KURS STATYSTYKI
DLA STUDENTÓW KIERUNKÓW PRZYRODNICZYCH

UNIWERSYTETU TECHNOLOGICZNO-PRZYRODNICZEGO
W BYDGOSZCZY

Autorstwa
Anna Wenda-Piesik
Lech Gałęzewski

Bydgoszcz 2020

ROZDZIAŁ I **RACHUNEK PRAWDOPODOBIEŃSTA**

Spis treści

1.Pojęcia zdarzenia i przestrzeni zdarzeń

Rachunek prawdopodobieństwa jest działem matematyki zajmującym się badaniem prawidłowości w zakresie doświadczeń losowych, zwanych także zjawiskami przypadkowymi.

Doświadczenie losowe – to takie doświadczenie, które może być powtarzane wiele razy w tych samych warunkach i którego wyników nie można jednoznacznie przewidzieć (rzut monetą, rzut kostką sześcienną, losowanie toto-lotka, rozdanie kart w brydża, strzelanie do tarczy), doświadczenie losowe naukowe- pomiar określonej wielkości fizycznej, np. zawartości białka w nasionach zbóż.

Zdarzenie elementarne – pojęcie pierwotne (nie definiuje się go); jest to wynik (każdy z wyników) pewnego doświadczenia, zwykle takiego, w którym pewne właściwości tego wyniku nie są znane z góry. Wszystkie możliwe zdarzenia elementarne e_i tworzą zbiór zdarzeń elementarnych E – przestrzeń zdarzeń elementarnych.

Zdarzeniem losowym (zdarzeniem) nazywamy dowolny podzbiór A zbioru zdarzeń elementarnych. Zdarzenie losowe składa się zatem z pewnej liczby zdarzeń elementarnych. O zdarzeniach elementarnych składających się na zdarzenie A mówimy, że sprzyjają zdarzeniu A.

Szczególnym zdarzeniem losowym jest

1. zdarzenie niemożliwe, tzn. takie, któremu nie sprzyja żadne ze zdarzeń elementarnych (jest zbiorem pustym $\emptyset$)
2. zdarzenie pewne, tzn. takie, któremu sprzyjają wszystkie zdarzenia ze zbioru zdarzeń elementarnych E.
3. zdarzenia przeciwne. Dla każdego zdarzenia A zdarzenie E-A, będące dopełnieniem zdarzenia A do zdarzenia pełnego, nazywamy zdarzeniem przeciwnym do zdarzenia A i oznaczamy $\bar{A}$.
4. zbiory jednoelementowe, składające się z jednego zdarzenia elementarnego

Jeśli przestrzeń zdarzeń elementarnych E ma n elementów, to zdarzeń losowych jest 2^n (łącznie ze zdarzeniem pewnym i niemożliwym)

2.Borelowskie ciało zdarzeń i relacje pomiędzy zdarzeniami

Borelowskim ciałem (σ-ciałem) zdarzeń nazywamy zbiór S, do którego należą zdarzenia:

- zdarzenie pewne E, zdarzenie niemożliwe $\emptyset$ (ZBIORY NIEWŁAŚCIWE)

- oraz w którym dla każdych zdarzeń losowych A_1, A_2, ... należących do zbioru S należą do niego także zdarzenia:
- suma zdarzeń $A_1 \cup A_2$, - zdarzenie składające się z tych wszystkich zdarzeń elementarnych, które należą do co najmniej jednego ze zdarzeń A_1, A_2
- iloczyn zdarzeń $A_1 \cap A_2$, - zdarzenie składające się z tych wszystkich zdarzeń elementarnych, które należą do każdego ze zdarzeń A_1, A_2
- różnica zdarzeń A_1 - A_2. – zdarzenie składające się z tych wszystkich zdarzeń elementarnych, które należą do A_1 i nie należą do A_2

Inne relacje między zdarzeniami

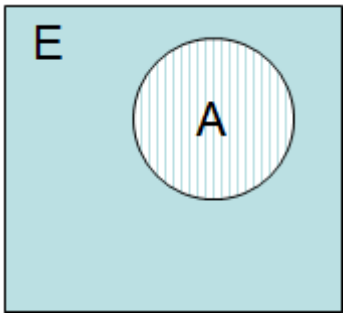
Zdarzenia przeciwne do zdarzenia A- zdarzenie składające się z tych wszystkich zdarzeń elementarnych, które nie należą do A (lecz należą do zbioru Ω), nazywamy je symbolem $\bar{A}$ i zachodzi ono wtedy, gdy nie zachodzi zdarzenie A.

Zdarzenie A_1 pociągające za sobą zdarzenie A_2 (implikujące) – jeśli każde zdarzenie elementarne należące do A_1 należy także do A_2. zapisujemy je w postaci $A_1 \subset A_2$.

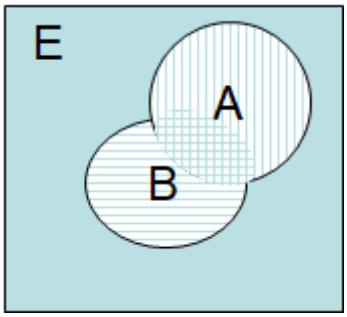
Wykluczające się zdarzenia A_1, A_2- jeśli nie mają one wspólnych zdarzeń elementarnych, tzn. iloczyn zdarzeń A_1 i A_2 jest zbiorem pustym $A_1 \cap A_2 = \varnothing$. Zdarzenia te wykluczają się, gdy nie mogą zajść łącznie.

3.Wykresy Eulera

Graficzna ilustracja działań na zdarzeniach przedstawiają wykresy Eulera, gdzie przestrzeń zdarzeń elementarnych E symbolizuje kwadrat a zdarzenia A lub B – koła w tym kwadracie.



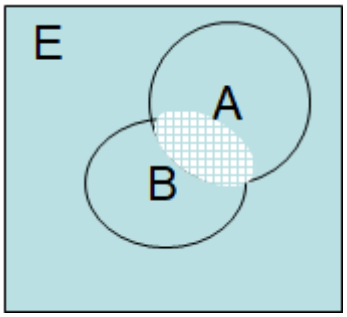
zdarzenie A w przestrzeni E



Sumowanie

$A \cup B$

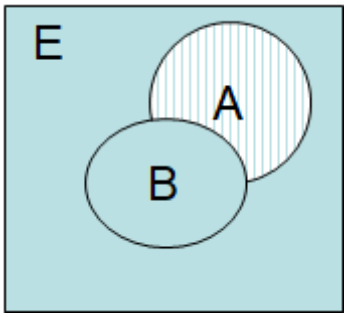
lub



Iloczyn

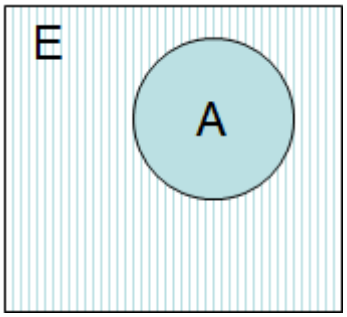
$A \cap B$

i



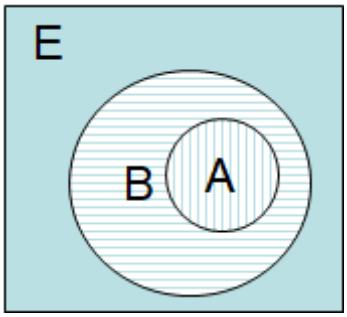
Różnica

$A - B$ (/)



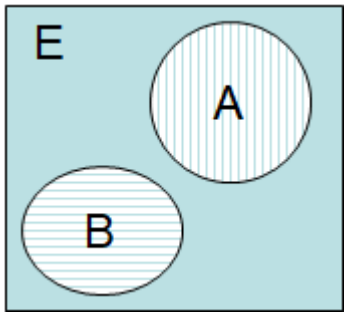
Zdarzenie przeciwne

$\overline{A}$



Implikacja

$A \subset B$



Zdarzenia wykluczające się

$A \cap B = \varnothing$

Jest przestrzeń składająca się z 2 zdarzeń elementarnych: $E = \{ e_1 \ e_2 \}$, może to być wyrzucenie orła O bądź reszki R na monecie. Ze zbioru można utworzyć $2^2 = 4$ zdarzenia :

$A_1 = \varnothing$

$A_2 = \{e_1\}$

$A_3 = \{e_2\}$

$A_4 = \{e_1, e_2\}$

Zbiór ten nazywamy CIAŁEM ZDARZEŃ „S”

Działania na tych zdarzeniach:

1. $A_1 \cap A_4 = \varnothing = A_1$

2. $A_2 \cap A_4 = A_2 = \{e_1\}$

3. $A_3 \cap A_4 = A_3 = \{e_2\}$

4. $A_2 \cup A_3 = A_4 = \{e_1, e_2\}$

5. $A_2 \cup A_4 = A_4 = \{e_1, e_2\}$

6. $A_1 \cup A_4 = A_4 = \{e_1, e_2\}$

7. $A_4 - A_3 = A_2 = \{e_1\}$

8. $A_2 - A_3 = A_2 \{e_1\}$

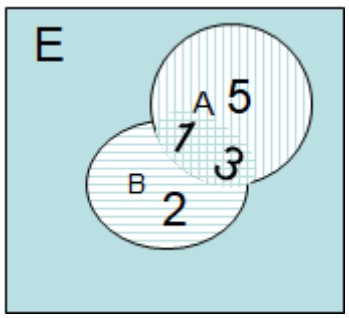
9. $\bar{A}_1 = A_4$

10. $\bar{A}_2 = A_3$

11. $A_1 \cap A_2 \cap A_3 = A_4 = \{e_1, e_2\}$

Ciałem zdarzeń nazywamy taki zbiór zdarzeń, w którym możliwe jest tworzenie sum, iloczynów, różnic, zdarzeń przeciwnych, pewnych i niemożliwych dla wszystkich zdarzeń należących do tego zbioru.

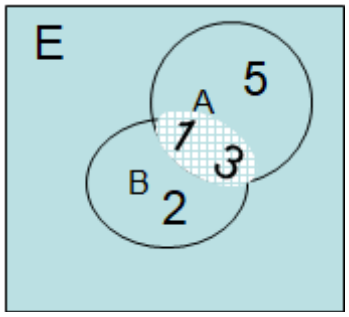
Przykład 1.
Doświadczenie – jednokrotny rzut kostką do gry E $\{ e_1,e_2,e_3,e_4,e_5,e_6\}$
zdarzenie A – liczba oczek nieparzysta $A\{e_1,e_3,e_5\}$
zdarzenie B – liczba oczek mniejsza od 4 $B\{e_1,e_2,e_3,\}$



Wynik sumowania jest zbiorem zawierającym zdarzenia elementarne zbiorów A lub B więc zbiór liczb nieparzystych lub mniejszych od 4

$A \cup B = \{ e_1,e_2,e_3,e_5\}$

Wynik mnożenia jest zbiorem zawierającym wspólne zdarzenia elementarne zbiorów A i B więc zbiór liczb nieparzystych i (jednocześnie) mniejszych od 4



$A \cap B = \{e_1, e_3,\}$

Przykład 2.
Doświadczenie – jednoczesny rzut dwiema monetami

Ciało zdarzeń dla zbioru: E $\{e_1, e_2, e_3\}$ co odpowiada: OO, RR, OR

Ze zbioru można utworzyć $2^3 = 8$ zdarzeń losowych:

A1 Ø, A2$\{e_1\}$, A3$\{e_2)$, A4$\{e_3\}$, A5$\{e_1, e_2\}$, A6$\{e_1, e_3\}$, A7$\{e_2, e_3\}$, A8$\{e_1, e_2, e_3\}$

Przykłady działań na tych zdarzeniach:

$A6 \cap A7 = A4$

$A4 \cap A5 = A1$

$A2 \cup A7 = A8$

$A5 \cup A6 = A8$

$A1 \cup A2 \cup A3 = A5$

$A6 - A2 = A4$

$A3 - A6 = A3$

$\overline{A7} = A2$

$\overline{A8} = A1$

$\overline{A1} = A8$

$A1 \cap A2 \cap A3 = A1$

Przykłady zadań na określanie zdarzeń losowych.

1) Student zalicza matematykę, fizykę i statystykę. Interesuje nas, które przedmioty zaliczy:
   a) Określ przestrzeń zdarzeń elementarnych
   b) Zapisz, jako podzbiory przestrzeni zdarzeń elementarnych następujące zdarzenia losowe: A – student zaliczył wszystkie przedmioty, B – student nie zaliczył tylko matematyki, C – student zaliczył tylko statystykę, D- student zaliczył dokładnie 2 przedmioty, E – student zaliczył co najmniej 2 przedmioty, F – student zaliczył fizykę, G – student zaliczył co najwyżej 2 przedmioty.
2) Z partii towaru zawierającej sztuki dobre i wadliwe wylosowano 3 sztuki towaru. Interesuje nas liczba wylosowanych sztuk dobrych.

a) Określ przestrzeń zdarzeń elementarnych

b) Zapisz, jako podzbiory przestrzeni zdarzeń elementarnych następujące zdarzenia losowe: A – wylosowano 3 sztuki dobre, B – wylosowano co najmniej jedną sztukę dobrą, C- wylosowano co najwyżej jedną sztukę dobrą.

c) Co oznaczają zdarzenia: A', B', C', $A \cup B$, $B \cup C$, $A \cap B$, $A \cap B' \cap C'$

3) Dwukrotnie strzelamy do celu. Interesuje nas, w którym strzale cel zostanie trafiony. Określamy zdarzenia: A – trafienie w 1 strzale, B – trafienie dokładnie raz, C – trafienie dokładnie dwa razy.

a) Rozpisz przestrzeń zdarzeń elementarnych

b) Zapisz, jako podzbiory przestrzeni zdarzeń elementarnych następujące zdarzenia losowe: A, B, C, A', $A \cup B$, $A \cap B$, $A \cap B \cap C$, $A' \cup B \cup C$

4.Kombinatoryka

<u>Zastosowanie wariacji z powtórzeniami</u>

Zadanie: ile 3.nutowych (k) melodii utworzyć ze zbioru (n) nut {c,d,e,f,g,a,h}

Jest istotna kolejność elementów, cde, ced, dce,…itd. są istotne - wchodzą do wyniku

Elementy zbioru mogą się powtarzać, np.{c,c,g}.

Wzór na wariacje z powtórzeniami:

$$W_n^k = n^k$$

$W_7^3 = 7^3$ Można utworzyć 343 różnych melodii

<u>Zastosowanie wariacji bez powtórzeń</u>

Zadanie: ile 3.nutowych (k) melodii utworzyć ze zbioru (n) nut {c,d,e,f,g,a,h), ale nuty nie mogą się powtarzać

Jest istotna kolejność elementów, cde, ced, dce,…itd. są istotne - wchodzą do wyniku

Elementy zbioru nie mogą się powtarzać np.{c,c,g}

Wzór na wariacje bez powtórzeń:

$$V_n^k = \frac{n!}{(n-k)!}$$

$$V_7^3 = \frac{7!}{(7-3)!}$$ Można utworzyć 210 różnych melodii

<u>Zastosowanie kombinacji</u>

Zadanie: ile 3.nutowych (k) melodii utworzyć ze zbioru (n) nut {c,d,e,f,g,a,h), raz użyta nuta nie może się powtarzać i nie jest ważna kolejność nut

Nie jest istotna kolejność elementów, cde, ced, dce,…itd. są nieistotne - wchodzą do wyniku jako 1 kombinacja

Elementy zbioru nie mogą się powtarzać np.{**c**,**c**,g}

Liczba kombinacji C^k_n wyrażana jest wzorem

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

$$C_7^3 = \frac{7!}{3!(7-3)!}$$ Można utworzyć 35 różnych melodii

<u>Zastosowanie permutacji</u>

Zadanie: ile można utworzyć melodii 7.nutowych ze zbioru 7 nut {c,d,e,f,g,a,h)

Jest istotna kolejność elementów

Elementy zbioru nie mogą się powtarzać

Liczba kombinacji P wyrażana jest wzorem P = n!

Mając siedem nut utworzymy 7! = 5040 różnych melodii

1) Ile liczb czterocyfrowych o niepowtarzających się cyfrach można otrzymać z cyfr : 0,1,3,5? Wypisać te liczby. (Użyj wzoru na permutacje i uwzględnij, że liczby nie mogą się zacząć od 0).
2) Dany jest zbiór 3 różnych cyfr {5,6,7}. Ile różnych liczb naturalnych 1 cyfrowych, dwucyfrowych i trzycyfrowych o niepowtarzających się cyfrach można utworzyć z elementów tego zbioru? Wymień te liczby. (Wzór na wariacje bez powtórzeń).
3) Gracz w brydża otrzymuje 13 kart spośród 52 kart. Ile jest możliwych rozdań, w których gracz otrzyma: a) dokładnie 10 kierów, b) 8 blotek? (Wzór na kombinację).
4) W Toto-lotku piłkarskim typuje się wyniki 13 meczów piłkarskich. Jeśli w danym meczu typuje się zwycięstwo gospodarzy, to należy do kuponu wpisać cyfrę 1, jeśli remis to wpisujemy X, jeśli zwycięstwo gości to cyfrę 2. Ile jest sposobów typowania? (wykorzystaj wzór na wariację z powtórzeniami).
5) W szpitalu zatrudnionych jest 8 lekarzy. Podczas dyżuru nocnego obecnych jest 4 lekarzy. Ile jest możliwych wariantów ustawienia dyżuru nocnego w tym szpitalu? (wykorzystaj wzór na kombinację).

5.Prawdopodobieństwo i jego własności

Jeżeli na zdarzenie pewne Ω składa się n jednakowo możliwych i wzajemnie się wykluczających zdarzeń elementarnych, spośród których m sprzyja zdarzeniu losowemu A, to prawdopodobieństwem zdarzenia A nazywamy liczbę $P(A) = n/k$.

Aksjomatyczna definicja prawdopodobieństwa:

Prawdopodobieństwem zdarzenia losowego A nazywamy liczbę P(A) przypisaną w sposób jednoznaczny dowolnemu zdarzeniu A i spełniającą warunki:

I. $0<P(A)<1$,

II. prawdopodobieństwo zdarzenia pewnego $P(\Omega)=1$,

III. prawdopodobieństwo sumy dowolnych, parami wykluczających się zdarzeń A_1, A_2, ... jest równe sumie ich prawdopodobieństw: $P(A_1 \cup A_2...) = P(A_1) + P(A_2) + ...$.

Własności prawdopodobieństwa:

1. $P(\emptyset)=0$ – prawdopodobieństwo zdarzenia niemożliwego jest równe zeru
2. jeśli zdarzenia A1...An wykluczają się parami, to prawdopodobieństwo sumy zdarzeń jest równe sumie ich prawdopodobieństw $P(A1 \cup A2 \cup A3... \cup An) = P(A1)+ P(A2)+...P(An)$
3. Jeśli zdarzenie A pociąga zdarzenie B $A \subset B$, to: $P(A) < P(B)$, $P(B-A) = P(B) – P(A)$

4. Prawdopodobieństwo sumy dwóch dowolnych zdarzeń jest równe sumie prawdopodobieństw tych zdarzeń zmniejszonej o prawdopodobieństwo ich iloczynu :

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

5. Prawdopodobieństwo zdarzenia A jest równe różnicy jedności i prawdopodobieństwa zdarzenia przeciwnego do A: $P(A) = 1- P(A')$

Prawdopodobieństwo P jest zatem funkcją $P: \Omega \rightarrow <0,1>$.

Przykłady zadań na wykorzystanie klasycznej definicji prawdopodobieństwa.

1. Dwukrotnie rzucamy kością do gry. Rozważ zdarzenia losowe: A – suma wyrzuconych oczek jest równa 6, B – przynajmniej w jednym rzucie wypadną 2 oczka. Oblicz prawdopodobieństwa zdarzeń: A, B, $A \cup B$, $A \cap B$, A – B
2. Trzy razy rzucamy monetą. Oblicz prawdopodobieństwo, że orzeł wypadnie:
   a) Dwa razy
   b) Co najmniej dwa razy
   c) Co najwyżej dwa razy
3. Z tali kart (52) wyciągnięto 1 kartę. Oblicz prawdopodobieństwo, że jest ona asem lub pikiem.

Bardzo często do wyliczania liczby k – zdarzeń sprzyjających i liczby n – ogólnej liczby zdarzeń posługujemy się wzorami kombinatorycznymi.

Zadanie: rzucono 5 razy kością do gry, określ P(A) że w każdym rzucie otrzymamy inną liczbę oczek. Przestrzeń zdarzeń elementarnych to E { $e_1, e_2, e_3, e_4, e_5, e_6$}

Należy rozważyć, według jakich wzorów kombinatorycznych da się wyliczyć liczbę k i n.

Skoro mamy 6 wariantów i rzucamy kością dwa razy, to liczba wszystkich możliwych wyników (n) będzie pochodziła od wariacji z powtórzeniami. Natomiast sprzyjająca liczba sukcesów k, że w każdym rzucie będzie inna liczba oczek podlega wyliczeniu według wariacji bez powtórzeń. Tabelka poniżej zamieszczona ma nam pomóc w zapamiętaniu warunków losowania.

Sposób losowania	Kolejność wyrazów	Wariant liczenia
bez zwracania (bez powtórzeń)	istotna	wariacja bez powtórzeń V
	nieistotna	kombinacja C

ze zwracaniem (z powtórzeniami)	istotna	wariacja z powtórzeniami W

$$P(A) = \frac{k}{n}$$

k – to liczba zdarzeń sprzyjających

$$V_6^5 = \frac{6!}{(6-5)!} = \frac{6!}{1!}$$

$$= 720$$

n – to liczba zdarzeń możliwych

$$W_6^5 = 6^5$$

$$= 7776$$

P(A) = 720/7776 = 0,093 Prawdopodobieństwo tego zdarzenia losowego wynosi niespełna 10% (9,30%).

9.Przykłady na zastosowanie kombinatoryki do obliczania prawdopodobieństwa według klasycznej definicji.

1) W urnie znajduje się 5 kul białych i 3 czerwone. Wyciągnięto losowo 2 kule. Jakie jest prawdopodobieństwo, że są to kule różnokolorowe?
2) Obliczyć i porównać prawdopodobieństwo osiągnięcia głównej wygranej w dwóch grach liczbowych:
   a) Duży- Lotek (trafienie 6 liczb z wylosowanych 6 spośród 49)
   b) Multi-Lotek (trafne skreślenie 10 spośród 20 premiowanych, wylosowanych z 80 liczb).
3) 20 osobowa grupa studencka, w której jest 12 studentek, otrzymała 5 biletów do kina. Bilety rozdziela się drogą losową. Jakie jest prawdopodobieństwo tego, że wśród posiadaczy biletów znajdzie się dokładnie dwóch studentów?
4) Student potrafi odpowiedzieć na 15 spośród 20 pytań. Oblicz prawdopodobieństwo tego, że student odpowie na 2 spośród wylosowanych 3 pytań.
5) Oblicz prawdopodobieństwo, że gracz w brydża wśród 13 kart otrzyma dokładnie 1 asa (w talii są 52 karty i 4 asy).

6.Przestrzeń probabilistyczna i własności prawdopodobieństwa

Trójkę (E,S,P) nazywamy przestrzenią probabilistyczną: E - przestrzeń zdarzeń elementarnych, S - ciało zdarzeń, oraz określone na tych zdarzeniach prawdopodobieństwo P.

Zadania na własności prawdopodobieństwa:

Prawdopodobieństwo sumy zdarzeń wykluczających się

Zadanie: Oblicz prawdopodobieństwo P(C) wyrzucenia kostką do gry albo A - 1oczko albo B- 2oczka

Jest to przykład na sumowanie prawdopodobieństwa (albo 1 albo 2). Są to zdarzenia niezależne **wykluczające** się, ponieważ jeśli wykulam jedynkę to nie wykulam dwójki. Prawdopodobieństwo sumy tych dwóch zdarzeń niezależnych, wykluczających się obliczymy:

$$P(C) = P(A \cup B) \qquad P(C) = \frac{1}{6} + \frac{1}{6}$$

Dwa zdarzenia A i B nazywa się niezależnymi, jeżeli zajście jednego z nich nie ma wpływu na zajście drugiego zdarzenia, tzn. P(A) = P(A|B) oraz P(B) = P(B|A).

Prawdopodobieństwo iloczynu dwóch zdarzeń: Jeżeli zdarzenia A i B są zdarzeniami niezależnymi, to $P(A \cap B) = P(A) \times P(B)$.

Przykład. Oblicz prawdopodobieństwa wyrzucenia dwóch jedynek w dwóch kolejnych rzutach kością do gry.

P(A) = 1/6, P(B) =1/6

$P(A \cap B) = P(A) \times P(B) = 1/6 \times 1/6 = 1/36$.

Prawdopodobieństwo zdarzenia B w sytuacji, gdy zaszło zdarzenie A nazywamy prawdopodobieństwem warunkowym zdarzenia B i oznaczamy P(B|A).

Zadanie: w urnie jest 10 kul – 7 białych i 3 czarne

A – wylosowanie kuli białej w pierwszym ciągnięciu

B– wylosowanie kuli białej w drugim ciągnięciu

Jakie jest prawdopodobieństwo wylosowania kuli białej w pierwszym ciągnięciu?

$$P(A) = \frac{7}{10}$$

Jakie jest prawdopodobieństwo wylosowania kuli białej w drugim ciągnięciu jeśli wrzuciliśmy z powrotem wylosowaną wcześniej kulę – losowanie z zwracaniem?

$P(B) = \frac{7}{10}$ Zdarzenia niezależne $\Rightarrow P(A) = P(B)$

Jakie jest prawdopodobieństwo wylosowania kuli białej w drugim ciągnięciu jeśli nie wrzuciliśmy z powrotem wylosowanej wcześniej kuli – losowanie bez zwracania?

1) w pierwszym ciągnięciu wylosowano kulę białą: $P(B) = \frac{6}{9}$

2) w pierwszym ciągnięciu wylosowano kulę czarną $P(B) = \frac{7}{9}$

Są to zdarzenia zależne (warunkowe, względne) $\Rightarrow P(A) \neq P(B)$

Prawdopodobieństwo zdarzenia A przy założeniu zajścia zdarzenia B $P(A|B)$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B) > 0$$

Prawdopodobieństwo zdarzenia B przy założeniu zajścia zdarzenia A $P(B|A)$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad P(A) > 0$$

Jakie jest prawdopodobieństwo zdarzenia C
P(C) wylosowania w pierwszym i drugim rzucie kuli białej?

$$P(A \cap B) = P(A) \times P(B|A) = P(B) \times P(A|B)$$

$$P(C) = \frac{7}{10} \times \frac{6}{9} = 0{,}47$$

Zadanie: wylicz prawdopodobieństwo P(C) wyciągnięcia dwóch asów z rzędu z tali 52 kart

A – as w pierwszym ciągnięciu

B – as w drugim ciągnięciu

$$P(A) = \frac{4}{52} = 0{,}076923$$

$$P(B) = \frac{3}{51} = 0{,}058824$$

$$P(A \cap B) = P(A) \times P(B|A) = 0{,}004525$$

Wzór można uogólnić na większą (dowolną, skończoną) liczbę zdarzeń zależnych:

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1) \times P(A_2|A_1) \times P(A_3|A_1 \cap A_2) \times \ldots \times P(A_n|A_1 \cap A_2 \cap \ldots A_{n-1})$$

Wzór dla trzech zdarzeń zależnych:

$$P(A \cap B \cap C) = P(A) \times P(B|A) \times P(C|A \cap B)$$

Przykład do wykonania. Produkt przechodzi kolejno przez 3 próby kontrolne i jest odrzucany po wykryciu wady w dowolnej próbie. Prawdopodobieństwo odrzucenia produktu w 1 próbie wynosi 0,1 w drugiej, jeśli przeszedł pierwszą 0,3, i w trzeciej, jeśli przeszedł poprzednie 0,2. Oblicz prawdopodobieństwo, że produkt przejdzie przez trzy próby.

Zdarzenia niezależne zespołowo

Zadanie: Wylosowałem zestaw kopert z pytaniami ze statystyki a na każde z nich do wyboru 4 odpowiedzi a,b,c,d poprawna jest tylko jedna. Nie mam zielonego pojęcia o statystyce więc będę strzelał. Jakie jest prawdopodobieństwo D, że odpowiem poprawnie na trzy pytania z rzędu i P(E), że nie odpowiem na trzy pytania z rzędu i P(F) co najmniej 1 raz odpowiem dobrze.

P(A) – poprawna odpowiedź na pierwsze pytanie $P(A) = \frac{1}{4}$ $P(D) = \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} =$ 0,016

P(B) – poprawna odpowiedź na drugie pytanie $P(B) = \frac{1}{4}$

P(C) – poprawna odpowiedź na trzecie pytanie $P(C) = \frac{1}{4}$

Prawdopodobieństwo łącznego zajścia n zdarzeń niezależnych zespołowo jest równe iloczynowi prawdopodobieństwa tych zdarzeń

$P(\overline{A})$ – błędna odpowiedź na pierwsze pytanie $P(\overline{A}) = 1 - P(A) = 0{,}75$

$P(\overline{B})$ – błędna odpowiedź na drugie pytanie $P(\overline{B}) = 1 - P(B) = 0{,}75$

$P(\overline{C})$ – błędna odpowiedź na trzecie pytanie $P(\overline{C}) = 1 - P(C) = 0{,}75$

$$P(E) = 0{,}75 \times 0{,}75 \times 0{,}75 = 0{,}42$$

Prawdopodobieństwo P(F) zajścia co najmniej 1 z niezależnych zespołowo zdarzeń A,B,C równe jest różnicy jedności i iloczynu prawdopodobieństwa zdarzeń przeciwnych

$$P(F) = 1 - P(\overline{A}) \times P(\overline{B}) \times P(\overline{C})$$

Jeśli $P(\overline{A_1}) = q_1$ **a** $q_1 = q_2 = q_n$ **to** $P(F) = 1 - q^n$

$$P(F) = 1 - 0{,}75 \times 0{,}75 \times 0{,}75 = 1 - 0{,}75^3 = 1 - 0{,}422 = 0{,}578$$

Przykład do wykonania. Prawdopodobieństwo pomyślnego wykonania ćwiczeń przez jednego sportowca wynosi 0,6. Dwaj sportowcy wykonują to ćwiczenie kolejno, każdy z nich po 2 razy. Sportowiec, który pierwszy pomyślnie wykona to ćwiczenia otrzyma nagrodę (zdarzenie A). Znaleźć prawdopodobieństwo otrzymania nagrody przez sporowców.

Prawdopodobieństwo sumy dwóch dowolnych zdarzeń:

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Zadanie: Egzamin z statystyki (A) zdają 7 na 10 studentów, natomiast egzamin z biochemii (B) zaledwie 4 na 10 studentów. Oblicz prawdopodobieństwo zdania obu egzaminów (C) oraz zdania albo jednego albo drugiego (D). Zdarzenia niezależne ale rozłączne gdyż nie ma powiązania miedzy jednym a drugim egzaminem.

$$P(A) = 0{,}7 \qquad P(B) = 0{,}4 \qquad P(C) = 0{,}7 \times 0{,}4$$

Zdarzenia się nie wykluczają prawdopodobieństwo sumy dwóch zdarzeń A i B niewykluczających się jest równie sumie prawdopodobieństwa tych zdarzeń pomniejszonej o prawdopodobieństwo ich iloczynu.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(D) = 0{,}7 + 0{,}4 - 0{,}7 \times 0{,}4 = 0{,}82$$

Zadanie: oblicz P(C), że wylosowana z 52 karta jest Asem (A) albo treflem (B) – nie wkluczają bo karta może być asem a może być trefl ale nie jest rozłączne bo może być asem trefl.

$$P(A) = \frac{4}{52} \qquad P(B) = \frac{13}{52} \qquad P(C) = \frac{13}{52} + \frac{4}{52} - \frac{1}{13}$$

1) W pierwszej urnie są dwa losy wygrywające i osiem przegrywających, w drugiej urnie – cztery wygrywające i sześć przegrywających. Rzucamy kostką do gry. Jeśli wypadnie liczba oczek podzielna przez 3 losujemy z 1 urny, w przeciwnym razie losujemy z 2 urny. Obliczyć prawdopodobieństwo tego, że wyciągnięty los jest wygrywający.

2) Oblicz prawdopodobieństwo tego, że rzucając trzykrotnie kostką do gry wyrzucimy za pierwszym, za drugim i za trzecim razem szóstkę.

3) 40% Polaków to blondyni. Jakie jest prawdopodobieństwo, że w sześcioosobowej rodzinie wszyscy są blondynami?

4) Na osiedlu znajdują się dwa sklepy spożywcze. Prawdopodobieństwo zamknięcia każdego z nich wynosi 0,5. Jakie jest prawdopodobieństwo tego, że przynajmniej jeden z nich będzie otwarty? Jakie jest prawdopodobieństwo, że oba sklepy będą otwarte?

5) W każdym z trzech pudełek znajduje się po 10 detali W pierwszym pudełku jest 8, w drugim 7, a w trzecim 9 detali standardowych. Z każdego pudełka pobieramy losowo po 1 detalu. Obliczyć prawdopodobieństwo, że wszystkie 3 detale będą standardowe.

6) Dwóch strzelców strzeliło do jednego celu. Pierwszy trafia do celu z prawdopodobieństwem 0,9, zaś drugi z prawdopodobieństwem 0,8. Obliczyć prawdopodobieństwo, że cel został trafiony co najmniej raz.

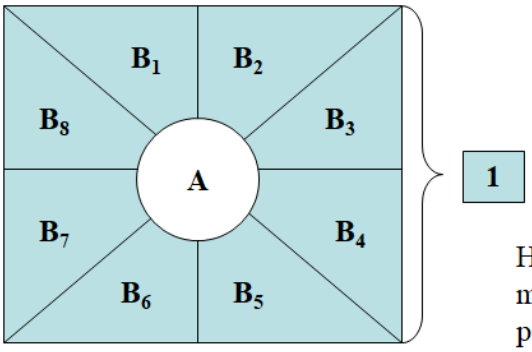
7) Robotnik obsługuje 3 maszyny. Prawdopodobieństwo, że w czasie T maszyny nie wymagają obsługi wynosi: 0,9 dla pierwszej, 0,8 dla drugiej i 0,85 dla trzeciej. Maszyny

te pracują niezależnie od siebie. Oblicz prawdopodobieństwo, że w czasie T: a) żadna z maszyn nie wymaga obsługi, b) wszystkie maszyny wymagają obsługi.

8) Prawdopodobieństwo spotkania w parku psa na smyczy wynosi 0,8 a kota na smyczy 0,1. Jakie jest prawdopodobieństwo spotkania w parku pupila na smyczy?

7.Prawdopodobieństwo całkowite i wzór Bayes'a

Na zaistnienie zdarzenia A ma wpływ wiele przyczyn zdarzeń wzajemnie wykluczających się B



A - skutek

$B_{1\text{-}n}$ – przyczyny, inaczej hipotezy

Hipoteza to przyczyna, która może ale nie musi wywołać danego skutku – zachodzi z pewnym prawdopodobieństwem.

Jeśli wystąpił dany skutek (zdarzenie A) to suma prawdopodobieństw wszystkich hipotez B_1-B_n, które mogły ją wywołać = 1

Stąd wzór na **prawdopodobieństwo całkowite P(A)** zdarzenia skutkowego A ma postać:

$$P(A) = P(B_1) \times P(A|B_1) + \ldots + P(B_n) \times P(A|B_n) = \sum_{i=1}^{n} P(B_i) \times P(A|B_i)$$

Zadanie. Sporządzamy mieszankę nasion B_1 - jęczmienia i B_2 - owsa o równym udziale ziarniaków obu komponentów. Oznaczono zdolność kiełkowania w laboratorium dla (A/B_1) jęczmienia na 90% a dla (A/B_2) owsa na 96%.

