie 3.20.

W pewnym doświadczeniu medycznym bada się czas snu pacjentów po zastosowaniu pewnego leku. Przyjmuje się, że czas snu (min) na rozkład $N(m, 67)$. Zmierzono czas snu u 16 losowo wybranych pacjentów i otrzymano: 435, 533, 393, 458, 525, 481, 324, 433, 503, 383, 395, 416, 555, 500, 488. Sprawdzić testem bootstrapowym hipotezę, że średni czas snu wynosi 7 godz, a następnie porównać rezultat z wynikiem z zadania 3.23.

Zadanie 3.21. [7]

W celu porownania wyników studiow studentów dwóch wydziałów pewnej uczelni wylosowano dwie próby studentów z obu wydziałów. Dla wydziału A otrzymano następujące przeciętne oceny uzyskane w sesji egzaminacyjnej: 3,64 4,20 3,25 4,50 4,63 488 2,65 4,00 3,24 4,15 4,73 5,00 2,60 2,94 3,05, a dla wydziału B oceny: 4,30 4,05 2,80 3,00 3,75 3,00 4,55 3,45 3,66 2,50 3,75 3,50 4,28 4,07 3,90. Na poziomie istotności $\alpha = 0,05$ należy zweryfikować hipotezę, że studenci obu wydziałów uzyskali takie same wyniki w tej sesji egzaminacyjnej.

Zadanie 3.24.

W celu porownania wyników studiow studentów dwóch wydziałów pewnej uczelni wylosowano dwie próby studentów z obu wydziałów. Dla wydziału A otrzymano następujące przeciętne oceny uzyskane w sesji egzaminacyjnej: 3,64 4,20 3,25 4,50 4,63 488 2,65 4,00 3,24 4,15 4,73 5,00 2,60 2,94 3,05, a dla wydziału B oceny: 4,30 4,05 2,80 3,00 3,75 3,00 4,55 3,45 3,66 2,50 3,75 3,50 4,28 4,07 3,90. Na poziomie istotności $\alpha = 0,05$ wykorzystując testy bootstrapowe należy zweryfikować hipotezę, że studenci obu wydziałów uzyskali takie same wyniki w tej sesji egzaminacyjnej.

Rozdział 4

Weryfikacja hipotez statystycznych w typowych doświadczeniach

Podczas wykonywania badań statystycznych, gdy już mamy zebrane dane, istotne są dwie kwestie:

— odpowiednie sformułowanie hipotez,

— dobór odpowiedniego testu statystycznego.

Zwykle, aby sformulować hipotezy należy posiadać dogłębna wiedzę poza statystyczną o przedmiocie badań.

Przykład 4.1:

W grupie 20 mężczyzn i 20 kobiet przeprowadzono badanie poziomu hemoglobiny we krwi. Uzyskano następujące dane (w g/dl):

mężczyźni	15,9	14,0	16,0	16,9	15,7	16,3	16,0	14,9	15,5	14,5
	16,1	14,1	14,2	14,4	12,3	16,7	15,6	14,5	16,6	13,5
kobiety	14,7	13,5	13,4	14,8	12,4	13,9	14,1	13,0	13,5	12,1
	12,1	13,4	14,0	15,8	12,6	13,4	13,2	11,3	12,8	11,5

Mając takie dane można zadać kilka pytań:

— Pytanie 1: Jakiego typu rozkładami są rozkłady wartości poziomu hemoglobiny we krwi w poszczególnych grupach?

— Pytanie 2: Czy rozkłady wartości poziomu hemoglobiny we krwi w poszczególnych grupach są takie same?

— Pytanie 3: Czy rozrzut wartości poziomu hemoglobiny we krwi w grupie mążczyzn jest taki sam ja w grupie kobiet?

— Pytanie 4: Czy średnia wartość poziomu hemoglobiny we krwi w grupie mążczyzn równa się średniaj wartości w grupie kobiet?

— itp.

W zależności od tego co chcemy badać, czyli jaką informację chcemy uzyskać, musimy umieć postawić odpowiednie pytanie, a następnie sformułować właściwe hipotezy. Zatem dla powyższych pytań:

— Pytanie 1:

Dla grupy mężczyzn:

$$H_0: \quad F_M(x) = F_0(x) \qquad\qquad H_1: \quad F_M(x) \neq F_0(x)$$

gdzie $F_0(x)$ – dystrybuanta rozkładu normalnego.

Dla grupy kobiet:

$$H_0: \quad F_K(x) = F_0(x) \qquad\qquad H_1: \quad F_K(x) \neq F_0(x)$$

gdzie $F_0(x)$ – dystrybuanta rozkładu normalnego.

W obu przypadkach hipoteza zerowa, mówi o tym, że rozkłady są zbliżone do rozkladu normalnego. Jest tak, ponieważ z doświadczenia wiemy, że ta cecha zazwyczaj ma taki rozkład. A hipoteza altrernatywna zawiera wszystkie pozostale przypadki.

— Pytanie 2:

$$H_0: \quad F_M(x) = F_K(x) \qquad\qquad H_1: \quad F_M(x) \neq F_K(x)$$

W przypadku pytania drugiego możliwość wyboru hipotez jest ograniczona, ponieważ albo rozkłady są jednakowe (H_0), albo rózne (H_1).

— Pytanie 3:

$$H_0: \quad \sigma_M = \sigma_K \qquad\qquad H_1: \quad \sigma_M \neq \sigma_K$$

Podobnie jak w przypadku pytania drugiego, możliwości wyboru są ograniczone. Zatem o tym, że odchylenia standardowe są sobie równe mówi hipoteza zerowa, a że są różne – hipoteza alternatywna.

