Statystyka w analizie i planowaniu eksperymentu
lista nr 2

1 Wprowadzenie

- **populacja** - zbiorowość będąca przedmiotem badania statystycznego;
- **parametry** - stałe liczbowe charakteryzujące poszczególne populacje;
- **próba** - część populacji na podstawie, której staramy się wnioskować o całej populacji.
- **statystyki** - oszacowania parametrów populacji na podstawie próby.

Sumę (iloczyn) wszystkich elementów zbioru liczb $A = \{X_1, X_2, \ldots, X_n\}$ możemy zapisać w postaci:

$$X_1 + X_2 + \ldots X_n,\ X_1 \cdot X_2 \cdot \ldots \cdot X_n$$

ale gdy liczb jest dużo wypisywanie tych x-ów nie ma sensu i dlatego w zamian stosujemy znak sumowania (iloczynu)

$$\sum_{i=1}^{n} X_i,\ \Pi_{i=1}^{n} X_i$$

Zadanie 1 *Oblicz wartości następujących sum:*

a) $\sum_{i=3}^{7} i$ b) $\sum_{i=1}^{3} 3^i$ c) $\sum_{i=0}^{2} (i+1)^i$ d) $\sum_{i=0}^{5} 4$

e) $\sum_{i=3}^{7} 6$ f) $\sum_{i=0}^{7} c$ g) $\sum_{i=1}^{100} i$ h) $\sum_{i=1}^{\infty} \frac{1}{2^i}$

2 Miary położenia

2.1 Średnia arytmetyczna

Średnia arytmetyczna jest pewną charakterystyką zbioru (próby) dla danych liczb $X_1, X_2, \ldots, X_n$ średnia arytmetyczna wyraża się wzorem:

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

Zadanie 2 *Oblicz średnie arytmetyczne dla następujących zbiorów danych*

a) $55.1, 56.55, 61.55, 66.55, 71.55, 76.55, 81.55, 86.55,$ odpowiedź: 69.49375
b) $12.64, 395.85, 553.95, 665.5, 787.05, 995.15, 244.65, 173.1,$ odpowiedź: 478.48625

2.2 Średnia ważona

Odnosi się zwykle do sytuacji, gdy obliczamy średnią arytmetyczną z już policzonych średnich dla podgrup zadanego zbioru.

Zadanie 3 *Trzech studentów leśnictwa mierzyło pierśnicę drzew w jednym oddziale leśnym. Pierwszy zmierzył* 5 *drzew i otrzymał średnią* 120.0 *cm, drugi* 20 *drzew i średnią* 70.0*cm, trzeci natomiast* 50 *drzew i średnią* 40.0 *cm*

1. Oblicz zwykła średnią arytmetyczną. Dlaczego uzyskany wynik jest niewiarygodny?
2. Oblicz średnią ważoną posługując się wzorem:

$$\bar{X}_w = \frac{\sum w_i \bar{X}_i}{\sum w_i}$$

Zadanie 4 *W grupie* 12 *kobiet i* 6 *mężczyzn oznaczono zawartość glukozy w* $mg/100ml$. *Oblicz średnią wartość:*

(a) dla kobiet;

(b) dla mężczyzn;

(c) w grupie połączonej.

K) $123, 65, 72, 105, 95, 110, 82, 115, 80, 99, 117, 77$
M) $55, 67, 80, 51, 47, 72$

2.3 Średnia geometryczna

Dla zadanego zbioru danych $X_1, X_2, \ldots, X_n$ średnia geometryczna wyraża się wzorem:

$$M_G = \sqrt[n]{\Pi_{i=1}^{n} X_i}.$$

Średnią geometryczną oblicza się wówczas, gdy wyniki w trakcie badań zmieniają się w przybliżeniu w postępie geometrycznym. Średniej tej nie stosuje się, gdy mamy wartości ujemne lub równe zero. Zauważmy, że

$$\log(M_G) = \frac{1}{n} \sum_{i=1}^{n} \log(X_i).$$

Bardzo często w badaniach biologicznych interesuje nas zmiana pewnej wielkości (np. przyrost masy) w stosunku do okresu poprzedniego. Możemy obliczyć średni przyrost względny dla danego okresu.

