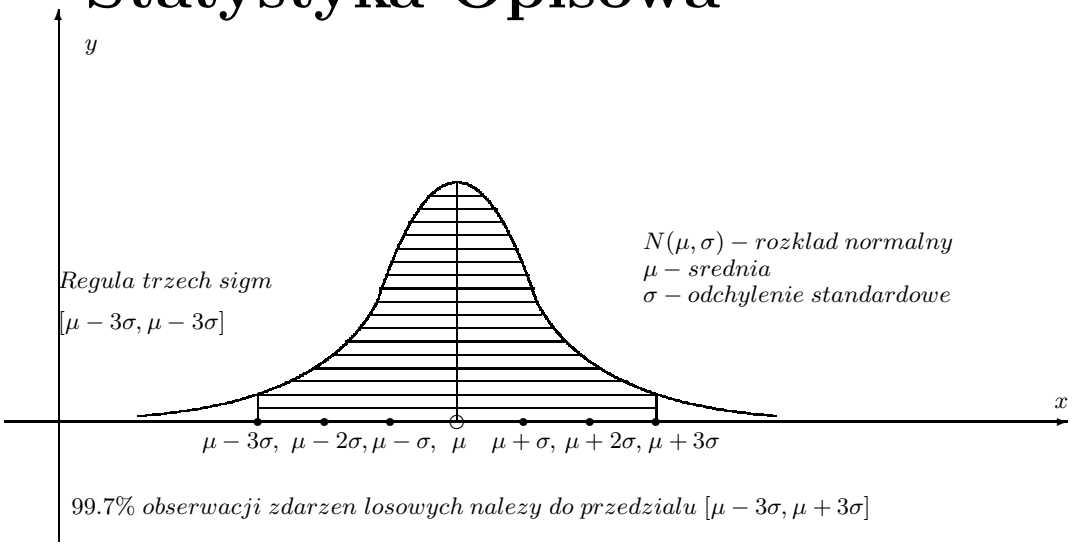
Chapter 1

Statystyka Opisowa



1.1 Wstęp

Pierwszym i ważnym etapem opracowań statystycznych jest zbieranie i prezetacja danych. Najważniejsze dane statystyczne podawane są w każdym roku przez Główny Urząd Statystyczny (GUS) z siedzibą w Warszawie. Dotyczą one informacji o ludności w Polsce, dane o wzorście w przemysłe i rolnictwie, w ekonowmi i finansach. Te dane stanowią ważną informację dla planowania i administracji państwa. Oprócz tego dane statystyczne zbierane są w ankietach z pytaniami o szczególnym znaczeniu. Na przykład w sondażach i prognozach w wyborach do sejmu i w ważnych decyzjach administrcji w których głos społeczeństwa ma istotne znaczenie. Zebrane dane statystyczne przedstawiamy w tabelach i ilustrujemy na diagramach. Stosowane są różne formy diagramów. Najbardziej powszechne diagramy są w formie słupków lub koła z zaznaczeniem kolorów lub danych liczbowych lub w procentach. Zatem diagramy są prostym i ważnym sposobem prezetacji danych statystycznych.

1.2 Dane Statystyczne. Diagramy

Dane statystyczne zapisujemy w tablicach z opisem ich znaczenia wartości liczbowych.