— Pytanie 4:

$$H_0: \quad \mu_M = \mu_K \qquad\qquad H_1: \quad \mu_M > \mu_K$$

Inaczej przedstawia się sytuacja w przypadku pytania czwartego, ponieważ posiadając już pewną wiedzę na temat tego zjawiska, wiemy, że u mężczyzn poziom hemoglobiny jest zazwyczaj wyższy niż u kobiet. Dlatego też hipoteza zerowa mówi o równości średnich, a hipoteza alternatywna o tym, że poziom hemoglobiny u mężczyzn jest wyższy niż u kobiet.

Gdy już postawimy właściwe hipotezy musimy przejść do wyboru testu statystycznego, którym będziemy weryfikować postawione hipotezy. Pomocne w wyborze odpowiedniego testu mogą okazać się poniższe tabele. W pierwszej znajduje się zestawienie testów nieparametrycznych wraz z ich podstawowymi założeniami, w drugiej tabeli zestawienie testów parametrycznych służących do weryfikacji własności populacji jednowymiarowej, a w trzeciej – testów parametrycznych służących do weryfikacji własności populacji dwuwymiarowej.

Testy zgodności dla populacji jednowymiarowej		
Test	własności dystrybuant	liczebność próby
zgodności rozkladu χ^2	$F(x)$ – dystrybuanta rozkładu populacji, $F_0(x)$ – dowolna	dowolna
λ Kołmogorowa	$F(x)$ – dystrybuanta rozkładu ciągłego, $F_0(x)$ – dowolna dystrybuanta rozkładu ciągłego	dowolna
Shapiro-Wilka	$F(x)$ – dystrybuanta badanej cechy, $F_0(x)$ – dystrybuanta rozkładu normalnego	$n \leqslant 50$
Testy zgodności dla dwóch prób niezależnych		
Test	własności dystrybuant	liczebność próby
Kołmogorowa-Smirnowa	$F_1(x)$ i $F_2(x)$ – ciągłe dystrybuanty	$n_1, n_2 > 30$
Walda-Wolfowitza	$F_1(x)$ i $F_2(x)$ – ciągłe dystrybuanty	n_1 i n_2 – dowolne
U Manna-Whitneya	$F_1(x)$ i $F_2(x)$ – ciągłe dystrybuanty	n_1 i n_2 – dowolne
Testy zgodności dla dwóch prób zależnych		
Test	własności dystrybuant	liczebność próby
znaków	$F_1(x)$ i $F_2(x)$ – ciągłe dystrybuanty	$n_1 = n_2$ – dowolne
Wilcoxona	$F_1(x)$ i $F_2(x)$ – ciągłe dystrybuanty	$n_1 = n_2$ – dowolne

Testy parametryczne dla populacji jednowymiarowej				
Testy istotności dla wartości średnej				
Model	typ badanego rozkładu	wartość oczekiwana	odchylenie standardowe	liczebność próby
1	$N(\mu, \sigma)$	nieznana	znane	dowolna
2	$N(\mu, \sigma)$	nieznana	nieznane	$n \leqslant 30$
3	$N(\mu, \sigma)$	nieznana	nieznane	$n > 30$
Testy istotności dla wariancji				
Model	typ badanego rozkładu	wartość oczekiwana	odchylenie standardowe	liczebność próby
1	$N(\mu, \sigma)$	nieznana	nieznane	$n \leqslant 30$
2	$N(\mu, \sigma)$	nieznana	nieznane	$n > 30$

Testy parametryczne dla dwóch populacji				
Model	typy badanych rozkładów	wartości oczekiwane	odchylenia standardowe	liczebności prób
Testy istotności dla dwóch wartości średnich				
1	$N(\mu_1, \sigma_1)$ $N(\mu_2, \sigma_2)$	nieznane	σ_1, σ_2 – znane	n_1, n_2 – dowolne
2	$N(\mu_1, \sigma_1)$ $N(\mu_2, \sigma_2)$	nieznane	$\sigma_1 = \sigma_2$ – nieznane	$n_1, n_2 \leqslant 30$
3	$N(\mu_1, \sigma_1)$ $N(\mu_2, \sigma_2)$	nieznane	$\sigma_1 \neq \sigma_2$ – nieznane	$n_1, n_2 \leqslant 30$
4	$N(\mu_1, \sigma_1)$ $N(\mu_2, \sigma_2)$	nieznane	σ_1, σ_2 – nieznane	$n_1, n_2 > 30$
Testy istotności dla dwóch wariancji				
1	$N(\mu, \sigma)$	nieznane	nieznane	n_1, n_2 – dowolne

Przykład 4.2:

Odpowiedzmy na pytanie 4 z porzykładu 4.1. Załużmy, że dla danych z przykładu 4.1. nieznamy odchyleń standardowy obu grup, ale wiemy, że:

$$\sigma_1 = \sigma_2$$

oraz rozkłady badanej cechy są zgodne z rozkładami normlnymi. Teraz korzystając z tabeli dla testów parametrycznych dla dwóch populacji zauwazamy, że musimy skorzystać z Modelu 2 testu istotności dla wartości średnich.
Zatem w modelu 2 sprawdzamy hipotezy:

$$H_0: \quad \mu_M = \mu_K \qquad\qquad H_1: \quad \mu_M > \mu_K$$

Ustalmy poziom istotności $\alpha = 0,05$.

$$\bar{x}_M \approx 15,19 \qquad \bar{x}_K \approx 13,28$$

$$S_M^2 \approx \frac{28,05}{20} \approx 1,40 \qquad S_K^2 \approx \frac{23,82}{20} = 1,19$$

$$t = \frac{15,19 - 13,28}{\sqrt{\frac{20 \cdot 1,40 + 20 \cdot 13,28}{20+20-2}\left(\frac{1}{20} + \frac{1}{20}\right)}} \approx 5,17$$

Z tabeli wartości krytycznych rozkładu t Studenta odczytujemy wartość krytyczną:

$$t_{0,05,38} = 2,0244$$

Zatem mamy: $t \geqslant t_{0,05,38}$, czyli odrzucamy hipotezę zerową mówiącą, że wartości średnie poziomu hemoglobiny u mężczyzn i u kobiet są takie równe. Natomiast przyjmujemy hipotezę alternatywną i stwierdzamy, że na poziomoe istności 0,05 mężczyźni mają wyższy poziom hemoglobiny we krwi niż kobiety.