Prawdopodobieństwo całkowite P(A) wzejścia nasion

$$P(A) = 0{,}5 \times 0{,}9 + 0{,}5 \times 0{,}96 = 0{,}93$$

Prawdopodobieństwa A/B_1 i A/B_2 znane są góry ***a priori***

Prawdopodobieństwa hipotez $P(B_1)$ i $P(B_2)$ oraz prawdopodobieństwa warunkowe $P(A/B_1)$ i $P(A/B_2)$ są to prawdopodobieństwa z nadania, aprioryczne. Uznajemy, że są nam znane przed doświadczeniem (przed faktem wykiełkowania nasion). Prawdopodobieństwo całkowite P(A) czyli prawdopodobieństwo kiełkowalności wynosi 0,93.

Jeśli po fakcie, po doświadczeniu wykiełkowania nasion, patrzymy na siewki i zastanowimy się, jakie jest prawdopodobieństwo, że obserwowana siewka jest jęczmieniem, a jakie że jest owsem? To tak, jakbyśmy odwrócili zdarzenie przyczynowe ze skutkowym. Chodzi teraz o prawdopodobieństwo aposterioryczne, czyli po fakcie, inaczej mówiąc po doświadczeniu.

Wzór Bayes'a

Jeżeli zdarzenie A zawiera się w sumie zdarzeń B_1, B_2, ..., B_n parami wyłączających się, to:

$$P(B_1 / A) = \frac{P(B_1) \times P(A|B_1)}{P(A)}$$

Prawdopodobieństwa szukane ***a posteriori***

.

prawdopodobieństwo, że siewka jest jęczmieniem

$$P(B_1|A) = \frac{0{,}9 \times 0{,}5}{0{,}93} = 0{,}48$$

prawdopodobieństwo, że siewka jest owsem

$$P(B_2|A) = \frac{0{,}96 \times 0{,}5}{0{,}93} = 0{,}52$$

Przykłady na zastosowanie prawdopodobieństwa całkowitego i wzoru Bayes'a.

1) Zakłady metalowe kooperują z trzema odlewniami. Z poszczególnych odlewni pochodzi odpowiednio: 10%, 30% i 60% odlewów. Z założenia (a priori) wiadomo, że odlewy dostarczane z pierwszej odlewni zawierają 2% wad, z drugiej – 10%, a z trzeciej -4%. Stwierdzono, że pewien odlew posiada wadę ukrytą. Z której odlewni najprawdopodobniej on pochodzi?

2) W magazynie znajdują się żarówki pochodzące z dwóch fabryk. 60% pochodzi z fabryki I. Wśród żarówek z fabryki I jest 1% wadliwych, a z pośród żarówek z II fabryki 2% wadliwych. Z magazynu pobrano losowo 1 żarówkę, która okazała się wadliwa. Jakie jest prawdopodobieństwo tego, że ta żarówka pochodziła z II fabryki?

3) Spośród 100 mężczyzn 5 nie rozróżnia kolorów, a spośród 10000 kobiet 25 to daltonistki. Z grupy o jednakowej liczbie mężczyzn i kobiet wybrano osobę, która okazała się dotknięta tą wadą wzroku. Jakie jest prawdopodobieństwo, że wylosowana osoba jest mężczyzna?

4) Wysyłany jest sygnał binarny 0 lub 1. Prawdopodobieństwo wysłania sygnału 0 wynosi 0,3, zaś sygnału 1 wynosi 0,7. Prawdopodobieństwo zniekształcenia sygnału 0 wynosi 0,4 a sygnału 1 wynosi 0,2. A) Oblicz prawdopodobieństwo, że wysłany sygnał został zniekształcony. B). Wiadomo, że sygnał został zniekształcony oblicz, że był to sygnał 1.

5) Fabryka samochodów kooperuje z czterema producentami uszczelek silnikowych. Ich udziały w zaopatrzeniu fabryki wynoszą: 25%, 25%, 40% i 10%. Kontrola jakości uszczelek wykazała następujące odsetki wybrakowanej produkcji: I producent -5%, II producent – 3%, III producent – 2% i IV producent – 6%. Właściciel zakupionego samochodu złożył reklamację z powodu wady uszczelki. A) Który z kooperantów jest najbardziej prawdopodobnym producentem wadliwej uszczelki? B) Który z kooperantów jest najmniej prawdopodobnym producentem wadliwej uszczelki?

ROZDZIAŁ II **STATYSTYKA OPISOWA**

Spis treści

1.Dokładność pomiarów i dokładność liczbowa

Dokładność inaczej czułość, jest własnością przyrządu, którym posługuje się badacz w czasie pomiaru. Czynność ta wiąże się z zastosowaniem przyrządu, którym może być prosty przymiar (np. linijka o długości 50cm) lub skomplikowana aparatura (np. chromatograf gazowy). Zadaniem mierzenia jest uzyskiwanie wyników, tj. wartości liczbowych opisujących cechę ilościową badanego przedmiotu. Ponieważ liczby są swoistymi komunikatorami dla badacza odnośnie wartości interesującej go cechy, należy umiejętnie się nimi posługiwać, już na etapie zbierania danych. Nazywamy je wówczas **pierwotnym materiałem liczbowym, albo danymi źródłowymi**.

Wartości bezpośrednich pomiarów cech ilościowych nazywamy **liczbami absolutnymi**. Towarzyszą im jednostki miar podstawowych, zawarte w Międzynarodowym Układzie Jednostek Miar (w skrócie Układ SI) lub jednostki spoza układu SI uznawane w biometrii za jednostki legalne (tabela 1 i 2).

Tabela 1. Podstawowe jednostki SI

Wielkość		Jednostka	
nazwa	Symbol	nazwa	Symbol
Długość	l – długość b – szerokość h – wysokość	Metr	m
Masa	M(M)	Kilogram	kg
Czas	t (T)	Sekunda	s
Objętość	V	metr sześcienny	m^3

Tabela 2. Legalne jednostki miar, nie należące do układu SI

Nazwa wielkości	Jednostka		Jednostka podstawowa.
	Nazwa	symbol	
Masa	Tona decytona	t dt	1000kg 100kg
Czas	Minuta godzina	min h	60s 3600s
Powierzchnia	Hektar ar	ha a	$1000m^2$ $100m^2$
Objętość, pojemność	Litr	L	dm^3
Temperatura	Stopień Celsjusza	^0C	273^0K

Jednym z podstawowych przyrządów w laboratorium biometrycznym jest elektroniczna waga laboratoryjna. Posługujemy się nią, jeśli na przykład chcemy dowiedzieć się, jaką masę ma odliczonych 1000 ziarniaków pszenicy jarej. Na tabliczce znamionowej wagi znajdziemy kilka informacji o tym przyrządzie. Najpierw o zakresie możliwych pomiarów: Min = 2g, Max = 2000g. Oznacza to, że masa nasion cięższych od 2kg nie zostanie wyświetlona (wyświetlacz pokaże „błąd”), natomiast masa lżejszych od 2g będzie wynikiem niepewnym (mruganie na wyświetlaczu). Ponadto jest informacja, która mówi o

dopuszczalnym błędzie pomiaru (e = 1g). Oznacza to, że producent zastrzega, iż odczyty na tym przyrządzie mogą się różnić od odczytów na innych wagach o ±1g.

Następnie jest podana **dokładność**, z jaką pomiar zostanie wykonany (wyświetlony) d = 0,1g. Mówimy, że jest to właśnie czułość tej wagi. Oznacza to, że jeśli masa 1000 nasion będzie wynosiła 51g, to my ją odczytamy na wadze np. jako 51,2g. Ale, czy taka dokładność wystarczy, czy może powinniśmy posłużyć się wagą, która mierzy z dokładnością do 0,01g? Uzyskalibyśmy wówczas wyniki np. 51,15g lub 51,25g. Jeśli nie mamy pewności, co do tego, z jaką dokładnością powinniśmy dokonywać pomiarów, to stosujemy zasadę, że liczba jednostek pomiędzy wartością największą a najmniejszą w naszych pomiarach powinna się mieścić w przedziale od 30 do 300.

W naszym przykładzie masa 1000 nasion pszenicy jarej mieści się w zakresie 45 – 54g, to znaczy że jednostek wyrażonych w gramach od wartości największej do najmniejszej jest 9, a to w myśl powyższej zasady zbyt mała dokładność. Jeśli więc posłużymy się ową wagą z dokładnością do 0,1g, to otrzymamy zakres np. 45,1 – 54,1g, a to daje nam 90 jednostek różnicy wyrażonej w 0,1g. Jest to więc wystarczająca dokładność. Mówimy, że uzyskaliśmy wartość liczbową, z **dokładnością** do jednego miejsca po przecinku, z trzema miejscami znaczącymi. W biometrii często stosujemy zasadę dokonywania pomiarów w taki sposób, aby liczba wynikająca z pomiaru miała w zapisie 3 cyfry znaczące. W rozważanym tu przykładzie dokładność masy 1000 nasion pszenicy nie musi być wyrażona liczbą z dwoma miejscami po przecinku (0,01g). Z kolei masy 1000 nasion grochu, która mieści się w zakresie 120 - 230g, nie potrzebujemy ważyć z dokładnością do 0,1g, bowiem liczba jednostek pomiędzy wartością największą i najmniejszą wynosi 110.

Dla przypomnienia, cyfry od 1 do 9 są zawsze znaczące, natomiast cyfra 0 jest znacząca w zależności od pozycji w liczbie; np. w liczbie 50,4 są 3 cyfry znaczące (5,0,4) i zero przed przecinkiem jest znaczące, natomiast liczba 0,04 ma 1 cyfrę znaczącą (4) na drugim miejscu po przecinku, a zera przed nią są nieznaczące. Zatem zer początkowych, ani zer końcowych napisanych w wyniku zaokrąglenia lub w celu zapełnienia miejsca nie zaliczamy do cyfr znaczących. Zaokrąglenie do N cyfr znaczących polega na takim zaokrągleniu liczby, aby w efekcie miała N cyfr znaczących.

Dla przykładu zaokrąglenie do 3 cyfr znaczących liczby 10,08 ≈ 10,1, natomiast do 2 cyfr znaczących ≈ 10 (otrzymaliśmy jedną dziesiątkę znaczącą, „0” nie jest tu znaczące).

Dokładność pomiarów zależy od możliwości aparatury, ale także od celu badań i zasad przestrzeganych w danej dyscyplinie naukowej. Rozpatrzmy tę kwestię na podstawie wielkości fizycznej zwanej „długością”.

W badaniach z fitopatologii, czasami potrzebne są pomiary długości zarodników w celu rozpoznania patogena. Pomiarów tych dokonuje się za pomocą mikroskopu z podziałką, gdzie jednostką jest µm – mikrometr, czyli 0,001mm (10^{-6}m), a dokładność skali wynosi 0,1µm (tj. 100nm). Jeśli wiemy, że długość zarodników dla rdzy karłowej jęczmienia mieści się w zakresie 19 – 22µm, to wyniki naszych pomiarów zapisujemy jako np. 20,3µm ewentualnie możemy zapisać 20,35µm, jeśli chcemy cyfrę 5 uznać za niepewną, bo widzimy „na oko", że długość jest pomiędzy 20,3 a 20,4µm. Nie należy jednak stosować jednostek zbyt małych lub zbyt dużych. W tym przykładzie byłyby to zapisy 20 350nm lub 0,02035mm. Pierwszy zapis ma za dużo cyfr znaczących (5), drugi zaś za dużą dokładność po przecinku (5 cyfr, z czego 4 znaczące). Dlatego też stosuje się odpowiednie jednostki podwielokrotne lub wielokrotne dla jednostek podstawowych (tabela 3).

Tabela 3. Przedrostki i symbole do tworzenia jednostek wielokrotnych i podwielokrotnych

Nazwa mnożnika	Mnożnik	Przedrostek	
		nazwa	Symbol
Bilion	10^{12}= 1000 000 000 000	tera	T
Miliard	10^{9} = 1000 000 000	giga	G
Milion	10^{6} = 1000 000	mega	M
Tysiąc	10^{3} = 1000	kilo	K
Sto	10^{2} = 100	hekto	H
Dziesięć	10^{1} = 10	deka	Da
jednostka	-	-	-
Dziesiąta	10^{-1} = 0,1	decy	D
Setna	10^{-2} = 0,01	centy	C
Tysięczna	10^{-3} = 0,001	mili	M
Milionowa	10^{-6} = 0,000 001	mikro	µ
Miliardowa	10^{-9} = 0,000 000 001	nano	N

Pomiary plonów płodów rolnych (nasion zbóż, bulw ziemniaka, korzeni buraka itp.) dokonujemy tak, aby móc je zapisać w postaci decyton np. 43,2 [dt x ha^{-1}], lub tej samej wartości w tonach 4,32 [t x ha^{-1}] dla ziarna zbóż, a w przypadku korzeni marchwi w postaci np. 851 [dt x ha^{-1}] lub 85,1 [t x ha^{-1}]. Ważna jest więc kwestia doboru przyrządu do mierzenia masy plonu z powierzchni mniejszych niż 1 ha. Na przykład z poletka doświadczalnego o powierzchni 25 m^2 zebrano rzepak ozimy i wymłócono jego nasiona. Ich masa mieści się w zakresie 5-9kg. Wiemy, że odczyt powinien mieć 3 cyfry znaczące, więc sięgamy po wagę, która waży z dokładnością do 0,01kg. Otrzymany wynik np. 7,28kg, który przeliczymy na tony z hektara w następujący sposób:

Plon [t x ha^{-1}] = $\frac{10}{25} \times 7{,}28 = 2{,}912$. Ten wynik jest jednak dla nas „zbyt" dokładny, więc zaokrąglamy go do setnych wartości tony 2,912 ≈ 2,91 t x ha^{-1}.

Posługując się danymi pierwotnymi wykonujemy szereg obliczeń, np. liczymy średnią arytmetyczną, odchylenie przeciętne lub odchylenie standardowe. Przetworzenie wyników uzyskanych z pomiarów prowadzi do uzyskania **wtórnej informacji liczbowej.** Te wartości przeliczone podajemy z dokładnością o jeden rząd większą niż dokładność pomiaru – mówimy tu o **dokładności prezentacji wyników**.

Dokładność prezentacji wyników wiąże się z zasadami zaokrąglania liczb.

Plony pszenicy jarej pewnej odmiany badano w 2006 i 2007 roku w trzech stacjach doświadczalnych. Należy więc zaprezentować średni plon tej odmiany pszenicy w badanych latach:

Tabela 4. Zasady prezentacji średnich po zaokrągleniu

Rok	Wyniki pomiarów plonów [t x ha^{-1}]	Średnia arytmetyczna	Prezentacja średnich plonów [t x ha^{-1}]
2006	4,31 4,32 4,35	$\bar{x} = \frac{4{,}31 + 4{,}32 + 4{,}35}{3} = \frac{12{,}98}{3} = 4{,}32(6)$	4,327
2007	4,42 4,15 4,31	$\bar{x} = \frac{4{,}42 + 4{,}15 + 4{,}31}{3} = \frac{12{,}88}{3} = 4{,}29(3)$	4,293

Średnie w obydwu latach otrzymaliśmy jako wartości w postaci ułamka dziesiętnego nieskończonego okresowego. Dla danych z 2006 roku średnią z okresem (6), przedstawimy z dokładnością do 3 miejsc po przecinku, zaokrąglając liczbę 4,32(6) do 4,327. Zastosowano tutaj zasadę, która mówi, że jeśli za cyfrą, do której należy zaokrąglić (tutaj cyfra 6) jest cyfra większa od 5 tzn. 6,7,8 lub 9 (tutaj cyfra 6), to zaokrąglamy tę cyfrę o 1 w górę. Dla danych z 2007 roku średnią wynoszącą 4,29(3) po zaokrągleniu do trzech miejsc po przecinku zapiszemy jako 4,293. Jeśli bowiem po cyfrze, do której zaokrąglamy (tutaj cyfra 3) jest cyfra mniejsza od 5, tzn. 0,1,2,3 lub 4 (tutaj 3), to zostawiamy tę cyfrę bez zmian (tabela 4).

A jak należy postąpić w zaokrąglaniu wartości liczbowych, jeśli po cyfrze do której zaokrąglamy jest cyfra 5? Rozważmy to na innym przykładzie.

Wykonano 2 serie pomiarów liczby dni, w których rośliny chryzantem utrzymywały się w fazie kwitnienia:

I seria: 15, 16, 18, 20 $\bar{x} = \frac{69}{4} = 17{,}25$ [dni]

II seria: 21, 17, 18, 19 $\bar{x} = \frac{75}{4} = 18{,}75$ [dni]

Średnia arytmetyczna I serii wynosi 17,25, a nam potrzebna jest dokładność o jeden rząd większa od wartości pomiarów (do jednego miejsca po przecinku), więc zaokrąglamy ją do 17,2 dni. Średnią II serii, która wyniosła 18,75, zaokrąglimy do liczby 18,8 dni. Przyjmujemy tu zasadę, że jeśli po cyfrze, do której zaokrąglamy jest 5 a po 5 nie ma żadnej

innej cyfry znaczącej, to cyfry parzyste (0,2,4,6,8) zostawiamy bez zmian, a cyfry nieparzyste (1,3,5,7,9) zaokrąglamy o 1 w górę. Jeśli natomiast po cyfrze, do której zaokrąglamy jest cyfra 5, a po niej na dowolnym jeszcze miejscu będzie cyfra znacząca (większa od 0), to cyfrę tę zaokrąglamy o jeden w górę, np. liczbę 124,503 chcemy zaokrąglić do liczby z 3 cyframi znaczącymi, to otrzymamy liczbę całkowitą 125.

W pracy badawczej zagadnienia dokładności pomiarów wiążą się także z pojęciem **precyzji pomiarów**. O ile w mowie potocznej te dwa słowa są używane zamiennie, np. mówimy, że praca jubilera jest dokładna lub precyzyjna, przez co wyrażamy w jaki sposób rzemieślnik ten wykonuje bardzo drobne detale w zdobieniach, to w pomiarach naukowych precyzję i dokładność rozumiemy na zasadzie przeciwstawień. Precyzja bowiem w badaniach oznacza powtarzalność otrzymywanych wyników z pomiarów. Otrzymywanie niewielkich różnic w wielokrotnych pomiarach jednej próby tym samym przyrządem i w takich samych warunkach przy danej dokładności pomiaru świadczy o dużej precyzji. Stąd, im większa dokładność pomiarów, tym mniejsza ich precyzja i odwrotnie. Nie należy stosować zbyt czułych przyrządów (hiperdokładnych), skoro nie będą dostarczały nam wyników precyzyjnych.

Pozostaje jeszcze odpowiedzieć sobie na pytanie: Czy pomiar zawsze musi się wiązać z zastosowaniem jakiegoś przyrządu, np. wagi?
Otóż nie koniecznie, w badaniach rolniczych czynność polegająca na obserwacji za pomocą wzroku, bez zastosowania sprzętu, jest także pomiarem. Może to być zliczanie, tak jak w przykładzie liczby dni, w których kwitły chryzantemy. Otrzymujemy wówczas wartości ze zbioru liczb naturalnych, które nazywamy realizacjami zmiennej skokowej.

2.Skale pomiarowe

Przedmiotem badania statystycznego jest zbiorowość statystyczna, nazywana inaczej populacją generalną. Weźmy jako przykład przedmiotu badań gatunek pszenica jara (*Triticum aestivum* sp. *vulgare)*. Zbiorowość składa się z **jednostek statystycznych,** które w badaniach rolniczych nazywamy osobnikami lub pojedynkami (np. pojedyncze rośliny pszenicy są badane pod względem wydzielania związków lotnych z zielonego liścia). Specyfiką doświadczeń rolniczych jest stosowanie tzw. **jednostki zbiorczej**, np. poletka czy wazonu, na których uprawianych jest kilkadziesiąt a nawet kilka tysięcy roślin przedmiotu badań, i to oznacza dla nas jednostkę statystyczną badaną na przykład ze względu na wielkość plonu nasion.

Badacza interesują określone własności przedmiotu badań (odmiana, typ użytkowy, wysokość pędu, zawartość % białka w nasionach, plon nasion). W zależności od zakresu prowadzonych badań, mówimy o populacjach jednocechowych (np. badamy tylko plon nasion), dwucechowych (plon nasion i zawartość % białka) oraz wielocechowych, jeśli w zakresie badań jest więcej niż dwie **własności, (syn. cechy, zmienne)** przedmiotu.

Wartości cech przypisywane są jednostkom statystycznym w wyniku pomiaru określonej własności. Własność mierzoną określa się za pomocą zestawu relacji empirycznych pomiędzy jednostkami, którym ona przysługuje. Sposób wyrażenia cechy zależy od konstrukcji **skali pomiarowej.** Ważnym jest zagadnienie jednoznaczności reprezentacji, które mówi, że w danej skali pomiarowej, w sposób mniej lub bardziej jednoznacznie określony, są przyporządkowane liczby jednostkom statystycznym. W zależności od rozstrzygnięcia tego zagadnienia ustala się, jakie relacje między liczbami są w danej skali spełnione oraz jakie przekształcenia liczbowe są dopuszczalne w danej skali pomiarowej. Im więcej relacji między liczbami jest w danej skali spełnionych a mniej przekształceń dopuszcza skala pomiarowa, tym jest ona mocniejsza. Skale mocne posiadają własności skal słabszych, a ponadto dodatkowe pewne własności. Prześledzimy najważniejsze typy skal pomiarowych, od najsłabszej do najmocniejszej na przykładach stosowanych w rolnictwie.

Skala nominalna

W tej skali możemy zastosować tylko słowny opis własności jednostek przedmiotu badań, czyli stworzyć tzw. kategorie nominalne. Na przykład obserwując wybarwienie miąższu w bulwach ziemniaka (cechę jakościową) stworzymy kategorie barwy: biała, kremowa, jasno-żółta, ciemno-żółta itp. Badania organoleptyczne w zakresie przetwórstwa często bazują na klasyfikacji rodzaju smaków (gorzki, kwaśny, słodki itp.). W fitopatologii skala nominalna jest wykorzystywana do opisu wyglądu kultur grzybowych, gdzie grzybnia jest opisywana pod względem barwy, struktury (luźna, zbita) i obfitości. Dla pszenicy jarej przykładami cech jakościowych, wyrażanych w skalach nominalnych są: odmiana, typ użytkowy, wypiekowość. Kategorie niektórych cech można zapisać za pomocą liczby, jak to się stosuje np. w oznaczaniu rodów hodowlanych (R112, R012). Specyficzną odmianą tej skali jest skala **nominalna dychotomiczna**, która wyróżnia tylko dwie grupy danej cechy, np. podział bakterii na Gram+ i Gram-, płeć osobników (♀, ♂). Kategorie w skali nominalnej służą tylko do oznaczania oraz identyfikacji i klasyfikowania jednostek statystycznych. Nie posiadają jednostki pomiaru oraz nie można ich uporządkować, tzn. powiedzieć, w jakiej kolejności powinny występować. Możemy jedynie określić relacje typu =, ≠ (gorzkie ≠

kwaśne). Matematycznie jest to najsłabsza skala, która nie dopuszcza nawet prostych obliczeń arytmetycznych. Jedyną dopuszczalną operacją na liczbach jest tu zliczanie osobników w danej kategorii.

Skala porządkowa

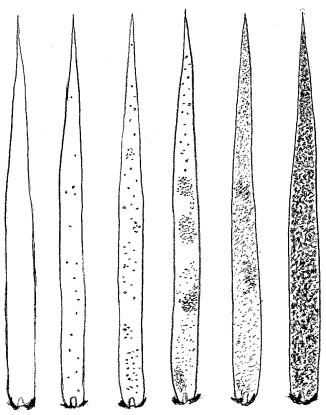
Stosowana jest do wyrażania takich cech, na których nie można przeprowadzić bezpośrednich pomiarów, lecz jednostki statystyczne można uporządkować według pewnej zasady. Jest to zasada, która przydziela liczby ze względu na stopień nasilenia cechy. Bardzo często jest to zasada umowna, mająca zastosowanie w wąskim zakresie badań, lub na użytek pewnego tematu badań. Na przykład, po przeczytaniu tego podręcznika chcielibyśmy przeprowadzić pomiar wśród jego czytelników, chcąc się dowiedzieć, na ile książka ta okazała się przydatna do praktycznych zastosowań. Moglibyśmy ustalić zakres skali porządkowej od 5 do 1, gdzie 5 oznacza, że książka znakomicie przygotowała czytelnika do praktycznych zastosowań metod statystycznych zaś 1 oznacza kompletny brak przydatności tej książki. Zwróćmy uwagę, że stopniom liczbowym możemy nadać opis werbalny, czyli zamienić skalę porządkową na nominalną, np. 5 -znakomita, 4 -bardzo dobra, 3 -czasami pomocna, 2 -mało przydatna, 1 -nieprzydatna.

Rangowanie to ustalanie hierarchii jednostek ze względu na natężenie cechy, np. chcemy ustalić, która cecha podręcznika jest najważniejsza dla czytelnika. Pytamy o 5 cech: cenę książki, objętość książki, język przekazu, przykłady zadań, zakres wiedzy. Czytelnik ma ocenić poszczególne jej elementy w skali od 1 do 5, w taki sposób, że najważniejsza cecha ma się znaleźć na 1 miejscu.

W Polsce do oceny jakości gleb ornych stosuje się skalę bonitacyjną 8. stopniową (klasy: I, II, IIIa, IIIb, Iva, Ivb, V i VI), w której najlepsze gleby orne są zaliczane do I klasy, natomiast najsłabsze do VI klasy, czyli wraz ze wzrostem stopni spada jakość gleb ornych. Ponieważ jakości gleb ornych nie można bezpośrednio „zmierzyć" w całej rozciągłości tej cechy, musimy posługiwać się umownymi klasami, które w opisie zawierają określone własności gleb ornych. Jest to więc pomiar wskaźnikowy. W badaniach z zakresu ochrony roślin wykorzystuje się tzw. skale porażenia (lub uszkodzenia) przez określone gatunki agrofagów, np. skalę 3. stopniową Poncheta (0,1,2) dla oceny stopnia występowania chorób korzeni i podstawy źdźbła na zbożach. W standardach Europejskiego Towarzystwa Ochrony Roślin (EPPO) znajdziemy skale oceny występowania np. rdzy na liściach zbóż (rys. 1). Zauważmy, że odległości między stopniami w tej skali są umowne i co ważne niejednakowe.

Pomiędzy rangami 1 - 2 a 4 - 5 jest zupełnie inne natężenie plam (pokrycie liścia przez mączniaka).

Wspólną cechą wszystkich skal porządkowych jest brak jednostki pomiaru oraz tzw. punktu zerowego. Natomiast kolejność rang (stopni, klas) mówi nam, które jednostki statystyczne mają mniejsze, a które większe natężenie badanej cechy. W porównaniu więc do skali nominalnej skala ta charakteryzuje się dodatkowymi własnościami, bowiem, oprócz =, ≠ wnosi relacje <, >. O te właściwości jest od niej mocniejsza. Nie dopuszcza się do wykonywania działań matematycznych na wartościach w skali porządkowej (tym bardziej statystycznych opartych na wyliczeniu sum kwadratów odchyleń). Skale porządkowe pozwalają na użycie statystyk opartych na centylach, wyznaczanie kwartyli, mediany oraz metod wnioskowania w oparciu o testy nieparametryczne.



Rys.1. Graficzna skala do oceny stopnia porażenia zbóż przez *Puccinia* spp. (EPPO Standards 1997).

Skala przedziałowa

Skale przedziałowe (interwałowe) stosowane są do wyrażania mierzalnych cech jednostek przedmiotu badań takich jak długość pędu, zawartość azotu w glebie, masa nasion z 1 rośliny szarłatu, liczba rozgałęzień rzepaku.

Cechy te, nazywane także ilościowymi, dzielą się zasadniczo na dwie grupy:

Pierwsza grupa to cechy **skokowe** (dyskretne), które przyjmują określone wartości ze skończonego zbioru liczb, np. liczba rozgałęzień I stopnia na łodydze rzepaku, może przyjmować wartości np. od 10 do 30, ale nie może przyjąć wartości 2,5. Wartości dyskretne (liczby naturalne) pochodzą ze zliczania i mówiąc o nich stosujemy sformułowanie „liczba", w odniesieniu do dni, rodzin, kwiatów itp. Jednostką podziału (interwałem) w tej skali jest 1 a jednostką pomiaru może być np. sztuka.

Druga grupa cech mierzalnych to cechy w skali przedziałowej **ciągłej**, które przyjmują dowolne wartości liczbowe w określonym przedziale, np. temperatura gleby w maju może mieścić się w przedziale od 1,5^{0}C do 10,0^{0}C, może więc wynosić 3,0^{0}C lub 3,2^{0}C. Wartości liczbowe w skali ciągłej pochodzące z bezpośrednich pomiarów są liczbami absolutnymi, ale interwały (jednostki pomiaru) mogą być dowolnie przyjmowane np. temperaturę można mierzyć w stopniach Celsjusza lub Fahrenheita (pamiętając, że „0" w tych skalach nie jest zerem absolutnym, a jedynie umownym). Jednostka pomiaru w skali przedziałowej jest ustalana arbitralnie. W przypadku tych cech mówimy o „ilości"- stopni, gramów, milimetrów, promili, w zależności w jakich jednostkach dokonujemy pomiaru. Co do interwału, pamiętajmy, że ma on taką wartość, jak dokładność dokonanego pomiaru, np. 0,01 t.

Skala przedziałowa umożliwia porównywanie różnic pomiędzy jednostkami statystycznymi ze względu na mierzoną własność, np. możemy wyliczać różnicę dla liczby dni wegetacji dwóch odmian grochu, lub masy 1000 nasion tych odmian. Na wynikach w tej skali możemy stosować wszystkie techniki statystyczne właściwe dla skali nominalnej i porządkowej, ponadto wyliczać średnią arytmetyczną, wariancję i odchylenie standardowe oraz stosować metody wnioskowania parametrycznego, pod warunkiem, że zmienne te reprezentują rozkład normalny.

Skala stosunkowa (ilorazowa)

Jest to najmocniejsza skala pomiarowa, która ma wszystkie cechy skali przedziałowej oraz dodatkową własność taką, że posiada ustalony punkt zerowy (tzw. 0 bezwzględne, czyli absolutny brak mierzonej wielkości). Skala stosunkowa dotyczy np. temperatury absolutnej mierzonej w stopniach Kelvina, podczas gdy temperatura w skalach Celsjusza i Fahrenheita jest wyrażona w skali przedziałowej (zero umowne). Takie cechy przedmiotu badań, jak: długość, masa, liczba dni, plon, są wyrażone w skali stosunkowej, a to pozwala nam na stosowanie tych wszystkich obliczeń jak dla skali przedziałowej, a ponadto na wyliczanie stosunków liczbowych pomiędzy jednostkami (np. stosunek długości łodygi grochu do jęczmienia wynosi 1:0,8). Statystyczne miary opisu, które wymagają skali stosunkowej to: średnia geometryczna i średnia harmoniczna oraz współczynnik zmienności. Pozostałe statystyki są dozwolone, jak dla w skali przedziałowej, a metody wnioskowania parametrycznego pod rygorem rozkładu normalnego zmiennej.

Zastosowanie różnych skal pomiarowych w badaniach.

Na podstawie przykładu badań obserwacyjnych nt. „Występowanie grzybów patogenicznych i endofitycznych na trawach w różnych siedliskach" zaklasyfikujemy badane cechy do odpowiednich skal pomiarowych.

- Rodzaj siedliska (naturalne, półnaturalne, agrocenozy pół uprawnych) - NOMINALNA
- Oznaczone gatunki traw w siedlisku (np. rajgras, tymotka, kupkówka) - NOMINALNA
- Liczba oznaczonych gatunków traw – PRZEDZIAŁOWA SKOKOWA
- Rozpoznane choroby traw (np. mączniak rzekomy traw, plamistość obwódkowa, rdze, głownie) - NOMINALNA
- Oznaczone gatunki patogenów grzybowych traw (np. *Puccinia graminis, Urocystis agropyri*) - NOMINALNA
- Stopień porażenia poszczególnych gatunków traw przez grzyby patogeniczne - PORZĄDKOWA
- Liczba porażonych roślin danego gatunku trawy przez określonego patogena – PRZEDZIAŁOWA SKOKOWA
- Indeks porażenia (%) trawy określonego gatunku przez patogena – PRZEDZIAŁOWA CIĄGŁA
- Gatunki wyizolowanych endofitów - NOMINALNA
- Liczba gatunków endofitów wyizolowanych na 1 gatunku trawy – PRZEDZIAŁOWA SKOKOWA
- Stosunek liczby patogenów do liczby endofitów na poszczególnych trawach – ILORAZOWA
- Wskaźnik bioróżnorodności grzybowej - ?

3.Sposoby przekształceń danych liczbowych

Dane liczbowe w badaniach przyrodniczych to przede wszystkim wyniki **pomiarów bezpośrednich**, tzn. takich, które są dokonywane na badanym osobniku za pomocą różnych przyrządów, lub dotyczą zliczania tych osobników pod względem jakiejś cechy.

Bardzo często jednak musimy posługiwać się tzw. **pomiarami pośrednimi.** Chcąc dowiedzieć się, jaka jest zawartość białka ogólnego w świeżej masie roślin grochu musimy

najpierw w odpowiedniej metodzie bezpośredniej oznaczyć zawartość azotu N (g x kg^{-1}) którą mnożymy przez 6,25 i otrzymamy zawartość białka ogólnego (g x kg^{-1}), w zielonce.
Są jednak zagadnienia w badaniach rolniczych takie, gdzie natężenia cechy nie da się określić za pomocą pomiaru bezpośredniego, ani też pośredniego. Klasycznym przykładem są walory rolniczej przestrzeni produkcyjnej. Do oceny rolniczej przestrzeni produkcyjnej wykorzystujemy tzw. **pomiary wskaźnikowe**. Wskaźnik rozumiemy jako liczbę wyrażającą poziom danego zjawiska (cechy jakościowej); wskaźnik jakości i przydatności rolniczej gleb, agroklimatu, warunków wodnych, rzeźby terenu. Wszystkie elementy tej oceny składają się na ogólny wskaźnik jakości rolniczej przestrzeni produkcyjnej, którego rozpiętość wynosi od 40 do 100 punktów.

Wskaźnik Margalefa to inaczej indeks bioróżnorodności; liczy się go ze wzoru:

$$D = \frac{S}{\log N}$$

gdzie:
S – liczba wszystkich gatunków (taksonów),
N – liczebność wszystkich osobników ze wszystkich gatunków.

W odniesieniu do wartości tego indeksu można ustalić np. klasy czystości wód (wg. Rozporządzenia Ministra Środowiska z 11 lutego 2004):

I klasa: D > 5,50; bardzo czyste wody
II klasa: D = 4,0 – 5,49; czyste wody
III klasa: D = 2,50 – 3,99; wody nieznacznie zanieczyszczone
IV klasa: D = 1,0 – 2,49; wody zanieczyszczone
V klasa: D < 1,0; wody bardzo zanieczyszczone

Do oceny bioróżnorodności stosuje się wiele różnych miar i wskaźników na Świecie. Ciekawy ich przegląd znajdzie czytelnik w publikacji Jadwigi Sienkiewicz pt. „Koncepcje bioróżnorodności – ich wymiary i miary w świetle literatury" (Ochrona Środowiska i Zasobów Naturalnych nr 45, 2010 r.)