Niech $x_1, x_2, \ldots x_n$ będą to wartości danej wielkości w poszczegolnych okresach czasu. Średnia względna zmiana tej wielkości wyraża się wzorem:

$$\sqrt[n]{\frac{x_2}{x_1} \cdot \frac{x_3}{x_2} \cdot \ldots \cdot \frac{x_n}{x_{n-1}}} = \sqrt[n]{\frac{x_n}{x_1}}.$$

Zadanie 5 *W doświadczeniu nad tuczem bydła otrzymano następujące średnie masy zwierząt w kolejnych siedmiu miesiącach:* $209, 236, 277, 283, 330, 335, 340$ *kg. Oblicz średni miesięczny przyrost masy tej rasy bydła.*

2.4 Średnia harmoniczna

$$M_H = \frac{N}{\sum \frac{1}{X_i}}$$

Zadanie 6 *W kwadracie o boku* 100 *km dokonywany jest przelot samolotem z różną prędkością. Pierwszy bok kwadratu pokonany został z prędkością* 100 *m/s, drugi bok kwadratu z prędkością* 200 *m/s, trzeci bok kwadratu został pokonany z prędkością* 300 *m/s a czwarty z prędkością* 400 *m/s. Jaka jest średnia prędkość przelotu tego samolotu?*

2.5 Moda (dominanta)

Moda M_0 (wartość modalna, dominanta) jest to wartość występująca najczęsciej w danej zbiorowości z wykluczeniem wartości skrajnych (x_{min}, x_{max}). Kiedy każdy pomiar pojawia się taką samą liczbę razy w próbie, wtedy brak jest mody. W przypadku, gdy wartości występują w próbie z różną częstością modą będzie wartość pojawiająca się najczęściej (próba jednomodalna). Kiedy dwie wartości pojawiają się tę samą ilość razy (a ich częstotliwość jest większa od pozostałych, próba ma dwie mody (bimodalna). Jeżeli tych wartości jest więcej próbę nazwiemy wtedy **wielomodalną**.

Zadanie 7 *Określ wartość modalną dla podanych czterech ciągów:*

a) $12, 16, 15, 14, 16, 19, 16, 21, 11, 16$

b) $5, 6, 4, 6, 5, 6, 4, 3, 6, 9, 7, 6, 5$

c) $11, 10, 18, 19, 25, 22, 17, 13, 14,$

d) $2, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 9$

2.6 Mediana

Mediana (Me, me) dzieli podany, uporządkowany rosnąco ciąg liczbowy na połowe. (jest to taka liczba od której połowa liczb w tym ciągu jest mniejsza). Przy parzystej liczbie obserwacji medianą będzie umownie średnia arytmetyczna dwu środkowych obserwacji o numerach $n/2$ i $n/2 + 1$

Zadanie 8 *Zmierzono zawartość glukozy u 9 pacjentów. Otrzymano wyniki:*
$110, 78, 52, 59, 127, 90, 135, 110, 93$. *Oblicz:*

1. średnią zawartość glukozy;
2. medianę;
3. modę.

2.7 Kwantyle (kwartyle, decyle, centyle)

Podobnie jak mediana dzielą one uporządkowany ciąg liczbowy na części o jednakowej liczności. Kwartyl pierwszy Q_1 jest to wartość poniżej, której znajduje się 1/4 wszystkich jednostek, podczas gdy kwartylem trzecim Q_3 nazywamy taką wartość poniżej, której znajduje sie 3/4 wartości liczbowych.

Decyle dzielą uporządkowany ciąg liczb na 10 tak samo licznych części.

Centyle dzielą uporządkowany ciąg liczb na 100 tak samo licznych części.

3 Wskaźniki rozproszenia

Zadanie 9 *Oblicz srednią arytmetyczną dla zadanych ciągów liczbowych:*

1. $26, 26, 26, 27, 27, 27, 28, 29$;
2. $12, 13, 21, 27, 31, 32, 38, 42$;
3. $3, 4, 10, 14, 27, 50, 51, 57$;

Porównując średnie tak bardzo różniących sie pomiarów. Widać dobitnie potrzebę wprowadzenia kolejnych charakterystyk opisujących zachowanie obserwowanej próby.