Ćwiczenia nr 9

Zadanie 4.1.

U 20 studentów II roku przeprowadzono badanie krwi. Badano liczbę trombocytów w mm^3 krwi i uzyskano następujące dane (w tys/mm^3):

221 359 319 293 308 187 289 406 302 298

244 293 311 324 269 279 435 139 449 318

Dla powyższych danych zaproponuj i wykonaj odpowiednie testy satystyczne (co najmniej 2).

Zadanie 4.2.

U 40 studentów II roku przeprowadzono badanie krwi. Badano liczbę retikulocytów w mm^3 krwi i uzyskano następujące dane (w $10^9/mm^3$):

77 53 85 91 95 59 86 77 94 88

96 72 72 99 25 63 45 66 90 94

51 64 78 68 82 84 53 71 25 102

60 47 69 41 75 62 126 102 16 85

Dla powyższych danych zaproponuj i wykonaj odpowiednie testy satystyczne (co najmniej 2). Dla powyższych danych zaproponuj i wykonaj odpowiednie testy satystyczne (co najmniej 2).

Zadanie 4.3.

W grupie 40 mężczyzn i 40 kobiet przeprowadzono badanie krwi. Badano liczbę erytrocytów w mm^3 krwi i uzyskano następujące dane (w mln/mm^3):

kobiety	5,2	4,9	4,9	5,2	4,6	5,0	5,0	4,7	4,9	4,5
	4,5	4,8	5,0	5,5	4,6	4,8	4,8	4,3	4,7	4,3
	5,0	4,6	4,8	4,7	4,9	4,6	4,9	5,0	5,2	4,8
	4,3	4,6	5,1	4,5	5,0	4,8	5,2	4,3	4,7	4,5
mężczyźni	6,0	5,4	5,6	4,9	5,1	5,1	5,5	5,1	5,4	4,6
	5,1	5,0	4,8	5,3	5,3	5,2	4,8	5,5	5,3	5,1
	5,2	5,1	4,7	5,1	5,0	4,9	4,9	5,0	4,6	5,5
	5,3	5,2	5,2	5,0	5,5	5,2	5,0	5,6	5,1	5,0

Dla powyższych danych zaproponuj i wykonaj odpowiednie testy satystyczne (co najmniej 2).

Zadanie 4.4.

W grupie 35 mężczyzn i 35 kobiet przeprowadzono badanie krwi. Badano liczbę leukocytów w mm^3 krwi i uzyskano następujące dane (w tys/mm^3):

mężczyżni	7,0	6,4	6,1	9,8	8,9	7,6	6,0
	6,4	7,1	8,0	5,9	4,9	5,8	7,6
	7,6	7,7	7,1	7,4	6,8	8,1	5,9
	7,7	6,4	6,9	7,8	8,3	6,1	8,5
	7,9	7,2	7,1	7,0	6,9	7,3	7,3
kobiety	7,4	6,2	8,2	7,6	7,4	7,8	7,5
	6,3	7,1	7,1	6,7	8,9	6,4	6,8
	6,5	6,1	7,1	7,0	8,8	7,0	6,2
	8,8	8,5	7,0	5,8	6,8	7,1	7,5
	7,0	7,7	7,6	6,0	6,3	6,5	6,7

Dla powyższych danych zaproponuj i wykonaj odpowiednie testy satystyczne (co najmniej 2).

Rozdział 5

Analiza wariancji

W rozdziale 3 zostały opisane testy pozwalające porównywać wartości średnie dla dwóch populacji. Obecnie pozostaje do rozstrzygnięcia sytuacja, w której chcemy porównać średnie dla więcej niż dwóch populacji. Naturalnym spostrzeżeniem jest to, że można porownywać średnie parami, lecz takie postępowanie prowadzi do dużych błędów I rodzaju. Załużmy, że chcemy porównać wartości średnie dla pięciu populacji na poziomie istotności $0,05$. Wówczas prawdopodobieństwo, że nie pomylimy się porównując dwie śrenie wynosi $0,95$. Jeśli mamy parami porównać 5 wartosci średnich musimy dokonać 10 porównań, a więc prawdopodobieństwo, że ani raz się nie pomylimy jest równe $0,95^{10} \approx 0,599$. Natomiast prawdopodobieństwo co najmniej jednej pomyłki jest równe $1 - 0,95^{10} \approx 0,401$. Zatem do porównywania wielu wartości śrenich należy zastosować inne metody.

Taką metodą jest *analiza wariancji* (ANOVA – Analysis Of Variance) zaproponowana przez angielskiego biologa R. A. Fishera. Pozwala ona zweryfikować hipotezę mówiącą o równości średnich dla kilku prób. Zazwyczaj w praktyce występują dwa typy analizy wariancji:

— analiza wariancji dla klasyfikacji pojedyńczej – badany jest wpływ tylko jednego czyynnika klasyfikującego,

— wieloczynnikowa analiza wariancji – badany jest wpływ wielu czyynników klasyfikujących.

Poniżej zostanie omowiona jedynie analiza wariancji dla klasyfikacji pojedyńczej.

5.1. ANOVA dla klasyfikacji pojedyńczej

W analizie wariancji istotne są dwa założenia:

— wszystkie badane próby mają rozkłady normalne $N(\mu_i, \sigma_i)$ dla $i = 1, 2, \ldots, k$,

— wariancje we wszystkich k próbach są równe, czyli $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$.