Dane liczbowe z dwóch pomiarów bezpośrednich można poddawać przekształceniom poprzez dzielenie ich wartości. Otrzymamy wówczas **stosunek liczbowy**, np. stosunek liczby dni w pełni słonecznych do dni z częściowym zachmurzeniem w miesiącu wrześniu może wynosić 1:3. Różne informacje możemy odczytywać ze stosunków. Stosunek węgla organicznego w glebie do azotu (C:N) powinien wynosić 10-17:1, co oznacza, że węgla w glebie powinno być od 10 do 17 razy więcej niż azotu, aby nie doszło do tzw. procesu zbiałczania gleby. Ustalanie

długości oświetlenia w badaniach laboratoryjnych, tzw. fotoperiodu również podaje się jako stosunek liczby godzin dnia do nocy (D:N) np. 16:8 oznacza, że w laboratorium faza oświetlania będzie 2x dłuższa od fazy ciemnej. Stosunki liczbowe można także podawać dla zmiennych ciągłych, np. plonu. Ze stosunku plonu nasion grochu do plonu ziarna jęczmienia w uprawie mieszanej tych dwóch gatunków wynoszącym 0,81:2,83 t x ha^{-1}, dowiadujemy się, że ziarna jęczmienia jest 3,5 razy więcej (2,83:0,81 ≈ 3,49 ≈ 3,5) niż nasion grochu w plonie całkowitym mieszanki. Zauważmy, że stosunek liczb z bezpośrednich pomiarów dobrze jest zamienić na stosunek do liczby 1. Jeśli za 1 przyjmiemy plon grochu, stosunek wynosi 1,0:3,5, a gdy 1 oznaczymy jako plon jęczmienia, to stosunek wyniesie 0,29:1,0.

Proporcje (frakcje) obliczamy jako udział części jednostek (n_i) w stosunku do wszystkich jednostek (*N*). Oznaczymy proporcję literą f.

$$f = \frac{n_i}{N},$$

Chcemy się dowiedzieć, jaka jest proporcja jałówek na początku roku w stadzie bydła. Stado liczy N=48 sztuk bydła, w tym jałówek n_i=12. Dzieląc 12 przez 48 uzyskamy 0,25, co oznacza, że ¼ (ćwierć) stada to jałówki. Pod koniec roku, w tym samym stadzie liczącym 48 sztuk, mamy 8 jałówek. Ich proporcja wynosi teraz 8/48 = 0,1(6). Jeżeli chcemy zaprezentować udział jałówek w stadzie w liczbach względnych, to zamieniamy proporcje na **procenty,** mnożąc proporcję x 100**.** Oznaczymy je literą W.

$$W = \frac{n_i}{N} \times 100,$$

w naszym przykładzie uzyskamy 25% jałówek na początku roku i w zaokrągleniu 17% jałówek na koniec roku. Wartości tych procentów powinniśmy jednak zapisywać z dokładnością odpowiadającą zapisowi liczb ułamkowych, czyli 25,00% i 16,67% (zgodnie z zasadą zaokrąglania liczb). Przypomnijmy sobie podstawowe działania na procentach. Chcą wiedzieć, o ile procent zmniejszyła się liczba jałówek pod koniec roku w stosunku do początku roku stosujemy obliczenie:

$$\frac{10-8}{10} \times 100 = \frac{2}{10} \times 100 = 20\%$$

Jeśli pytanie sprecyzujemy, o ile procent była większa liczba jałówek na początku roku od liczby na koniec roku postąpimy tak:

$$\frac{10-8}{8} \times 100 = \frac{2}{8} \times 100 = 25\%$$

Jeśli natomiast chcemy powiedzieć, jaka jest różnica pomiędzy procentem jałówek na początku i na końcu roku to musimy użyć jednostki zwanej **punkt procentowy** (***pkt.%***). Różnica ta wynosi 8,33 *pkt.%* (25,00–16,67).

Pamiętajmy jednak, że udziały procentowe mają swój magiczny urok, który może wprowadzać czytelnika w błąd (w sposób zamierzony lub niezamierzony). Chodzi o podstawę, czyli liczbę, w stosunku do której odnosi się wyliczany %. Jeżeli N jest niewielkie (np. badaliśmy tylko 6 gospodarstw), to mówiąc, że 33,33% gospodarstw (2 z 6) stosuje prawidłowe dawki nawozów mineralnych, wyobrażamy sobie ich udział w ogóle, ale pamiętajmy, że to bardzo mała próba i nie powinniśmy wysuwać takich wniosków ogólnych na jej podstawie. Ponadto, wyliczanie % na podstawie mało liczebnych prób daje nam przerwy pomiędzy wartościami w skali %. Dla N=6, jedynie możliwe wartości % to: 0,00; 16,67; 33,33; 50,00; 66,67; 83,33 i 100,00 (brak wartości w przedziałach np. od 0,00 do 16,67). Dlatego też, bardzo często musimy transformować dane procentowe, aby uzyskać ich rozkład normalny.

W badaniach przyrodniczych posługujemy się także **indeksami.** W przypadku obliczania indeksu porażenia (lub uszkodzenia) rośliny uprawnej przez agrofaga, którego natężenie oceniono w skali porządkowej (np. 0 – 5) stosujemy wzór Townsenda-Heurbergera:

$$IP(\%) = \frac{\sum (n_i \times v_i)}{N \times V} \times 100, \text{ gdzie}$$

v_i – stopień porażenia
n_i – liczba porażonych roślin (lub ich części) w stopniu v_i
N - liczba badanych roślin (lub ich części)
V – najwyższy stopień porażenia

Przykład dotyczy oceny porażenia 150 roślin pszenicy określonej odmiany przez patogena rdzy źdźbłowej wg skali EPPO:

Stopień porażenia v_i	Liczba roślin n_i	n_i x v_i
0*	45	0
1	23	23
2	35	70
3	22	66
4	15	60
5 (V)	10	50
suma	N=150	269

*0 w skali porządkowej z zasady nie występuje, nie ma bowiem zerowego odniesienia do kolejności występowania wariantów.

Jednak skale *quasi*-porządkowe, jak te stosowane w ochronie roślin mają umowne 0, które tu oznacza całkowity brak objawów porażania przez patogena. Ponadto, zwróćmy uwagę, że gdyby skala ta zaczynała się 1 a kończyła na 6 stopniu to wartość IP w % byłaby wyższa (46,56% zamiast prawidłowej 35,87%).

$$IP(\%) = \frac{\sum (n_i \times v_i)}{N \times V} \times 100 = \frac{269}{150 \times 5} = 35{,}86(6) \approx 35{,}87\%$$

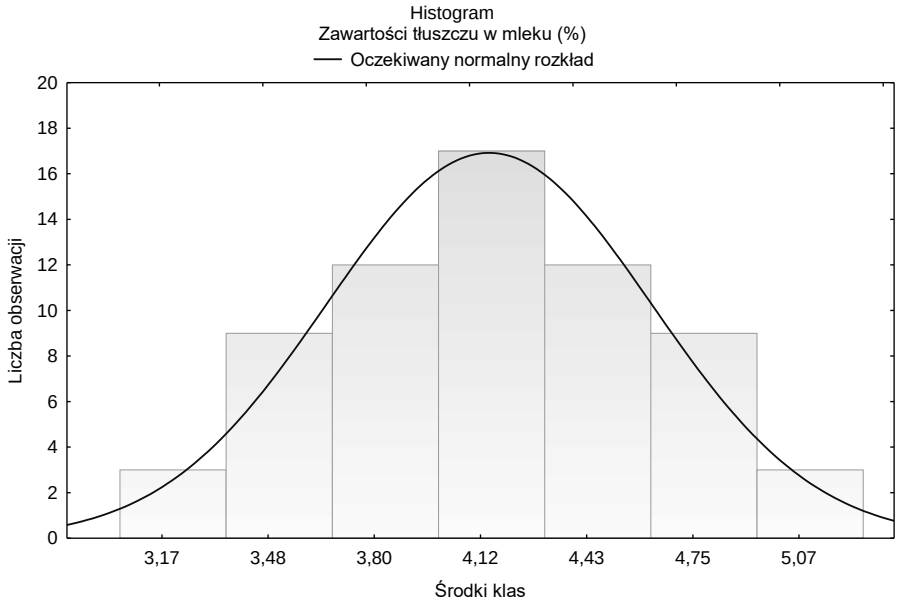
IP możemy nazwać średnią ważoną procentową porażenia wszystkich badanych jednostek (roślin lub ich części).

4.Transformacje danych liczbowych

Transformacje danych to inaczej przekształcenia algebraiczne mające służyć poprawieniu ich rozkładu do jak najbliżej podobnego rozkładowi normalnemu. Odnosimy się tu właściwości krzywej Gaussa, która ma wygląd krzywej dzwonowej symetrycznej z osią symetrii przebiegającą przez punkt wartości oczekiwanej w populacji generalnej E(x). Symetryczny rozkład oznacza, że w tym punkcie znajduje się również wartość najczęściej powtarzana w rozkładzie (modalna) oraz wartość środkowa z uporządkowanego szeregu danych (mediana). Wartość współczynnika asymetrii (As) wynosi 0 w sytuacji rozkładu symetrycznego. Drugą ważną własnością rozkładu normalnego jest odpowiednia koncentracja wyników wokół wartości oczekiwanej. Jej miarą jest kurtoza, która będzie świadczyła albo o normalnej koncentracji (K wynosi 3), o nadmiernej koncentracji ($K > 3$) lub o zbyt małej koncentracji ($K < 3$).

Rozważymy kilka typów rozkładów odbiegających od rozkładu normalnego.

4.1. Rozkład symetryczny, normalnie skoncentrowany

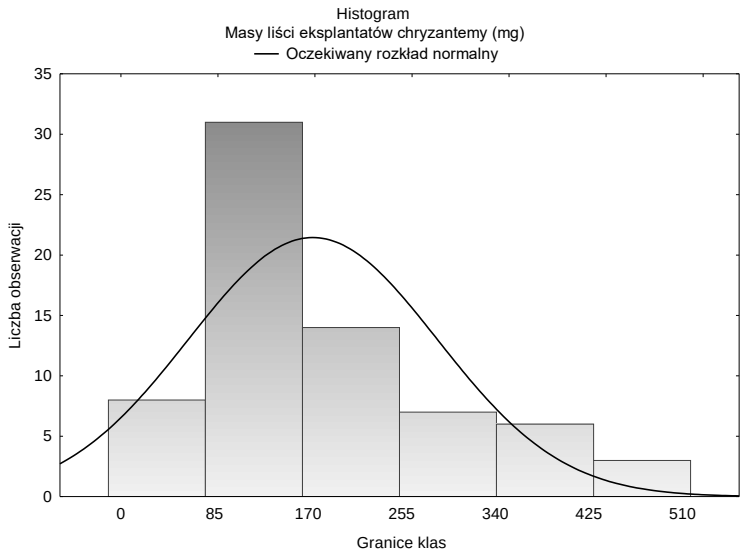


$$M_o = M_e = \bar{x}$$
$$4{,}00 \approx 4{,}14 \approx 4{,}15$$
$$As = 0{,}05 \; K = 2{,}3$$

Dane dotyczące zawartości tłuszczu w mleku krowim mają trzy miary centralne bardzo do siebie zbliżone. Widzimy, że różnica pomiędzy średnią a medianą to 0,01 %, a modalna odbiega od nich o 0,1 %. To nie zaburza symetrii rozkładu (As = 0,05). Kurtoza wynosi 2,3, co świadczy o lekkim spłaszczeniu (platykurtyczności rozkładu). Nie stanowi to jednak problemu ze zgodnością tego rozkładu z rozkładem normalnym, o czy będzie mowa w rozdziale poświęconym testom zgodności.

4.2. Rozkład asymetryczny – prawostronnie skośny

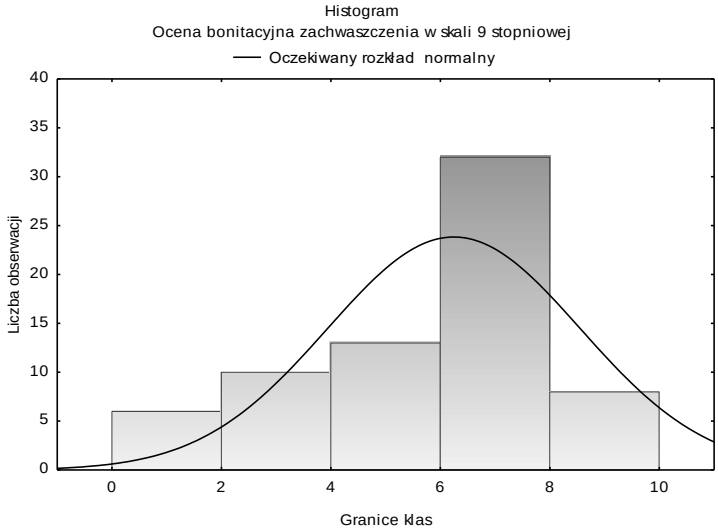
Rozkład masy liści eksplantatów chryzantemy prezentują wyraźnie prawostronnie asymetryczny rozkład, to znaczy, że więcej wyników jest poniżej średniej arytmetycznej (*Mo* wynosi 90,5 mg), a średnia 174,3 mg. Współczynnik asymetrii *As* jest dodatni = 1,14, to na wykresie objawia się wyciągnięciem prawego skrzydła rozkładu. Koncentracja wokół średniej jest nieco powyżej 3, co świadczy o lekkiej leptokurtyczności.



$$M_o < M_e < \bar{x}$$
$$90{,}5 < 140{,}2 < 174{,}3$$
$$As = 1{,}14 \; K = 3{,}55$$

4.3. Rozkład asymetryczny – lewostronnie skośny

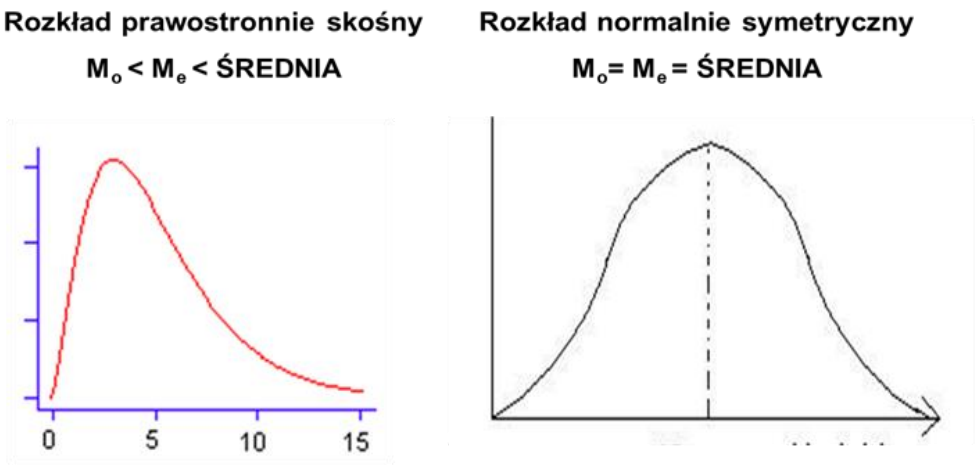
Rozkład ocen w skali bonitacyjnej dla zachwaszczenia plantacji wykazuje natomiast asymetrię lewostronną, ponieważ większość danych jest powyżej średniej arytmetycznej, natomiast nieliczne plantacje nam tę średnią zaniżają. W tej sytuacji zawsze Mo jest największa, a średnia najmniejsza. Współczynnik asymetrii będzie więc ujemny (*As* = -1,0). Koncentracja wyników wokół średniej nieznacznie odbiega od wartości 3.



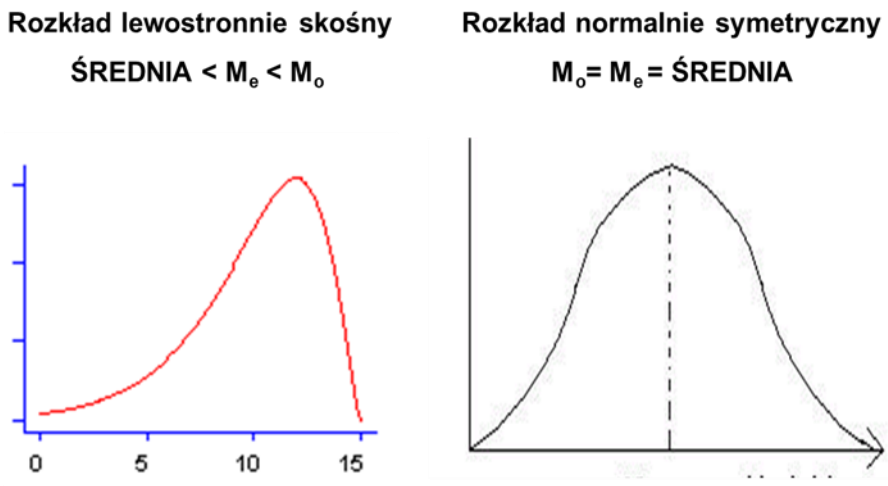
$$\bar{x} < Me < Mo$$
$$6{,}2 < 7{,}0 < 8{,}0$$
$$As = -1{,}0 \; K = 2{,}78$$

Jak należy transformować dane w sytuacji asymetrycznych rozkładów?

Transformacje jako przekształcenia algebraiczne wyników nie zmieniają relacji między liczbami ale wpływają na odległości między nimi. Zmieniają przez to na kształt rozkładu, tak, aby uzyskać rozkład najbardziej zbliżony do rozkładu normalnego.



Prawostronnie skośne rozkłady transformujemy poprzez: pierwiastek kwadratowy, jeśli to nie pomaga to przez pierwiastek trzeciego stopnia, a przy bardzo dużej skośności poprzez logarytm naturalny lub dziesiętny



Lewostronnie skośne rozkłady transformujemy poprzez: podnoszenie do kwadratu, do sześcianu, a przy dużej skośności stosujemy funkcję wykładniczą: $Y = e^x$ gdzie x jest to pomiarem oryginalnym, Y – wynik po transformacji, e – to podstawa logarytmu naturalnego, która wynosi w przybliżeniu 2,72.

W niektórych badaniach należy stosować gotowe procedury w celu normalizacji danych.

Zasady transformacji danych mających rozkład Poissona.

1. Jeśli zmienne dyskretne przyjmują wartości w szerokim zakresie 0 - 1000..., np. liczba kolonii bakterii na płytce Petriego, liczba jaj owada na liściu, liczba owadów na parcelce.

Stosujemy przekształcenie logarytmiczne wartości obserwowanej „x". Jeśli x < 10 przekształcamy $y = log\ (x+1)$, przy pozostałych $y = log\ x$

2. Jeśli zmienne dyskretne wykazują małe liczebności, a prawdopodobieństwo ich wystąpienia jest trudne do określenia, np. Liczba kolonii bakterii określonego szczepu na płytce Petriego, liczba roślin porażonych określoną chorobą, liczba chwastów danego gatunku na poletku, liczba parazytoidów w mszycy itp.

Stosujemy przekształcenie poprzez pierwiastkowanie wartości obserwowanej „x". Jeśli x < 10 przekształcamy $y = \sqrt{x + 0{,}5}$ przy pozostałych $y = \sqrt{x}$

3. Jeśli zmienne dyskretne wykazują małe liczebności, lecz prawdopodobieństwo ich wystąpienia jest całkowicie pewne (bliskie jedności), np. liczba kwiatów na roślinie, liczba owoców, liczba rozgałęzień rzepaku itp. pomiary biologiczne.

Wystarczające przybliżenie rozkładu normalnego daje wyliczenie średniej arytmetycznej z próby składającej się z 10-30 elementów. Należy pamiętać o losowym pobieraniu próby!

Zasady transformacji danych pochodzących z rozkładu Bernoulliego (0,1) a zamienionych na proporcje.

1. Jeśli wartości zmiennych mieszczą się w granicach od 30 do 70 % - MOŻNA NIE STOSOWAĆ PRZEKSZTAŁCEŃ

2. Jeśli wartości zmiennych mieszczą się w granicach 0 < X < 30% lub 70 < X < 100 % (ale nie w obu naraz) - STOSUJEMY PIERWIASTKOWANIE

3. Jeśli wartości zmiennych mieszczą się w granicach 0 < X < 100% - STOSUJEMY PRZEKSZTAŁCENIA TRYGONOMETRYCZNE:

a. na stopnie BLISSA - Jeśli proporcja wynika z bardzo wielu obserwacji lub ich dokładna liczba nie jest zdefiniowana (np.. Przezimowanie roślin na polu, stopień wylegania) $y = \arcsin\sqrt{x}$

b. na stopnie FREEMANA – TUKEYA Jeśli proporcja wynika z obserwacji w próbie n < 50

$$y = \frac{1}{2}\left[\arcsin\sqrt{\frac{k_i}{n_i+1}} + \arcsin\sqrt{\frac{k_i+1}{n_i+1}}\right]$$

gdzie k - liczba sukcesów, n-liczność próby

5.Skalowanie danych liczbowych

5.1. Skala standaryzowana czyli skala wartości unormowanych.

Standaryzacja polega na sprowadzeniu dowolnego rozkładu normalnego o danych parametrach μ i σ do **rozkładu standaryzowanego** (modelowego) o wartości oczekiwanej μ = 0 i odchyleniu standardowym σ = 1. Zmienną losową X zastępujemy **zmienną**

$$z = \frac{x_i - \mu}{\sigma}$$

standaryzowaną Z, która ma **rozkład N(0,1).**

Daje nam ona wyniki mieszczące się w granicach od – 3 do + 3, jednak zakres zależy od zmienności danych podlegających standaryzacji i może być węższy. Poniżej -3 i powyżej +3 wyniki standaryzowane traktujemy jako patologiczne, ekstremalne, lub wątpliwe i po diagnozie możemy je wykluczyć z populacji.

Prześledzimy dwa przykłady o różnym współczynniku zmienności.

Przykład 1.
Mamy 10 danych z pomiaru zawartości błonnika w wybranych produktach zbożowych (wyrażonych w g na 100 g produktu)

n_i	1	2	3	4	5	6	7	8	9	10
x_i	2,00	2,05	2,25	2,40	2,85	3,00	3,65	3,85	3,90	4,00
z_i	-1,25	-1,19	-0,94	-0,75	-0,19	0,00	0,81	1,06	1,13	1,25

Ponieważ nie znamy parametrów μ i σ dla zawartości błonnika w populacji produktów zbożowych, wyliczamy $\bar{x}$ = 3,00 i s = 0,80 dla badanej próby i standaryzujemy każdą wartość x_i według wzoru:

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

Stąd, wartości pomiarów mniejsze od średniej będą miały ujemny znak po unormowaniu, np.

$$z_1 = \frac{x_1 - \bar{x}}{s} = \frac{2{,}00 - 3{,}00}{0{,}80} = -1{,}25$$

zaś wartości powyżej średniej będą miały znak dodatni, np.

$$z_9 = \frac{x_9 - \bar{x}}{s} = \frac{3{,}90 - 3{,}00}{0{,}80} = 1{,}13$$

W tym przykładzie zmienność zawartości błonnika była stosunkowo duża, o czym przekonamy się wyliczając współczynnik zmienności v-Pearsona:

$$v = \frac{s}{\bar{x}} \times 100 = \frac{0{,}80}{3{,}00} \times 100 = 26{,}67\%$$

natomiast rozpiętość wyników po standaryzacji jest stosunkowo mała, bo 2,5 jednostki standaryzowanej (od -1,25 do + 1,25).

Prześledzimy następny przykład danych o znacznie mniejszej zmienności.

Przykład 2.
Zawartości % tłuszczu w mleku krowim ze średnią 3,21 % i odchyleniem standaryzowanym 0,026 %.

n_i	1	2	3	4	5	6	7	8	9	10
x_i	3,16	3,19	3,20	3,20	3,21	3,21	3,21	3,22	3,23	3,26
z_i	-1,92	-0,77	-0,38	-0,38	0,00	0,00	0,00	0,38	0,77	1,92

Posłużymy się współczynnikiem zmienności v-Pearsona do wyliczenia zmienności względnej:

$\bar{x} = 3{,}21$ i $s = 0{,}026$ stąd

$$v = \frac{s}{\bar{x}} \times 100 = \frac{0{,}026}{3{,}21} \times 100 = 0{,}81\%$$

Jak więc widzimy zakres danych standaryzowanych wynosi 3,84 jednostki, od -1,92 do + 1,92.

Jest to właściwość skali standaryzowanej, o której warto pamiętać, jeżeli chcemy wykorzystywać ją w wielocechowych analizach, a mamy do czynienia z cechami o różnej zmienności. Decyduje o tym wartość odchylenia standardowego.
Zasada ogólna, którą musimy pamiętać mówi, że tym większa rozpiętość skali standaryzowanej, im mniejsza zmienność wyników, czyli mniejsze odchylenie standardowe.

5.2. Skala centylowa

Centyle są to wartości cechy dzielące szereg na 100 równolicznych części. Pierwszy centyl C_1 jest wartością x_i, od której mniejszych jest 1% obserwacji. Ogólnie p-centyl (centyl rzędu p)

C_p jest taką wartością x_i, od której mniejszych jest p%, a większych (100 – p)% obserwacji. Wartość p nazywamy rangą centylową (np. centyl 80 (C_{80}) jest to wartość x_i, poniżej której znajduje się 80% pomiarów. Teoretyczną wartość poszczególnych centyli dla rozkładu normalnego obliczamy ze wzoru:

$$x_p = \bar{x} + Z_p \cdot s$$

Gdzie: x_p – wartość cechy na poziomie p-tego centyla, Z_p – wartość zmiennej standaryzowanej Z dla prawdopodobieństwa (%) odpowiadającemu danemu centylowi.

5.3. Skala zunitaryzowana

Unitaryzacja prowadzi do stałego, jednostkowego zakresu zmienności cech znormalizowanych. Polega ona na dzieleniu przez rozstęp odległości danej wartości od zaobserwowanej wartości „najgorszej".

Cechy typu **stymulanty** – duże wartości cechy mają korzystne znaczenie dla procesów (im większe natężenie wartości tej cechy tym lepiej),

Cechy typu **de stymulanty** – małe natężenie cechy sprzyjają procesowi (im mniejsze wartości liczbowe, tym lepiej). Kryterium doboru, co jest lepsze, czyli co ma znaczenie dla procesu, zależy od badacza (np. kryterium jakości płodów rolnych).

Dla stymulant wzór na unitaryzację $x_u = \frac{x_i - x_{min}}{R}$ gdzie: x_u – wartość zmiennej po unitaryzacji, x_i – wartość zmiennej, R – rozstęp, tj. różnica pomiędzy x_{max} i x_{min}.

Przykład dotyczy zawartości tłuszczu w mleku (%): $x_{min} = 3{,}16$, $R = 0{,}1$

n_i	1	2	3	4	5	6	7	8	9	10
x_i	3,16	3,19	3,2	3,2	3,21	3,21	3,21	3,22	3,23	3,26
x_u	0,0	0,3	0,4	0,4	0,5	0,5	0,5	0,6	0,7	1,0

Dla de stymulant unitaryzacja: $x_u = \frac{x_{max} - x_i}{R}$

Przykład dotyczy zawartości zanieczyszczeń w partii nasion gryki: $x_{max} = 8{,}5$, $R = 7{,}4$

n_i	1	2	3	4	5	6	7	8	9	10
x_i	1,1	2,4	2,5	2,8	3	4	5,6	6,2	7,1	8,5
x_u	1,0	0,8	0,8	0,8	0,7	0,6	0,4	0,3	0,2	0,0

Zakres wartości zunitaryzowanych zawsze wynosi 1.

6.Sposoby prezentacji danych liczbowych

6.1. Szeregi statystyczne

Szereg statystyczny to uporządkowany lub uporządkowany i pogrupowany według wariantów cechy pierwotny materiał statystyczny.

Ze względu na formę i objętość materiału wyróżniamy szeregi:

- proste
- jednopunktowe
- przedziałowe

Z uwagi na kryterium treści wyróżniamy szeregi:

- strukturalne
- czasowe
- przestrzenne

a. Szereg prosty

To, w jaki sposób zaprezentujemy materiał pierwotny z pomiarów, zależy od objętości tegoż materiału. Dla kilku lub kilkunastu wyników możemy zastosować uporządkowanie ich w szereg prosty. Percepcja jednej lub dwóch linijek danych liczbowych jest dość dobra, czego nie można już powiedzieć o kilkudziesięciu wynikach zajmujących na przykład połowę strony formatu A4.

Jeżeli prezentowana zmienna jest według badacza cechą pozytywną (na przykład rolnicze kryterium rozpatrywania), to nazywamy ją *stymulantą*, a jej natężenie możemy zaprezentować w szeregu malejącym.

Przykład dotyczy wyników 15 pomiarów zawartości białka (%) w nasionach grochu:

20,9 22,5 22,0 20,8 21,4 21,9 21,8 21,0 21,5 22,6 21,3 19,8 22,5 20,8 21,7.

Z rolniczego punktu widzenia, im więcej białka w nasionach grochu tym „lepiej”, stąd cechę tę uznajemy za stymulantę i wyniki prezentujemy od największej do najmniejszej wartości:

<u>Szereg prosty malejący</u> dla zawartości białka w nasionach grochu (%):

22,6 22,5 22,5 22,0 21,9 21,8 21,7 21,5 21,4 21,3 21,0 20,9 20,8 20,8 19,8.

Jeżeli zmienna jest cechą negatywną, oznacza to dla nas, że im mniejsze jej natężenie tym „lepiej” w merytorycznym rozpatrywaniu (rolniczym). Nazywamy ją *destymulantą,* a wyniki prezentujemy w <u>szeregu prostym rosnącym</u>. Zanieczyszczenia w materiale siewnym to cecha negatywna, ponieważ im większa ich zawartość, tym gorsza jakość materiału, aż do dyskwalifikacji tegoż materiału. Przykład dotyczy pomiarów udziału (%) zanieczyszczeń w

materiale siewnym, w 10 próbach nasion gryki: 2,4 2,8 3,0 5,6 7,1 1,1 8,5 2,5 6,2 4,0. Po uszeregowaniu w szereg prosty rosnący otrzymamy.

Szereg prosty rosnący dla zanieczyszczeń materiału siewnego gryki (%):
1,1 2,4 2,5 2,8 3,0 4,0 5,6 6,2 7,1 8,5

Jest jeszcze trzecia kategoria cech, tzw. *nominanty*, odnosząca się do zmiennych, których „najlepsze" natężenie jest w środkowych zakresach, i ani zbyt duże ani zbyt małe wartości, nie są korzystne. Przykładem takiej cechy w rolnictwie jest suma opadów w określonym miesiącu wegetacji.

Wyniki dotyczą sumy opadów w lipcu (mm) notowanych w latach 1996-2001 w Stacji Doświadczalnej Oceny Odmian w Chrząstowie, woj. kujawsko-pomorskie: 138,5 110,7 100,6 36,8 93,0 114,3.

Dla roślin wcześnie schodzących z pola, na przykład groch, najbardziej optymalną sumą opadów w lipcu są wartości w przedziale 90-100 mm (dwa lata w tej stacji), natomiast bardzo niska oraz zbyt duża suma opadów w lipcu jest dla grochu niekorzystna. Wyniki takich rozkładów można prezentować w postaci szeregu rosnącego lub malejącego z zaznaczeniem wartości najkorzystniejszych.

Szereg prosty rosnący dla sumy opadów w lipcu (mm):
36,8 <u>93,0 100,6</u> 110,7 114,3 138,5.

b. Szereg jednopunktowy

Mając do czynienia ze zmienną typu skokowego, przy dużej liczbie obserwacji (powyżej 30 wyników), stosujemy prezentację w postaci szeregu jednopunktowego. Szereg taki składa się z dwóch rubryk, tj. z kolumn i wierszy. W pierwszej kolumnie zamieszczamy wartości zmiennej skokowej (tzw. skoki - x_i), a w drugiej kolumnie liczbę obserwacji skoków - n_i. Z kolei w pierwszym wierszu zawarta jest treść kolejnych pozycji, według cechy porządkującej materiał z badań. W następnych wierszach (ze względu na edycję można ich nie eksponować siatką podziału) zamieszcza się wartości liczbowe przypadające na warianty cechy. Na samym dole, w oddzielnym wierszu, można zamieścić podsumowanie liczb wszystkich wariantów cechy.

Przykład dotyczy danych odnośnie liczby uprawianych odmian ziemniaka w gospodarstwach rolnych na terenie powiatu płockiego. W gospodarstwach tych uprawiano od 1 do 5 odmian ziemniaka, a częstotliwość tej cechy zaprezentowano w tabeli 5.

Tabela 5. Zestawienie gospodarstw według liczby uprawianych odmian ziemniaka w roku 2006, w powiecie płockim, woj. mazowieckie.

Liczba uprawianych odmian ziemniaka w gospodarstwie (x_i)	Liczba gospodarstw (n_i)
1	42
2	36
3	14
4	8
5	5
Ogółem	105

*- źródło – badania własne

c. Szereg przedziałowy (klasowy) i strukturalny

Zmienne typu ciągłego, w przypadku dużej zbiorowości statystycznej, prezentujemy w postaci szeregu rozdzielczego, którego nazywamy także szeregiem klasowym. Klasa, to inaczej przedział wartości cechy, którą chcemy zaprezentować w całej jej rozciągłości. Szereg składa się z „k" liczby przedziałów o szerokości klasy „c". Najbardziej użytecznym do celów analizy statystycznej jest szereg klasowy o równych przedziałach, zamknięty dołem i górą, co oznacza, że zarówno dolną granicę pierwszego przedziału jak i górną granicę ostatniego przedziału w nim podajemy.

Zasady tworzenia **szeregu przedziałowego zamkniętego górą i dołem, o jednakowej rozpiętości klas** zostaną omówione na przykładzie 90 wyników pomiarów długości łodygi narcyza odmiany Ice Flower. Pomiarów tych dokonano za pomocą linijki z dokładnością do 1mm i otrzymano w kolejności następujące wyniki.

Przykład 1. Długość łodygi narcyza (cm) odmiany Ice Flower.

18,0 17,1 16,8 18,9 18,2 15,2 16,4 16,6 14,5 17,9 17,3 21,4 22,5 19,2 20,2 17,2 15,4 16,7 18,8 22,0 19,7 20,0 20,7 21,1 17,0 16,2 15,5 18,6 18,1 18,4 19,9 16,9 15,5 14,3 18,8 19,0 20,4 19,6 19,1 21,0 17,9 15,9 15,3 18,4 16,9 18,1 18,8 16,7 **14,1** 22,4 19,8 21,3 15,6 18,0 17,1 18,5 21,6 19,4 14,6 14,6 18,7 18,0 20,1 20,0 19,2 20,6 16,6 18,4 **22,8** 20,9 18,3 18,1 17,9 15,4 15,7 16,0 16,8 16,3 17,6 16,5 17,2 20,5 19,0 19,8 19,3 20,3 19,6 19,1 19,1 19,5

Wśród 90 wyników zaznaczono najmniejszą oraz największą wartość z pomiaru, co jednak przy tak dużej liczbie wyników, wcale nam nie ułatwia ich percepcji.

Uporządkujmy wyniki w szereg prosty malejący, można bowiem uznać, że im dłuższa łodyga narcyza, tym większa jego wartość (na przykład estetyczna).