3.1 Rozstęp

Najprostszą miarą rozproszenia jest rozstęp, równy różnicy między największą a najmniejszą wartością obserwacji

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

Zadanie 10 *Oblicz rozstęp we wszystkich wymienionych w zadaniu* 8 *przykładach.*

3.2 Rozstęp międzykwartylowy

Rozstęp międzykwartylowy

$$q = Q_3 - Q_1$$

jest to różnica między 3 i 1 kwartylem

3.3 Odchylenie przeciętne

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

3.4 Wariancja i odchylenie standardowe w populacji

Wariancja populacji skończonej σ^2 jest miarą rozproszenia wartości indywidualnych w populacji wokół średniej w populacji μ. Wariancja to średnia arytmetyczna kwadratów odchyleń cechy populacji od jej wartości średniej:

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

Natomiast odchylenie standardowe w populacji to

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

3.5 Wariancja i odchylenie standardowe próby

Wariancja próby

$$S_{n-1}^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Odchylenie standardowe próby

$$S_{n-1} = \sqrt{S_{n-1}^2}$$

Zadanie 11 *Oblicz wariancję (S_{n-1}^2) oraz odchylenie standardowe (S_{n-1}) dla próby:*

$$110, 78, 52, 59, 127, 90, 135, 93, 110$$

3.6 Odchylenie standardowe średniej arytmetycznej

Średnia arytmetyczna próby ($\bar{x}$) jest przybliżoną wartością średniej populacji μ. Wykonanie analogicznych pomiarów w tych samych warunkach da prawie zawsze różne wartości średnich arytmetycznych ($\bar{x_1}, \bar{x_2}, \ldots, \bar{x_n}$). Zrozumiałe jest zatem, że rozrzut (wariancja) średnich arytmetycznych z prób jest znacznie mniejszy niż pojedynczych wyników. Miarą rozrzutu średnich arytmetycznych jest odchylenie standardowe średnich arytmetycznych ($S_{\bar{X}}$), czyli **błąd średniej arytmetycznej** (**błąd standardowy**):

$$S_{\bar{X}} = \frac{S_{n-1}}{\sqrt{n}}$$

Zadanie 12 *Dla danych z zadania* 10 *obliczyć odchylenie standardowe średniej arytmetycznej.*

Zadanie 13 *Z populacji mężczyzn, celem określenia ich masy wybrano losowo próbę złożoną z* 58 *osób. Ich masę określono z dokadnością do* 0.1 *kg. Otrzymano następujące dane liczbowe:*

$$49.1, 54.5, 63.0, 64.6, 69.5, 74.4, 79.4, 85.8, 53.2, 55.4, 74.9, 75.0, 75.6$$
$$61.5, 65.0, 70.0, 75.0, 82.1, 87.1, 54.0, 54.1, 62.2, 65.6, 70.4, 75.9, 83.8$$
$$63.4, 66.7, 71.6, 75.2, 58.4, 60.9, 67.4, 72.7, 76.2, 59.0, 64.0, 68.3, 73.3$$
$$63.4, 66.7, 71.6, 75.2, 58.4, 60.9, 67.4, 72.7, 76.2, 59.0, 64.0, 68.3, 73.3,$$
$$63.4, 66.7, 71.6, 75.2, 58.4, 60.9, 67.4, 72.7, 76.2, 59.0, 64.0, 68.3, 70.9, 71.9, 60.7$$
$$73.3, 56.3, 76.5, , 57.7, 68.9, 74.0, 78.2, 61.0, 69.0, 72.6, 78.7, 62.8, 67.0,$$
$$73.3, 76.5, , 57.7, 68.9, 74.0, 78.2, 61.0, 69.0, 72.6, 78.7, 62.8, 67.0, 73.1, 78.1, 66.8,$$

Oblicz wariancję i odchylenie standardowe Odpowiedzi: $S^2_{n-1} = 61.3936125419933$*;*
$S_{n-1} = 7.83540761811364$*;*

3.7 Wariancja grupowa, wewnątrzgrupowa i międzygrupowa

Niech wartości liczbowe cechy mierzonej z dowolnej populacji będą podzielone na k- prób (grup). Rozpatrzmy każdą grupę oddzielnie, wariancja grupowa to wariancja wartości z danej grupy (próby) względem średniej z tej grupy.