ANOVA jest testem na podstawie, którego można zweryfikować następujące hipotezy:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

$$H_1 : \exists_{i \neq j} \mu_i \neq \mu_j$$

Te rozważania można zapisać w postaci matematycznego modelu, w którym badane wartości są sumą ogólnej średniej z populacji μ, wpływu i-tego czynnika eksperymentalnego α_i oraz składnika losowego ϵ_{ij} o rozkładzie $N(0, \sigma)$:

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

W zależności od interpretacji współczynnika α_i wyróżniamy dwa modele:

— model I rodzaju – współczynnik α_i opisuje wpływ zabiegu eksperymentalnego, jest pod naszą kontrolą,

— model II rodzaju – współczynnik α_i jest zmienną losową o rozkładzie normalnym i nie kontrolujemy go.
Wybór modelu wpływa na interpretację wyników, lecz nie na algorytm obliczeń statystyki testowej.

Aby wyznaczyć wartość statystyki testowej umożliwiającej weryfikację hipotez, należy najpierw obliczyć *średnie grupowe* i *średnią ogólną*. *Średnimi grupowymi* $\bar{x}_i$ nazywamy średnie arytmetyczne wyników x_{ij} dla każdej n_i-elementowej próby ($i = 1, 2, \ldots, k$):

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}.$$

A *średnią ogólną* $\bar{\bar{x}}$ wyznaczamy korzystając ze wzoru:

$$\bar{\bar{x}} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij},$$

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

Następnie obliczamy *wariancję wewnątrz grup* ($\hat{s}_1^2$) i *wariancję między grupami* ($\hat{s}_2^2$). *Wariancję wewnątrz grup* wyznaczamy ze wzoru:

$$\hat{s}_1^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{n - k}.$$

Liczba $n - k$ jest liczbą stopni swobody odpowiadającą wariancji wewnątrz grup. Natomiast *wariancję międzygrupową* wyznaczamy ze wzoru:

$$\hat{s}_2^2 = \frac{\sum_{i=1}^{k} (\bar{x}_i - \bar{\bar{x}})^2 n_i}{k - 1},$$

gdzie $k - 1$ jest liczbą stopni swobody dla wariancji międzygrupowej. Statystykę testową F wyznaczamy jako stosunek wariancji międzygrupowej do wariancji wewnątrz grup:

$$F = \frac{\hat{s}_2^2}{\hat{s}_1^2}.$$

Obliczona statystyka F ma rozkład F Snedecora o $k - 1$ i $n - k$ stopniach swobody. Dla ustalonego poziomu istotności α oraz stopni swobody $k - 1$ i $n - k$ z tablicy rozkładu F Snedecora odczytujemy wartość krytyczną. Gdy $F < F_\alpha$, to uznajemy, ze nie ma podstaw do odrzucenia hipotezy H_0. Hipotezę zerową odrzucamy, jeśli $F \geqslant F_\alpha$. Wówczas przyjmujemy hipotezę alternatywną H_1, mówiacą, że spośród testowanych wartości średnich co najmniej dwie są rożne.

Przykład 5.1:

W poniższej tabeli zostaly przedstawione wyniki dla trzeczh grup badanych pod względem pewnej cechy. Rozkład cech w kszedj grupie jest zgodny z rozkładem normalnym o odchyleniu standardowym σ. Sprawdzić na poziomie istotości $\alpha = 0,05$ czy wartości średnie we wszystkich grupach są takie same.

grupa 1	grupa 2	grupa 3
3	4	8
5	6	6
6	5	5
7	7	8
8	3	3
5	2	6
2	2	7
4	6	5
3	5	5
10	5	3

W poniższej tabeli zostaną przedstawione kolejne obliczenia pozwalające wyznaczyć statystykę testową. Wartość krytyczna dla $\alpha = 0,05$ i stopni swobody 2 oraz 27: $F_\alpha = 3,3541$.

	grupa 1	grupa 2	grupa 3
$\bar{x_i}$	5,3	4,5	5,6
$\bar{\bar{x}}$	$5\frac{1}{3}$		
$\hat{s}_1^2$	$\frac{56,1+26,5+28,4}{3\cdot 10-3} = 4\frac{1}{9}$		
$\hat{s}_2^2$	$\frac{0,278+4,011+2,178}{3-1} \approx 3,233$		
F	$0,7865$		

Mamy, że $F < F_\alpha$, zatem nie ma podstaw do stwierdzenia, iż wartości średnie we wszystkich grupach nie są równe.

Ćwiczenia nr 10

Zadanie 5.1. [6,7]

Do pewnych dalszych doświadczen chodowane są cztery grupy królików jednorodnych pod względem wagi. Wybrano losowo po 5 królików z poszczegolnych grup i otrzymano następujące wagi (w kg): Na poziomie istotności 0,05 zweryfikować

grupa 1	grupa 2	grupa 3	grupa 4
2,95	3,20	3,05	3,00
2,80	3,05	3,30	3,30
3,10	2,90	3,15	2,75
3,00	3,05	3,20	2,85
3,15	3,05	2,80	3,10

hipotezę, że średnie wagi we wszystkich grupach są takie same, przy założeniu, że wariancje wagi we wszystkich grupach są takie same.

Zadanie 5.2. [6]

Przeprowadzono pomiary erytrocytów we krwi szczurów w czterech grupach (3 grupy badane oraz grupa kontrolna) i otrzymano następujące wyniki:

grupa 1	grupa 2	grupa 3	grupa 4
4,4	4,1	4,1	4,9
4,2	4,3	4,2	5,2
4,3	4,5	4,6	5,1
4,5	4,5	4,5	4,9
4,4	4,1	4,7	4,8
4,6	4,3	4,4	5,0
4,8	4,8	4,8	5,1
4,5	4,9	4,5	5,4
4,4	4,6	4,6	5,2
4,2	4,7	4,2	5,4

Na poziomie istotności 0,05 zweryfikować hipotezę, że średnie wagi we wszystkich grupach są takie same, przy założeniu, że wariancje wagi we wszystkich grupach są takie same.