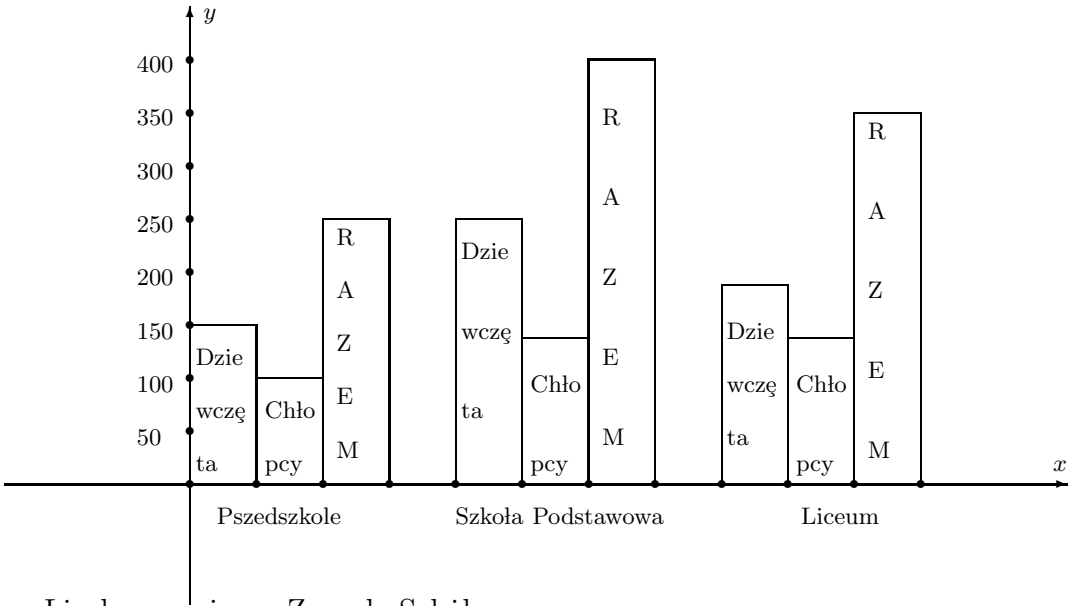
Przykład 1.1 *W zespole szkół było Przedszkole, Szkoła Podstawowa i Liceum. W tabeli zebrano informacje dotyczące liczby uczniów*

Rodzaj Szkoły	Liczba dziewczyn	Liczba chłopców	RAZEM
Przedszkole	150	100	250
Szkoła Podstawowa	250	150	400
Liceum	200	150	350

W niżej przedstawionych diagramach w formie słupków i koła podane są wykresy dziewczyn, chłopców i wykresy razem uczniów w Przedszkolu, w Szkole Podstawowej i w Liceum.

Diagram w postaci słupków.

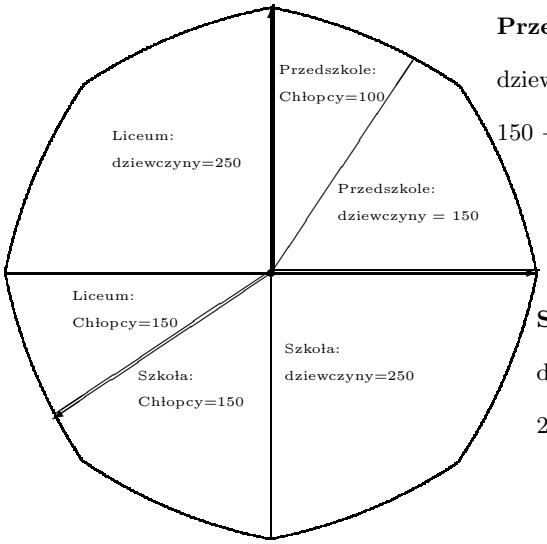
Legenda:
Dziewczyny słupek pierwszy,
Chłopcy słupek drugi,
Liczba uczniów razem słupek trzeci.
Trzy słupki są powtórzone dla każdej z trzech szkół.



Liczba uczniów w Zespole Szkół:
Liczba dziewczyn = 600
Liczba chłopców = 400
Razem dziewczyny+ chłopcy = 600+400 = 1000

Diagram w postaci koła.
Diagram w postaci koła zawiera następujące sekcje:

Sekcja Przedszkole : dziewczyny i chłopcy,
Sekcja Szkoła : dziewczyny i chłopcy,
Sekcja Liceum : dziewczyny i chłopcy



Przeszkole:
dziewczyny+Chłopcy
150 + 100 = 250

Liceum:
dziewczyny+Chłopcy
200+150 = 350

Szkoła:
dziewczyny+Chłopcy
250 + 150 = 400

1.3 Wartość Średnia i Mediana

Ważnymi parametrami danych statystycznych są wartość średnia i mediana.
Średnia Arytmetyczna. Wartością średnią arytmetyczną danych n liczb

$$a_1, a_2, ..., a_n$$

nazywamy liczbę

$$\mu = \frac{a_1 + a_2 + \cdots + a_n}{n} \tag{1.1}$$

Średnia Arytmetyczna Ważona. Bardziej ogólnym pojęciem średniej jest pojęcie średniej arytmetycznej ważonej. Mianowicie, niech wagami bedą liczby dodatnie $\rho_1, \rho_2, \cdots, \rho_n$ takie, że suma

$$\rho_1 + \rho_2 + \cdots + \rho_n = 1, \quad \rho_i > 0, \quad i = 1, 2, ..., n.$$

Wtedy średnią ważoną danych

$$a_1, a_2, ..., a_n$$

nazywamy następującą sumę iloczynów

$$\mu_\rho = \rho_1 a_1 + \rho_2 a_2 + \cdots + \rho_n a_n$$

W przypadku szczególnym, gdy wagi są równe

$$\rho_1 = \rho_2 = \cdots = \rho_n = \frac{1}{n}$$

wtedy średnia arytmytyczna ważona jest poprostu średnią arytmetyczną.