3.7.1 Wariancja wewnątrzgrupowa

to średnia ważona wariancji grupowych

$$S^2_w = \frac{\sum S^2_i \cdot n_i}{N}$$

gdzie $N = \sum n_i$ liczność wszystkich grup
S^2_i wariancja grupy i
n_i liczność grupy i

3.7.2 Wariancja międzygrupowa

to wariancja średnich grupowych $\bar{x}_i$ względem średniej ogólnej $\bar{X}$

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

Zadanie 14 *Dla ustalenia liczby ziaren w kłosach badanej odmiany pszenicy zebrano losowo z poletka doświadczalnego określoną liczbę kłosów* 52*, a uzyskane dane zestawiono poniżej Oblicz wariancję i odchylenie standardowe.*

Liczba ziaren w kłosie	40	44	47	51	55	62	68	70
liczba kłosów	3	5	4	7	16	6	9	2

3.8 Współczynnik zmienności

$$V = \frac{S_{n-1}}{\bar{X}} \cdot 100\%, \text{ gdy } \bar{X} \neq 0$$

Współczynnik zmienności pozwala ocenić, jakim procentem ze średniej z danej proby jest odchylenie standardowe próby.

Zadanie 15 *Oblliczyć rozstęp, średnią arytmetyczną , odchylenie standardowe, odchylenie standardowe średniej oraz współczynnik zmienności dla następujących prób*

a) $26, 26, 26, 27, 27, 27, 28, 29$
b) $12, 13, 21, 27, 31, 32, 38, 42$
c) $3, 4, 10, 14, 27, 50, 51, 57$

4 Graficzne przedstawienie danych

Tablica 1: Masa ciała poczwarek mącznika

Numer pomiaru	Masa ciała
1	148
2	148
3	136
4	152
5	142
6	130
7	176
8	152
9	150
10	140
11	123
12	113
13	133
14	117
15	126
16	129
17	219
18	156
19	160
20	123

Praktycznym sposobem aby uswiadomić sobie jak kształtuje się zmienność tej próby, jest graficzne przedstawienie pomiarów. najprostszy polega na zaznaczeniu każdego pomiaru na poziomej lub pionowej osi, na której każdy pomiar reprezentowany jest przez kreskę.

Zadanie 16 *Na podstawie danych z tabeli wykonaj taki wykres.*

Opisana powyżej procedura jest przydatna gdy mamy do czynienia z kilkoma bądź kilkunastoma pomiarami. W naszym przypadku trzy wartości masy ciala powtórzyły się dwukrotnie i w tych trzech przypadkach kreska oznaczająca długość pomiaru jest dwukrotnie dłuższa. Gdyby pomiarów było znacznie więcej, więcej też byłoby powtarzających się wartości. To stwarzałoby duży problem w jasnym i klarownym opisie.