Zadanie 5.3. [6]

Badano czterema różnymi metodami czas krzepnięcia osocza krwi dziesięciu losowo wybranym pacjentom i otrzymano następujące wyniki:

metoda 1	metoda 2	metoda 3	metoda 4
9,1	10,0	10,0	10,9
8,9	10,2	9,9	11,1
8,4	9,8	9,8	12,2
12,8	11,6	12,9	14,4
8,7	9,5	11,2	9,8
9,2	9,2	9,9	12,0
7,6	8,6	8,5	8,5
8,6	10,3	9,8	10,9
8,9	9,4	9,2	10,4
7,9	8,5	8,2	10,0

Na poziomie istotności 0,05 zweryfikować hipotezę, że średnie wagi we wszystkich grupach są takie same, przy założeniu, że wariancje wagi we wszystkich grupach są takie same.

Zadanie 5.4. [7]

Przeprowadzono badania w celu porównania czterech metod leczenia pewnej choroby. Pobrano pięcioelementowe próby losowe spośród chorych na daną chorobę, których leczono odpowiednio metodą I, II, III, IV. Wyniki terapii oceniono w specjalnym teście. Wartości testu w umownej punktacji przedstawiono w tabeli.
Na poziomie istotności 0,05 zweryfikować hipotezę, że wszystkie metody leczenia dają jednakowe wyniki w leczzeniu choroby.

metoda I	metoda II	metoda III	metoda IV
57	74	63	62
58	66	68	63
67	65	59	66
50	72	59	71
62	68	68	70

Zadanie 5.5. [2]

Zmierzono długości czasów świecenia trzech typów żarówek, otrzymując (w h):

typ I	typ II	typ III
1802	1664	1877
1992	1755	1710
1854	1823	1882
1880	1862	1720
1761		1950
1900		

Na poziomie istotności 0,05 zweryfikować hipotezę, że wartości przecientne czasów świecenia żarówek tych typów są jednakowe, wobec hipotezy alternatywnej, że te wartosci przeciętne nie są jednakowe.

Zadanie 5.6.

Napisz w Matlabie funkcję, wyznaczającą wartść statystyki testowej F dla dowolnej zadanej wartości k i n_i.

Rozdział 6

Korelacja i regresja

Zagadnienia przedstawione we wcześniejszych rozdziałach dotyczyły analizy statystycznej jednej zmiennej losowej. Jednak we wielu przypadkach, aby opisać populację, konieczna jest analiza kilku zmiennych losowych. Ponadto chcemy wiedzieć, czy te zmienne losowe są ze sobą powiązane. W tym celu wykorzystywane są metody korelacji i regresji.

Metody korelacji pozwalają szacować wartości oczekiwane zmiennej losowej zależnej na podstawie pojedynczych reprezentacji zmiennej losowej niezależnej. Istotne jest to, że skorelowanie powinno być badane pomiędzy zmiennymi losowymi, z których jedna jest przyczyną drugiej. Istotne jest aby na etapie analizy merytorycznej ustalić ten związek przyczynowo skutkowy. Czasami można zaobserwować zależność między badanymi cechami, lecz nie są one ze sobą powiązane np. między kolorem włosów, a IQ. Metody korelacji pozwalają badać siłę współzależności dwóch zmiennych.

Metody regresja pozwalają określić kształt zależności pomiędzy cechami. Wyróżniamy dwa typy regresji:

— regresję liniową – graficzną reprezentacją jest prosta,

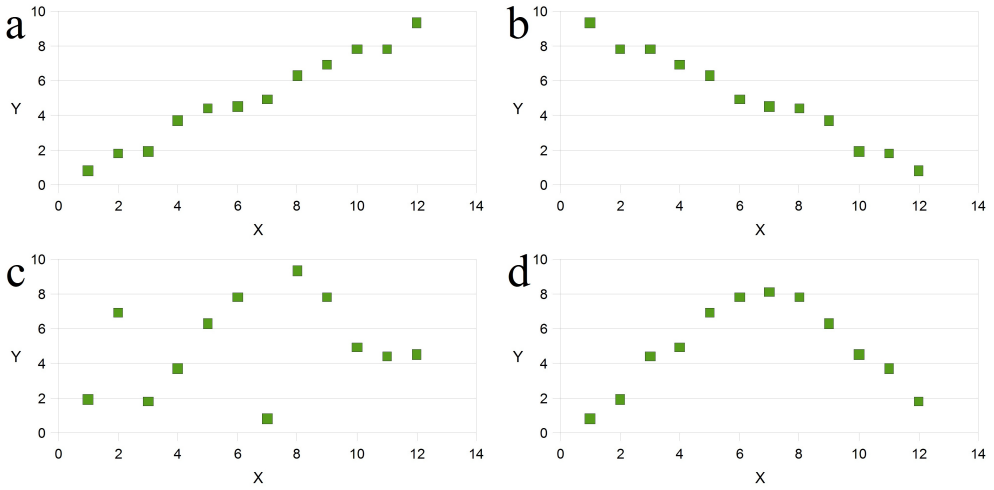
— regresję nieliniową – graficzną reprezentacją jest krzywa wyższewgo rzedu np. parabola.

Poniżej opisany zostanie jedynie najprostszy przypadek korelacji i regresji prostoliniowej dwóch zmiennych.