Mediana. Dla danych statystycznych znajdujemy ich mediane to znaczy wartość, która leży w środkowej pozycji danych. Mianowicie, w pierwszej kolejności sortujemy dane porządkując je od najmiejszej do największej lub od największej do najmieszej. Wtedy liczba, która leży na pozycji w równej odległości od początku i od końca uporządkowanych danych nazywa się medianą. Może zdażyć się że nie ma takiej jednej liczby, natomiast są dwie liczby obok siebie, które leżą w tej samaej odległości pierwsza od początku a druga od końca. Wtedy medianą jest ich średnia arytmetyczna.
Niżej, wyjaśniamy to na przykładach.

Przykład 1.2 *Rozpatrzmy następujące dane:*

$$(i) \quad 2, 1, 6, 8, 3, 2, 10, 12, 11$$

$$(ii) \quad 9, 4, 2, 7, 5, 1, 3, 10, 15, 17, 16$$

Rozwiązanie (i). Dane $2, 1, 5, 8, 3, 2, 10, 12, 11$ porządkujemy w kierunku rosnącym od najmieszej do największej

$$1, 2, 2, 3, 6, 8, 10, 11, 12$$

Zauważamy, że liczba 6 jest odległa od początku o cztery pozycje i od końca również o cztery pozycje. Zatem liczba 6 jest medianą danych (i).
Rozwiązanie (ii). Dane $0, -1, 9, 4, 2, 7, 5, 1, 3, 10, 15, 17, 16$ porządkujemy w kierunku rosnącym od najmieszej do największej

$$-1, 0, 1, 2, 3, 4, 5, 7, 9, 15, 16, 17$$

Zauważamy, że liczba 4 jest odległa od początku o pięć pozycji, a liczba 5 jest odległa od końca również o pięć pozycji. Zatem mamy dwie liczby w środku danych 4 i 5. Wtedy medianą jest ich średnia arytmetyczna, to znaczy

$$mediana \ = \ \frac{4+5}{2} = 4.5$$

1.3.1 Correlacja

Rozpatrzmy dwa ciągi danych

$$a = \{a_1, a_2, ..., a_n\}, \qquad b = \{b_1, b_2, ..., b_n\},$$

o tej samej liczbie elementów n.

Definition 1.1 *Correlację danych*

$$a = \{a_1, a_2, ..., a_n\}, \qquad b = \{b_1, b_2, ..., b_n\},$$

określamy następującym wzorem:

$$Cor(a,b) = \frac{a_1\, b_1 + a_2\, b_2 + \cdots + a_n\, b_n}{\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}\sqrt{b_1^2 + b_2^2 + \cdots + b_n^2}} = \frac{(a,b)}{|a| * |b|}, \tag{1.2}$$

gdzie iloczyn skalarny

$$(a,b) = a_1\, b_1 + a_2\, b_2 + \cdots + a_n\, b_n$$

oraz długość danych $a,\ b$

$$|a| = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}, \qquad |b| = \sqrt{b_1^2 + b_2^2 + \cdots + b_n^2}$$