22,8 22,5 22,4 22,0 21,6 21,4 21,3 21,1 21,0 20,9 20,7 20,6 20,5 20,4 20,3 20,2 20,1 20,0 20,0 19,9 19,8 19,8 19,7 19,6 19,6 19,5 19,4 19,3 19,2 19,2 19,1 19,1 19,1 19,0 19,0 18,9 18,8 18,8 18,8 18,7 18,6 18,5 18,4 18,4 18,4 18,3 18,2 18,1 18,1 18,1 18,0 18,0 18,0 17,9 17,9 17,9 17,6 17,3 17,2 17,2 17,1 17,1 17,0 16,9 16,9 16,8 16,8 16,7 16,7 16,6 16,6 16,5 16,4 16,3 16,2 16,0 15,9 15,7 15,6 15,5 15,5 15,4 15,4 15,3 15,2 14,8 14,6 14,5 14,3 14,1

Dalej jednak niewiele możemy powiedzieć o długości łodygi, chociażby tego, w jakim przedziale jest najwięcej roślin. Dlatego wyniki te musimy pogrupować w klasy.

Obliczenia potrzebne do utworzenia szeregu klasowego:

Liczebność zbiorowości: N = 90

Rozstęp: $R = x_{max} - x_{min} = 22{,}8\text{cm} - 14{,}1\text{cm} = 8{,}7\text{cm}$

Aby ustalić liczbę klas „k” w naszym szeregu możemy skorzystać z zasady, że liczba ta powinna się mieścić w przedziale od $\frac{1}{2}\sqrt{N}$ do $\sqrt{N}$, W tym przypadku liczba „k” mieści się pomiędzy $\frac{\sqrt{90}}{2} = \frac{9{,}5}{2} = 4{,}7 \approx 5$ a $\sqrt{N} = 9{,}5 \approx 10$. Tutaj wybieramy dowolną wartość z przedziału 5 - 10, możemy więc wybrać wartość środkową 7 klas.

Mając rozstęp i liczbę klas do szeregu obliczmy teraz, jaką szerokość „c” będzie miał przedział. Szerokość przedziału będzie taka sama dla wszystkich klas i obliczymy ją dzieląc rozstęp „R” przez liczbę klas „k”:

$$c = \frac{R}{k} = \frac{8{,}7}{7} = 1{,}2428571\text{cm}$$

Wynik ilorazu jest ułamkiem nieskończonym i chcąc go zaokrąglić do dwóch miejsc po przecinku otrzymamy liczbę 1,24, co jednak daje pewną niewielką stratę. Szereg okazałby się zbyt „ciasny”, gdybyśmy przyjęli c=1,24 (1,24 x 7 = 8,68). Ponadto, 1,24cm jest wartością „niewygodną” do konstruowania szeregu, bowiem łatwiej jest dodawać wartość zakończoną na 0 lub 5, na przykład 1,25 (1,25 x 7 = 8,75). W tym przypadku jest uzasadnione zaokrąglanie w górę, bowiem, przy tej rozpiętości szeregu nie ma ryzyka, że wartości największe się w nim nie zmieszczą.

Każda klasa ma granicę dolną G_d i granicę górną G_g. Kolejnym etapem jest ustalenie, z jaką dokładnością liczbową będziemy prezentować granice klas w szeregu. Skoro szerokość klas ma dokładność o jeden rząd większą niż wartości pomiarów, konsekwentnie konstruujemy szereg o granicach też o jeden rząd dokładniejszych niż wyniki pomiarów. Posłużymy się tutaj zasadą, że dokładność granic przedziałów „d" wynosi 0,5 razy dokładność wyników pomiarów (d = 0,5 x 0,1cm = 0,05 cm).

Możemy teraz przystąpić do wyliczania G_{dI} - dolnej granicy pierwszej klasy.

$$G_{dI} = x_{\min} - d = 14{,}1 - 0{,}05 = 14{,}05$$

Teraz wyliczamy G_{gI} – górną granicę pierwszej klasy

$$G_{gI} = G_{dI} + c = 14{,}05 + 1{,}25 = 15{,}30$$

Przyjmujemy, że górna granica pierwszej klasy jest dolną granicą drugiej klasy $G_{gI} = G_{dII}$. Natomiast górną granicę klasy drugiej G_{gII} utworzymy dodając do dolnej granicy szerokość przedziału „c" (15,30 + 1,25 = 16,55). Tak postępujemy do ostatniej, u nas do 7 klasy.

Najprostszy szereg przedziałowy składa się z dwóch kolumn (tabela 6), w pierwszej prezentujemy granice klas (x_i), w drugiej kolumnie wpisujemy liczebność osobników w danej klasie (n_i). Pamiętamy przy tym, że wyniki równe granicy klas są zaliczane do klasy niższej. U nas są to przypadki trzech wyników 15,3 20,3 i 22,8. Liczbę 15,3 zaliczamy do pierwszej klasy, bowiem klasa ta zawiera wyniki większe od $14{,}05 < x_i$ i $x_i \leq 15{,}30$. Znak $\leq$ oznacza mniejsze i równe 15,30. Wynik 20,3 należy do klasy piątej, ponieważ klasa ta zawiera wyniki $19{,}50 < x_i \leq 20{,}3$.

Tabela 6. Zestawienie narcyzów odmiany Ice Flower według długości łodygi

Długość łodygi narcyza odmiany Ice Flower (x_i) – cm	Liczba roślin (n_i)
14,05 – 15,30	7
15,30 – 16,55	12
16,55 – 17,80	15
17,80 – 19,05	23
19,05 – 20,30	19
20,30 – 21,50	9
21,50 – 22,80	5
Ogółem	90

*- źródło - Katedra Ogrodnictwa, Zakład Roślin Ozdobnych, UT-P Bydgoszcz

Szeregi mogą być w różny sposób konstruowane. Pewną modyfikacją szeregu przedziałowego jest **szereg otwarty dołem lub górą, albo jednocześnie dołem i górą.** Zaprezentowany w kolejnej tablicy szereg dotyczy struktury gospodarstw pod względem powierzchni użytków rolnych. W pierwszej kategorii, poniżej i równe 5ha, znajdują się wszystkie gospodarstwa, których udział w zbiorowości nie przekracza 5% i dlatego zwyczajowo klasę tę zostawiamy otwartą dołem. Natomiast w ostatniej kategorii zawarto gospodarstwa, w których powierzchnia użytków rolnych była większa lub równa 40,1ha, tzn., że znalazły się tam gospodarstwa z powierzchnią UR 50ha oraz 100ha. Ponieważ ich udział wynosi 5% można by te klasę zamknąć, ale nie jest to wymóg konieczny.

Jest to także przykład szeregu z niejednakową szerokością klas. Otóż 1 i 2 klasa mają mniejszą szerokość (c=5ha), zaś klasy 3,4 i 5 szerokość c=10ha.

Taki sposób prezentacji wyników daje dobrą ich wizualizację, natomiast w opisie statystycznym ma o wiele mniejsze zastosowanie niż szereg zamknięty dołem i górą o jednakowej szerokości klas.

Tabela 7. Struktura powierzchni użytków rolnych w gospodarstwach produkujących ziemniaki przemysłowe

Powierzchnia użytków (ha)	Liczebność n_i	Udział n_i / N	n_i cum
≤ 5,0	4	0,04	4
5,1 - 10,0	20	0,20	24
10,1 - 20,0	33	0,33	57
20,1 - 30,0	19	0,19	76
30,1 - 40,0	15	0,15	91
≥ 40,1	5	0,05	100
Ogółem	100	x	x

*- źródło – badania własne

Omawiane powyżej szeregi są przykładem **szeregu strukturalnego**, ze względu na kryterium treści (struktura długości pędów, struktura powierzchni użytków rolnych). Drugi szereg jest bardziej rozbudowany od poprzedniego szeregu o 2 kolejne kolumny (tabela 7).

W 3 kolumnie zawarto informację o udziale gospodarstw w kategorii wielkości, co możemy prezentować w liczbach bezwzględnych, wyliczając frakcje dla poszczególnych klas według poznanego już wzoru: $f = \frac{n_i}{N}$, albo w liczbach względnych, tj. w procencie $W = \frac{n_i}{N} \times 100$.

W 4 kolumnie zaprezentowano liczebność skumulowaną, którą uzyskujemy poprzez dodawanie liczebności klas poprzednich do klasy bieżącej. Na przykład, liczebność

skumulowana (cum) do 3 klasy wynosi 4 + 20 + 33 = 57. Wartość ta informuje nas, ile jest gospodarstw o powierzchni UR mniejszej od 20ha. W naszym przykładzie skumulowane liczebności równają się wartościom skumulowanych udziałów procentowych. Jest to zaleta szeregów dla zbiorowości o liczebności równej 100. Wartości skumulowane będą nam potrzebne do obliczania statystyk centralnych, którymi zajmiemy się w kolejnym rozdziale.

d. Szereg przestrzenny i czasowy

W opracowaniu wielu zagadnień przyrodniczych przydatne okazują się szeregi przestrzenne, w których treść, jaką zamierzamy zaprezentować jest rozpatrywana wraz z rozmieszczeniem na określonym obszarze, na przykład w województwach, powiatach, albo w krajach jakiegoś regionu Świata. Ludność rolnicza w danym kraju, została przedstawiona w Polsce, na tle krajów sąsiadujących oraz na tle Świata (tabela 8). Taka prezentacja pozwala nam na szybką interpretację zjawiska. Mianowicie, na tle innych krajów sąsiadujących z Polską, w roku 2002, zajmowaliśmy trzecie miejsce (po Rosji i Ukrainie) pod względem liczby ludności rolniczej i drugie (po Rosji) ze względu na ludność aktywną zawodowo w rolnictwie. Prezentowany tutaj szereg przestrzenny zawiera dwie cechy, jest więc przykładem szeregu złożonego.

Tabela 8. Ludność rolnicza i ludność aktywna zawodowo w rolnictwie w Polsce i w krajach sąsiadujących w roku 2002.

Kraj	Ludność rolnicza (w tys.)	Ludność aktywna zawodowo w rolnictwie (w tys.)
Polska	**14 409**	**4 159**
Białoruś	3 032	651
Litwa	1 086	200
Ukraina	15 636	3 411
Niemcy	9 957	923
Republika Czeska	2 606	442
Rosja	39 090	7 773
Słowacja	2 286	255
Świat	3 233 565	1 333 329

*- źródło – GUS

Z kolei szereg czasowy służy do zaprezentowania zmienności cechy w czasie (w latach, kwartałach, miesiącach itp.).

Na przykładzie rozkładu cen skupu ziarna dwóch podstawowych zbóż w Polsce (tabela 9), możemy porównać ich zmienność w 10 latach. Jest to także szereg złożony, bo dotyczy dwóch cech (dwa gatunki zbóż).

Tabela 9. Wykaz średnich cen skupu ziarna pszenicy i żyta w Polsce w latach 1995-2004.

Lata	Cena skupu ziarna pszenicy (zł/1 dt)	Cena skupu ziarna żyta (zł/1 dt)
1995	35,36	22,54
1996	57,19	35,93
1997	50,85	37,12
1998	46,83	32,08
1999	42,98	30,13
2000	50,82	36,15
2001	50,45	36,46
2002	43,61	33,19
2003	45,51	35,35
2004	48,15	36,34

*- źródło – badania własne

6.2.Tabela

Tabela statystyczna jest formą prezentacji dla wtórnej informacji liczbowej (przetworzonego materiału empirycznego), umożliwiającą czytelnikowi łatwe poznanie zbiorowości statystycznej pod względem jednej lub kilku cech.

Tabela w sposób syntetyczny podaje liczby względne albo bezwzględne potrzebne do analizy badanego zjawiska. Najbardziej dopracowaną tabelą jest **tabela wynikowa,** która w odróżnieniu od tabeli **roboczej** jest prezentacją gotową do zamieszczenia w pracy naukowej i powinna być jak najbardziej przejrzysta dla czytelnika.

Ze względu na liczbę cech zawartych w tabeli wyróżniamy tabele proste i złożone. W tabeli prostej prezentujemy jedną cechę, np. relację ceny skupu żyta do ceny skupu ziarna pszenicy w latach (tabela 10). Podczas gdy w tabeli złożonej możemy zaprezentować naraz dwie lub więcej cech, na przykład relacje cen skupu kilku produktów rolniczych zmieniające się w czasie dekady lat (tabela 11).

Tabela 10. Relacja średnich cen skupu ziarna żyta do ceny skupu ziarna pszenicy w latach 1995-2004 w Polsce.

Lata	zł/1 dt żyta / zł/1 dt pszenicy
1995	0,64
1996	0,63
1997	0,73
1998	0,69
1999	0,70
2000	0,71
2001	0,72
2002	0,76
2003	0,78
2004	0,78

*- źródło – badania własne

Tabela 11. Relacje średnich cen skupu niektórych produktów rolniczych względem ceny skupu ziarna pszenicy na przełomie lat 1995 – 2004 w Polsce.

Lata	Produkt		
	ziarno żyta	mleko krowie	ziemniak jadalny
1995 – 1999	0,64 – 0,70	1,26 – 1,42	0,68 – 0,52
2000 - 2004	0,71 – 0,78	1,53 – 1,58	0,46 – 0,54

*- źródło – badania własne

Tabele tzw. kombinowane zawierają co najmniej dwa szeregi statystyczne opisywane ze względu na kilka cech. Przykładem takiej tabeli jest tabela 12, w której różne jednostki statystyczne (państwa, m. in. Polska oraz Świat) zostały scharakteryzowane pod względem dwóch cech ilościowych (% ludności rolniczej w ogólnej liczbie ludności oraz % ludności aktywnej zawodowo w rolnictwie). Natężenie tych cech w wartościach względnych jest ponadto rozpatrywane w ujęciu czasowym (dwa lata: 1995 i 2002). Jest to kombinacja czasowo-przestrzennego szeregu, sposób zaprezentowania zmian na przestrzeni 8 lat w udziale ludności rolniczej w Polsce i w krajach sąsiednich, w odniesieniu do danych dla całego Świata. Względne wartości tych cech pozwalają nam porównać kraje między sobą, czego nie mogliśmy zrobić na danych rzeczywistych w szeregu przestrzennym (tabela 12).

Tabela 12. Zmiany w ludności rolniczej w Polsce i w poszczególnych krajach sąsiadujących w latach 1995 i 2002 w odniesieniu do ludności rolniczej na Świecie.

Kraj	Rok	Ludność rolnicza	
		w % ogółu ludności	aktywna zawodowo w rolnictwie w % ogółu ludności
Polska	1995	38,4	12,3
	2002	37,3	10,8
Białoruś	1995	31,2	8,5
	2002	30,5	6,5
Litwa	1995	31,9	7,7
	2002	31,3	5,8
Ukraina	1995	32,6	8,6
	2002	32,0	7,0
Niemcy	1995	13,5	1,6
	2002	12,1	1,1
Republika Czeska	1995	25,4	5,2
	2002	25,4	4,3
Rosja	1995	27,1	6,3
	2002	27,1	5,4
Słowacja	1995	43,0	5,5
	2002	42,3	4,7
Świat	1995	54,7	22,5
	2002	51,9	21,4

*- źródło – GUS

Forma tabeli powinna być zgodna z zasadami ogólnymi dla naukowych tabel. Jest to kilka zasad, które po kolei zostaną omówione.

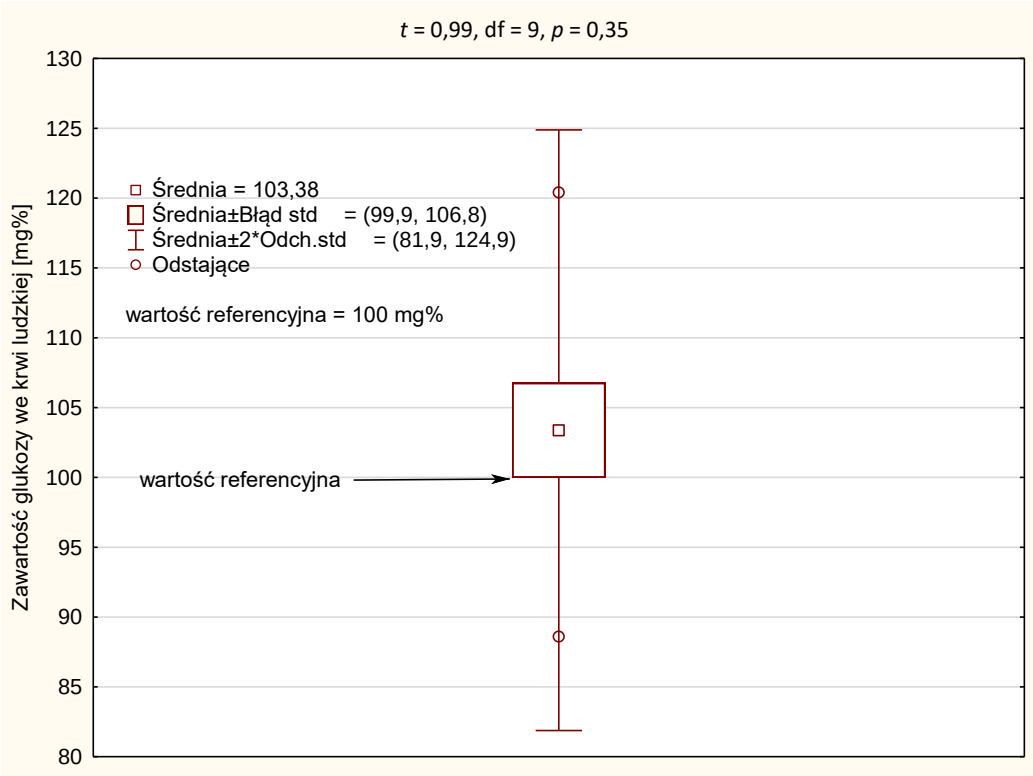
1. Każda tabela powinna mieć numer kolejny i tytuł zawarty nad tabelą.
2. Tytuł powinien być zwarty, sformułowany jasno i zwięźle a zarazem zawierać wyczerpujące informacje o przedmiocie badań, który należy opisać w sensie rzeczowym, czasowym i przestrzennym.

6.3.Wykres

Wykres to graficzna forma prezentacji wyników, w myśl powiedzenia „*lepszy jeden obrazek niż tysiąc słów*” może dać natychmiastowe wyobrażenie o zakresie danych liczbowych, ich rozkładzie i frekwencji oraz wielu innych informacji statystycznych. Przed rozpoczęciem wizualizacji danych należy dobrać układ współrzędnych, osi X (odciętych) i osi

Y (rzędnych. Na ogół wykres wykonuje się w I ćwiartce układu osi X,Y, w zakresie wartości dodatnich. Odpowiednio trzeba wyskalować obydwie osie w jednostkach pomiaru zmiennej zależnej (cechy), którą chcemy zobrazować graficznie. W tym celu stosuje się skale liniowe (metryczne) o podziałce równomiernej z użyciem jednego z 5 modułów. Najczęściej stosujemy moduł „1" (0,1,2,3,…,10), moduł „2" (0,2,4,6,…10), moduł „4" (0,4,8,12..20), moduł „5" (0,5,,10,20…100) lub „10" (0,10,20,40 50..100). Podziałki skali wykresu należy tak dobrać, aby odczytanie dowolnego punktu nie sprawiło trudności. Wykres 1 przedstawia zawartość glukozy we krwi ludzkiej w jednostkach mg%, na podstawie wyliczeń danych z próby statystycznej 10 osób. Jest to wykres typu WĄS – PUDEŁKO (BOX PLOT), który służy do zobrazowania położenia średniej, jej błędu (błąd standardowy średniej) oraz przedziału średnia +/- 2 x odchylenie standardowe, który jest przedziałem normy statystycznej (tzn. że pokrywa 95% populacji). Na tym wykresie podano również wartość referencyjną = 100 mg%, w odniesieniu do której wykonano porównanie średniej dla badanej grupy. Powyżej wykresu zamieszczono informację o teście t-Studenta (t_{obl} = 0,99), przy liczbie stopni swobody *df* (degree of freedom) = n-1 = 9, poziom błędu pierwszego rodzaju (*p*-value) = 0,35. Wykres 2 i 3, to histogram, który służy do zobrazowania częstotliwości (frekwencji) zmiennej losowej typu skokowego lub ciągłego w dużej próbie (N co najmniej 100). Ten rodzaj wizualizacji ma na celu przedstawienie rozkładu, najczęściej w celu diagnozy zgodności rozkładu empirycznego (z próby) z zakładanym rozkładem teoretycznym (normalnym). Dla cechy o charakterze interwałowym najczęściej jej częstość (liczność) oznaczamy jako n_i (frekwencja). Stosunek liczności klasy do liczności całej próby, czyli udział f (f = n_i/N i udział w wartości względnej $W_i = n_i/N$ x 100) nazywamy częstością względną. Na podstawie tabeli 7, rozbudowanej o informacje odnośnie udziału gospodarstw w strukturze powierzchni użytków rolnych (kolumna f_i i W_i) a także o skumulowaną wartość udziału przedstawione zostały dwa rodzaje histogramów. Na obydwu histogramach oś X jest wyskalowana w wartościach ha, tak jak dla szeregu klasowego. Z kolei oś Y jest wyskalowana w liczbach odnoszących się do frekwencji ni (wykres 2) i do wartości % (wykres 3).

Pamiętajmy, że każdy wykres powinien mieć numer kolejny i zwarty tytuł mówiący o tym, czego on dotyczy, pod obrazkiem.

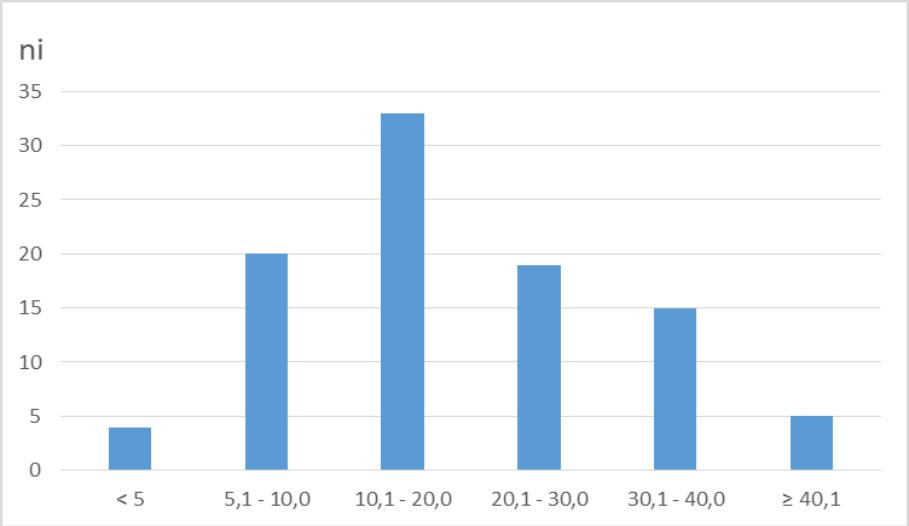


Wykres 1. Zawartość glukozy we krwi ludzkiej na podstawie próby n=10 osób w odniesieniu do wartości referencyjnej = 100 mg%.

Tabela 7. Struktura powierzchni użytków rolnych w gospodarstwach produkujących ziemniaki przemysłowe

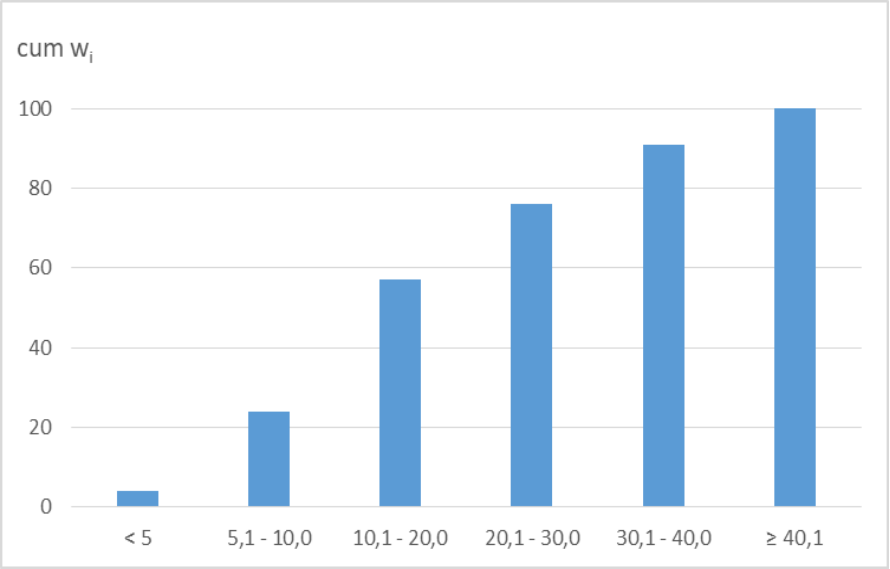
Powierzchnia użytków (ha)	Liczebność n_i	Udział f = n_i / N	W_i (%) = n_i / N x 100	cum n_i	cum W_i
≤ 5,0	4	0,04	4	4	4
5,1 - 10,0	20	0,20	20	24	24
10,1 - 20,0	33	0,33	33	57	57
20,1 - 30,0	19	0,19	19	76	76
30,1 - 40,0	15	0,15	15	91	91
≥ 40,1	5	0,05	5	100	100
Ogółem	100	x	100	x	x

HISTOGRAM CZĘSTOTLIWOŚCI



Wykres 2. Histogram częstotliwości (n_i) gospodarstw produkujących ziemniaki przemysłowe względem powierzchni użytków rolnych (ha).

HISTOGRAM SKUMULOWANYCH CZĘŚTOTLIWOŚCI WZGLĘDNYCH



Wykres 3. Histogram skumulowanych częstotliwości względnych (%) liczby gospodarstw produkujących ziemniaki przemysłowe względem powierzchni użytków rolnych (ha).

7.Miary statystycznego opisu (statystyki opisowe)

Wszystkie statystyki będą omówione na przykładzie próby małej zaprezentowanej w szeregu prostym (wzory w podpunkcie a) oraz dla próby dużej zaprezentowanej w szeregu przedziałowym (wzory w podpunkcie b).

7.1.Miary centralne położenia, pozycji

Do podstawowych miar centralnych zaliczamy **średnie klasyczne.**

W statystyce nazywamy je Momentem 1-rzędu (M1).

Średnia ARYTMETYCZNA

a. $$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Dla przykładu z 15 danymi pomiarów zawartości białka (%) w nasionach grochu: 20,9 22,5 22,0 20,8 21,4 21,9 21,8 21,0 21,5 22,6 21,3 19,8 22,5 20,8 21,7 utwórzmy szereg prosty do obliczeń statystyk.

n_i	x_i
1	22,6
2	22,5
3	22,5
4	22,0
5	21,9
6	21,8
7	21,7
8	21,5
9	21,4
10	21,3
11	21,0
12	20,9
13	20,8
14	20,8
15	19,8
suma	322,5

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{322{,}5}{15} = 21{,}50\%$$

Do wyliczenia średniej arytmetycznej dla próby niemałej (N powyżej 30) wykorzystamy dane dotyczące powierzchni gospodarstw w klasach zestawione w szereg przedziałowy:

Powierzchnia (ha)	Liczba gospodarstw (n_i)	Środek klasy x_i'	$x_i' \times n_i$
13,35 - 15,95	4	14,65	58,60
15,95 - 18,55	8	17,25	138,00
18,55 - 21,15	4	19,85	79,40
21,15 - 23,75	11	22,45	246,95
23,75 - 26,35	17	25,05	425,85
26,35 - 28,95	8	27,65	221,20
28,95 - 31,55	5	30,25	151,25
31,55 - 34,15	3	32,85	98,55
Ogółem	60	x	1419,8

b. $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$ $\bar{x} = \frac{\sum_{i=1}^{N} x_i' \times n_i}{N} = \frac{1419,8}{60} = 23,66$ ha

Inne średnie klasyczne:

Średnia WAŻONA

Jeśli mamy do czynienia z danymi o różnej częstotliwości, to powinniśmy zastosować średnią ważoną zamiast średniej arytmetycznej, według wzoru:

$\bar{x}_w = \frac{\sum_{i=1}^{w}(x_i' \times w_i)}{\sum w_i}$ gdzie, x_i' - średnia pierwotna o w_i częstotliwości (wadze).

Przykład dotyczy 5 grup obszarowych sołectw o różnej liczebności i wielkości powierzchni gospodarstw:

Grupa	Średnia powierzchnia gospodarstwa (ha) x_i'	Liczba sołectw w_i	$x_i' \times w_i$
1	10	20	200
2	15	16	240
3	20	10	200
4	35	2	70
5	50	2	50
suma	x	50	760

$$\bar{x}_w = \frac{\sum_{i=1}^{w}(x_i' \times w_i)}{\sum w_i} = \frac{760}{50} = 15,2\ ha$$

Gdybyśmy omyłkowo policzyli tutaj średnią arytmetyczną dla powierzchni gospodarstwa otrzymalibyśmy 26,0 ha. Oby nikt nie próbował w ten sposób „zawyżać” średniej powierzchni gospodarstwa.

Średnia HARMONICZNA

Jeśli mamy do czynienia z danymi dotyczącymi cech (bądź zjawisk), które wyrażamy w wartościach względnych, np. zagęszczenie populacji jako liczba osobników na powierzchni 1 km^2. Wyniki podane w przeliczeniu na stałą jednostkę innej zmiennej, czyli w postaci wskaźników natężenia, np. prędkość pojazdu w km/h.

Średnia harmoniczna (Moment harmonii) jest odwrotnością średniej arytmetycznej obliczonej z odwrotności wartości cechy:

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

Przykład dotyczy zagęszczenia populacji gąsienic motyla *Mamestra dissimilis* w 3 miejscach na użytkach zielonych, liczonych każdorazowo na powierzchni 100 m^2 gdzie otrzymano: w miejscu A – 10 osobników, B – 200, C - 120. Obliczymy średnie zagęszczenie populacji gąsienic na 100 m^2.

$$\bar{x}_H = \frac{3}{\sum_{i=1}^{n} \frac{1}{10} + \frac{1}{200} + \frac{1}{120}} = 26{,}5 \text{ osobników / 100 m}^2$$

Średnia GEOMETRYCZNA

Średnią geometryczną należy stosować jeśli mamy do czynienia z cechami o charakterze dynamicznych zjawisk, np. w badaniach średniego tempa zmian w populacjach. Wyliczamy ją jako pierwiastek *n*-tego stopnia z iloczynu wartości, które są pod pierwiastkiem:

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

Przykład będzie dotyczył zmian w liczbie biedronek na jednej roślinie jaśminu w ciągu 4 miesięcy. Dodajmy, że na jaśminie nie stosowano żadnych środków chemicznych do zwalczania mszyc, a mszyce wybitnie lubią zasiedlać się i żerować na jaśminie, stąd też ich

wrogowie naturalni – drapieżne biedronki tym liczniej się na jaśminie pojawiają. Najpierw podamy wyniki ze zliczania biedronek w 4 terminach: na początku miesiąca maja było 10 osobników na roślinie, na koniec miesiąca maja było ich 12, na koniec czerwca 15, na koniec lipca 20 a na koniec sierpnia 28. To oznacza, że względne miesięczne przyrosty liczby biedronek wyniosły: V - 20%, VI - 25%, VII - 33% i w VIII - 40%. Aby obliczyć średni przyrost z 4 miesięcy należy zastosować pierwiastek 4-stopnia z iloczynu wskaźników tego przyrostu: 1,2 1,25 1,33 1,4.

$$\bar{x}_G = \sqrt[4]{1{,}2 \times 1{,}25 \times 1{,}33 \times 1{,}4} = \sqrt[4]{2{,}793} = 1{,}29$$

Średni przyrost liczby biedronek z miesiąca na miesiąc wyniósł 29 %.

Moda (DOMINANTA)

W szeregu prostym M_o inaczej zwana D wyznaczana jest bezpośrednio z szeregu uporządkowanego rosnąco lub malejąco. W przykładzie z 15 danymi pomiarów zawartości białka (%) w nasionach grochu: 22,6 **22,5 22,5** 22,0 21,9 21,8 21,7 21,5 21,4 21,3 21,0 20,9 **20,8 20,8** 19,8 są dwie mody, tj. Mo = 22,5 i Mo = 20,8. Te dwie dane powtarzają się dwukrotnie, a żadna inna wartość nie powtarza się więcej razy.
Mogą być szeregi jednomodalne, wielomodalne i bez mody.

a. W szeregu klasowym wartość najczęściej występującą wyznaczamy na podstawie następującego wzoru:

$$M_o = x_o + c_o \frac{n_d - n_{d-1}}{(n_d - n_{d-1}) + (n_d - n_{d+1})},$$

gdzie x_0 – dolna granica przedziału najliczniejszego, w którym znajduje się moda
c_o – rozpiętość przedziału klasowego – H
n_d – liczebność przedziału najliczniejszego, w którym znajduje się moda
n_{d-1} – liczebność przedziału poprzedzającego przedział mody
n_{d+1} – liczebność przedziału następującego po przedziale mody

Wyznaczymy modę na podstawie danych dotyczących powierzchni gospodarstw

Powierzchnia (ha)	Liczba gospodarstw (n_i)	
13,35 - 15,95	4	
15,95 - 18,55	8	
18,55 - 21,15	4	
21,15 - 23,75	11	n_{d-1}
x_0 → 23,75 - 26,35	17	n_d
26,35 - 28,95	8	n_{d+1}
28,95 - 31,55	5	
31,55 - 34,15	3	
Ogółem	60	

Rozpiętość przedziału wynosi tutaj $c_0=2,6$, $x_0=23,75$, $n_{d-1}=11$, $n_d=17$ i $n_{d+1}=8$.

Podstawiając do wzoru wszystkie potrzebne dane otrzymamy:

$$M_o = x_o + c_o \frac{n_d - n_{d-1}}{(n_d - n_{d-1}) + (n_d - n_{d+1})} = 23{,}75 + \left(2{,}6 \times \frac{17-11}{(17-11)+(17-8)}\right) = 24{,}79$$

Najczęściej powtarzana powierzchnia gospodarstwa wynosi 24,79 ha.

Mediana

a. Wyznaczenie wartości środkowej tj. mediany (symbol Me), która dzieli uporządkowany szereg liczbowy dokładnie na dwie równe części zależy od tego, czy mamy szereg parzysty, czy nieparzysty. Dla szeregu parzystego medianę wyznaczymy za pomocą wzoru: $M_e = \frac{x_k + x_{k+1}}{2}$, gdzie $k = N/2$ podczas gdy dla szeregu nieparzystego zastosujemy wzór: $Me = x_{\frac{N+1}{2}}$

Szereg prosty rosnący dla zanieczyszczeń materiału siewnego gryki (%) ma parzystą liczbę $N = 10$:

Me

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1,1	2,4	2,5	2,8	3,0	4,0	5,6	6,2	7,1	8,5

Według wzoru medianę wyliczymy jako średnią wartość z 5 i 6 pomiaru, tj. z 3,0 i 4,0, a to oznacza, że Me = *3,5*.

Dla przykładu z 15 pomiarami zawartości białka:

							Me							
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
19,8	20,8	20,8	20,9	21,0	21,3	21,4	21,5	21,7	21,8	21,9	22,0	22,5	22,5	22,6

Me wyznaczymy następująco:

$$Me = x_{\frac{N+1}{2}} = x_{\frac{15+1}{2}} = x_8 = 21{,}5$$

b. Wartość środkową Me dla szeregu klasowego wyznaczamy na podstawie wzoru:

$$M_e = x_o + (\frac{N}{2} - \sum_{i=1}^{k-1} n_i)\frac{c_o}{n_o}$$

gdzie x_0 – dolna granica przedziału klasowego mediany
$^N/_2$ – połowa sumy liczebności, która wskazuje na numer mediany
c_o – rozpiętość przedziału mediany
n_o – liczebność przedziału klasowego, w którym znajduje się mediana
$\sum_{i=1}^{k-1} n_i$ - skumulowana liczebność poprzedzająca przedział klasowy mediany

Dla szeregu klasowego z powierzchnią gospodarstw mamy następujące dane: $^N/_2$ = 60 / 2 = 30 Pytanie, w którym przedziale mieści się x_{30} ? Do tego potrzebna jest nam kolumna z liczebnościami skumulowanymi Cum ni. Jak widzimy w wierszu 5 mieszczą się wartości od 27 do 44, a więc x o kolejności 30 jest w przedziale (klasie) 5.