6.1. Diagram korelacyjny

Analizę zależności między badanymi cechami (X oraz Y) zaczynamy od sporządzenia *diagramu korelacyjnego*, czyli wykresu punktowego utworzonego przez punkty o współrzędnych (x_i, y_i), gdzie x_i wartości cechy X, y_i są odpowiadającymi im realizacjami cechy Y. W tym wypadku cecha X jest zmienną losową niezależną, a cecha Y zmienną losową zależną. Na rysunku 6.1 zostały przedstawione przykładowe diagramy korelacyjne.

Jeżeli punkty (x_i, y_i) leżą dokładnie wzdłuż prostej, to mówimy, że mamy pełną korelację. Jest to jednak niezwykle rzadko spotykane. Częściej te punkty leżą mniej wiecej na pewnej krzywej (nierzadko jest to prosta). Jeżeli punkty układają się wzdłuż prostej rosnącej to mamy *korelację dodatnią* (rysunek 6.1 a), czyli wzrostowi jednej zmiennej towarzyszy wzrost drugiej zmiennej. Z koleji jeśli wzdłuż prostej malejącej to mamy *korelację ujemną* (rysunek 6.1 b), to znaczy wzrostowi jednej zmiennej towarzyszy spadek drugiej zmiennej. Natomiast gdy punkty są całkowicie rozrzucone to nie mamy korelacji (rysunek 6.1 c), oznacza to, że ze zmieny wartości jednej zmiennej nie możemy określić zmieny wartości drugiej zmiennej. A jeśli punkty układają się wzdłuż innej krzywej to mamy *korelację krzywoliniową* (rysunek 6.1 d), to innymi słowy zmiana wartości zmiennej zależnej jest opisana pewną funkcją zależną od wartości zmiennej nieazleżnej.



Rysunek 6.1. Diagramy korelacyjne: a – korelacja dodatnia, b – korelacja ujemna, c – brak korelacji, d – korelacja krzywoliniowa

6.2. Tablica korelacyjna

Inną formą przedstawienia zależności pomiedzy dwiema zmiennymi losowymi X i Y jest *tablica korelacyjna*. Tą tablicę budujemy wykonując następujące czynności:

— dla każdej zmiennej losowej tworzymy k przedziałów klasowych, tak jak przy szeregach rozdzielczych,

— następnie tworzymy tablicę $k \times k$, numer kolumny odpowiada numerowi klasy dla zmiennej X, a numer wiersza numerowi klasy dla zmiennej Y.

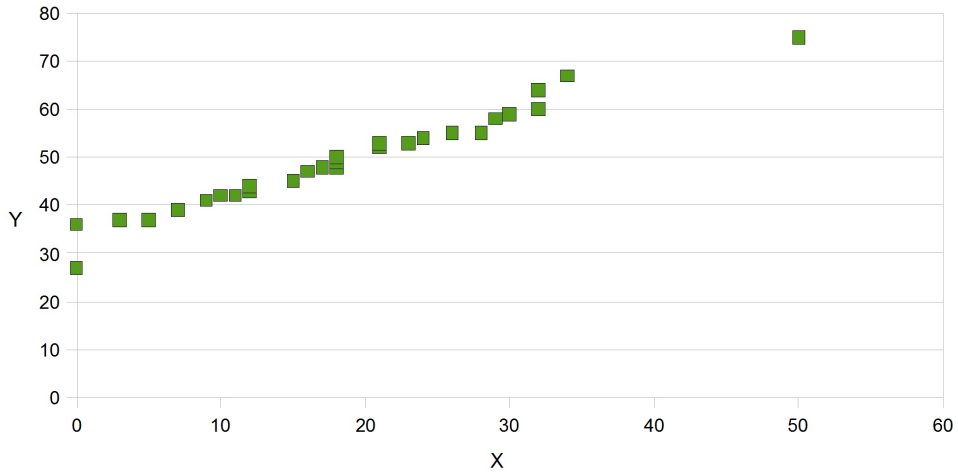
— nastepnie zliczamy ile par (x_k, y_k) spełnia taki warunek, że wartości x_k należą do i-tej klasy dla zmiennej X oraz y_k należy do j-tej klasy zmiennej Y, nastepnie wpisujemy tę liczbę w odpowiednią komórkę tablicy korelacyjnej.

Przykład 6.1:

W pewnym badaniu badono dwie cechy zależne. Wyniki przedstawia poniższa tabela. Narysyj diagram korelacyjny oraz utwórz tablicę korelacyjną.

X	Y
0	27
3	37
7	39
10	42
12	43
15	45
17	48
18	49
18	50
21	52
23	53
26	55
29	58
32	60
34	67

X	Y
0	36
5	37
9	41
11	42
12	44
16	47
18	48
18	49
18	50
21	53
24	54
28	55
30	59
32	64
50	75



Rysunek 6.2. Diagramy korelacyjny dla zmiennych X i Y

Tablica korelacyjna:

		x_i				
		$\langle 0,10)$	$\langle 10,20)$	$\langle 20,30)$	$\langle 30,40)$	$\langle 40,50\rangle$
y_i	$\langle 26,36)$	1	–	–	–	–
	$\langle 36,46)$	5	5	–	–	–
	$\langle 46,56)$	–	7	6	–	–
	$\langle 56,66)$	–	–	1	3	–
	$\langle 66,76)$	–	–	–	1	1

6.3. Współczynnik korelacji liniowej Pearsona

Współzależność dwóch zmiennych losowych (X i Y) może zostać wyrażona liczbowo. Istnieje wiele mierników współzależności zmiennych, lecz najpopularniejszym jest *współczynnik korelacji liniowej Pearsona* (r_{XY}), który przyjmuje wartości z przedziału $\langle -1, 1\rangle$. Wyznaczanie współczynnika korelacji liniowej Pearsona ma sens jedynie, gdy spełnione są następujące założenia:

— zmienne losowe X i Y są mierzalne,
— mają rozkłady zbliżone do normalnego,
— zależność jest prostoliniowa.