Dane zapisujemy również w ich unormowanej formie. Mianowicie, niech

$$\begin{aligned} \hat{a} &= \{\hat{a}_1, \hat{a}_2, ..., \hat{a}_n\} = \frac{\{a_1, a_2, ..., a_n\}}{\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}}, \\ \hat{b} &= \{\hat{b}_1, \hat{b}_2, ..., \hat{b}_n\} = \frac{\{b_1, b_2, ..., b_n\}}{\sqrt{b_1^2 + b_2^2 + \cdots + b_n^2}}, \end{aligned} \tag{1.3}$$

gdzie

$$\begin{aligned} \hat{a}_1 &= \frac{a_1}{\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}}, \quad \hat{b}_1 = \frac{b_1}{\sqrt{b_1^2 + b_2^2 + \cdots + b_n^2}}, \\ \hat{a}_2 &= \frac{a_1}{\sqrt{a_2^2 + a_2^2 + \cdots + a_n^2}}, \quad \hat{b}_2 = \frac{b_2}{\sqrt{b_1^2 + b_2^2 + \cdots + b_n^2}}, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \hat{a}_n &= \frac{a_n}{\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}}, \quad \hat{b}_n = \frac{b_n}{\sqrt{b_1^2 + b_2^2 + \cdots + b_n^2}} \end{aligned}$$

Zauważamy, że dane (1.3) w unormowanej formie spełniają następujące warunki:

$$\hat{a}_1^2 + \hat{a}_2^2 + \cdots + \hat{a}_n^2 = 1, ; \quad \hat{b}_1^2 + \hat{b}_2^2 + \cdots + \hat{b}_n^2 = 1$$

Wtedy correlacja pomiędzy danymi a i b oraz correlacja pomiędzy danymi unormowanymi $\hat{a}$ i $\hat{b}$ jest ta sama

$$\begin{aligned} &Cor(a,b) = Cor(\hat{a}, \hat{b}), \\ &Cor(\hat{a}, \hat{b}) = \hat{a}_1 * \hat{b}_1 + \hat{a}_2 * \hat{b}_2 + \cdots + \hat{a}_n * \hat{b}_n \end{aligned}$$

Przykład 1.3 *Oblicz correlację pomiędzy danymi*

$$a = \{2, 1, 5, 8\}, \qquad b = \{4, 3, 9, 3\}$$

Rozwiązanie. Podstawiając do wzoru dane (1.2)

$$a_1 = 2, \quad a_2 = 1, \quad a_3 = 5, \quad a_4 = 8,$$

$$b_1 = 4, \quad b_2 = 3, \quad b_3 = 9, \quad b_4 = 3$$

obliczamy współczynnik correlacji dla $n = 4$

$$\begin{aligned} Cor(a,b) &= \frac{a_1\, b_1 + a_2\, b_2 + \cdots + a_n\, b_n}{\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}\sqrt{b_1^2 + b_2^2 + \cdots + b_n^2}}, \\ &= \frac{2*4 + 1*3 + 5*9 + 8*3}{\sqrt{2^2 + 1^2 + 5^2 + 8^2}\sqrt{4^2 + 3^2 + 9^2 + 3^2}} = 0.769444, \end{aligned}$$

Przykład 1.4 *W klasie czwartej zmierzono i zważono 5 dziewczynek i 5 chłopców. Otrzymane wyniki pomiarów zapiasano w tabeli*

wzrost dziewczynek cm	*waga dziewczynek kg*	*wzrost chłopców cm*	*waga chłopców kg*
140	*35*	*142*	*40*
135	*30*	*145*	*38*
132	*33*	*150*	*45*
140	*35*	*142*	*40*
125	*30*	*135*	*37*

(*i*) *Oblicz współczynnik corelacji pomiędzy wzrostem i wagą dla dzewczynek*
(*ii*) *Oblicz współczynnik corelacji pomiędzy wzrostem i wagą dla chłopców*
(*iii*) *Oblicz współczynnik corelacji pomiędzy wzrostem i wagą dla dzewczynek i chłopców razem.*

Rozwiązanie (*i*)
Współczynnik corelacji dla dziewczynek obliczamy podstawiając do wzoru

$$Cor(a,b) = \frac{(a,b)}{|a| * |b|},$$

dane dziewczynek

$$a = \{140, 135, 132, 140, 125\}, \qquad b = \{35, 30, 33, 35, 30\}$$

gdzie iloczyn skalarny

$$\begin{aligned} (a,b) &= a_1 * b_1 + a_2 * b_2 + a_3 * b_3 + a_4 * b_4 + a_5 * b_5 \\ &= 140 * 35 + 135 * 30 + 132 * 33 + 140 * 35 + 125 * 30 \\ &= 21956 \end{aligned}$$

oraz długośċ danych $a,\ b$

$$\begin{aligned} |a| &= \sqrt{140^2+135^2+132^2+140^2+125^2} = \sqrt{90474} = 300.956 \\ |b| &= \sqrt{35^2+30^2+33^2+35^2+30^2} = \sqrt{5339} = 73.0685 \end{aligned}$$