4.1 Histogram

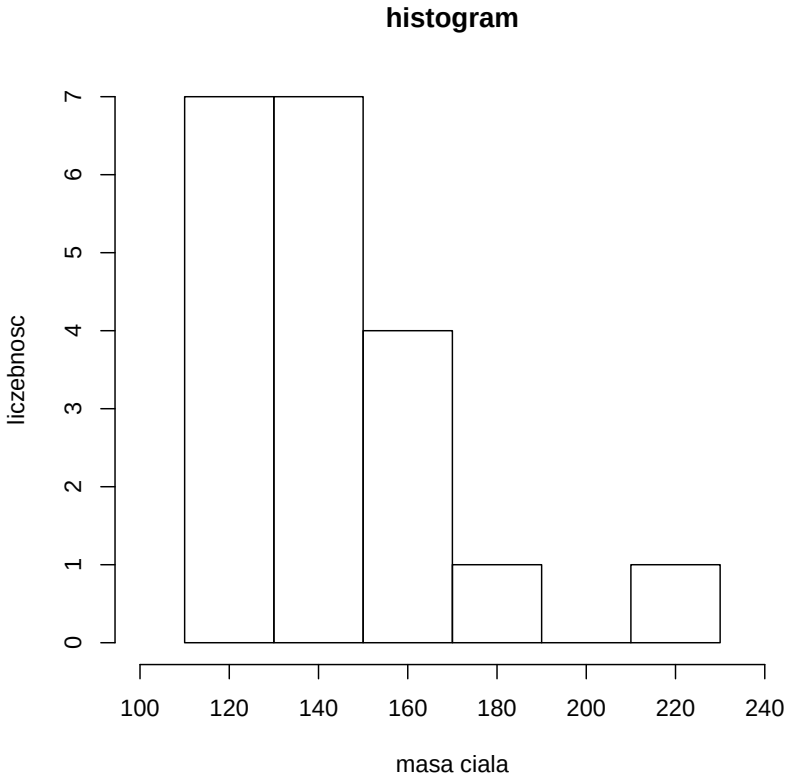
Gdy pomiarów jest dużo warto zrezygnować z dokładności i pogrupować dane w klasy tworząc tak zwany **szereg rozdzielczy**. Można na przykład wszystkie pomiary od 110 do 119 mg umieścić w jednej klasie (analogicznie dla 120 do 129 i tak dalej aż obejmiemy cały zakres) a następnie obliczyć częstości i liczebność wystąpień danych w zadanej klasie otrzymamy wtedy

Tablica 2: Masa ciała poczwarek mącznika

Granice przedziału	Liczebność	Częstość
$110 - 129$	6	0.30
$130 - 149$	7	0.35
$150 - 169$	5	0.25
$170 - 189$	1	0.05
$190 - 209$	0	0.00
$210 - 229$	1	0.05

Dzięki stworzeniu przedziałów klasowych możemy utworzyć wykres rozkładu pomiarów Przy konstruowaniu rozkładu liczebności trzeba podjąć decyzję, jak duży powinien być zakres każdego przedziału klasowego, a tym samym ile wprowadzić klas. Gdy pomiarów jest mniej nż 10 konstruowanie przedziałów klasowych nie ma sensu. Przy kilkunastu pomiarach stosujemy z reguły $4 - 5$ przedziałów klasowych, przy setkach pomiarów $8 - 10$ przedziałów, zaś przy tysiącach i więcej 12 przedziałów klasowych.

Zadanie 17 *Dla danych z zadania 12 utwórz histogram.*



4.2 Dystrybuanta empiryczna

Często zamiast badania liczby przypadków znajdujacych się pomiędzy dwiema wartościami badanej cechy (granice klasy), istnieje potrzeba poznania liczby przypadków n_x znajdujących się poniżej danej wartości x_i. w takim przypadku sporządza się **skumulowany szereg rozdzielczy**. Przykład takiego szeregu został zestawiony poniżej

Tablica 3: Masa ciała poczwarek mącznika

Granice przedziału	Liczebność skumulowana	Częstość skumulowana
$110 - 129$	6	0.30
$130 - 149$	13	0.65
$150 - 169$	18	0.90
$170 - 189$	19	0.95
$190 - 209$	19	0.95
$210 - 229$	20	1

4.3 Skośność - współczynnik skośności

Skośność jest miarą asymetrii obserwowanych wyników. Informuje nas o tym jak wyniki dla danej zmiennej kształtują się wokół średniej. Czy większość zaobserwowanych wyników jest z lewej strony średniej, blisko wartości średniej czy z prawej strony średniej? Innymi słowy, czy w naszym zbiorze obserwacji więcej jest wyników, które są niższe niż średnia dla całej grupy, wyższe czy równe średniej?