Wartość współczynnika korelacji liniowej Pearsona bliska zeru nie oznacza braku jakiejkolwiek korelacji, a jedynie brak korelacji liniowej.

Współczynnik korelacji liniowej Pearsona r_{XY} dla następujących realizacji zmiennych losowych X i Y:

$$x_1, x_2, \ldots, x_n \qquad oraz \qquad y_1, y_2, \ldots, y_n$$

dany jest wzorem:

$$r_{XY} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

gdzie: $\bar{x} = \sum_{i=1}^{n} x_i$ oraz $\bar{y} = \sum_{i=1}^{n} y_i$.

Wartość bezwzględna współczynnika korelacji liniowej Pearsona informuje nas o *sile korelacji*, natomiast znak współczynnika – o *kierunku korelacji*. W literaturze spotykane są rózne umowne określenia stopnia zależności dwóch zmiennych, poniżej została przedstawiona jedna z stosowanych skal:

— $r_{XY} = 0$ – brak korelacji,
— $0 < r_{XY} < 0,1$ – niska korelacja,
— $0,1 \leqslant r_{XY} < 0,3$ – niska korelacja,
— $0,3 \leqslant r_{XY} < 0,5$ – słaba korelacja,
— $0,5 \leqslant r_{XY} < 0,7$ – przeciętna korelacja,
— $0,7 \leqslant r_{XY} < 0,9$ – wysoka korelacja,
— $0,9 \leqslant r_{XY} < 1$ – bardzo wysoka korelacja,
— $r_{XY} = 1$ – całkowita korelacja.

Ważnym współczynnikiem powiązanym z współczynnikiem korelacji liniowej Pearsona jest *współczynnik determinacji* R^2. Jest on określony wzorem:

$$R^2 = r_{XY}^2$$

Współczynnik determinacji R^2 informuje o tym w jakim stopniu zmiany jednej zmiennej losowej są wyjaśniane przez zmiany drugiej zmiennej losowej.

Przykład 6.2:

Dla danych z przykładu 6.1 wyznacz współczynnik korelacji liniowej Pearsona oraz współczynnik determinacji.
Współczynnik korelacji liniowej Pearsona: $r_{XY} = 0,98$.
Współczynnik determinacji: $R^2 = 0,97$.

6.4. Regresja liniowa

Funkcja regresji jest metodą, która pozwala estymować wartości oczekiwane zmiennej zalażnej Y przy zadanych wartościach zmninnej niezależnej X. Wybor pomiędzy regresją liniawą i nieliniową nie jest łatwy. Aby go dokonać powinno się wstępnie przeanalizować materiał statystyczny, narysować diagram korelacyjny oraz zbadać materiał pozastatystyczny z danej dziedziny badań opisującej zjawisko. Jednak bardzo często stosowaną regresją jest regresja liniowa. Wówczas krzywą regresji są prosta regresji dana wzorem:

$$\hat{y} = a_y x + b_y$$

W powyższym równaniu współczynniki a_y oraz b_y szacuje się metodą najmniejszych kwadratów, czyli w taki sposób, aby równanie:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a_y x_i - b_y)^2 \qquad (\hat{y}_i = a_y x_i + b_y)$$

osiągało minimum, gdzie x_i oraz y_i ozaczają zrealizowane wartości zmiennych losowych X oraz Y. Zatem współczynniki a_y oraz b_y dane są wzorami:

$$a_y = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$b_y = \bar{y} - a_y \bar{x}$$

gdzie: $\bar{x} = \sum_{i=1}^{n} x_i$,a $\bar{y} = \sum_{i=1}^{n} y_i$. Współczynnik a_y nazywany jest także parametrem regresji, a współczynnik b_y wyrazem wolnym regresji. Oceną oszacowania wartości $\hat{y}_i = a_y x_i + b_y$ dla wartości próby losowej jest *wariancja składnika resztkowego* dana wzorem:

$$s_r^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$

Pierwiastek z s_r^2 nazywany jest błędem szacunku i określa średnie odchylenie wartości empirycznych od wartości teoretycznych.

Dopasowanie prostej refresji do punktów empirycznych określa *współczynnik determinacji* R^2 przyjmujący wartości z przedziału $\langle 0, 1 \rangle$. Im wartości współczynnika determinacji są bliższe jedności tym dopasowanie prostej regresji do punktów empirycznych jest lepsze. Współczynnik determinacji dla redresji liniowej jest równy współczynnikowi korelacji liniowej, zatem dany jest wzorem

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

gdzie: $\bar{x} = \sum_{i=1}^{n} x_i$ oraz $\bar{y} = \sum_{i=1}^{n} y_i$.

Przykład 6.3:

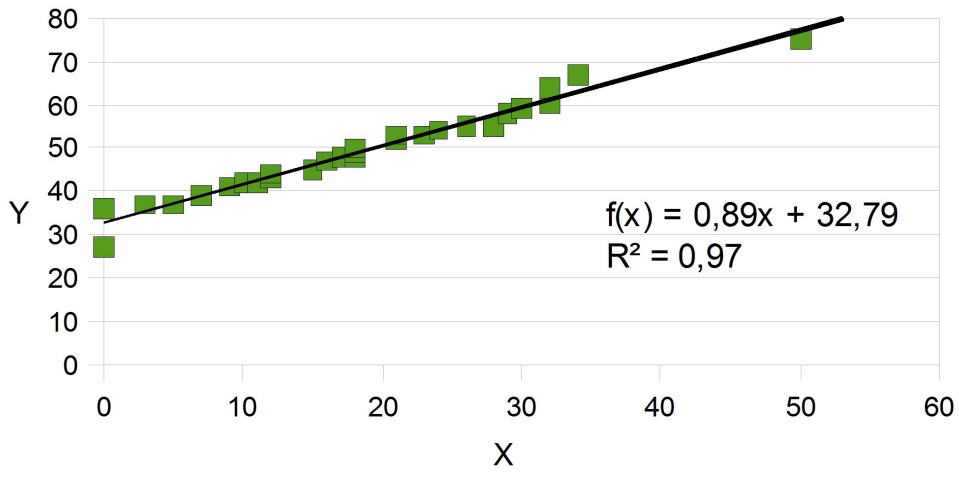
Wyznaczyć parametry prostej regresji, błąd szacunku oraz współczynnik determinacji dla danych z przykładu 6.1.