Skąd obliczamy współczynnik corelacji pomiędzy wzrostem i wagą dla dziewczynek.

$$Cor(a,b) = \frac{(a,b)}{|a| * |b|} = \frac{21956}{\sqrt{90474} * \sqrt{5339}} = 0.998991$$

Zadanie 1.1 *Oblicz współczynnik korelacji pomiędzy wzrostem i wagą dla chłopców dla danych z powyższej tabeli wzorując się na rozwiązaniu (i).*

1.4 Wariancja i Odchylenie Standardowe

Wariancja. Wariancja σ^2 danych statystycznych

$$a = \{a_1, a_2, ..., a_n\},$$

związana jest z ich średnią arytmetyczną

$$\mu = \frac{a_1 + a_2 + \cdots + a_n}{n} \tag{1.4}$$

Mianowicie wariancje danych

$$a = \{a_1, a_2, ..., a_n\},$$

określamy następującym wzorem:

$$\sigma^2 = \frac{(a_1-\mu)^2 + (a_2-\mu)^2 + \cdots + (a_n-\mu)^2}{n} \tag{1.5}$$

[1] **Odchylenie Standardowe** σ jest pierwiastkiem kwadratowym z wariancji

$$\sigma = \sqrt{\sigma^2} \tag{1.6}$$

Przykład 1.5 *Oblicz wariancje i odchylenie standardowe następujących danych:*

$$(i)\quad a = \{3, -1, 8, 4\}, \quad (ii)\quad b = \{12, 4, 8, 6\}.$$

Rozwiązanie (i). Rozwiązanie jest prostym i bezpośrednim podstawieniem danych do wzorów. Najpierw obliczamy wartość średnią podstawiając do wzoru (1.4) dane (i) $n = 4$

$$\mu = \frac{a_1 + a_2 + \cdots + a_n}{n} = \frac{3 - 1 + 8 + 4}{4} = 3.5$$

[1] Litera grecka σ, czytamy sigma

następnie obliczamy wariancję podstawiając do wzoru (1.5) $\mu = 3.5$ i dane (i) $n = 4$

$$\begin{aligned}\sigma^2 &= \frac{(a_1-\mu)^2+(a_2-\mu)^2+\cdots+(a_n-\mu)^2}{n} \\ &= \frac{(3-3.5)^2+(-1-3.5)^2+(8-3.5)^2+(4-3.5)}{4} = 10.31\end{aligned}$$

oraz odchylenie standardowe

$$\sigma = \sqrt{\sigma^2} = \sqrt{10.31} = 3.21131$$

Rozwiązanie (ii). Podobnie jak rozwiązanie (i), rozwiązanie (ii) jest prostym i bezpośrednim podstawieniem danych do wzorów. Najpierw obliczamy wartość średnią

$$\mu = \frac{a_1+a_2+\cdots+a_n}{n} = \frac{12+4+8+6}{4} = \frac{30}{4} = 7.5$$

następnie obliczamy wariancję

$$\begin{aligned}\sigma^2 &= \frac{(a_1-\mu)^2+(a_2-\mu)^2+\cdots+(a_n-\mu)^2}{n} \\ &= \frac{(12-7.5)^2+(4-7.5)^2+(8-7.5)^2+(6-7.5)^2}{4} = 8.75\end{aligned}$$