Lp	Powierzchnia (ha)	Liczba gospodarstw (n_i)	Cum ni
1	13,35 - 15,95	4	4
2	15,95 - 18,55	8	12
3	18,55 - 21,15	4	16
4	21,15 - 23,75	11	27
5	23,75 - 26,35	17	44
6	26,35 - 28,95	8	52
7	28,95 - 31,55	5	57
8	31,55 - 34,15	3	60
	Ogółem	60	x

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

x_0 n_0

$$M_e = x_o + (\frac{N}{2} - \sum_{i=1}^{k-1} n_i)\frac{c_o}{n_o} = 23{,}75 + (30 - 27) \times \frac{2{,}6}{17} = 24{,}21$$

Wartość, która dzieli całą populacje powierzchni gospodarstw na dwie równe części to 24,21 ha. Oznacza to, że poniżej 24,21 ha jest 50% i powyżej 24,21 ha też 50% gospodarstw.

W szeregach łagodnie asymetrycznych zachodzi relacja pomiędzy modalną a medianą:

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

Kwartyle (ćwiartki)

Kwartyle pierwszy i trzeci (symbol Q_1 i Q_3), dzielą uporządkowany szereg liczbowy dokładnie w ¼ i ¾ jego długości, co oznacza, że 25% wyników znajduje się poniżej Q_1, a 75% danych jest powyżej Q_1, z kolei poniżej Q_3 jest 75% danych, a powyżej Q_3 jest 25% danych.

a.

$$Q_1 = x_{k,gdzie}\, k = \frac{N+1}{4}$$

$$Q_3 = x_{k,gdzie}\, k = \frac{3(N+1)}{4}$$

Dla przykładu z 15 pomiarami zawartości białka:

			Q_1								Q_3			
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
19,8	20,8	20,8	20,9	21,0	21,3	21,4	21,5	21,7	21,8	21,9	22,0	22,5	22,5	22,6
25%				50%								25%		

$$Q_1 = x_{k,gdzie}\, k = \frac{N+1}{4} = \frac{15+1}{4} = 4$$

co oznacza, że kwartyl 1 to 20,9% zawartości białka

$$Q_3 = x_{k,gdzie}\, k = \frac{3(N+1)}{4} = \frac{3(16)}{4} = 12$$

co oznacza, że kwartyl 3 to 22,0% zawartości białka

Interpretując obydwie ćwiartki powiemy, że poniżej zawartości 20,9% białka mamy ¼ badanych nasion, a powyżej 22,0% zawartości białka również ¼ nasion grochu.

Dla szeregu klasowego wyznaczamy kwartyle według następujących wzorów

b. $Q_1 = x_o + (\frac{N}{4} - \sum_{i=1}^{l-1} n_i)\frac{c_o}{n_o}$ $Q_3 = x_o + (\frac{3N}{4} - \sum_{i=1}^{l-1} n_i)\frac{c_o}{n_o}$, gdzie

x_0 – dolna granica przedziału klasowego kwartylu
$^N/_4$ – ¼ sumy liczebności, która wskazuje na numer 1 kwartylu
$^{3N}/_4$ – ¾ sumy liczebności, która wskazuje na numer 3 kwartylu
c_o – rozpiętość przedziału
n_o – liczebność przedziału klasowego, w którym znajduje się kwartyl
$\sum_{i=1}^{k-1} n_i$ - skumulowana liczebność poprzedzająca przedział klasowy kwartylu

	Powierzchnia (ha)	Liczba gospodarstw (n_i)	Cum ni	
	13,35 - 15,95	4	4	
	15,95 - 18,55	8	12	$\sum_{i=1}^{k-1} n_i$
Q_1 x_0	18,55 - 21,15	4	16	n_o
	21,15 - 23,75	11	27	
	23,75 - 26,35	17	44	$\sum_{i=1}^{k-1} n_i$
Q_3 x_0	26,35 - 28,95	8	52	n_o
	28,95 - 31,55	5	57	
	31,55 - 34,15	3	60	
	Ogółem	60	x	

$$Q_1 = x_o + (\frac{N}{4} - \sum_{i=1}^{l-1} n_i)\frac{c_o}{n_o} = 18{,}55 + (15 - 12) \times \frac{2{,}6}{4} = 20{,}50$$

$$Q_3 = x_o + (\frac{3N}{4} - \sum_{i=1}^{l-1} n_i)\frac{c_o}{n_o} = 26{,}35 + (45 - 44) \times \frac{2{,}6}{8} = 26{,}68$$

7.2.Miary zmienności (rozproszenia, rozrzutu)

Rozstęp i rozstęp ćwiartkowy

$$R = x_{\max} - x_{\min} = 22{,}6 - 19{,}8 = 2{,}8$$

x_{min} Q_1 Q_3 x_{max}

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
19,8	20,8	20,8	20,9	21,0	21,3	21,4	21,5	21,7	21,8	21,9	22,0	22,5	22,5	22,6

$$q = Q_3 - Q_1 = 22{,}0 - 20{,}9 = 1{,}1$$

Wariancja

Wariancja należy do Momentu 2-rzędu, ponieważ do jej wyliczenia potrzebna jest suma kwadratów odchyleń każdego pomiaru od wartości średniej $(x_i - \bar{x})^2$. Z tego też względu wariancja nie posiada jednostki fizycznej.

a. $s^2 = \dfrac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}{N}$, dla próby małej $s^2 = \dfrac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}$ prostszy wzór: $s^2 = \dfrac{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}{n-1}$

Dla przykładu z 15 danymi pomiarów zawartości białka (%) w nasionach grochu utworzymy szereg prosty do obliczeń odchyleń prostych, kwadratów odchyleń ora kwadratów wartości x.

n_i	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	x_i^2
1	22,6	1,10	1,21	510,76
2	22,5	1,00	1,00	506,25
3	22,5	1,00	1,00	506,25
4	22,0	0,50	0,25	484,00
5	21,9	0,40	0,16	479,61
6	21,8	0,30	0,09	475,24
7	21,7	0,20	0,04	470,89
8	21,5	0,00	0,00	462,25
9	21,4	-0,10	0,01	457,96
10	21,3	-0,20	0,04	453,69
11	21,0	-0,50	0,25	441,00
12	20,9	-0,60	0,36	436,81
13	20,8	-0,70	0,49	432,64
14	20,8	-0,70	0,49	432,64
15	19,8	-1,70	2,89	392,04
suma	322,5	0,00	8,28	6942,03

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{8{,}28}{14} = 0{,}59$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{6942{,}03 - \frac{(322{,}5^2}{15}}{14} = \frac{8{,}28}{14} = 0{,}59$$

Wariancję dla danych w szeregu klasowym wyliczymy na analizowanym przykładzie z powierzchnią gospodarstw, w tabeli będą nam potrzebne odchylenia środków klas od średniej oraz kwadraty odchyleń środków klas od średniej pomnożone przez liczności w klasach.

Powierzchnia (ha)	Liczba gospodarstw (n_i)	Środek klasy x_i'	$x_i' \times n_i$	$(x_i' - \bar{x})$	$(x_i' - \bar{x})^2$	$(x_i' - \bar{x})^2 \times n_i$
13,35 - 15,95	4	14,65	58,60	-9,01	81,18	324,72
15,95 - 18,55	8	17,25	138,00	-6,41	41,09	328,70
18,55 - 21,15	4	19,85	79,40	-3,81	14,52	58,06
21,15 - 23,75	11	22,45	246,95	-1,21	1,46	16,1051
23,75 - 26,35	17	25,05	425,85	1,39	1,93	32,8457
26,35 - 28,95	8	27,65	221,20	3,99	15,92	127,3608
28,95 - 31,55	5	30,25	151,25	6,59	43,43	217,1405
31,55 - 34,15	3	32,85	98,55	9,19	84,46	253,3683
Ogółem	60	x	1419,8	x	x	1358,31

b. $s^2 = \frac{\sum_{i=1}^{k}(x_i' - \bar{x})^2 \times n_i}{N}$

$$s^2 = \frac{\sum_{i=1}^{k}(x_i' - \bar{x})^2 \times n_i}{N} = \frac{1358{,}31}{60} = 22{,}64$$

Jeśli jest niewielka liczba klas (K < 12) to stosujemy poprawkę Shepparda

Poprawka Shepparda: $\frac{1}{12}H^2$, gdzie – H to szerokość klas $\quad s^2_{popr} = s^2 - \frac{1}{12}H^2$

$$s^2_{popr} = s^2 - \frac{1}{12}H^2 = 22{,}64 - \frac{2{,}6^2}{12} = 22{,}64 - 0{,}56 = 22{,}08$$

Do dalszych obliczeń należy wykorzystywać wariancję poprawioną.

Odchylenie standardowe

Dla danych w szeregu prostym i w szeregu klasowym wartość odchylenia standardowego liczy się tak samo. Jest to pierwiastek drugiego stopnia z wariancji. Tym samym, wartość odchylenia standardowego jest wyrażona w jednostkach fizycznych, w których dokonano pomiaru badanej cechy.

a, b. $s = \sqrt{s^2}$

$$s = \sqrt{s^2} = \sqrt{0{,}59} = 0{,}77\ \%$$

Wartość odchylenia standardowego dla zawartości białka w nasionach grochu wynosi
s = 0,77 %

$$s = \sqrt{s^2} = \sqrt{22{,}08} = 4{,}70\ ha$$

Wartość odchylenia standardowego dla powierzchni gospodarstw wynosi 4,70 ha.

TYPOWY OBSZAR ZMIENNOŚCI

Mając wyliczoną średnią oraz odchylenie standardowe możemy policzyć obszar zmienności typowej dla analizowanej cechy.

$\overline{X} - S < X_{TYP} < \overline{X} + S$ - w tym obszarze mieści się około 2/3 wszystkich jednostek badanej zbiorowości, tj. 68% danych reprezentujących cechę.

Dla zawartości białka w nasionach grochu przy $\bar{x}$ = 21,50% i s = 0,77% otrzymamy zakres typowej zawartości:

$$21{,}50 - 0{,}77 < X_{TYP} < 21{,}50 + 0{,}77$$
$$20{,}73 < X_{TYP} < 22{,}27\ (\%)$$

Dla powierzchni gospodarstw przy $\bar{x}$ = 23,66 ha i s = 4,70 ha otrzymamy zakres typowej powierzchni:

$$23{,}66 - 4{,}70 < X_{TYP} < 23{,}66 + 4{,}70$$
$$18{,}96 < X_{TYP} < 28{,}36\ (ha)$$

OBSZAR NORMY STATYSTYCZNEJ

Norma statystyczna pokrywa 95% wartości analizowanej cechy. Obszar dla danych normy tworzy się za pomocą formuły: $\bar{x}$ – (2 x s) < X_{NORMY} < $\bar{x}$ + (2 x s)
Jest to przedział dwa razy dłuży od przedziału wartości typowych.

Dla zawartości białka w nasionach grochu przy $\bar{x}$ = 21,50% i s = 0,77% otrzymamy zakres normatywnej zawartości:

$$21{,}50 - (2 \times 0{,}77) < X_{NORMY} < 21{,}50 + (2 \times 0{,}77)$$
$$19{,}96 < X_{NORMY} < 23{,}04 \ (\%)$$

Dla powierzchni gospodarstw przy $\bar{x}$ = 23,66 ha i s = 4,70 ha otrzymamy zakres normatywnej powierzchni:

$$23{,}66 - (2 \times 4{,}70) < X_{TYP} < 23{,}66 + (2 \times 4{,}70)$$
$$14{,}27 < X_{TYP} < 33{,}06 \ (\text{ha})$$

Współczynnik zmienności względnej

Na podstawie średniej i odchylenia standardowego można wyliczyć współczynnik zmienności względnej, wyrażony w %. Interpretuje on udział średniej w odchyleniu standardowym i daje obraz jak duża jest zmienność badanej populacji. Zakresy zmienności zostały opisane w poniższej tabeli.

a,b. $v = \frac{s}{\bar{x}} \times 100$

v (%)	Opis zmienności
Do 5	Mała
5,1-10,0	Umiarkowana
10,1-20,0	Średnia
20,1-50	Duża
> 50,1	Bardzo duża

Dla zawartości białka w nasionach grochu przy $\bar{x}$ = 21,50% i s = 0,77% otrzymamy współczynnik zmienności:

$$v = \frac{s}{\bar{x}} \times 100 = \frac{0,77}{21,50} x100 = 3,58\ \%$$

Powiemy, że badana próba nasion grochu cechuje się małą zmiennością pod względem zawartości białka (poniżej 5%).

Dla powierzchni gospodarstw przy $\bar{x}$ = 23,66 ha i s = 4,70 ha otrzymamy współczynnik zmienności:

$$v = \frac{s}{\bar{x}} \times 100 = \frac{4,70}{23,66} x100 = 19,86\ \%$$

Współczynnik zmienności dla powierzchni gospodarstw mieści się w zakresie zmienności średniej (od 10,1 do 20,0%).

Odchylenie standardowe średniej - błąd średniej

Każda średnia ma swoje odchylenie standardowe, które nazywamy jej błędem. Zależy ono od wielkości próby, z której pochodzi. Im większa próba tym mniejszy błąd średniej. Dla prób o małej liczebności przy tej samej średnie błąd będzie zawsze większy. Wynika to ze wzoru, bowiem liczność próby pod pierwiastkiem jest w mianowniku wzoru na błąd średniej:

a,b. $s_{\bar{x}} = \frac{s}{\sqrt{n}}$

Dla zawartości białka w nasionach grochu przy s = 0,77% i n = 15 otrzymamy błąd średniej:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{0,77}{\sqrt{15}} = 0,20\ \%$$

Dla powierzchni gospodarstw przy s = 4,70ha i n = 60 otrzymamy błąd średniej:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{4,7}{\sqrt{60}} = 0,61\ ha$$

7.3.Miary asymetrii

Dla próby dużej można wykonać diagnozę rozkładu cechy pod kątem kształtu histogramu i diagramu (te zagadnienia były szczegółowo omówione w Rozdziale 4).

Do diagnozy rozkładu pod względem odstęp od symetryczności rozkładu służą miary asymetrii, które pochodzą od Momentu 3-rządu.

Momentem 3-tego rzędu nazywamy średnią arytmetyczną z odchyleń poszczególnych wartości zmiennej od średniej arytmetycznej podniesionych do 3-potęgi.

Moment trzeci :

$$M_3 = \frac{\sum_{i=1}^{N} \left(x'_i - \bar{x}\right)^3 \times n_i}{N}$$

Współczynnik asymetrii:

$$As = \frac{M_3}{s^3}$$

W rozkładzie symetrycznym zachodzi relacja:

$$\bar{x} = M_e = M_o$$

As = 0 – rozkład jest symetryczny

As > 0 – rozkład jest asymetryczny, prawostronnie skośny

As < 0 – rozkład jest asymetryczny, lewostronnie skośny

As od 0 do 2 - asymetria nie jest zbyt silna

Wskaźnik skośności:

$$W_s = \bar{x} - M_o$$

Ws = 0 – rozkład jest symetryczny

Ws > 0 – prawostronnie skośny

Ws < 0 – lewostronnie skośny

Klasyczno-pozycyjny współczynnik asymetrii:

$$A_s = \frac{3(\bar{x} - M_e)}{s}$$

Powierzchnia (ha)	Liczba gospodarstw (n_i)	Środek klasy x_i'	$x_i' \times n_i$	$(x_i' - \bar{x})$	$(x_i' - \bar{x})^3$	$(x_i' - \bar{x})^3 \times n_i$
13,35 - 15,95	4	14,65	58,60	-9,01	-731,43	-2925,73
15,95 - 18,55	8	17,25	138,00	-6,41	-263,37	-2107,00
18,55 - 21,15	4	19,85	79,40	-3,81	-55,31	-221,23
21,15 - 23,75	11	22,45	246,95	-1,21	-1,77	-19,49
23,75 - 26,35	17	25,05	425,85	1,39	2,69	45,66
26,35 - 28,95	8	27,65	221,20	3,99	63,52	508,17
28,95 - 31,55	5	30,25	151,25	6,59	286,19	1430,96
31,55 - 34,15	3	32,85	98,55	9,19	776,15	2328,45
Ogółem	60	x	1419,8	x	x	-960,21

$$M_3 = \frac{\sum_{i=1}^{N}(x'_i - \bar{x})^3 \times n_i}{N} = \frac{-960{,}21}{60} = -16{,}00$$

Moment 3-rzędu jest wartością ujemną, co wskazuje już na ewentualną skośność lewostronną (patrz podrozdział 4.3.)

$$As = \frac{M_3}{s^3} = \frac{-16{,}00}{4{,}70^3} = -0{,}15$$

Klasyczny parametr asymetrii As jest mniejszy od 0 co mówi nam, że rozkład jest asymetryczny, lewostronnie skośny. Jest to jednak bardzo niewielka asymetria, ze względu na wartość bezwzględną poniżej 1.

Klasyczno – pozycyjny parametry asymetrii jest również ujemny i jego wartość -0,3 wskazuje podobnie na niewielką asymetrię lewostronną.

$$A_s' = \frac{3 \times (\bar{x} - M_e)}{s} = \frac{3 \times (23{,}66 - 24{,}21)}{4{,}70} = -0{,}3$$

Ws = $\bar{x}$ - Mo = 23,66 - 24,79 = - 1,13

Wskaźnik skośności, który bierze pod uwagę różnice pomiędzy średnia a wartością modalną jest w tym przykładzie również wartością ujemną.

7.4.Miara koncentracji wokół średniej - kurtoza

Momentem 4-tego stopnia nazywamy średnią arytmetyczną z odchyleń poszczególnych wartości zmiennej od średniej arytmetycznej podniesionych do 4-potęgi.

Moment czwarty wyliczamy:

$$M_4 = \frac{\sum_{i=1}^{N} \left(x'_i - \bar{x}\right)^4 \times n_i}{N}$$

Moment 4 służy do wyliczenia Kurtozy, która jest miarą koncentracji wyników wokół wartości średniej

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

Opis koncentracji rozkładu zależy od wartości i znaku K

K = 3 – rozkład jest normalny, o koncentracji wokół średniej normalnej

K < 3 – rozkład jest spłaszczony (platykurtyczny) o skupieniu słabszym od normalnego

K > 3 – rozkład jest wysmukły (leptokurtyczny) o skupieniu silniejszym od normalnego

Wartość K odbiegająca o +/- jedną jednostkę nie stanowi poważnego zagrożenia dla rozkładu.

Powierzchnia (ha)	Liczba gospodarstw (n_i)	Środek klasy x_i'	x_i' x n_i	$(x_i' - \bar{x})$	$(x_i' - \bar{x})^4$	$(x_i' - \bar{x})^4$ x n_i
13,35 - 15,95	4	14,65	58,60	-9,01	6590,21	26360,83
15,95 - 18,55	8	17,25	138,00	-6,41	1688,23	13505,86
18,55 - 21,15	4	19,85	79,40	-3,81	210,72	842,87
21,15 - 23,75	11	22,45	246,95	-1,21	2,14	23,58
23,75 - 26,35	17	25,05	425,85	1,39	3,73	63,46
26,35 - 28,95	8	27,65	221,20	3,99	253,45	2027,60
28,95 - 31,55	5	30,25	151,25	6,59	1886,00	9430,00
31,55 - 34,15	3	32,85	98,55	9,19	7132,83	21398,50
Ogółem	60	x	1419,8	x	x	73652,69

$$M_4 = \frac{\sum_{i=1}^{N} (x'_i - \bar{x})^4 \times n_i}{N} \frac{73652,7}{60} = 1227,5$$

$$K = \frac{M_4}{s^4} = \frac{1227,5}{4,70^4} = 2,52$$

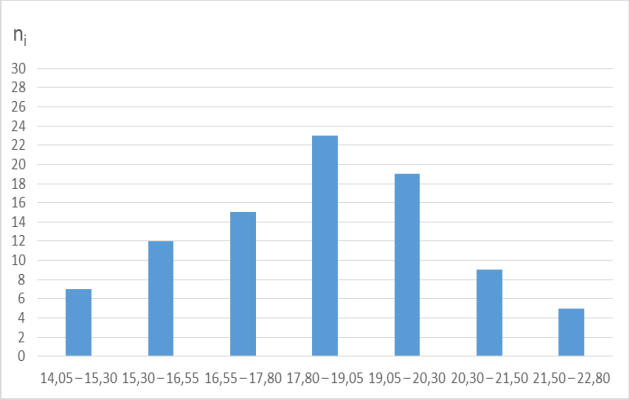
Na podstawie wyliczonej kurtozy = 2,52 stwierdzimy, że mamy do czynienia z rozkładem lekko platykurtycznym, to znaczy, że rozproszenie powierzchni gospodarstw wokół średniej arytmetycznej jest o 0,5 stopnia za duże w odniesieniu do rozkładu normalnego.

8.Kompleksowa analiza danych do opisu statystycznego

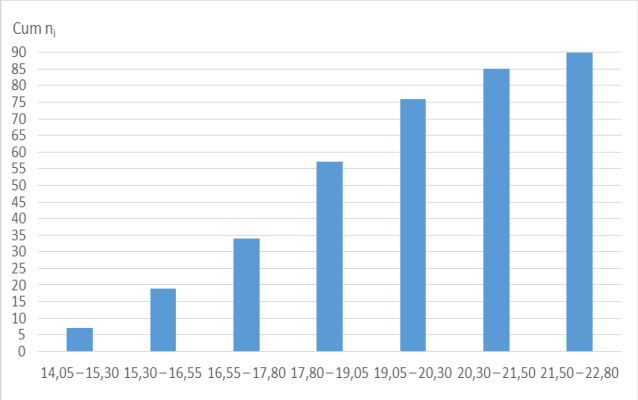
Przykład, który został już zaprezentowany wcześniej (str. 26-28) wykorzystamy do kompleksowej analizy opisu statystycznego wraz z wizualizacja danych. Przypomnijmy, że dotyczy pomiarów 90 roślin długości łodygi narcyza odmiany Ice Flower w cm.

Pierwszym krokiem w opisie statystycznym badanej populacji jest przygotowanie danych w szeregu przedziałowym do zobrazowania rozkładu. Korzystamy z kolumn 2,3,5 i 6.

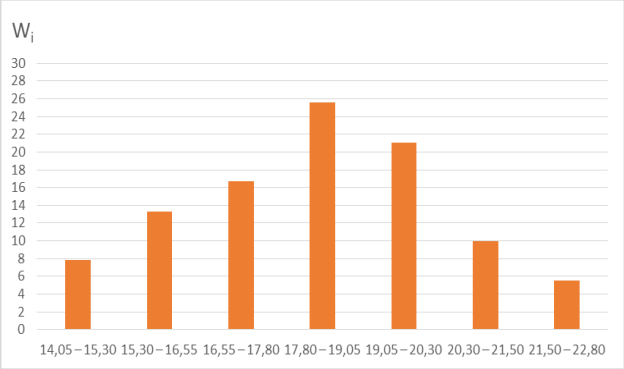
Długość łodygi narcyza odmiany Ice Flower [cm] x_i w zakresach	Liczba roślin (n_i)	Cum n_i	n_i / N	W_i n_i / N x 100	Cum W_i
1	2	3	4	5	6
14,05 – 15,30	7	7	0,078	7,8	7,8
15,30 – 16,55	12	19	0,133	13,3	21,1
16,55 – 17,80	15	34	0,167	16,7	37,8
17,80 – 19,05	23	57	0,256	25,6	63,4
19,05 – 20,30	19	76	0,211	21,1	84,5
20,30 – 21,50	9	85	0,100	10,0	94,5
21,50 – 22,80	5	90	0,055	5,5	100,0
Ogółem	90	x	x	100,0	x



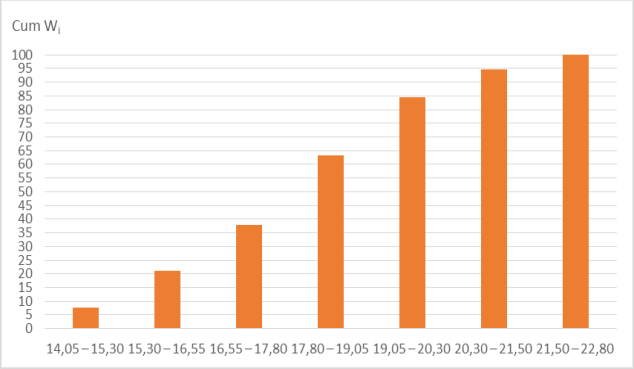
Wykres 4. Histogram liczności łodyg narcyza w klasach długości (cm).



Wykres 5. Histogram skumulowanej liczności łodyg narcyza w klasach długości (cm).



Wykres 6. Histogram udziału % łodyg narcyza w klasach długości (cm).



Wykres 7. Histogram skumulowanych udziałów % łodyg narcyza w klasach długości (cm).

Moment statystyczny 1 to miary centralnego położenia

Moda

Długość łodygi narcyza odmiany Ice Flower [cm] x_i w zakresach	Liczba roślin (n_i)	
14,05 – 15,30	7	
15,30 – 16,55	12	
16,55 – 17,80	15	n_{d-1}
x_0 17,80 – 19,05	23	n_d
19,05 – 20,30	19	n_{d+1}
20,30 – 21,50	9	
21,50 – 22,80	5	
Ogółem	90	

Rozpiętość przedziału wynosi tutaj $c_0 = 1{,}25$.
Podstawiając do wzoru wszystkie potrzebne dane otrzymamy:

$$M_o = x_o + c_o \frac{n_d - n_{d-1}}{(n_d - n_{d-1}) + (n_d - n_{d+1})} = 17{,}80 + (1{,}25 \times \frac{23 - 15}{(23 - 15) + (23 - 19)} = 18{,}63$$

Najczęściej powtarzana wartość, czyli Moda wynosi 18,63 cm.

Mediana

Dla szeregu klasowego z długością łodygi narcyza mamy następujące dane:

$^{N}/_{2} = 90 / 2 = 45$

Pytanie, w którym przedziale mieści się x_{45}? Do tego potrzebna jest nam kolumna z liczebnościami skumulowanymi Cum ni. Jak widzimy w wierszu 4 mieszczą się wartości od 34 do 57, a więc x_{45} jest w tym wierszu.

Długość łodygi narcyza odmiany Ice Flower [cm] x_i w zakresach	Liczba roślin (n_i)	Cum ni	
14,05 – 15,30	7	7	
15,30 – 16,55	12	19	
16,55 – 17,80	15	34	$\sum_{i=1}^{k-1} n_i$
x_0 17,80 – 19,05	23	57	n_0
19,05 – 20,30	19	76	
20,30 – 21,50	9	85	
21,50 – 22,80	5	90	
Ogółem	90	-	

$$M_e = x_o + (\frac{N}{2} - \sum_{i=1}^{k-1} n_i)\frac{c_o}{n_o} = 17{,}80 + (45 - 34) \times \frac{1{,}25}{23} = 18{,}40$$

Kwartyle Q_1 i Q_3

$$Q_1 = x_o + (\frac{N}{4} - \sum_{i=1}^{l-1} n_i)\frac{c_o}{n_o} \qquad Q_3 = x_o + (\frac{3N}{4} - \sum_{i=1}^{l-1} n_i)\frac{c_o}{n_o}, \text{ gdzie}$$

x_0 – dolna granica przedziału klasowego kwartylu
$^N/_4$ – ¼ sumy liczebności, która wskazuje na numer 1 kwartylu, tj. 90/4 = $x_{22,5}$
$^{3N}/_4$ – ¾ sumy liczebności, która wskazuje na numer 3 kwartylu, tj. 3 x 90/4 = $x_{67,5}$
c_o – rozpiętość przedziału
n_o – liczebność przedziału klasowego, w którym znajduje się kwartyl
$\sum_{i=1}^{k-1} n_i$ - skumulowana liczebność poprzedzająca przedział klasowy kwartylu

	Długość łodygi narcyza odmiany Ice Flower [cm] x_i w zakresach	Liczba roślin (n_i)	Cum ni
	14,05 – 15,30	7	7
	15,30 – 16,55	12	19
przedział Q1	16,55 – 17,80	15	34
	17,80 – 19,05	23	57
Przedział Q3	19,05 – 20,30	19	76
	20,30 – 21,50	9	85
	21,50 – 22,80	5	90
	Ogółem	90	-

$$Q_1 = x_o + (\frac{N}{4} - \sum_{i=1}^{l-1} n_i)\frac{c_o}{n_o} = 16{,}55 + (22{,}5 - 19) \times \frac{1{,}25}{15} = 16{,}84$$

$$Q_3 = x_o + (\frac{3N}{4} - \sum_{i=1}^{l-1} n_i)\frac{c_o}{n_o} = 19{,}05 + (67{,}5 - 57) \times \frac{1{,}25}{19} = 19{,}75$$

Długość łodygi narcyza odmiany Ice Flower [cm] x_i w zakresach	Liczba roślin (n_i)	Środek klasy x_i'	$x_i' \times n_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \times n_i$
1	2	7	8	9	10	11
14,05 – 15,30	7	14,675	102,725	-3,635	13,213	92,493
15,30 – 16,55	12	15,925	191,100	-2,385	5,688	68,259
16,55 – 17,80	15	17,175	257,625	-1,135	1,288	19,323
17,80 – 19,05	23	18,425	423,775	0,115	0,013	0,304
19,05 – 20,30	19	19,675	373,825	1,365	1,863	35,401
20,30 – 21,50	9	20,925	188,325	2,615	6,838	61,544
21,50 – 22,80	5	22,175	110,875	3,865	14,938	74,691
Suma	90	x	1648,25	x	x	352,015

Średnia arytmetyczna:

Do wyliczenia średniej arytmetycznej potrzebne są kolumny 2,7,8.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i' \times n_i}{N} = \frac{1648{,}25}{90} = 18{,}31 cm$$

W szeregach łagodnie asymetrycznych zachodzi relacja pomiędzy modalną a medianą:

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

17,97 = 18,31 – 3 x (18,31 – 17,93)

17,97 = 17,17

Moment statystyczny 2 to miary zmienności
Wariancja:
Do wyliczenia Momentu-2 rzędu potrzebna jest kolumna 11.

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

$$s^2 = \frac{\sum_{i=1}^{k} (x_i' - \bar{x})^2 \times n_i}{N} = \frac{352{,}015}{90} = 3{,}91$$

$$s_{popr}^2 = s^2 - \frac{1}{12} H^2 = 3{,}91 - \frac{1{,}25^2}{12} = 3{,}91 - 0{,}13 = 3{,}78$$

Odchylenie standardowe:

$$s = \sqrt{s^2} = \sqrt{3{,}78} = 1{,}94\ cm$$

Współczynnik zmienności względnej:

$$v = \frac{s}{\bar{x}} \times 100 = \frac{1,94}{18,31} x100 = 10,60\ \%$$

Błąd średniej:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{1,94}{\sqrt{90}} = 0,20\ cm$$

Dla $\bar{x}$ = 18,31cm i s = 1,94cm otrzymamy zakres typowej długości pędów narcyza:

$$18,31 - 1,94 < X_{TYP} < 18,31 + 1,94$$

$$16,37 < X_{TYP} < 20,25\ (cm)$$

Obszar dla danych normy tworzy się za pomocą formuły: $\bar{x} - (2 \times s) < X_{NORMY} < \bar{x} + (2 \times s)$

$$18,31 - (2 \times 1,94) < X_{NORMY} < 18,31 + (2 \times 1,94)$$

$$14,43 < X_{NORMY} < 22,19\ (cm)$$

Moment statystyczny 3 – to miary asymetrii

Do wyliczenia Momentu 3-rzędu oraz miar asymetrii są potrzebne kolumny nr 12 i 13.

Długość łodygi narcyza odmiany Ice Flower [cm] x_i w zakresach	Liczba roślin (n_i)	Środek klasy x_i'	$x_i' \times n_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^3 \times n_i$
1	2	7	8	9	12	13
14,05 – 15,30	7	14,675	102,725	-3,635	-48,03	-336,21
15,30 – 16,55	12	15,925	191,100	-2,385	-13,57	-162,80
16,55 – 17,80	15	17,175	257,625	-1,135	-1,46	-21,93
17,80 – 19,05	23	18,425	423,775	0,115	0,00	0,03
19,05 – 20,30	19	19,675	373,825	1,365	2,54	48,32
20,30 – 21,50	9	20,925	188,325	2,615	17,88	160,94
21,50 – 22,80	5	22,175	110,875	3,865	57,74	288,68
Suma	90	-	1648,250	-		-22,96

$$M_3 = \frac{\sum_{i=1}^{N}(x'_i - \bar{x})^3 \times n_i}{N} = \frac{-22,96}{90} = -0,26$$

$$As = \frac{M_3}{s^3} = \frac{-0,26}{1,94^3} = -0,035$$

$$A_s' = \frac{3 \times (\bar{x} - M_e)}{s} = \frac{3 \times (18,31 - 18,40)}{1,94} = -0,14$$

Moment statystyczny 4 – to miara koncentracji

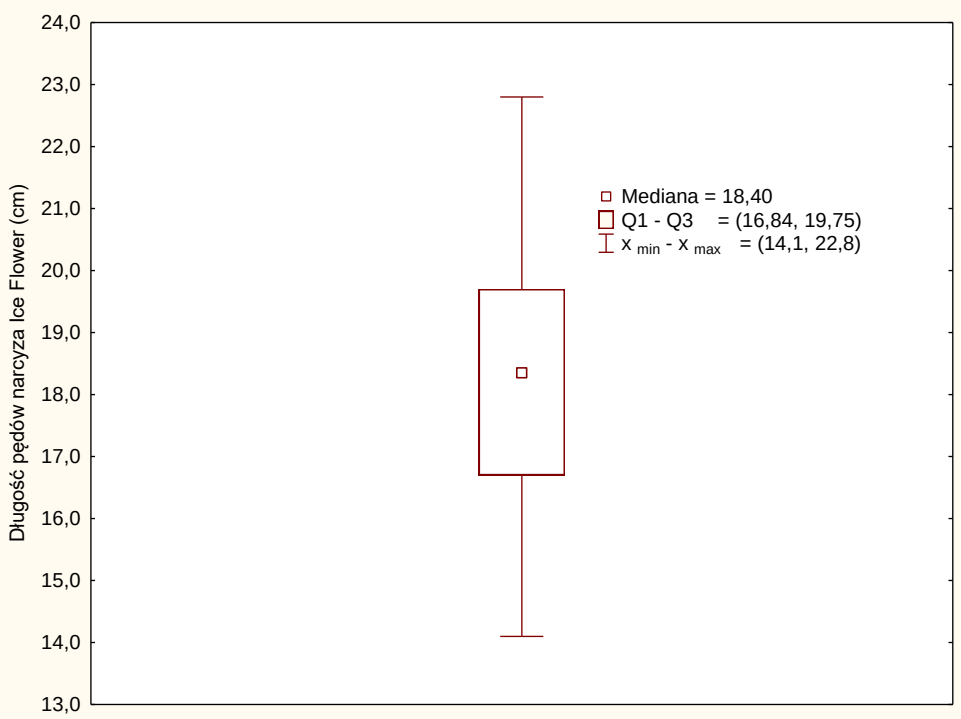
Do wyliczenia Momentu 4-rzędu i kurtozy potrzebne są kolumny nr 14 i 15.