$$a_y \approx \frac{107,1}{120,45} \approx 0,89$$

$$b_y \approx 49,30 - 0,89 \cdot 18,57 \approx 32,79$$

$$s_r^2 \approx \frac{93,48}{30-2} \approx 3,34$$

$$R^2 \approx 0,97$$



Rysunek 6.3. Prosta regresji dla zmiennych X i Y

Ćwiczenia nr 11

Zadanie 6.1. [6]

W klinice laryngologicznej badano związek pomiędzy wymiarami krtani a wiekiem niemowląt. Spośród lecznonych wylosowano 10 dzieci, dla których otrzymano nastepujace dane (x_i – wiek w miesiącach, y_i – długość szpary głośni w mm): Zbadać

x_i	5	2	3	4	7	8	9	12	11	5
y_i	6,0	3,5	4,0	7,0	10,0	9,0	10,0	10,5	10,0	7,0

za pomocą współczynnika korelacji liniowej siłę i kierunek zależności pomiedzy wiekiem niemiowlat a długością szpary głośni.

Zadanie 6.2. [6]

W badaniach nad wpływem zmian temperatury na myszy, określono szubkość utraty wody przez grupę myszy w efekcie działania różnych temperatur. Pomiary wykonano przez absorbcję wydalanej wody w okreslonym czasie. Otrzymano nastepujace wyniki:

temperatyra (w ^oC)	15	20	25	30	35
wydalana woda (w mg)	2794	2924	3175	3340	3576

Zbadać za pomocą współczynnika korelacji liniowej zależność pomiedzy ilością wydalanej wody od temperatury.

Zadanie 6.3. [6]

Wykorzystując dane zgrupowane w tablicy korelacji wyznaczyć współczynniki korelacji oraz proste regresji:

		x_i					
		1	3	5	7	9	11
y_i	2	24	–	–	–	–	–
	6	36	27	8	–	–	–
	10	–	48	55	22	14	–
	14	–	13	21	48	24	–
	18	–	–	–	5	15	14
	12	–	–	–	–	–	26

]

Zadanie 6.4. [7]

Wśród 10 pracowników huty przeprowadzono badanie zależności między wielkością uszkodzenia słuchu a ich stażem pracy. Dane przedstawiono w tabeli:

Lp.	Uszczerbek słuchu (w %)	Staż pracy (w latach)
1	3	2
2	2	2
3	3	5
4	5	4
5	5	6
6	8	10
7	6	5
8	5	6
9	10	15
10	13	15

Określić siłę i kierunek korelacji związku między badanymi cechami przy pomocy współczynnika korelacji liniowej. Napisać równanie reglesji liniowej. Naszkicować wykres prostej regresji. Jakiego ubytku słuchu należy spodziewać się w pracownika z 12-letnim stażem pracy?

Zadanie 6.5.

Napisać w Matlabie funkcję, która dla n par (x_i, y_i) będących realizacjami zmiennych losowych X i Y :

— wyznacza współczynnik korelacji liniowej,
— określa stopień zależności pomiędzy dwoma zmiennymi,
— wyznacza parametr regresji,
— oblicza wyraz wolny regresji,
— określa błąd szacunku.

Bibliografia

[1] W. Krysicki, J. Bartos, W. Dyczka, K. Królikowska, M. Wasilewska, *Rachunek prawdopodobienstwa i statystyka matemaycza w zadaniach, cz. 1. Statystyka matematyczna*, Wydawnictwo Naukowe PWN, Warszawa, 2007

[2] W. Krysicki, J. Bartos, W. Dyczka, K. Królikowska, M. Wasilewska, *Rachunek prawdopodobienstwa i statystyka matemaycza w zadaniach, cz. 2. Statystyka matematyczna*, Wydawnictwo Naukowe PWN, Warszawa, 2007

[3] A. Stanisz, *Przystępny kurs ststystykiz zastosowaniem STATISTICA PL na przykladach z medycyny. Tom 1 Statystyki podstawowe*, StstSoft, Kraków, 2006

[4] A. Łomnicki, *Wprowadzenie do statystyki dla przyrodników*, Wydawnictwo Naukowe PWN, Warszawa, 2003

[5] S. M. Kot, J. Jakubowski, A. Sokołowski, *Statystyka. Podręcznik dla studiów ekonomicznych*, Delfin, Warszawa, 2007

[6] A. Lemńczyk, *Statystyka . . . w pigułce*, Wydawnictwo Naukowe Uniwersytetu Medycznego im. Karola Marcinkowskiego w Poznaniu, Poznań, 2008

[7] M. Szewczyk, *Statystyka - materiały do wykładów i ćwiczeń : Skrypt dla studentów*, Politechnika Opolska, Opole, 2007

[8] Cz. Domański, K. Pruska, *Nieklasyczne metody statystyczne*, Polskie Wydawnictwo Ekonomiczne, Warszawa, 2000

[9] P. J. Durka, *Wstęp do współczesnej statystyki*, Wydawnictwo Adamantan, Warszawa, 2003

[10] S. Brandt, *Analiza danych : metody statystyczne i obliczeniowe*, Wydawnictwo Naukowe PWN, Warszawa, 1998

[11] http://white.stanford.edu/ knk/Psych216A/