oraz odchylenie standardowe

$$\sigma = \sqrt{\sigma^2} = \sqrt{10.31} = 2.95804$$

Reguła Trzech Sigm dla normalnego rozkładu $N(\mu, \sigma)$ określa przedziały

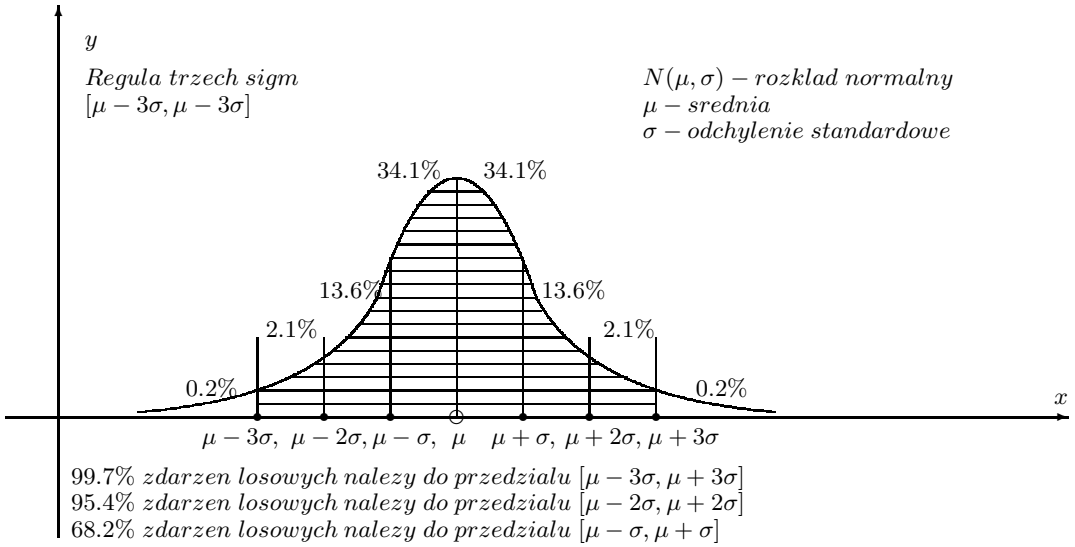
$$[\mu-3\sigma, \mu+3\sigma],$$

$$[\mu-2\sigma, \mu+2\sigma],$$

$$[\mu-\sigma, \mu+\sigma],$$

do których należy 99.7% wszystkich obserwowacji zadarzenia losowego. Wyniki obserwacji zdarzenia losowego poza przedziałem $[\mu-3\sigma, \mu+3\sigma]$ pojawiają się bardzo rzadko.

Na podanym wykresie zostały zaznaczone wszystkie obserwacje zdarzenia losowego w procentach.



Przykład 1.6 *Pracownia krawiecka planowała uszycie 1000 mundurków dla dziewczyn uczenic szkół podstawowych.*
W tym celu pracownia wykonała pomiary wzrostu 10 dziewczyn w wieku 7 lat. Wyniki pomiarów wzrostu w centymetrach zostały zapisane w postaci listy

$$dane = \{140, 131, 132, 138, 145, 135, 141, 135, 143, 130\}$$

(i) Oblicz średnią arytmetyczną i odchylenie standarodowe wzrostu dziewczyn.

(ii) Stosując regułę trzech sigm oblicz ile mundurków dla dziewczyn powinna uszyć pracownia krawiecka w każdym z niżej podanych przedziałów.

$$[\mu - 3\sigma, \mu + 3\sigma],$$

$$[\mu - 2\sigma, \mu + 2\sigma],$$

$$[\mu - \sigma, \mu + \sigma],$$

Rozwiązanie (i).
Wartośc średnią μ dla dziewczyn obliczamy podstawiając do wzoru (1.1) dane z tabeli dla $n = 10$

$$\mu = \frac{140 + 131 + 132 + 138 + 145 + 135 + 141 + 135 + 143 + 130}{10} = 137$$

Podobnie obliczamy odchylenie standardowe podstawiając do wzorów (1.5), (1.6) dane z tabeli

$$\sigma =$$

$$= \sqrt{\frac{(140-137)^2+(131-137)^2+(132-137)^2+(138-137)^2+(145-137)^2+(135-137)^2+(141-137)^2+(135-137)^2+(143-137)^2+(130-}{10}}$$

$$= \sqrt{\frac{(-7)^2+(-6)^2+(-5)^2+(1)^2+(8)^2+(-2)^2+(4)^2+(-2)^2+(6)^2+(-7)^2}{10}}$$

$$= 4.93964$$

Dla dużej ilości danych $n \geq 100$ obliczenia należy wykonać w aplikacji *Exel* lub w innych językach programowania jak *Pascal*, $C++$ lub *Mathematica*. Tutaj obliczenia wykonamliśmy w systemie *Mathematica* stosując proste instrukcje

$$Mean[dane];$$
$$Variance[dane];$$
$$StandardDeviation[dane]$$

dla danych w postaci listy

$$dane = \{140, 131, 132, 138, 145, 135, 141, 135, 143, 130\}$$

Rozwiązanie (ii).
Z reguły trzech sigm wiemy, że do przedziału $[\mu-\sigma, \mu+\sigma]$ należy $34.134.1\% = 68.2\%$ wartości zdarzeń losowych. Zatem pracownia krawiecka powinna uszyć