Długość łodygi narcyza odmiany Ice Flower [cm] x_i w zakresach	Liczba roślin (n_i)	Środek klasy x_i'	$x_i' \times n_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^4$	$(x_i - \bar{x})^4 \times n_i$
1	2	7	8	9	14	15
14,05 – 15,30	7	14,675	102,725	-3,635	174,59	1222,13
15,30 – 16,55	12	15,925	191,100	-2,385	32,36	388,27
16,55 – 17,80	15	17,175	257,625	-1,135	1,66	24,89
17,80 – 19,05	23	18,425	423,775	0,115	0,00	0,00
19,05 – 20,30	19	19,675	373,825	1,365	3,47	65,96
20,30 – 21,50	9	20,925	188,325	2,615	46,76	420,85
21,50 – 22,80	5	22,175	110,875	3,865	223,15	1115,75
Suma	90	-	1648,250	-		3237,86

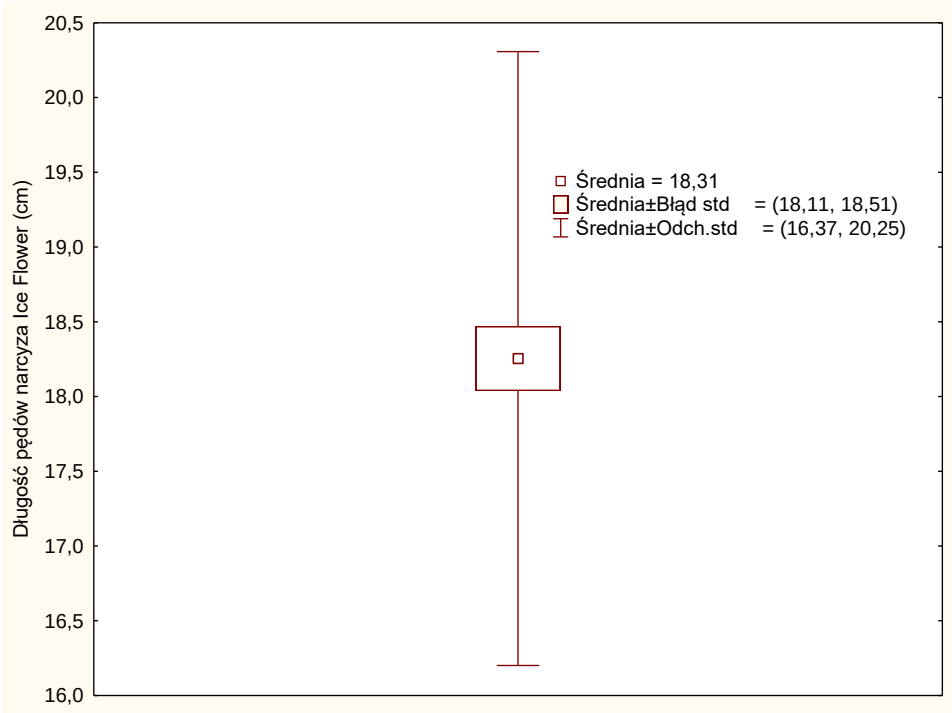
$$M_4 = \frac{\sum_{i=1}^{N}(x'_i - \bar{x})^4 \times n_i}{N} = \frac{3237{,}86}{90} = 35{,}98$$

$$K = \frac{M_4}{s^4} = \frac{35{,}98}{1{,}94^4} = 2{,}54$$

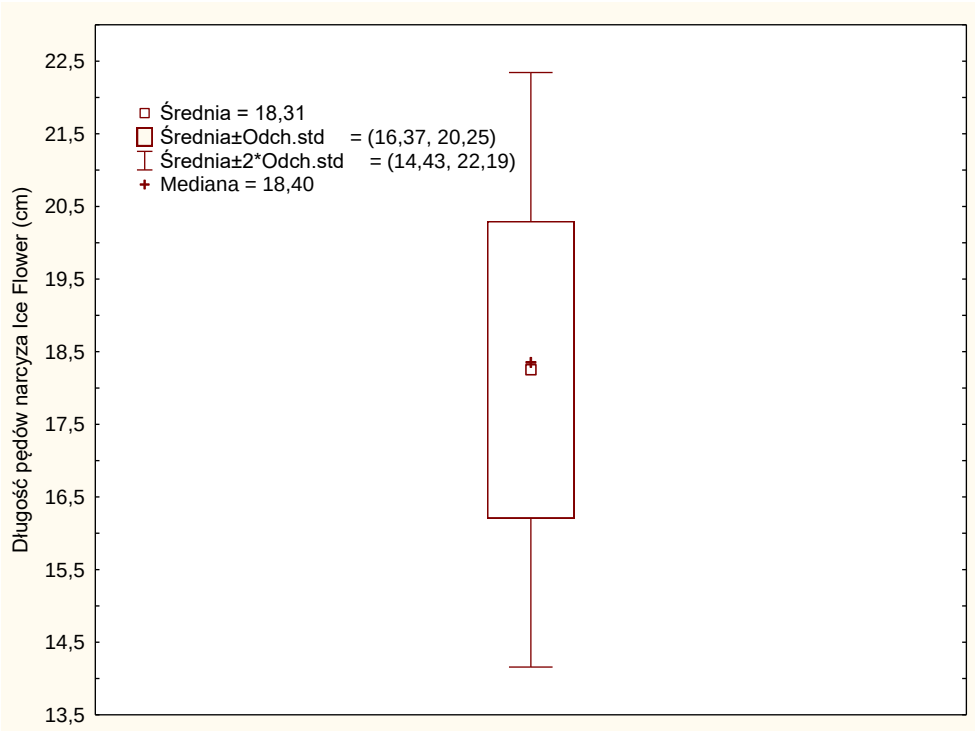
Wizualizacja opisu statystycznego



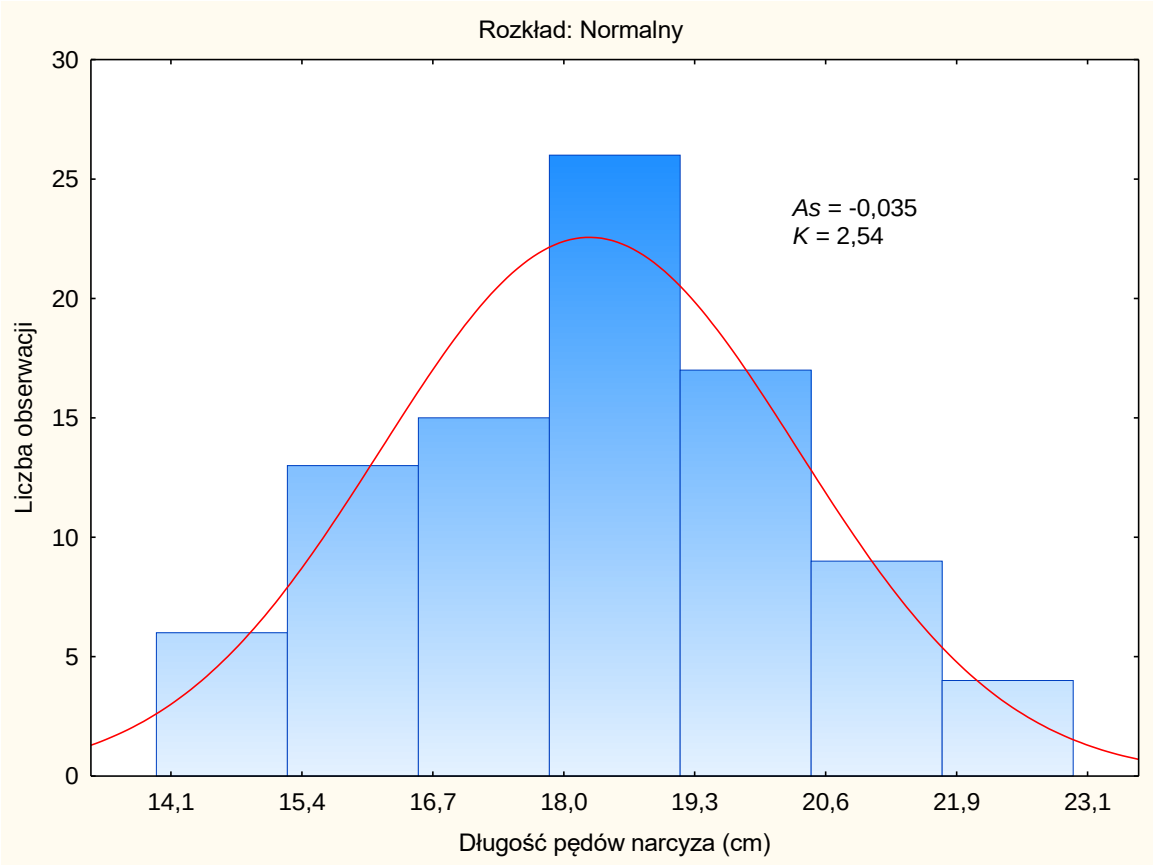
Wykres 8. Rozstęp całkowity, rozstęp ćwiartkowy oraz mediana dla długości pędów narcyza (cm).



Wykres 9. Zakresy zmienności średniej oraz typowej zmienności dla długości pędów narcyza (cm).



Wykres 10. Zakresy zmienności typowej zmienności oraz normy dla długości pędów narcyza (cm).



Wykres 11. Histogram liczebności oraz diagram rozkładu normalnego dla długości pędów narcyza (cm).

ROZDZIAŁ III **STATYSTYKA MATEMATYCZNA**

Spis treści

1.Badanie częściowe i zagadnienia próby statystycznej

Na początku należy wyjaśnić pojęcia, którymi posługujemy się w statystyce. Poza samymi definicjami przytoczonymi poniżej uwzględnione zostały objaśnienia pomocne dla ich zrozumienia.

Populacja przedmiotowa (zbiorowość statystyczna, masa statystyczna) - Zbiorowość stanowiąca przedmiot badań, do której odnoszą się wnioski.

Zbiorowość ta obejmuje dowolne elementy podobne pod względem określonych cech. Populację przedmiotową określa cel badań - należy zatem jednoznacznie sprecyzować jakie elementy będą wchodziły w skład tej populacji. Dla przykładu jeśli celem badań będzie rozpoznanie reakcji roślin pszenicy ozimej na nawożenie, to populacją przedmiotową będą wszystkie rośliny pszenicy ozimej uprawiane aktualnie na całym świecie. Realizacja takiego celu wymagałaby bardzo szeroko zakrojonych badań. Zważając na możliwości i potrzebę, cel badań odnosi się zwykle do mniejszych populacji przedmiotowych np. poznanie reakcji pszenicy ozimej odmiany Pilgrim na nawożenie azotowe w warunkach gleb bardzo słabych – wówczas populacją przedmiotową będą wszystkie rośliny tej odmiany pszenicy uprawiane na glebach bardzo słabych. Możemy oczywiście jeszcze zawężać tą populacje do określonych warunków klimatycznych itd. W ramach wyjaśnienia pszenicę uprawia się generalnie na glebach dobrych i bardzo dobrych, ale niektóre odmiany dedykowane są do uprawy w stanowiskach gorszych.

Populacja generalna - Ogół wszystkich możliwych wartości opisywanej jednorodnej cechy w populacji przedmiotowej.

Poszczególne elementy populacji przedmiotowej można scharakteryzować na podstawie różnych, dających się określić cech – nazywanych w statystyce „**zmiennymi**”. Przykładową pszenicę można opisać za pomocą cech morfologicznych takich jak: wysokość źdźbła, liczba międzywęźli, cech produkcyjnych takich jak plon ziarna, plon słomy, czy cech technologicznych jak zawartość białka, skrobi itd. Populację generalną stanowią wartości konkretnej, jednej wybranej zmiennej wszystkich elementów wchodzących w skład populacji przedmiotowej. W omawianym przypadku populacją generalną jest np. liczba międzywęźli każdego źdźbła pszenicy ozimej odmiany Pilgrim uprawianej w warunkach gleb słabych – czyli bardzo duży zbiór wartości liczbowych nazywanych w statystyce „**przypadkami**”.

Badania całościowe - badanie obejmujące wszystkie przypadki populacji generalnej.

Badanie całościowe ma bardzo dużą wartość dowodową – wynikiem badań są twierdzenia – twierdzenie to pewnik równy randze aksjomatowi. Takie twierdzenie bardzo trudno poddać pod wątpliwość. Jedyną możliwością podważenia twierdzeń płynących z badań całościowych jest wykazanie nierzetelności pomiarów, niewłaściwej metody ich wykonania, lub błędów obliczeniowych. Niestety populacje generalne to zazwyczaj bardzo wielkie zbiory, których nie sposób jest uzyskać. Dla przykładu nie możliwym jest określenie liczby międzywęźli wszystkich źdźbeł pszenicy ponieważ od początku kłoszenia do zbioru jest zbyt mało czasu aby takie badania wykonać, pomijając problemy techniczne i koszty ich realizacji. Ponadto niektóre badania mają charakter destrukcyjny – jeśli chcemy sprawdzić odsetek zapalających się zapałek produkowanych przez daną fabrykę, to po badaniach całościowych ta fabryka nie będzie miała czego sprzedać. Rozwiązaniem tego problemu są badania reprezentatywne.

Badania reprezentatywne - badanie obejmujące tyko pewną, reprezentatywną część populacji generalnej.

Takie badania pozwalają nam na wyciąganie wniosków o interesujących nas parametrach populacji generalnej. Wnioski mają mniejsza wartość dowodową niż twierdzenia. Podstawą poddania pod wątpliwość wniosków może być zarzut niereprezentatywności **populacji próbnej**. Reprezentatywność populacji próbnej jest bezwzględnym warunkiem wiarygodnej estymacji (oszacowania) **parametrów** populacji generalnej na podstawie **statystyk** próby. W terminologii statystycznej nie posługujemy się terminem „szacowanie” tylko „estymacja”.

Parametr vs statystyka

Parametr to prawdziwa (rzeczywista) wartość miary opisującej cechę populacji generalnej.

Taką miarą jest np. średnia i odchylenie standardowe wyliczone z wartości wszystkich przypadków całej populacji generalnej. Parametr średni w statystyce nazywany jest również wartością oczekiwaną i oznaczany symbolem μ (mi), (w niektórych opracowaniach spotyka się oznaczenie „m”). Odchylenie standardowe populacji generalnej, oznaczane symbolem σ (sigma). Jeśli z populacji generalnej wyłonimy pewną część przypadków (próbę) i z nich wyliczymy średnią to ta średnia jest nazywana **statystyką**, podobnie jak odchylenie standardowe i inne miary opisowe wyliczone z próby nazywać będziemy również

statystykami. Średnią wyliczona z próby oznaczmy symbolem: $\bar{x}$, a odchylenie standardowe próby oznaczamy symbolem: *S*

Populacja próbna - Określony zbiór przypadków (pomiarów), stanowiący część populacji generalnej (lub populacji przedmiotowej - jeśli badamy jednocześnie więcej niż jedną cechę populacji przedmiotowej).

Większość badań przeprowadza się na populacjach próbnych – w skrócie – próbach. Wynika to z oczywistych względów praktycznych. Jednak tak naprawdę nie interesują nas statystyki uzyskane z próby tylko estymowane na ich podstawie parametry populacji generalnej – wszak po to właśnie przeprowadzamy badanie reprezentatywne. Aby estymacja była wiarygodna próba musi spełniać warunek reprezentatywności.

Reprezentatywność - Własność populacji próbnej świadcząca o tym, że metoda doboru próby zachowała charakterystykę całej populacji przedmiotowej pod względem wybranych cech (lub populacji generalnej jeśli odnosimy się tylko do jednej cechy populacji przedmiotowej).

Oznacza to, że struktura próby odpowiada strukturze populacji generalnej. Próba, która nie spełnia tego warunku nazywana jest niereprezentatywną lub obciążoną. Jednymi z podstawowych kryteriów formalnych decydujących o reprezentatywności próby jest jej losowość i liczebność. Losowość oznacza, że prawdopodobieństwo wybrania do próby każdego przypadku populacji generalnej jest takie samo i jest różne od zera. Przy dostatecznie dużej próbie, prawdopodobieństwo, że **rozkład empiryczny (próby)** nie różni się od **rozkładu teoretycznego (populacji generalnej)** jest bliski jedności (twierdzenie Gliwienki - Cantellego). Zatem $\bar{x}, S = \mu, \sigma$. Dostępne są różne procedury służące określeniu liczebności populacji próbnej, można również znaleźć gotowe kalkulatory np. http://www.statystyka.az.pl/dobor/kalkulator-wielkosci-proby.php. Jedną z takich metod jest dwustopniowa metoda Steina:

Stopień pierwszy - losujemy najpierw niewielką (rzędu kilku, kilkunastu przypadków) próbę wstępną n_0 i wyznaczamy z niej wariancję S^2

Stopień drugi - określamy liczebność właściwej próby n, korzystając ze wzoru:

$\frac{t^2_{\alpha,n-1}S^2}{d^2}$ gdzie: *t* – wartość tablicowa, *d* – dopuszczalny błąd szacunku

W praktyce badawczej nauk przyrodniczych posługujemy się jednak próbami klasyfikowanymi pod względem liczebności jako: **próba mała** do 30 przypadków, **próba średnia** 31-120 przypadków i **próba duża** powyżej120 przypadków.

2.Estymacja przedziałowa

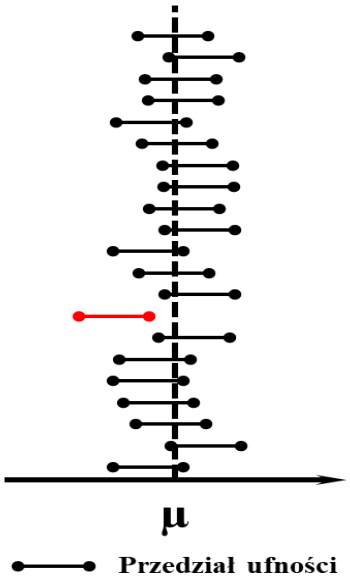
Załóżmy że w pewnej rodzinie rolników dwaj bracia i ojciec są bardzo ciekawi jaki uzyskają plon ziarna uprawianej przez nich kukurydzy. Ponieważ areał tej uprawy wynosi 100ha to rzetelność szacowania ma bardzo duże znaczenie przy podejmowaniu decyzji związanych z ekonomicznym aspektem zagospodarowania plonu.

Aby oszacować taki plon przed jego zebraniem za pomocą kombajnu trzeba dysponować następującymi danymi: obsadą roślin kukurydzy na jednostce powierzchni, liczbą kolb kukurydzy na jednej roślinie i masą ziarna w jednej kolbie. Oczywiście trzeba też przyjąć założenie, że wartości oznaczonych cech nie ulegną zmianie od momentu wykonania pomiarów do momentu zbioru. Populacją przedmiotową są rośliny kukurydzy uprawianej w tym gospodarstwie. Mamy w tym przykładzie trzy cechy do oznaczenia, a więc będziemy mieli trzy populacje generalne. Dla wyjaśnienia tematu estymacji przedziałowej skupimy się tylko na jednej z nich, a mianowicie na masie ziarna w kolbie – czyli populacją generalną są masy ziarna we wszystkich kolbach kukurydzy osobno na całym polu. Zważywszy, że na jednym hektarze w przybliżeniu znajduje się ok 80 000 roślin kukurydzy to populacja generalna jest bardzo liczna i należy przeprowadzić badanie reprezentatywne.

Zatem jeden z braci poszedł na pole pobrał 30 kolb kukurydzy, zważył ziarno z każdej z nich osobno i wyliczył średnią 208,7g. W praktyce badań statystycznych jest tak, że pobiera się tylko jedną populację próbną i na jej podstawie wykonuje się estymację parametru. Ale w naszym przykładzie drugi brat i ojciec dokonali bez konsultacji ze sobą własnych badań, wg. analogicznej metody. W trakcie rozmowy zainteresowanych plonem kukurydzy okazało się że drugi z braci uzyskał wynik 199,4g a ojciec 202,2g… Rozbieżność tych wyników wydaje się niewielka, ale przeliczając to już plon z całego areału to okazuje się, że te różnice przekładają się na dziesiątki ton z całego areału… Widomym jest, że próby o stosunkowo niewielkiej liczebności pobrane z tej samej populacji generalnej najprawdopodobniej będą się różniły pod względem wartości wyliczonych z nich statystyk. Oczywiście zwiększanie liczebności prób zmniejsza prawdopodobieństwo dużych rozbieżności, ale trzeba gdzieś postawić granicę, ponieważ zwiększanie liczebności prób wiąże się z wzrostem

pracochłonności badań. Tu właśnie potrzebna jest statystyka i pojęty temat przedziałowej estymacji parametrów populacji generalnej.

Na podstawie statystyk z próby nie da się oszacować parametru jako jednej liczby. Odnosząc się do przytoczonego powyżej przykładu nie da się na podstawie średniej z próby równej Xg oszacować, że średnia masa ziarna dla wszystkich kolb kukurydzy z całej plantacji wynosi Yg. Zastosowanie ma tu estymacja przedziałowa. Przedziałowa estymacja parametryczna polega na tym, że na podstawie statystyk z próby obliczamy przedział (a<średnia< b), w którym rzeczywisty parametr populacji generalnej powinien się „zmieścić" – w terminologii statystycznej używamy zwrotu przedział ufności pokrywa wartość parametru. Oczywiście istnieje pewne prawdopodobieństwo (α), że rzeczywista średnia i oszacowany przez nas przedział się nie pokryją. Jeśli w badaniach przyjmiemy *α=0,05* to w dużym uproszczeniu na 100 wylosowanych prób i przedziałów z nich estymowanych 5 nie będzie się pokrywało z rzeczywistą średnią populacji generalnej. Przedstawiono to w sposób graficzny na rysunku, przedział zaznaczony kolorem czerwonym nie pokrył się z estymowanym parametrem (rys. 1). Jak zaznaczono jest to tylko bardzo uproszczone wyjaśnienie, przedstawiające ideę zagadnienia, ponieważ w badaniach reprezentatywnych pobieramy tylko jedną próbę i szacujemy jeden przedział ufności. To wprowadzenie w tematykę pozwala na łatwiejsze zrozumienie rozpatrywanych poniżej definicji.



Rys. 1. Losowe przedziały ufności dla wartości oczekiwanej

Przedział ufności – losowy przedział wyznaczony za pomocą rozkładu estymatora, a mający tę własność, że z dużym, z góry ustalonym prawdopodobieństwem, pokrywa wartość szacowanego parametru. Przedział ufności zapisujemy:

$$\mathbf{P\ (a < \Theta < b) = 1 - \alpha} \quad (1.)$$

Θ – estymowany parametr np. średnia (μ) lub odchylenie standardowe (σ)

a i b –dolna i górna granica przedziału ufności

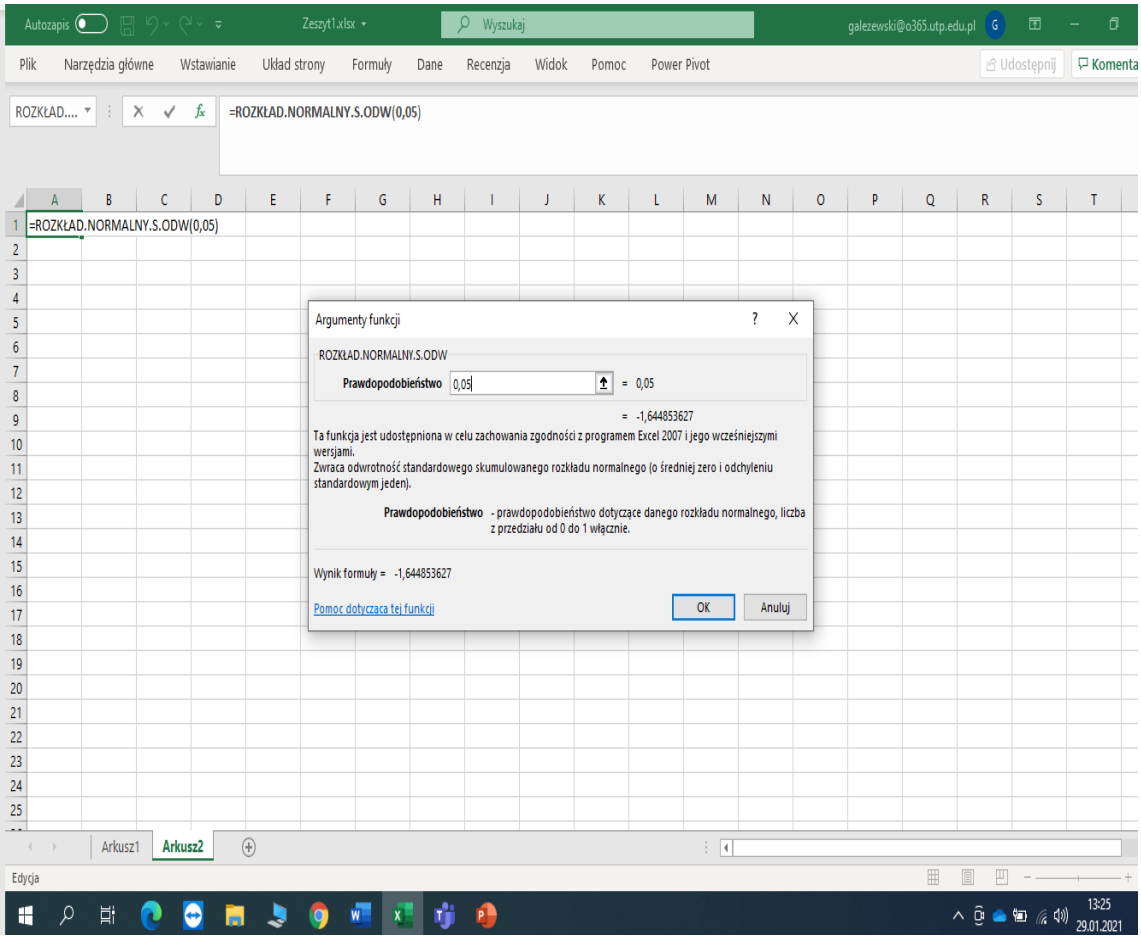
1- α - współczynnik ufności - prawdopodobieństwo, z jakim parametr Θ jest pokryty przedziałem ufności.

Najczęściej przyjmowanym współczynnikiem ufności jest 0,90; 0,95; 0,99 (tab. 1).

Tab. 1. Współczynniki ufności α i wartość $Z_{\alpha/2}$

1 - α	α	α/2	$z_{\alpha/2}$
0,90 (90%)	0,1	0,05	1,64
0,95 (95%)	0,05	0,025	1,96
0,99 (99%)	0,01	0,005	2,58

Wartości *Zα* odczytujemy z tablic dystrybuanty rozkładu normalnego N(0,1). Tablice takie dostępne są w większości podręczników ze statystyki, można znaleźć je również na stronach internetowych. Najprościej jednak skorzystać z formuły =ROZKŁAD.NORMALNY>S>ODW(…) w programie Excel. W miejsce trzech kropek w nawiasie przedstawionej formuły wstawiamy wartość szukanego α lub *1-α*. Jeśli w programie Excel tą formułę przywołamy z paska formuł klikając symbol f_x to wartość α (lub *1-α*) wpisujemy w przywołanym oknie (rys. 2).



Rys. 2. Okno programu Excel z formułą wyszukania wartości *Zα*

2.1.Przedział ufności dla średniej μ

Budowa przedziałów ufności dla średniej (wartości oczekiwanej) jest uzależniona od 3 założeń: 1. typ rozkładu 2. znajomość wariancji 3. wielkość próby.

MODEL I

Założenia:

1. Populacja generalna ma rozkład normalny $N(\mu,\sigma)$
2. Odchylenie standardowe σ (wariancja σ^2) jest znane
3. Z populacji wylosowano n elementową próbę

Zatem przyjęte założenia wskazują, że ten model ma zastosowanie tyko dla zmiennych cechujących się rozkładem normalnym. Wymagana jest w tym modelu znajomość odchylenia standardowego (lub wariancji) dla populacji generalnej. Liczebność populacji próbnej jest dowolna. Ten model jest stosowany rzadko ze względu na to, że tylko nieliczne populacje generalne są sparametryzowane, tzn. że rzadko kiedy znamy odchylenie standardowe populacji generalnej. Odchylenie standardowe σ może być jednak z góry założone np. jeśli producent konserw określa masę netto ich zawartości to przepisy prawa określają również dopuszczalne odchylenie (art. 8 ust. 1 ustawy o towarach paczkowanych). Wprawdzie ustawowo określa się tylko dopuszczalną niedowagę, ale zakładając, że rozkład normalny jest dwustronny to odchylenie od deklarowanej masy można dla potrzeb testowania zastosować dla nadwagi. Trzeba zaznaczyć, że dopuszczalne odchylenie określa tylko wartość minimalną, a więc wartość równą -3 σ, stąd aby otrzymać wartość parametru odchylenia standardowego trzeba dopuszczalne odchylenie podzielić przez trzy.

Przedział ufności dla tego modelu budujemy wg wzoru:

$$P\left\{\bar{x} - Z_\alpha \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_\alpha \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha \quad (2.)$$

Estymatorem parametru μ jest tutaj średnia z próby $\bar{x}$. Estymator ten ma rozkład

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad (3.)$$

Gdzie:

$\bar{x}$ – średnia z próby

Z_α– b wartość zmiennej losowej Z mającej rozkład normalny standaryzowany. Jest to wartość odczytywana z tablic dystrybuanty rozkładu normalnego N(0,1).

σ – odchylenie standardowe populacji generalnej

n – liczebność próby

Przedział ten buduje się najczęściej dla $\alpha = 0{,}1$ wówczas Zα=1,64 lub $\alpha = 0{,}05$ wówczas Zα=1,96 lub dla $\alpha = 0{,}01$ wówczas Zα=2,58

Z oczywistych względów zależy nam aby zbudowany przedział ufności był jak najwęższy. Dla przykładu, jeśli chcemy estymować średnią wysokość Polaków w wieku 18lat i zbudujemy przedział o rozpiętości dwóch metrów to tak naprawdę niczego się nie dowiemy, im przedział jest węższy tym bardziej wartościową informację otrzymujemy. Z wzoru (2.) wynika, że o szerokości przedziału decyduje średnia z próby $\bar{x}$, zmienna losowa Z, odchylenie standardowe σ i liczebność populacji próbnej n. σ jest wartością na którą nie mamy wpływu bo jest parametrem rzeczywistym populacji generalnej. $\bar{x}$ jest wartością otrzymaną z próby wyłonionej przez losowanie więc również nie mamy wpływu na jej wartość. Zα wynika z założeń określonych w danej dyscyplinie naukowej. Zatem tak naprawdę wpływ mamy na liczebność populacji próbnej n. Ze wzoru (2.) wynika że im większa liczebność próby tym przedział jest węższy, zatem tworząc metodykę badań należy unikać minimalizacji liczebności próby.

Przykład 1.

Masa netto chipsów w paczce wg. deklaracji producenta wynosi 100g, dopuszczalne przepisami prawa odchylenie to 4,5g, zatem $\sigma = 1{,}5$g. Wiadomo, że rozkład tej cechy jest normalny. Pobrano próbę 30 pacek i po przeprowadzeni badania obliczono średnią masę chipsów w paczce $\bar{x} = 96{,}0g$. Przyjmując współczynnik ufności 1-α = 0,95 zbudować przedział ufności dla nieznanej wartości oczekiwanej masy chipsów w paczce.

$$P\left\{96{,}0 - 1{,}96\frac{1{,}5}{\sqrt{30}} < \mu < 96{,}0 + 1{,}96\frac{1{,}5}{\sqrt{30}}\right\} = 1 - 0{,}05$$

Uzyskujemy wynik

$$P\{95{,}5 < \mu < 96{,}5\} = 0{,}95$$

Z obliczeń wynika, że rzeczywista średnia masa chipsów w paczkach od tego producenta z 95% prawdopodobieństwem mieści się w przedziale 95,5-96,5g.

Przykład 2.

W pewnym bardzo dużym zakładzie produkcyjnym postanowiono zbadać staż pracowników fizycznych. W tym celu z populacji tych pracowników wylosowano próbę o liczebności n=196 pracowników, z której obliczono, że średni staż = 6,9 lat. Wiadomo, że rozkład stażu jest normalny z odchyleniem standardowym σ = 2,8 lat. Przyjmując współczynnik ufności 1-α = 0,95, zbudować przedział ufności dla nieznanej wartości oczekiwanej stażu.

$$P\left\{6{,}9 - 1{,}96\frac{2{,}8}{\sqrt{196}} < \mu < 6{,}9 + 1{,}96\frac{2{,}8}{\sqrt{196}}\right\} = 1 - 0{,}05$$

Uzyskujemy wynik

$$P\{6{,}51 < \mu < 7{,}29\} = 0{,}95$$

Wniosek: średni staż pracowników fizycznych w objętym badaniami zakładzie pracy miesi się w przedziale 6,51-7,29 lat, przy $\alpha = 0{,}05$.

MODEL II

Założenia:

1. Populacja generalna ma rozkład normalny $N(\mu, \sigma)$
2. Odchylenie standardowe σ (wariancja σ^2) jest nieznane
3. Z populacji wylosowano n elementową próbę małą (do 30 elementów)

Ten model jest najczęściej stosowany w naukach przyrodniczych, ponieważ większość cech przyrodniczych ma rozkład normalny o nieznanych parametrach. Zazwyczaj również pobierane są próby małe, czyli o liczebności nie przekraczającej 30 przypadków. Ograniczenie liczebności próby wiąże się z tym, że badania naukowe, zwłaszcza te czynnikowe (z specjalnie wprowadzonym i kontrolowanym czynnikiem) obejmują wiele obiektów doświadczalnych i jednocześnie określa się wiele cech. Powoduje to, że pobieranie prób o większej liczebności jest zbyt pracochłonne i kosztowne. Trzeba również wziąć pod uwagę to, że wiele badań ma destrukcyjny lub inwazyjny charakter dla przypadków objętych testami; nabiera to szczególnego znaczenia przy pracy na organizmach żywych.