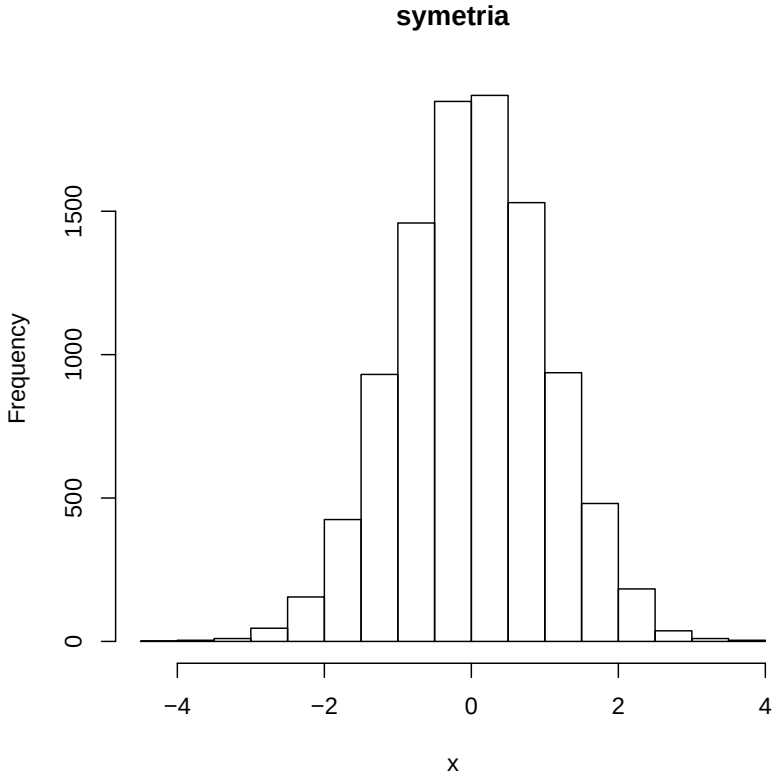
Współczynnik skośności gdy przyjmuje wartość bliską 0 świadczy o braku asymetrii wyników. Współczynnik skośności powyżej 0 świadczy o prawostronnej asymetrii rozkładu (inaczej nazywanym rozkładem dodatnioskośnym), a wyniki poniżej 0 świadczą o lewostronnej asymetrii rozkładu (inaczej nazwanym ujemnoskośnym rozkładem).

1. **symetria**

$$\bar{X} = Me = Mo$$

```
> x<-rnorm(10000)
> hist(x, main="symetria")
> library(e1071)
> skewness(x)

[1] -0.02759773
```

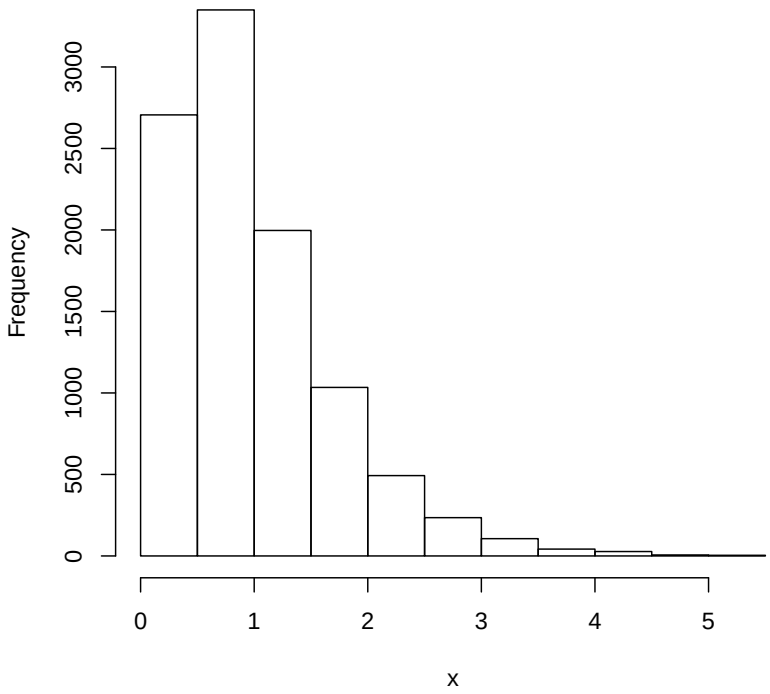


2. **asymetria prawostronna**

$$Mo < Me < \bar{X}$$

```
> x<-rgamma(10000,2,2)
> hist(x, main="asymetria prawostronna")
> library(e1071)
> skewness(x)

[1] 1.438639
```



3. **asymetria lewostronna**

$$\bar{X} < Me < Mo$$

```
> x<- -rgamma(10000,2,2)
> hist(x, main="asymetria lewostronna")
> library(e1071)
> skewness(x)

[1] -1.457087
```