- w przedziale od $\mu - \sigma$ do $\mu + \sigma$

$$1000 * 68.2\% = \frac{1000 * 68.2}{100} = 682$$

mundurków

- w przedziałach od $2\mu - \sigma$ do $\mu - \sigma$ i od $\mu + \sigma$ do $\mu + 2\sigma$

$$1000 * 13.6\% + 1000 * 13.6\% = \frac{1000 * 13.6}{100} + \frac{1000 * 13.6}{100} = 136 + 136 = 272$$

mundurków.

- w przedziałach od $2\mu - \sigma$ do $\mu - \sigma$ i od $\mu + \sigma$ do $\mu + 2\sigma$

$$1000 * 2.1\% + 1000 * 2.1\% = \frac{1000 * 2.1}{100} + \frac{1000 * 2.1}{100} = 21 + 21 = 42$$

mundurków.

Razem pracownia krawiecka powinna uszyć

$$682 + 272 + 42 = 994$$

mundurki z przedziału od $\mu - 3\sigma$ do $\mu + 3\sigma$.

Pozostałe 6 mundurki pracownia krawiecka powinna uszyć z poza tego przedziału.

1.5 Zadania

Zadanie 1.2 *Oblicz średnią arytmetyczneną danych*

$$(i) \quad \{1,3,5,7,9\}, \qquad (ii) \quad \{2,4,6,8,10\}$$

Zadanie 1.3 *Oblicz średnią arytmetyczneną ważoną danych*

$$(i) \quad \{1,3,5,7,9,11\}, \qquad (ii) \quad \{2,4,6,8,10,12\}$$

dla wag

$$\rho_1 = \frac{8}{24}, \ \rho_2 = \frac{6}{24}, \quad \rho_3 = \frac{4}{24}, \quad \rho_4 = \frac{3}{24}, \quad \rho_5 = \frac{2}{24}, \quad \rho_6 = \frac{1}{24}$$

Zadanie 1.4 *Marysia i Tomek skończyli ośmio klasową szkołę podstawową z ocenami z języka polskiego i matematyki w klasach I-VIII zapiasane w następującej tabeli*

Klasa	*Marysia j. polski*	*Marysia matematyka*	*Tomek j. polski*	*Tomek matematyka*
I	*6*	*4*	*4*	*5*
II	*6*	*3*	*5*	*5*
III	*5*	*2*	*4*	*6*
IV	*6*	*3*	*4*	*5*
V	*5*	*3*	*4*	*5*
VI	*6*	*3*	*3*	*6*
VII	*6*	*3*	*4*	*5*
VIII	*6*	*4*	*3*	*6*

(*i*) *Oblicz współczynnik corelacji pomiędzy ocenęnami z języka polskiego i matematki dla Marysi*
(*ii*) *Oblicz współczynnik corelacji pomiędzy ocenami z języka polskiego i matematyki dla Tomka*
(*iii*) *Oblicz współczynnik corelacji pomiędzy oceną z języka polskiego dla Marysi i dla Tomka.*
(*iv*) *Oblicz współczynnik corelacji pomiędzy oceną z matematyki dla Marysi i dla Tomka.*

Zadanie 1.5 *Pracownia krawiecka planowała uszycie 1000 mundurkw dla chłopców uczniów szkół podstawowych.*
W tym celu pracownia wykonała pomiary wzrostu 10 chłopców w wieku 7 lat.
Wyniki pomiarów wzrostu w centymetrach zostały zapisane w postaci listy

$$dane = \{145, 151, 134, 138, 142, 149, 141, 135, 143, 132\}$$

(i) Oblicz średnią arytmetyczną i odchylenie standarodowe wzrostu chłopców

(ii) Stosując regułę trzech sigm oblicz ile mundurków dla chłopców powinna uszyć pracownia krawiecka w każdym z niżej podanych przedziałów wzrostu.

$$[\mu - 3\sigma, \mu + 3\sigma],$$

$$[\mu - 2\sigma, \mu + 2\sigma],$$

$$[\mu - \sigma, \mu + \sigma],$$