Przedział ufności dla tego modelu budujemy wg wzoru:

$$P\left\{\bar{x} - t_{\alpha,\nu}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha,\nu}\frac{s}{\sqrt{n}}\right\} = 1 - \alpha \quad (4.)$$

Gdzie:

$\bar{x}$ – średnia z próby

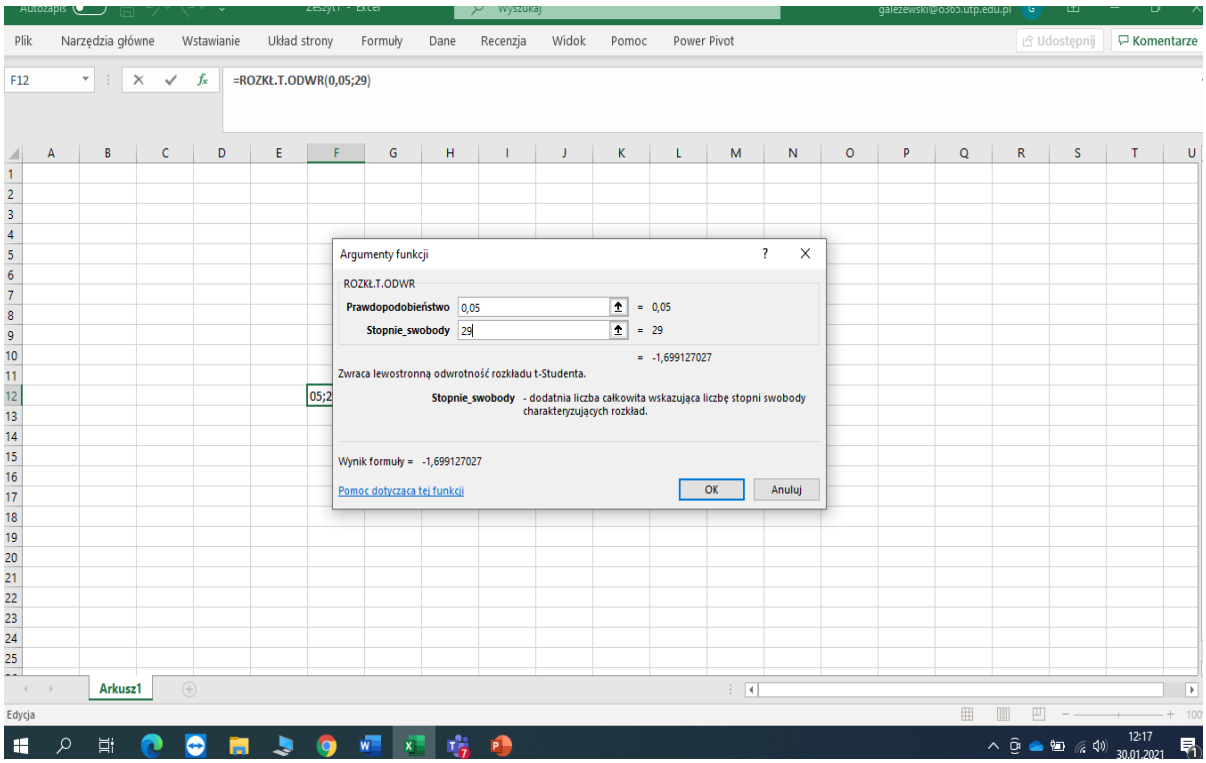
$t_{\alpha,\nu}$ – wartość t-Studenta odczytywana z tablic dla ustalonego α i liczby stopni swobody

$$\nu = n - 1$$

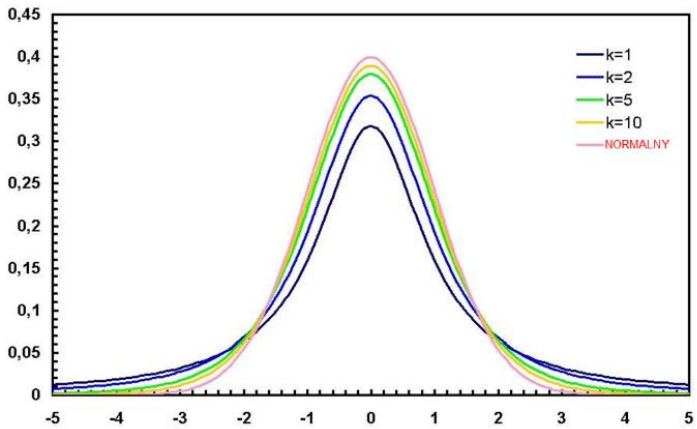
S – odchylenie standardowe populacji próbnej

n – liczebność próby

Przedział ten buduje się najczęściej dla $\alpha = 0{,}1\ lub\ 0{,}05\ lub\ 0{,}01$. Wartości t odczytujemy z tablic rozkładu t-studenta. Są to powszechnie dostępne tablice statystyczne. Można również uzyskać wartość t z programu Excel za pomocą formuły =ROZKŁAD.T.ODW(…; …), w miejsce kropek wpisując odpowiednio poziom ufności α i liczbę stopni swobody n-1 (rys.3)



Rys. 3. Okno programu Excel z formułą wyszukania wartości t- studenta



Rys. 4. Porównanie rozkładu normalnego z rozkładem t-studenta przy rożnej liczbie stopni swobody

Trzeba mieć świadomość, że dla takiego samego poziomu ufności α wartość t- studenta jest większa niż wartość Z. Stąd estymowany za pomocą modelu II przedział jest szerszy niż za pomocą modelu I. Im mniejsza próba tym rozkład t bardziej odbiega od rozkładu normalnego tj. ma szerszą podstawę - tzn. bardziej prawdopodobne są wartości mocno odbiegające od średniej niż w przypadku rozkładu normalnego, w miarę wzrostu liczebności rozkład t - Studenta jest zbieżny do rozkładu normalnego standaryzowanego (rys. 4).

Przykład 3.

Wyznaczyć przedział ufności (1-α =0,95 i 1-α =0,90) dla wartości oczekiwanej glukozy we krwi ludzkiej [mg%] na podstawie 6 przebadanych osób: 115; 110; 120; 118; 103; 125.

Wyleczona z próby średnia = 115,17 mg%, a odchylenie standardowe 7,78 mg%

$$P\left\{115{,}17 - 2{,}015\frac{7{,}78}{\sqrt{6}} < \mu < 115{,}17 + 2{,}015\frac{7{,}78}{\sqrt{6}}\right\} = 0{,}95$$

$$P\{108{,}77 < \mu < 121{,}57\} = 0{,}95$$

$$P\left\{115{,}17 - 1{,}476\frac{7{,}78}{\sqrt{6}} < \mu < 115{,}17 + 1{,}476\frac{7{,}78}{\sqrt{6}}\right\} = 0{,}90$$

$$P\{110{,}48 < \mu < 119{,}86\} = 0{,}90$$

Dla poziomu ufności 0,95 wartość oczekiwania glukozy we krwi mieści się w przedziale 107,77 – 121,57 mg%, a dla poziomu ufności 0,90 przedział ten jest węższy tj. 110,48-119,86. Wydawać się może, że skoro uzyskujemy węższy przedział przy mniejszym poziomie ufności to zasadnym jest jego zmniejszanie. Niestety wiąże się to z większym prawdopodobieństwem tego, że oszacowany przez nas przedział nie pokryje szacownego parametru.

Przykład 4.

Sprawdzano zawartość (%) białka w nasionach łubinu żółtego odmiany Teo. W tym celu z piętnastu losowo wybranych plantacji pobrano po jednej próbie do oznaczenia tej cechy. Po wykonaniu oznaczeń uzyskano następujące wyniki: 45,1; 46,1; 43,6; 44,4; 45,8; 44,4; 44,4; 43,4; 44,2; 44,8; 44,9; 48,7; 44,6; 46,3; 44,5. Wyznacz przedział ufności (1-α

=0,99) dla średniej populacji generalnej. Wyliczona z próby średnia = 45,0%, a odchylenie standardowe = 1,31%

$$P\left\{45{,}0 - 2{,}977\frac{1{,}31}{\sqrt{15}} < \mu < 45{,}0 + 2{,}977\frac{1{,}31}{\sqrt{15}}\right\} = 0{,}99$$

$$P\{43{,}99 < \mu < 46{,}01\} = 0{,}99$$

Zawartość białka w nasionach łubinu żółtego odmiany Teo mieści się w przedziale 43,99-46,01% przy α=0,01

MODEL III

Założenia:

1. Populacja generalna ma rozkład dowolny (normalny lub inny)
2. Odchylenie standardowe (wariancja) jest nieznane.
3. Z populacji wylosowano n elementową próbę niemałą (powyżej 30 elementów)

W badaniach gdzie parametryzacji podlega niewielka liczba obiektów zasadnym jest wykonywania oznaczeń na próbach średnich lub dużych. Niektóre populacje generalne charakteryzują się innym rozkładem niż rozkłada normalny, wtedy powinno się pobierać próby duże. Dla takich właśnie badań zastosowanie ma model III.

Przedział ufności dla tego modelu budujemy wg wzoru:

$$P\left\{\bar{x} - Z_{\alpha}\frac{s}{\sqrt{n}} < \mu < \bar{x} + Z_{\alpha}\frac{s}{\sqrt{n}}\right\} = 1 - \alpha \quad (5.)$$

Przedział budujemy podobnie, jak w modelu I, ale ponieważ wariancji i odchylenia standardowego nie znamy, musimy je obliczyć jak dla prób niemałych (średnich i dużych):

$$S^2 = \frac{\sum(\bar{x} - x_i)^2}{n} \quad (6.) \text{ lub } S^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n} \quad (7.)$$

Różnica we wzorach na wariancję dla prób małych i niemałych polega na tym, że w mianowniku dla prób małych mamy (n-1), a dla prób niemałych (n). Posługując się arkuszem kalkulacyjnym Excel wariancję dla prób małych obliczamy formułą =WARIANCJA, a dla prób niemałych =WARIANCJA.POP (w zależności od wersji Excel

może też być WARIANCJA.POPUL). Czasami dla prób dużych dane zbierane są w postaci szeregów rozdzielczych, wówczas wariancję wylicza się ze wzoru:

$$S^2 = \frac{\Sigma(\bar{x}-x_i')^2 \times n_i}{n} \quad (8.)$$

Wzory 6-7 to wzory na wariancję a we wzorze 5 potrzebujemy odchylenie standardowe. Oczywiście wystarczy z wariancji wyciągnąć pierwiastek kwadratowy aby otrzymać potrzebne odchylenie standardowe.

Przykład 5.

Badano średnicę pnia samosiewów sosny zwyczajnej na nieużytku rolnym po czterdziestu latach ugorowania. Ze względu na dużą przewagę drzew w pewnej grupie wiekowej rozkład tej cechy był asymetryczny (inny niż rozkład normalny). Pomierzono więc 120 pni losowo wybranych sosen w różnym wieku. Dla tej próby uzyskano statystyki: średnia = 22,7cm odchylenie standardowe = 4,23cm. Przyjmując współczynnik ufności 1-α = 0,95, zbudować przedział ufności dla nieznanej wartości oczekiwanej średnicy pni samosiewów sosen na badanym nieużytku rolnym.

$$\left\{22{,}7 - 1{,}96\frac{4{,}23}{\sqrt{120}} < \mu < 22{,}7 + 1{,}96\frac{4{,}23}{\sqrt{120}}\right\} = 0{,}95$$

$$P\{21{,}94 < \mu < 23{,}46\} = 0{,}95$$

Z prawdopodobieństwem P=0,95 średnica pni samosiewów sosen na badanym użytku rolnym mieści się w przedziale 21,94-23,46cm.

Przykład 6.

Przeprowadzono badania na populacji próbnej myszy o liczebności n=42, którym podawano pewien farmaceutyk. Oszacować średnią (wartość oczekiwaną) długość rozpadu komórek macierzystych u myszy dla P = 0,95 na podstawie uzyskanych wyników:

Długość rozpadu (dni)	Liczba przypadków
1 – 3	9
3 – 5	13
5 – 7	12
7 – 9	5
9 – 11	3

Dla rozwiązania tego przykładu będą miały zastosowanie wzory dla szeregów rozdzielczych.

x (dni)	n_i	x_i'	$x_i' n_i$	$(x_i' - \bar{x})^2$	$(x_i' - \bar{x})^2 n_i$
1 – 3	9	2	18	9,29	83,61
3 – 5	13	4	52	1,10	14,3
5 – 7	12	6	72	0,91	10,92
7 – 9	5	8	40	8,72	43,6
9 – 11	3	10	30	24,53	73,59
suma	42	-	212		226,02

$$\bar{x} = \frac{212}{42} = 5{,}047 dni$$

$$S^2 = \frac{226{,}02}{42} = 5{,}38$$

$$S = \sqrt{5{,}38} = 2{,}32\ dni$$

$$\left\{5{,}047 - 1{,}96\frac{2{,}32}{\sqrt{42}} < \mu < 5{,}047 + 1{,}96\frac{2{,}32}{\sqrt{42}}\right\} = 0{,}95$$

$$P\{4{,}34 < \mu < 5{,}75\} = 0{,}95$$

Średnia długość rozpadu komórek u badanej populacji myszy traktowanych testowanym farmaceutykiem mieści się w przedziale 4,34-5,75 dni (przy α=0,05).

2.2.Przedział ufności dla wariancji σ^2

Bardzo ważnym parametrem populacji generalnej jest również odchylenie standardowe czyli podstawowa miara rozproszenia przypadków wokół średniej. Również ten parametr możemy szacować na podstawie populacji próbnej. Dobór metody estymacji zależy od typu rozkładu i wielkości próby. Przedstawione zostaną dwa modele model I dla prób małych i model II dla prób średnich i dużych. W obu przypadkach przyjmujemy założenie normalności rozkładu populacji. Należy jednak zwrócić uwagę, że w modelu pierwszym nie szacujemy bezpośrednio odchylenia standardowego tylko wariancję, z której możemy

wyliczyć odchylenie standardowe. W modelu drugim estymujemy bezpośrednio odchylenie standardowe.

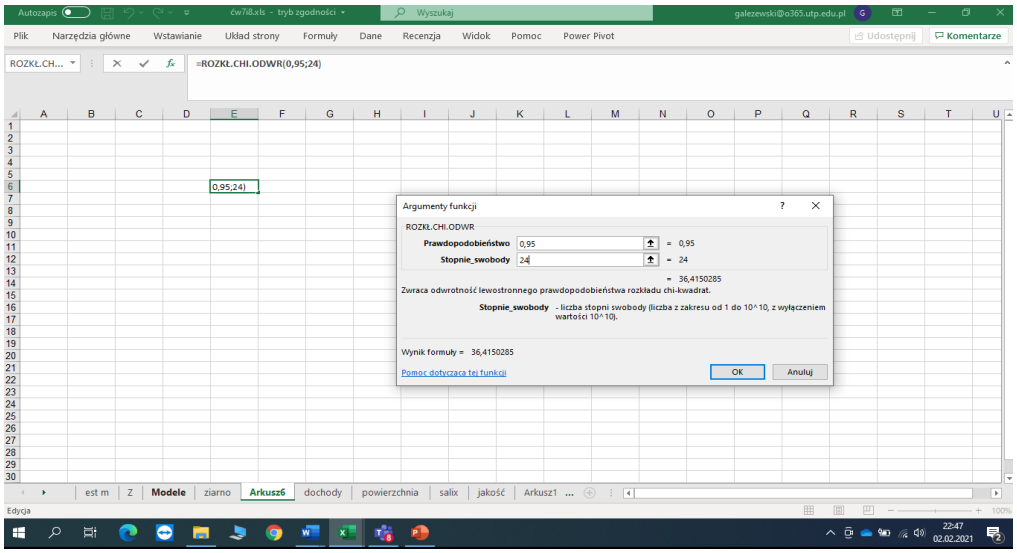
MODEL I

1. Populacja generalna ma rozkład normalny N(m, σ^2)
2. Nieznane są parametry m i σ^2
3. Z populacji wylosowano próbę małą (do 30 elementów)

Przedział ufności dla tego modelu budujemy wg wzoru:

$$P\left\{\frac{nS^2}{\chi^2_{\frac{\alpha}{2},n-1}} < \sigma^2 < \frac{nS^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right\} = 1-\alpha \quad (9.)$$

Zasada konstruowania przedziału ufności nie odbiega od wcześniej przedstawionych modeli estymacji wartości oczekiwanej. W tym modelu pojawił się jednak nie omawiana jeszcze wartość a mianowicie χ^2 (chi kwadrat) - odczytujemy z tablic rozkładu χ^2dla n-1 stopni swobody i $\frac{\alpha}{2} = 0{,}05$ oraz $1-\frac{\alpha}{2} = 0{,}95$ dla współczynnika ufności 0,90. Możemy oczywiście przyjąć inne poziomy współczynnika ufności. Dla większości wydrukowanych tablic statystycznych kłopotliwe bywa odczytanie wartości chi kwadrat dla innych poziomów niż 0,90 i trzeba posiłkować się interpolacją między podanymi w tablicy wartościami. Zazwyczaj stosuje się interpolacje prostą, ale w przypadku akurat rozkładu chi kwadrat nie jest ona właściwa. Można jednak uzyskać dowolną szukaną wartość tablicową rozkładu chi kwadrat korzystając z formuły Excel: =ROZKŁ.CHI.ODWR(…;…) w miejsce kropek wpisując odpowiednio szukane prawdopodobieństwo (poziom ufności) i liczbę stopni swobody (rys. 5).



Rys. 5. Okno programu Excel z formułą wyszukania wartości chi kwadrat

Przykład 7.

W pewnym obwodzie łowieckim w sezonie 2019/2020 pozyskano w drodze polowania 25 łań danieli (*dama dama*), i określono masę tuszy każdej sztuki. Wariancja masy tuszy wyliczona z tych 25 przypadków wynosiła 2,8kg. Przyjęto, że masa tuszy danieli cechuje się rozkładem normalnym. Na podstawie tych danych zbudować przedział ufności (1-α=0,90) dla odchylenia standardowego masy tusz populacji danieli w tym łowisku

$$P\left\{\frac{25 \times 2{,}8}{36{,}415} < \sigma^2 < \frac{25 \times 2{,}8}{13{,}848}\right\} = 0{,}90$$

$$P\{1{,}92 < \sigma^2 < 5{,}05\} = 0{,}90$$

$$P\{1{,}39 < \sigma < 2{,}25\} = 0{,}90$$

Odchylenie standardowe masy tuszy populacji danieli w badanym łowisku mieści się w przedziale 1,39-2,25kg dla P=0,90

Przykład 8.

Ocenić zróżnicowanie średnicy drzew sosnowych w całym lesie, jeśli w 30 elementowej próbie drzew z tego lasu otrzymano $\bar{x}$=37,3 cm i s^2 = 13,5. Przyjmujemy, że badana cecha posiada rozkład normalny.

$$\left\{\frac{30 \times 13{,}5}{45{,}722} < \sigma^2 < \frac{30 \times 13{,}5}{16{,}047}\right\} = 0{,}95$$

$$P\{8{,}858 < \sigma^2 < 25{,}238\} = 0{,}95$$

$$P\{2{,}976 < \sigma < 5{,}024\} = 0{,}95$$

MODEL II

1. Populacja generalna ma rozkład normalny N(m, σ^2) lub zbliżony do normalnego
2. Nieznane są parametry m i σ^2
3. Z populacji wylosowano próbę średnią lub dużą (powyżej 30 elementów),

Przedział ufności dla tego modelu budujemy wg wzoru:

$$P\left\{\frac{s}{1+\frac{Z_\alpha}{\sqrt{2n}}} < \sigma < \frac{s}{1-\frac{Z_\alpha}{\sqrt{2n}}}\right\} = 1-\alpha \quad (10.)$$

Ten przedział jak już wspomniano wyżej wyznaczamy bezpośrednio dla estymowanej wartości parametru odchylenia standardowego. Estymację przeprowadzamy w oparciu o wyliczone z próby odchylenie standardowe i wartość tablicową Z_α.

Przykład 9.

Prognozowano plon rzepaku ozimego na pewnej plantacji. Problem polega na tym, że pole cechuje się dużą zmiennością glebową i na każdym fragmencie pola plon jest inny. Ze względu na dużą wartość potencjalnego zbioru zachodzi konieczność rzetelnej oceny - szacowano więc nie tylko wartość oczekiwaną, ale również odchylenie standardowe plonowania na podstawie 60 prób. Pojedynczą próbę stanowił plon roślin z powierzchni $1m^2$. Odchylenie standardowe dla tej próby wyniosło $21{,}7g\cdot m^{-2}$. Dokonaj estymacji przedziałowej odchylenia standardowego plonowania rzepaku dla tej plantacji przyjmując P=0,99.

$$P\left\{\frac{21{,}7}{1+\frac{2{,}58}{\sqrt{60}}} < \sigma < \frac{21{,}7}{1-\frac{2{,}58}{\sqrt{60}}}\right\} = 1-0{,}01$$

$$P\{16{,}28 < \sigma < 32{,}54\} = 0{,}99$$

Zmienność plonowania rzepaku ozimego na przedmiotowej plantacji mierzona odchyleniem standardowym mieści się w przedziale $16{,}28\text{-}32{,}54g\cdot m^{-2}$

Przykład 10.

W badaniach nad wydatkami rodzinnymi, ponoszonymi na żywność w naszym kraju, posłużono się próbą wielką 632 gospodarstw domowych, na podstawie której ustalono, że średnia wydatków wynosi 1570 zł, a odchylenie standardowe tych wydatków 224 zł. Wyznaczyć przedział ufności (dla współczynnika 0,90) dla odchylenia standardowego wydatków na żywność.

$$P\left\{\frac{224}{1+\frac{1{,}64}{\sqrt{632}}} < \sigma < \frac{224}{1-\frac{1{,}64}{\sqrt{632}}}\right\} = 1-0{,}10$$

$$P\{210{,}3 < \sigma < 239{,}6\} = 0{,}90$$

Odchylenie standardowe wydatków ponoszonych przez rodzinę na żywność w naszym kraju mieści się w przedziale 210,3-239,6zł (przy P=0,90)

2.3.Przedział ufności dla wskaźnika struktury (%)

Jeśli w badaniach statystycznych opracowujemy cechy jakościowe, niemierzalne, czasem zachodzi potrzeba szacowania frakcji elementów posiadających daną cechę. Wskaźnikiem struktury (FRAKCJĄ) nazywamy wielkość W równą:

$W = \frac{k}{n}$ (11.), gdzie k – liczba jednostek statystycznych posiadających daną cechę,

n –liczba wszystkich jednostek statystycznych.

Przedział ufności dla wskaźnika struktury **p**, czyli dla frakcji elementów wyróżnionych w populacji generalnej na podstawie próby dużej (n>100, w której liczba k-posiada daną cechę).

$$P\left\{\frac{k}{n} - Z_\alpha\sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}} < p < \frac{k}{n} + Z_\alpha\sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}\right\} = 1 - \alpha \text{ (12.)}$$

Przykład 11.

W celu wyznaczenia przedziału ufności dla frakcji nasion fasoli zdolnych do kiełkowania wykonano próbę wysiania 800 nasion. Wykiełkowało z nich 728 fasoli. Dokonaj oszacowania dla współczynnika ufności = 0,95.

$$P\left\{\frac{728}{800} - Z_\alpha\sqrt{\frac{\frac{728}{800}\left(1-\frac{728}{800}\right)}{800}} < p < \frac{728}{800} + Z_\alpha\sqrt{\frac{\frac{728}{800}\left(1-\frac{728}{800}\right)}{800}}\right\} = 1 - \alpha$$

$$P\{0{,}89 < \sigma < 0{,}93\} = 0{,}95$$

Zdolność kiełkowania nasion fasoli miesi się w przedziale 89-93%.

Przykłady do opracowania własnego.

1. Aby ocenić ilość siana na obszarze łąk pokrywających 20 ha pobrano losowo 150 prób o powierzchni 1 m^2 każda. Plony siana zważono i otrzymano średnią = 218,2 g, z odchyleniem standardowym 38,0 g. W jakich granicach plon siana z łąk się waha, jeśli przyjmiemy współczynnik ufności 0,95?

2. Odłowiono 10 tysięcy motyli, w tym 5433 samic. Oblicz frakcję samic w próbie oraz 95% przedział ufności dla tej frakcji.

3. Wiadomo, że odchylenie standardowe populacji generalnej długości czaszki kozic alpejskich wynosi 15 mm. W zbiorach pewnego myśliwego znajduje się 40 czaszek kozic o średniej długości 202,5 mm. Podaj przedział ufności dla średniej długości czaszki w populacji generalnej kozic alpejskich przyjmując współczynnik ufności 0,90 i 0,95.

4. Dokonaj estymacji przedziałowej średniej zawartości białka w nasionach grochu odmiany Ramrod, na podstawie próby n=18, jeśli średnia dla próby wynosi 21,2 % a odchylenie standardowe 0,5 %.

5. W celu oceny zróżnicowania masy jaj pochodzących od pewnej rasy kur zważono 15 jaj otrzymując następujące wartości: 62, 70, 57, 58, 59, 67,65, 69, 55, 57, 60, 54, 72, 66, 74. Przy współczynniku 0,96 zbuduj przedział ufności dla wariancji masy jaj.

6. W celu ustalenia przeciętnej zawartości witaminy C w owocach dzikiej róży pobrano 45 próbek 100 gramowych, i ustalono, że zawartość witaminy C (w mg na 100 g miąższu) mieści się w granicach:

x_i	430-455	455-480	480-505	505-530
n_i	10	12	13	10

Przyjmując współczynnik ufności 0,95 zbuduj przedział ufności dla odchylenia standardowego zawartości witaminy C oraz dla średniej zawartości witaminy C.

3.TESTOWANIE STATYSTYCZNE

Omówiona w poprzednim rozdziale estymacja przedziałowa odnosi się do szacowania parametrów populacji generalnej. Wnioskowanie statystyczne polega na weryfikacji (testowaniu) postawionych hipotez.

Podstawowe pojęcia:

Hipoteza – przypuszczenie, teza, osąd

Hipoteza statystyczna – dowolne przypuszczenie o jakiejś właściwości populacji generalnej wydane bez jej badania całkowitego. Oczywiście to przypuszczenie odnosi się miar statystycznych opisujących tą populację.

Hipoteza parametryczna – hipoteza statystyczna precyzująca wartość parametru w rozkładzie populacji generalnej znanego typu. Zatem jest to hipoteza odnosząca się do konkretnego parametru populacji generalnej.

Hipoteza zerowa – podstawowa hipoteza statystyczna sprawdzana danym testem. Oznacza się ją symbolem $\mathbf{H_0}$. W hipotezie zerowej zakłada się brak różnic pomiędzy estymatorami i parametrami lub pomiędzy rozkładami empirycznymi (z prób) i rozkładami teoretycznymi (rozkładem populacji generalnej). To znaczy, że hipoteza ta zakłada barak dającej się udowodnić różnicy między np. średnią z próby $\bar{x}$, a średnią populacji generalnej μ. Można to zapisać matematycznie $\bar{x} = \mu$, czasami również można spotkać się z zapisem: $\bar{x} \approx \mu$. Ten drugi zapis ma uzasadnienie, w tym że nie tak właściwie średniej z próby nie porównujemy do parametru średniego jako jednej liczby a estymowanego przedziału tego parametru. W większości podręczników H_0 matematycznie zapisywana jest inaczej: $m = m_0$ gdzie: m_0 - wartości oczekiwana populacji generalnej μ, m - średnia z próby $\bar{x}$. Dla zachowania zgodności innymi opracowaniami dydaktycznymi będziemy posługiwać się tym właśnie zapisem.

Hipoteza alternatywna (robocza, merytoryczna) – hipoteza statystyczna przeciwna do zerowej hipotezy. Oznacza się ją jako $\mathbf{H_1}$. Jeśli na podstawie testu uzyskuje się podstawę do odrzucenia hipotezy zerowej, to oznacza, że jednocześnie przyjmuje się jako słuszną hipotezę alternatywną. W hipotezie alternatywnej zakłada się istnienie istotnych (udowodnionych) różnic pomiędzy estymatorami i parametrami lub pomiędzy rozkładami z prób i rozkładami teoretycznymi. Matematyczny zapis jest podobny: $m \neq m_0$

Test statystyczny – narzędzie statystyczne (metoda obliczeniowa), służące do weryfikacji hipotez statystycznych na podstawie wyników z próby.

Wnioskowanie statystyczne ma określone granice niepewności, a więc określone prawdopodobieństwo uzyskania rzetelnego wyniku. Wynika to z określonego prawdopodobieństwa przyjęcia prawdziwej lub odrzucenia fałszywej hipotezy czyli popełnienia błędu. Wyróżniamy dwie kategorie błędów weryfikacji hipotez:

Błąd I rodzaju (α) – możliwy do popełnienia błąd przy weryfikacji hipotezy statystycznej polegający na odrzuceniu prawdziwej hipotezy.

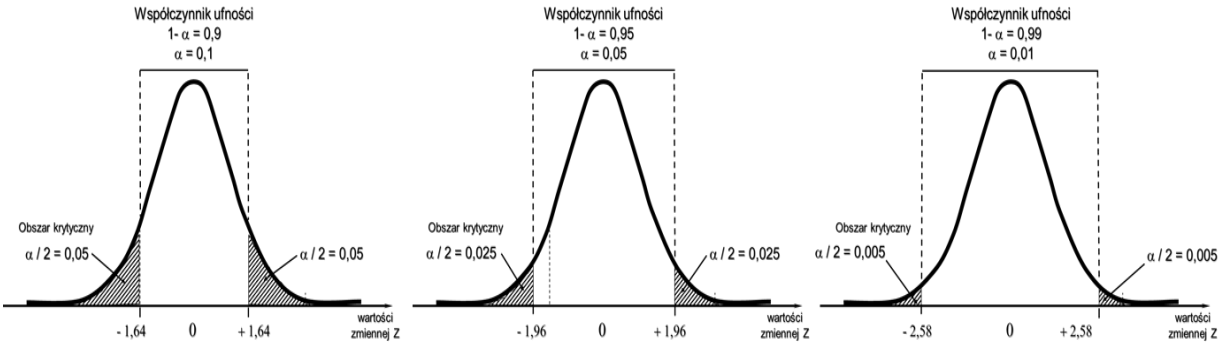
Błąd II rodzaju (β) – możliwy do popełnienia przy weryfikacji hipotez błąd polegający na przyjęciu fałszywej hipotezy.

Test istotności – najczęściej używany w praktyce statystycznej typ testu statystycznego, pozwalający na odrzuceniu hipotezy H_0 z małym ryzykiem popełnienia **błędu I rodzaju** **(** zwykle $\alpha = 0{,}05$**)** – jest to **prawdopodobieństwo** popełnienia pomyłki – odrzucenia hipotezy zerowej gdy rzeczywiście była ona właściwa. Prawdopodobieństwo to nazywamy **poziomem istotności.**

Parametryczny test istotności – test istotności weryfikujący hipotezę H_0 mówiącą o wartości parametru w ustalonym typie rozkładu populacji generalnej. Innymi słowy jest to procedura obliczeniowa sprawdzająca zasadność hipotezy zerowej, właściwa dla przyjętego modelu, odnosząca się do wartości konkretnego parametru populacji generalnej.

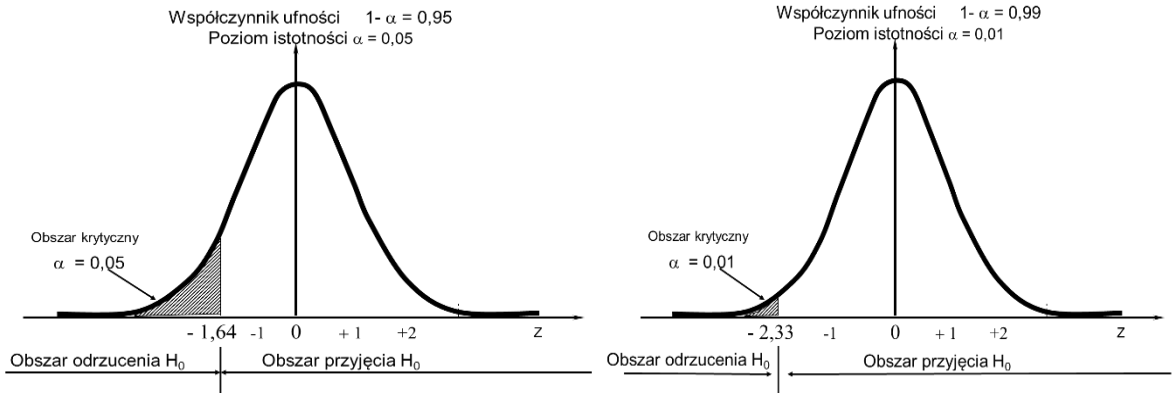
Obszar krytyczny testu – podzbiór przestrzeni próby o tej własności, że jeżeli otrzymamy w próbie punkt przestrzeni należący do podzbioru, to podejmuje się decyzję o odrzuceniu hipotezy zerowej. Podzbiory te na rysunkach 6,7 i 8 oznaczono zakreskowanym polem i nazywane są obszarem krytycznym.

Dwustronny obszar krytyczny testu – obszar krytyczny złożony z dwu rozłącznych podzbiorów przestrzeni próby, w rozkładzie odpowiedniej statystyki. Używa się go wówczas, gdy hipoteza alternatywna H_1 jest w postaci nierówności $\neq$. Czyli obszary krytyczne znajdują się po obu stronach rozkładu (rys. 6). Z tablic rozkładu *Z* odczytuje się wartości stanowiące granice tych obszarów. Na rysunku 6 graficznie przedstawiono wartości *Z* dla różnych poziomów istotności

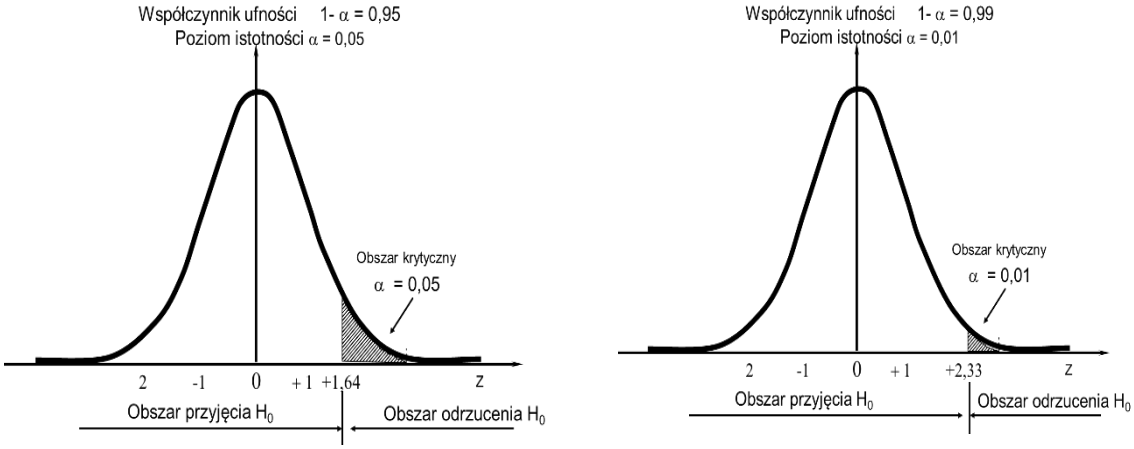


Rys. 6. Wartości zmiennej *Z* na tle rozkładu zmiennej standaryzowanej dla *α=0,1, α=0,05, α=0,01* dla testów dwustronnych - obszar krytyczny dwustronny: H_1: $m \neq m_0$

Jednostronny obszar krytyczny testu – jest to obszar krytyczny złożony z jednego podzbioru przestrzeni próby, wybranego z jednej strony w rozkładzie odpowiedniej statystyki. Używa się go wówczas, gdy hipoteza H_1 występuje w postaci nierówności typu <, > (rys. 7 i 8).



Rys. 7. Wartości zmiennej *Z* na tle rozkładu zmiennej standaryzowanej dla *α=0,05, α=0,01* dla testów jednostronnych - obszar krytyczny lewostronny: H_1: $m < m_0$



Rys. 8. Wartości zmiennej *Z* na tle rozkładu zmiennej standaryzowanej dla *α=0,05, α=0,01* dla testów jednostronnych - obszar krytyczny prawostronny: H_1: $m > m_0$

3.1.Testy istotności dla wartości oczekiwanej (średniej) - jednopróbkowy.

Testy te mają zastosowanie dla porównania średniej empirycznej, czyli średniej uzyskanej z próby $\bar{x}$ z wartością oczekiwaną μ czyli parametrem populacji generalnej. W przypadku tych testów parametr nie jest określony jako przedział, tak jako było pokazane w wcześniejszych rozdziałach traktujących o estymacji przedziałowej, tylko jako konkretna wartość liczbowa. Zatem testy te mają zastosowanie do porównania średniej z próby z konkretną ustaloną wartością parametru.

W tym rozdziale przedstawiono dwa modele, ale w każdym modelu rozpatrywane są testy dwustronne i jednostronne (lewostronny i prawostronny).

MODEL I

Założenia:

1. Populacja generalna ma rozkład normalny N(m, σ)
2. Znamy odchylenie standardowe σ

Ten model może mieć zastosowanie w kontroli zgodności deklarowanych parametrów produktu albo usługi z wynikami próby. Można również zastosować ten test do porówna wyniku jednego przypadku (wówczas liczebność próby n=1) z wynikiem jaki powinien być osiągnięty tzn. czy pojedynczy przypadek odbiega od normy czy tez się w niej mieści. W zależności od celu stosuje się testy dwustronne lub jednostronne.

TEST DWUSTRONNY

Z populacji tej pobrano n- elementową próbę w celu zweryfikowania hipotezy
H_0: $m = m_0$ WOBEC H_1: $m \neq m_0$,
Gdzie: m_0 - wartości oczekiwan populacji generalnej,
m - średnia z próby $\bar{x}$

Dla rozwiązania problemu, która hipoteza jest słuszna zastosowanie ma **test istotności** – sprawdza on słuszność hipotezy H_0. Statystyką testową jest wartość z_d

$$Z_d = \frac{m - m_0}{\sigma} \times \sqrt{n} \quad (13.)$$

Z tablic rozkładu normalnego N(0,1) wyznacza się wartość krytyczną $z_{\alpha/2}$, aby dla założonego z góry małego prawdopodobieństwa α (poziomu istotności) zachodziła zależność (nierówność):

$$P\left\{|Z_d| \geq Z_{\frac{\alpha}{2}}\right\} \quad (14.)$$

Zbiór wartości z określony nierównością $|z_d| \geq z_{\alpha/2}$ jest obszarem krytycznym tego testu, tzn. że jeśli z próby otrzymamy taką wartość z_d, że

$\left|Z_d\right| \geq Z_{\alpha/2}$ to H_0 odrzucamy,

$\left|Z_d\right| < Z_{\alpha/2}$ to nie ma podstaw do odrzucenia H_0.

Przykład 12.

Po przestudiowaniu rozmiarów pewnego chromosomu w dużej próbie ludzi zdrowych ustalono parametry stosunku dłuższego ramienia do krótkiego. Cecha ta ma rozkład normalny z wartością oczekiwaną = 1,77 i odchyleniem standardowym = 0,043. U pacjenta z podejrzeniem choroby genetycznej, stwierdzono, że stosunek ten ma wartość 1,62. Czy można zaliczyć tego pacjenta do populacji ludzi zdrowych? Przyjąć poziom istotności = 0,01. W tym przykładzie liczebność próby stanowi n=1. Nie zakładamy też z góry czy wynik pacjenta jest większy czy mniejszy od prawidłowego tylko czy się od niego rożni.

$$Z_d = \frac{162 - 1{,}77}{0{,}043} \times \sqrt{1}$$

$$Z_d = -3{,}49$$

$$Z_{\alpha/2} = 2{,}58$$

$$\left|Z_d\right| \geq Z_{\alpha/2}$$

Należy zwrócić uwagę, że z porównujemy wartości bezwzględne, zatem odrzucamy hipotezę zerową. Wniosek: pacjent jest dotknięty chorobą genetyczną.

Przykład 13.

W pewnej fabryce uruchamiano nowa linię technologiczną. Zgodnie z projektem na podstawie wyliczeń matematycznych procesu produkcyjnego dzienne zużycie wody tej linii produkcyjnej jest zmienną losową o rozkładzie normalnym N(1000, 20). Na podstawie obserwacji n = 196 dni w roku stwierdzono, że średnie zużycie wody wynosi 1025 m^3. Na poziomie istotności 0,05 zweryfikować hipotezę, że średnie dzienne zużycie wody różni się od teoretycznego.

Ponieważ stosujemy nie zakładamy z góry czy zużycie wody jest mniejsze czy większe, zastosowanie ma test dwustronny:

H_0: $\bar{x} = m_0$ (1000)

H_1: $\bar{x} \neq m_0$ – test dwustronny

$$Z_d = \frac{m - m_0}{\sigma} \times \sqrt{n} = \frac{1025 - 1000}{20} \times \sqrt{196} = 17{,}5$$

Z tablic dystrybuanty rozkładu normalnego N(0,1) odczytujemy wartość $z_{\alpha/2}$ w taki sposób, aby P = 1-α = 0,95, tj. $z_{\alpha/2}$ = 1,96. (patrz rys. 6 – środkowy rozkład)

Ponieważ $z_{\alpha/2} < z_d$, Hipotezę H_0 odrzucamy na korzyść hipotezy alternatywnej H_1.

Wniosek: Średnie dzienne zużycie wody różni się istotnie od teoretycznego 1000 l.

TESTY JEDNOSTRONNE

H_0: $\bar{x} = m_0$ wobec hipotezy alternatywnej

H_1: $\bar{x} > m_0$, *lub* $\bar{x} < m_0$ gdzie m_0 jest konkretną wartością oczekiwaną populacji generalnej

A. TEST PRAWOSTRONNY

H_0: $\bar{x} = m_0$ wobec H_1: $\bar{x} > m_0$

Statystyką testową jest wartość z_d – i obliczamy ją tak samo jak dla testu dwustronnego - wzór nr 13. Różnica w stosunku do testu dwustronnego polega na porównaniu $|Z_d|$ z inną wartością Z (tj. cały obszar krytyczny znajduje się po jednej stronie rozkładu - patrz rys. 8):

$|Z_d| \geq Z_{1-\alpha}$ to H_0 odrzucamy,

$|Z_d| < Z_{1-\alpha}$ to nie ma podstaw do odrzucenia H_0.

Przykład 14.

Zakładamy, że średnia plonów żyta w Polsce jest większa od 25 dt·ha^{-1}. Wiadomo, że odchylenie standardowe plonów żyta wynosi 4 dt·ha^{-1}. Na podstawie próby 100 gospodarstw otrzymano średni plon = 26,5 dt·ha^{-1}. Zweryfikować tę hipotezę na poziomie 5% błędu.

H_0: $\bar{x} = m_0$ (26,5 = 25) wobec H_1: $\bar{x} > m_0$ (26,5 > 25)

$$Z_d = \frac{m - m_0}{\sigma} \times \sqrt{n} = \frac{26{,}5 - 25}{4} \times \sqrt{100} = 3{,}75$$

5% błąd oznacza, że musimy przyjąć poziom istotności α=0,05 tj. $Z_{1-\alpha} = 0{,}95$ czyli 1,64 (patrz rys. 8 - rozkład po lewej stronie rysunku). Ponieważ z_d 3,75 jest > od $z_{1-\alpha}$ 1,64 mamy podstawię do przyjęcia hipotezy H_1 Wniosek: plon żyta w Polsce jest większy niż 25 dt·ha^{-1}.

Sprawdź, czy można uznać, że plon jest większy od 26 dt·ha^{-1}?

B. TEST LEWOSTRONNY

H_0: $\bar{x} = m_0$ wobec H_1: $\bar{x} < m_0$

JEŚLI

$|Z_d| \geq |Z_\alpha|$ to H_0 odrzucamy,

$|Z_d| < |Z_\alpha|$ to nie ma podstaw do odrzucenia H_0.

Przykład 15.

Pewien automat w fabryce czekolady wytwarza tabliczki czekolady o nominalnej masie 250 g. Wiadomo, że rozkład masy produkowanych tabliczek jest normalny N(m=250, σ=5). Kontrola w pewnym dniu pobrała próbę 16 tabliczek i otrzymała średnią = 244 g. Czy można twierdzić, że automat rozregulował się i produkuje tabliczki o mniejszej masie niż przewiduje norma? Zweryfikować tę hipotezę na poziomie α=0,05

H_0: $m = m_0$ (244=250) wobec H_1: $m < m_0$ (244< 250)

$$Z_d = \frac{m - m_0}{\sigma} \times \sqrt{n} = \frac{244 - 250}{5} \times \sqrt{16} = -4{,}8$$

$$z_{\alpha = 0{,}05} = -1{,}64$$

$$|-4{,}8| > |-1.64|$$

WNIOSEK: Z prawdopodobieństwem błędu mniejszym, niż 0,05 możemy twierdzić, że średnia masa produkowanych tabliczek czekolady jest za niska.

MODEL II

Założenia:

1. Populacja generalna ma rozkład normalny N(m, σ)
2. Odchylenie standardowe σ populacji jest nieznane
3. Z populacji pobieramy próbę małą (do 30 elementów)

Procedura obliczeniowa jest podobna jak w modelu pierwszym z tym, że statystyką testową jest wartość *t – Studenta*:

$$t_{obl} = \frac{m - m_0}{s} \times \sqrt{n-1} \quad (15.)$$

TEST DWUSTRONNY

H_0: $\overline{X} = m_0$ wobec H_1: $\overline{X} \neq m_0$

Jeśli $t_{obl} < t_{tab,\ dla}\ \alpha\ i\ \nu = n\text{-}1$ to H_0 przyjmujemy

Jeśli $t_{obl} \geq t_{tab,\ dla}\ \alpha\ i\ \nu = n\text{-}1$ to H_0 odrzucamy

Przykład 16.

Pobrano losowe próby gniazd lęgowych myszy polnej o liczności n = 8, i obliczono, że wariancja liczby myszy w gnieździe równa się 4, ze średnią = 6,25. Zweryfikować hipotezę zerową, że wartość oczekiwana populacji generalnej wynosi 5 myszy, przy poziomie istotności 0,05.

$$t_{obl} = \frac{m - m_0}{S} \times \sqrt{n-1} = \frac{6{,}25 - 5}{2} \times \sqrt{7} = 1{,}65$$

Wartość tablicowa t dla n-1=7 i α=0,05 (dwustronny) = 2,365.

Wniosek: brak podstaw do odrzucenia hipotezy zerowej, która mówi, że wartość oczekiwana osobników myszy w gnieździe jest równa 5.

Ten wynik i wniosek wymaga komentarza. Różnica między średnią empiryczną i teoretyczną jest całkiem spora (1,25 osobników) a jednak nie została potwierdzona statystycznie. Im mniejsza jest liczebność próby tym trudniej jest udowodnić istotność uzyskanej różnicy. Inaczej im mniejsza jest próba tym różnica musi być większa by ją udowodnić dotyczy to nie tylko tego testu ale praktycznie wszystkich testów statystycznych. Oczywiście w powyższych uwagach pominąłem rolę odchylenia standardowego – im jest ono mniejsze tym łatwiej jest udowodnić różnice, ale w praktyce zazwyczaj jest tak, że im mniejsza liczebność próby tym również odchylenie standardowe jest większe.

TEST PRAWOSTRONNY:

H_0: $\overline{X} = m_0$ wobec H_1: $\overline{X} > m_0$

Jeśli $t_{obl} < t_{tab, dla}$ α i $\nu = n-1$ to H_0 przyjmujemy

Jeśli $t_{obl} \geq t_{tab, dla}$ α i $\nu = n-1$ to H_0 odrzucamy

Rozpatrujemy przykład 16 testem prawostronnym:

Wartość tablicowa t dla n-1=7 i α=0,05 (jednostronny) = 1,895

Wniosek: w tym przypadku również nie ma podstaw do odrzucenia hipotezy zerowej, która mówi, że wartość oczekiwana osobników myszy w gnieździe jest równa 5.

TEST LEWOSTRONNY

H_0: $\overline{X} = m_0$ wobec H_1: $\overline{X} < m_0$

Przykład 17.

Plony pszenicy w gospodarstwach indywidualnych województwa kujawsko-pomorskiego maja rozkład normalny o nieznanych parametrach. Przypuszcza się, że plony są rzędu 45 dt z ha. Czy przypuszczenie to jest słuszne, jeśli w próbie złożonej z losowo wybranych 26 gospodarstw otrzymano średnią równą 38,9 dt·ha^{-1} oraz odchylenie standardowe = 3,5 dt·ha^{-1}? Przyjmij poziom istotności 0,05.

$$t_{obl} = \frac{m - m_0}{S} \times \sqrt{n-1} = \frac{38{,}9 - 45}{3{,}5} \times \sqrt{26} = -8{,}89$$

Wartość tablicowa t dla n-1=25 i α=0,05 (jednostronny) = 1,708

Oczywiście porównujemy wartości bezwzględne. Wartość t_{obl} jest większa niż tablicowa więc są podstawy do odrzucenia hipotezy zerowej. Wniosek: na podstawie przeprowadzonych badań stwierdzono, że plon pszenicy są mniejsze niż się powszechnie przypuszcza.

3.2.Testy istotności dla dwóch średnich

W badaniach przyrodniczych bardzo często zachodzi potrzeba porównania dwóch obiektów. Nie wystarczy pobrać próbę dla jednego obiektu i drugiego, wyliczyć średnie i je ze sobą porównać. Przecież do każdej próby poszczególne przypadki pobierane są z populacji

generalne w sposób całkowicie losowy (zrandomizowany). Zatem mamy tu do czynienia z tym samym dylematem jaki rozpatrywaliśmy w rozdziale drugim, każda próba pobrana z tej samej populacji da najprawdopodobniej inny wynik. Wynika z tego, że tym bardziej porównując dwie próby z dwóch różnych populacji uzyskany wynik w postaci różnicy średnich musi być udowodniony. Do tego właśnie służą testy istotności różnicy dwóch średnich. Wybór testu zależy od rozkładu, znajomości wariancji i liczebności próby.

Model I

1. Badane populacje mają rozkłady normalne $N(m_1, \sigma_1)$ oraz $N(m_2, \sigma_2)$
2. Wartości oczekiwanych nie znamy, ale σ_1 i σ_2 są znane.
3. Z każdej populacji losujemy próby o liczebności odpowiednio n_1 i n_2 z nich obliczmy średnie dla prób $\overline{x_1}$ i $\overline{x_2}$.

H_0: $m_1 = m_2$
wobec
H_1: $m_1 \neq m_2$ - test dwustronny lub H_1: $m_1 < m_2$ – test lewostronny

lub H_1: $m_1 > m_2$ - test prawostronny

$$Z_d = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (16.)$$

Ten model można też zastosować w przypadku kiedy nie znamy faktycznej wartości wariancji populacji generalnej, ale dysponujemy próbami niemałymi pod warunkiem, że $n_1 + n_2 > 120$. Trzeba zaznaczyć również że nie powinno być dużej dysproporcji między liczebnością obu prób. Wówczas wzór przyjmuje postać:

$$Z_d = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (17.)$$

Dla testu dwustronnego – jeśli:

$|Z_d| \geq Z_{\frac{\alpha}{2}}$ to H_0 odrzucamy,

$|Z_d| < Z_{\frac{\alpha}{2}}$ to nie ma podstaw do odrzucenia H_0

Dla testu lewostronnego – jeśli:

$|Z_d| \geq Z_\alpha$ to H_0 odrzucamy,

$|Z_d| < Z_\alpha$ to nie ma podstaw do odrzucenia H_0

Przykład 18.

Studenci dwóch równoległych lat zarządzania (1) i biotechnologii (2) uzyskali następujące średnie wyników nauczania:

$\overline{x_1} = 3{,}6$

$\overline{x_2} = 4{,}1$

Odchylenia standardowe wyników były następujące:

$S_1 = 2{,}1$

$S_2 = 1{,}8$

Liczba ocen dla 1 próby $n_1 = 200$

Liczba ocen dla 2 próby $n_2 = 280$

Sprawdź słuszność hipotezy zerowej versus hipotezy roboczej mówiącej o zróżnicowaniu średnich ocen oraz o lepszych wynikach studentów z kierunku biotechnologii dla poziomu istotności 0,05.

$$Z_d = \frac{3{,}6 - 4{,}1}{\sqrt{\frac{2{,}1^2}{200} + \frac{1{,}8^2}{280}}} = -2{,}73$$

Pamiętamy, że porównujemy zawsze wartość bezwzględne, dla poziomu istotności 0,05 testu dwustronnego jest to 1,96 a dla testu jednostronnego 1,64. A więc w obu przypadkach wartość obliczona jest większa od wartości krytycznej (tablicowe). Zatem mamy podstawę do przyjęcia hipotezy H1. Wniosek: Średnie ocen studentów porównywanych kierunków różnią się istotnie, studenci biotechnologii uzyskali lepsze wyniki.

Model II

Ten model ma najczęstsze zastosowanie w naukach przyrodniczych i prezentowany jest w różnych opracowaniach jako podstawowy test istotności różnicy średnich t-studenta. Trzeba mieć jednak świadomość, że jest więcej modeli i modyfikacji testów istotności różnicy średnich w zależności od tego czy próby są powiązane czy nie i czy ich wariancje są homogeiczne (jednorodne) czy nie. Dla celów dydaktycznych jednak prezentujemy tylko model podstawowy

1. Badane populacje mają rozkłady normalne $N(m_1, \sigma_1)$ oraz $N(m_2, \sigma_2)$
2. Wartości oczekiwanych nie znamy i σ_1 i σ_2 są nieznane.
3. Wariancje wyliczone z prób są jednorodne (homogeiczne)

4. Próby nie są powiązane (sparowane)

5. Z każdej populacji losujemy próby o liczebności odpowiednio n_1 i n_2 – próby małe do n = 30, z nich obliczmy średnie dla prób $\overline{x_1}$ i $\overline{x_2}$ i s^2_1 oraz s^2_2

H_0: $m_1 = m_2$

H_1: $m_1 \neq m_2$ - test dwustronny lub H_1: $m_1 < m_2$ – test lewostronny

lub H_1: $m_1 > m_2$ - test prawostronny

$$t_{obl} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (19.)$$

Jeśli $t_{obl} < t_{tab,\ dla}\ \alpha$ i ν = n-1 to H_0 przyjmujemy

Jeśli $t_{obl} \geq t_{tab,\ dla}\ \alpha$ i ν = n-1 to H_0 odrzucamy

Przykład 19.

W celu porównania (dla α=0,05) przeciętnego stażu pracy pracowników w dwóch zakładach wylosowano z każdego z tych zakładów grupę i zbadano pod względem liczby lat pracy. Otrzymano następujące wyniki:

Zakład A: $n_1 = 26, \overline{x_1} = 6{,}8, s_1 = 1{,}7$

Zakład B: $n_2 = 30, \overline{x}_2 = 8{,}2, s_2 = 2{,}5$

$$t_{obl} = \frac{6{,}8 - 8{,}2}{\sqrt{\frac{25 \times 1{,}7^2 + 29 \times 2{,}5^2}{26 + 30 - 2} \times \left(\frac{1}{26} + \frac{1}{30}\right)}} = -2{,}41$$

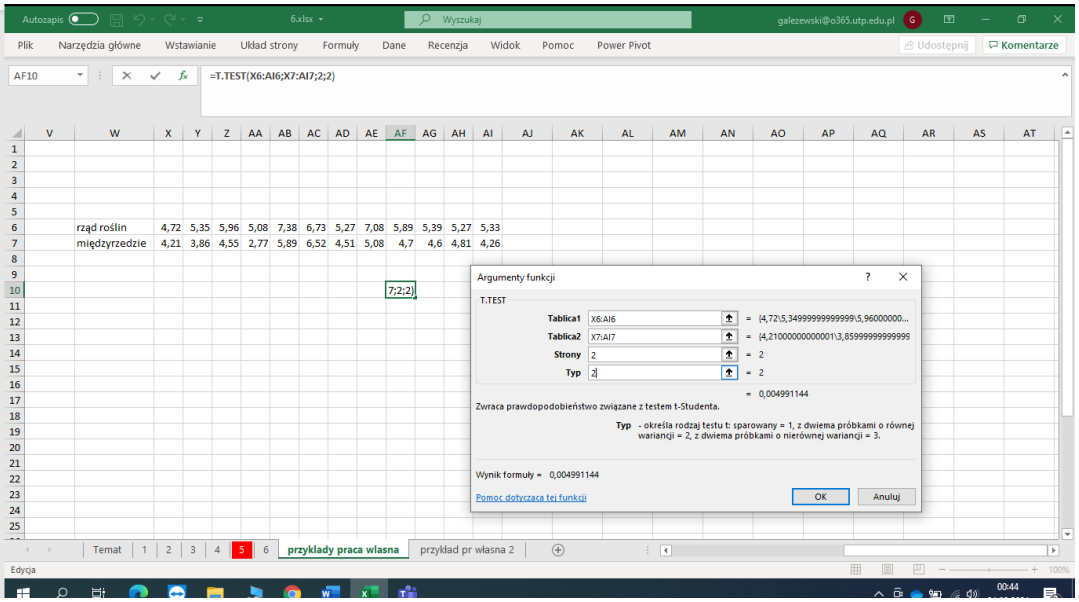
Wartość t_{tabl} dla α=0,05 i liczby stopni swobody ν=54 wynosi 1,67. Ponieważ wartość obliczona jest większa niż wartość tablicowa mamy podstawy do odrzucenia hipotezy zerowej i twierdzenia o istotności różnicy średniej stażu pracy pracowników porównywanych zakładów.

Przykład 20.

Porównywano wilgotność gleby mierzoną z pomocą miernika TDR (reflektometrycznego) w rzędzie roślin i w międzyrzędziu – czyli w strefie korzeniowej i poza nią. Pomiarów dokonan w dwunastu punktach dla każdego obiektu. Wyniki zestawiono w tabeli:

rząd roślin	4,72	5,35	5,96	5,08	7,38	6,73	5,27	7,08	5,89	5,39	5,27	5,33
międzyrzedzie	4,21	3,86	4,55	2,77	5,89	6,52	4,51	5,08	4,7	4,6	4,81	4,26

Ten przykład rozwiążemy używając formuły =T.TEST w programie Excel.



Rys. 9. Zastosowanie testu t-studenta w programie Excel

W polu: tablica1- należy wpisać zakres komórek dla próby pierwszej (adresy wpisują się automatycznie jeśli zaznaczy kursorem te dane)

W polu: tablica 2- należy wpisać zakres komórek dla próby drugiej

W polu: strony - należy wpisać „1” jeśli ma być to test jednostronny lub „2” dla testu dwustronnego

W polu: Typ proszę przyjąć, że wpisujemy „2” – w wersji polskojęzycznej jest pewna nieścisłość w widocznym na rys 9 objaśnieniu tego okna „2” opisano jako dla prób o „równej” wariancji – zamiast tego powinno być „jednorodnej”.

Po naciśnięciu „ok” uzyskamy wynik 0,00491144 – nie jest to wartość t_{obl} tylko wartość p-value czyli prawdopodobieństwo popełnienia błędu jeśli przyjmiemy za prawdziwą hipotezę H1. Ponieważ prawdopodobieństwo to jest bardzo małe (poniżej p=0,05) możemy przyjąć hipotezę H1 za prawdziwą.

Wniosek: wilgotność gleby w rzędzie roślin i w międzyrzędu są różne.

Przykłady do własnego opracowania przez studentów:

1. Obserwując liczbę kilometrów, jaką w ciągu roku przebywają prywatne samochody osobowe, otrzymano w losowej próbie 100 samochodów xśr=12500 km i s=2400 km. Na poziomie istotności a=0,05 zweryfikować hipotezę, że przeciętna liczba kilometrów przebytych rocznie przez prywatny samochód wynosi 12000 km.

2. W fabryce produkującej klej do ceramiki każdy worek tego produktu ma określony na opakowaniu ciężar 25 kg z tolerancją ±0,2 kg. Zaistniały przesłanki pozwalające przypuszczać, że pakowany klej nie odpowiada normom wagowym, co skłoniło dyrekcję fabryki do zlecenia odpowiednich badań testujących. Pobrano próbę złożoną ze 100 opakowań, zważono je, po czym, wyznaczono średnią masę 1 worka xśr=24,7 kg. Na poziomie istotności a=0,05 zbadać zasadność zastrzeżeń, zakładając, że rozkład masy kleju jest rozkładem normalnym.

3. Sklep spożywczy otrzymał dostawę maku w torebkach, z których każda powinna ważyć 500 gramów. Stosunkowo częste reklamacje spowodowały, że przeprowadzono wyrywkową kontrolę ich wagi. Podczas kontroli wylosowano, a następnie zważono 17 torebek uzyskując następujące wyniki: 500, 485, 480, 500, 480, 485, 465, 475, 480, 480, 491, 489, 503, 492, 475, 465, 500. Zakłada się normalność rozkładu masy zawartości torebki. Przyjmując poziom istotności a=0,05 zweryfikować hipotezę, że masa maku w torebkach jest zgodna z masą nominalną.

4. W wyniku przeprowadzonej ankiety wśród wybranych 2 sektorów przedsiębiorstw uważa się, że nakłady kapitałowe w tych przedsiębiorstwach wzrosły w dwóch kolejnych latach. W roku 1998 przepadano 20 przedsiębiorstw otrzymując średnią nakładów 21 670 PLN przy odchyleniu standardowym 8300 PLN, natomiast w roku 1999 przebadano 30 przedsiębiorstw otrzymując średnią 42 889 PLN i odchylenie standardowe 9302 PLN. Na poziomie istotności a=0,01 przeprowadź weryfikację przyjętej opinii.

5. Bank chciał sprawdzić, które źródła pozyskiwania funduszy są częściej wybierane: publiczne czy prywatne. W tym celu zbadał 90 firm, które zaciągnęły pożyczki publiczne (przeciętna wysokość pożyczki wynosi 12 500 PLN przy odchyleniu standardowym 3400 PLN) oraz 130 firm zaciągających pożyczki ze źródeł prywatnych (przeciętna wysokość pożyczki wyniosła 16 000 PLN przy odchyleniu standardowym 5000 PLN). Czy na poziomie istotności a=0,05 można wyciągnąć wniosek, że częściej wybierane jest prywatne źródło pozyskiwania środków?

6. Ekonomiści przypuszczają, że tygodniowe wydatki rodzin w Krakowie są przeciętnie wyższe od tygodniowych wydatków rodzin w Katowicach. Zweryfikować tę hipotezę na poziomie istotności a=0,05 na podstawie próby 12 rodzin z Krakowa, dla których otrzymano średnią 48,2 PLN i odchylenie standardowe 10,3 PLN oraz na podstawie próby 15 rodzin z

Katowic, dla których otrzymano średnią 45 PLN i odchylenie standardowe 8,6 PLN. Rozkład tygodniowych wydatków rodzin jest rozkładem normalnym.

7. Zawartość skrobi w ziemniakach odmiany Bila wynosiła dla 8 prób: 18,1; 16,8; 17,0; 17,5; 17,8; 18,1; 17,9; 18,0. Natomiast dla odmiany Bard w 10 próbach: 16,5; 16,8; 16,9; 17,0; 17,2; 17,4; 15,8; 16,5; 17,0; 16,7. Na poziomie istotności 0,05 zweryfikuj przypuszczenie o wyższej zawartości skrobi w odmianie Bila.

Rozkład t-studenta

obszar krytyczny jednostronny, α=	0.1	0.05	0.025	0.02	0.01	0.005	0.001	0.0005
obszar krytyczny dwustronny, α=	**0.2**	**0.1**	**0.05**	**0.04**	**0.02**	**0.01**	**0.002**	**0.001**
liczba stopni swobody n-1	3.07768	6.31375	12.7062	15.8945	31.8205	63.6568	318.306	636.627
2	1.88562	2.91999	4.30265	4.84873	6.96456	9.92484	22.3272	31.5990
3	1.63774	2.35336	3.18245	3.48191	4.54070	5.84091	10.2145	12.9240
4	1.53321	2.13185	2.77644	2.99853	3.74695	4.60409	7.17318	8.61031
5	1.47588	2.01505	2.57058	2.75651	3.36493	4.03214	5.89344	6.86884
6	1.43976	1.94318	2.44691	2.61224	3.14267	3.70743	5.20763	5.95880
7	1.41492	1.89458	2.36462	2.51675	2.99795	3.49948	4.78528	5.40787
8	1.39682	1.85955	2.30600	2.44898	2.89646	3.35539	4.50079	5.04130
9	1.38303	1.83311	2.26216	2.39844	2.82144	3.24984	4.29681	4.78092
10	1.37218	1.81246	2.22814	2.35931	2.76377	3.16927	4.14370	4.58691
11	1.36343	1.79588	2.20099	2.32814	2.71808	3.10581	4.02470	4.43697
12	1.35622	1.78229	2.17881	2.30272	2.68100	3.05454	3.92963	4.31779
13	1.35017	1.77093	2.16037	2.28160	2.65031	3.01228	3.85198	4.22083
14	1.34503	1.76131	2.14479	2.26378	2.62449	2.97684	3.78739	4.14045
15	1.34061	1.75305	2.13145	2.24854	2.60248	2.94671	3.73283	4.07276
16	1.33676	1.74588	2.11991	2.23536	2.58349	2.92078	3.68615	4.01500
17	1.33338	1.73961	2.10982	2.22385	2.56693	2.89823	3.64576	3.96512
18	1.33039	1.73406	2.10092	2.21370	2.55238	2.87844	3.61048	3.92164
19	1.32773	1.72913	2.09302	2.20470	2.53948	2.86094	3.57940	3.88341
20	1.32534	1.72472	2.08596	2.19666	2.52798	2.84534	3.55181	3.84952
21	1.32319	1.72074	2.07961	2.18943	2.51765	2.83136	3.52715	3.81927
22	1.32124	1.71714	2.07387	2.18289	2.50832	2.81876	3.50499	3.79214
23	1.31946	1.71387	2.06866	2.17696	2.49987	2.80734	3.48496	3.76762
24	1.31784	1.71088	2.06390	2.17154	2.49216	2.79694	3.46678	3.74539
25	1.31635	1.70814	2.05954	2.16659	2.48511	2.78744	3.45019	3.72514
26	1.31497	1.70562	2.05553	2.16203	2.47863	2.77871	3.43500	3.70660
27	1.31370	1.70329	2.05183	2.15783	2.47266	2.77068	3.42103	3.68959
28	1.31253	1.70113	2.04841	2.15393	2.46714	2.76326	3.40816	3.67391
29	1.31143	1.69913	2.04523	2.15033	2.46202	2.75639	3.39624	3.65941
30	1.31041	1.69726	2.04227	2.14697	2.45726	2.75000	3.38519	3.64596
31	1.30946	1.69552	2.03951	2.14383	2.45282	2.74404	3.37490	3.63345
32	1.30857	1.69389	2.03693	2.14090	2.44868	2.73848	3.36531	3.62180
33	1.30774	1.69236	2.03452	2.13816	2.44479	2.73328	3.35634	3.61091
34	1.30695	1.69092	2.03224	2.13558	2.44115	2.72840	3.34793	3.60072
35	1.30621	1.68957	2.03011	2.13316	2.43772	2.72381	3.34004	3.59115
36	1.30551	1.68830	2.02809	2.13087	2.43449	2.71948	3.33262	3.58215
37	1.30485	1.68709	2.02619	2.12871	2.43145	2.71541	3.32563	3.57367
38	1.30423	1.68595	2.02439	2.12667	2.42857	2.71156	3.31903	3.56568
39	1.30364	1.68488	2.02269	2.12474	2.42584	2.70791	3.31279	3.55811
40	1.30308	1.68385	2.02108	2.12291	2.42326	2.70446	3.30688	3.55096
41	1.30254	1.68288	2.01954	2.12117	2.42080	2.70118	3.30127	3.54418
42	1.30204	1.68195	2.01808	2.11952	2.41847	2.69807	3.29595	3.53774
43	1.30155	1.68107	2.01669	2.11794	2.41625	2.69510	3.29089	3.53162
44	1.30109	1.68023	2.01537	2.11644	2.41413	2.69228	3.28607	3.52580
45	1.30065	1.67943	2.01410	2.11500	2.41212	2.68959	3.28148	3.52025
46	1.30023	1.67866	2.01290	2.11364	2.41019	2.68701	3.27710	3.51496
47	1.29982	1.67793	2.01174	2.11233	2.40835	2.68456	3.27291	3.50990
48	1.29944	1.67722	2.01063	2.11107	2.40658	2.68220	3.26891	3.50507
49	1.29907	1.67655	2.00958	2.10987	2.40489	2.67995	3.26508	3.50045
50	1.29871	1.67590	2.00856	2.10872	2.40327	2.67779	3.26141	3.49601
55	1.29713	1.67303	2.00404	2.10361	2.39608	2.66822	3.24515	3.47640
60	1.29582	1.67065	2.00030	2.09936	2.39012	2.66028	3.23171	3.46020
65	1.29471	1.66864	1.99714	2.09578	2.38510	2.65360	3.22042	3.44659
70	1.29376	1.66691	1.99444	2.09273	2.38081	2.64790	3.21079	3.43502
75	1.29294	1.66543	1.99210	2.09008	2.37710	2.64298	3.20249	3.42503
80	1.29222	1.66412	1.99006	2.08778	2.37387	2.63869	3.19526	3.41634
85	1.29159	1.66298	1.98827	2.08574	2.37102	2.63491	3.18890	3.40870
90	1.29103	1.66196	1.98667	2.08394	2.36850	2.63157	3.18327	3.40194
95	1.29053	1.66105	1.98525	2.08233	2.36624	2.62858	3.17825	3.39590
100	1.29007	1.66023	1.98397	2.08088	2.36422	2.62589	3.17374	3.39049
110	1.28930	1.65882	1.98177	2.07839	2.36073	2.62127	3.16598	3.38118
120	1.28865	1.65765	1.97993	2.07631	2.35782	2.61742	3.15954	3.37346
130	1.28810	1.65666	1.97838	2.07456	2.35537	2.61418	3.15411	3.36695
140	1.28763	1.65581	1.97705	2.07306	2.35328	2.61140	3.14946	3.36137
150	1.28722	1.65508	1.97591	2.07176	2.35146	2.60900	3.14545	3.35657
160	1.28687	1.65443	1.97490	2.07063	2.34988	2.60691	3.14195	3.35237
170	1.28655	1.65387	1.97402	2.06963	2.34848	2.60506	3.13886	3.34868
180	1.28627	1.65336	1.97323	2.06874	2.34724	2.60342	3.13612	3.34540
190	1.28602	1.65291	1.97253	2.06794	2.34613	2.60195	3.13368	3.34246
200	1.28580	1.65251	1.97190	2.06723	2.34514	2.60063	3.13148	3.33983
210	1.28560	1.65214	1.97132	2.06658	2.34424	2.59944	3.12949	3.33746
220	1.28541	1.65181	1.97081	2.06600	2.34342	2.59836	3.12769	3.33530
230	1.28524	1.65151	1.97033	2.06546	2.34267	2.59737	3.12604	3.33333
240	1.28509	1.65123	1.96990	2.06497	2.34199	2.59647	3.12454	3.33153
250	1.28495	1.65097	1.96950	2.06452	2.34136	2.59564	3.12315	3.32987
260	1.28482	1.65074	1.96913	2.06410	2.34078	2.59487	3.12188	3.32834
270	1.28469	1.65052	1.96879	2.06372	2.34024	2.59416	3.12069	3.32692
280	1.28458	1.65031	1.96847	2.06336	2.33974	2.59350	3.11960	3.32561
290	1.28448	1.65013	1.96818	2.06303	2.33928	2.59289	3.11857	3.32439
300	1.28438	1.64995	1.96790	2.06272	2.33884	2.59232	3.11762	3.32326
350	1.28398	1.64922	1.96677	2.06143	2.33705	2.58995	3.11368	3.31854
400	1.28367	1.64867	1.96591	2.06047	2.33571	2.58818	3.11073	3.31502
450	1.28344	1.64825	1.96525	2.05972	2.33466	2.58680	3.10844	3.31227
500	1.28325	1.64791	1.96472	2.05912	2.33383	2.58570	3.10661	3.31009
600	1.28296	1.64740	1.96393	2.05822	2.33258	2.58405	3.10387	3.30682
700	1.28276	1.64703	1.96336	2.05758	2.33169	2.58287	3.10192	3.30448
800	1.28261	1.64676	1.96293	2.05710	2.33102	2.58199	3.10045	3.30273
900	1.28249	1.64655	1.96260	2.05673	2.33050	2.58130	3.09931	3.30137
1000	1.28240	1.64638	1.96234	2.05643	2.33008	2.58075	3.09840	3.30028
rozkład normalny ∞	1.28155	1.64485	1.95996	2.05375	2.32635	2.57583	3.09023	3.29053