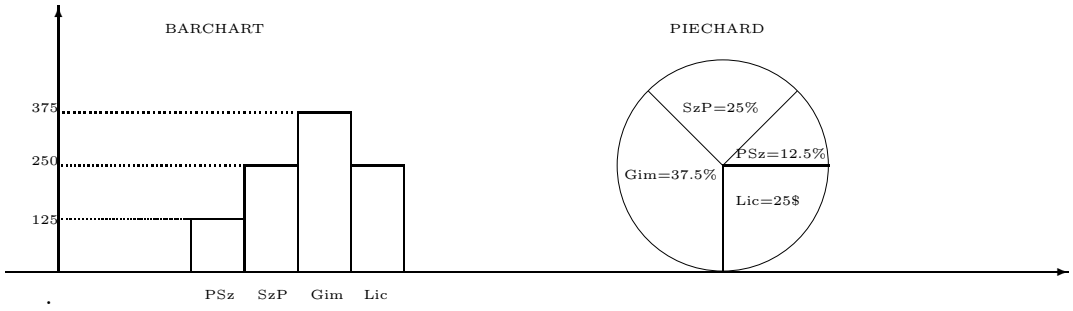
SZKOŁA PODSTAWOWA HELIANTUS
02-892 WARSZAWA
ul. BAŻANCIA 16

TEMAT 16.1: STATYSTYKA OPISOWA,

10 godzin lekcyjnych po 45 minut

Tadeusz STYŠ



Contents

Chapter 1

Statystyka opisowa

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 przemyśle 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.

1.1 Przykłady danych statystycznych i diagramów

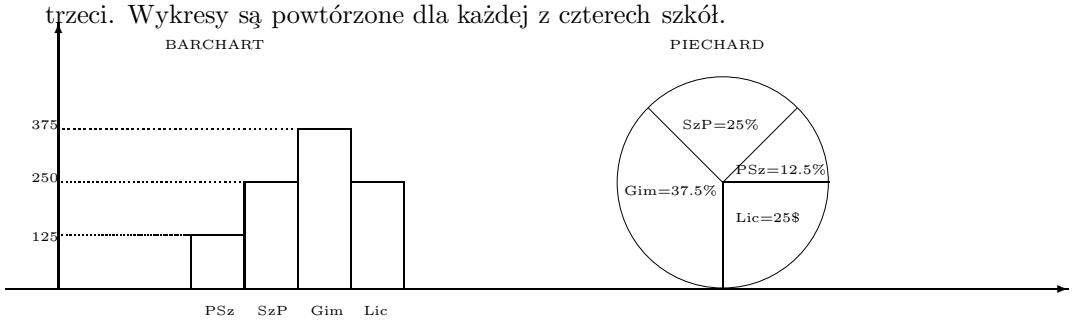
Dane statystyczne piszemy w tablicach z opisem ich znaczenia i wartości liczbowych.

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

Rodzaj Szkoły	Liczba uczniów	Część z całości	Procent
Przedszkole (PSz)	125	1/8 z 1000	12.5%
Szkoła Podstawowa (SzP)	250	1/4 z 1000	25%
Gimnazjum (Gim)	375	3/8 z 1000	37.5%
Liceum (Lic)	250	1/4 z 1000	25%
Razem	1000	$\frac{1}{8}+\frac{1}{4}+\frac{3}{8}+\frac{1}{4}=1$	100%

W niżej podanych diagramach w formie słupków i koła podane są wykresy dziewcząt, chłopców i razem uczniów w Przedszkolu (Psz), w Szkole Podstawowej (szP), w Gimnazjum (Gim), i w Liceuam (Lic).

Legenda: Dziewczęta słupek pierwszy, chłopcy słupek drugi i liczba uczniów raze słupek trzeci. Wykresy są powtórzone dla każdej z czterech szkół.



Legenda: Dziewczęta koło pierwsze, chłopcy koło drugie i liczba uczniów raze koło trzecie. Wykresy są powtórzone dla każdej z czterech szkół.

1.2 Wartość średnia i mediana

Ważnymi parametrami danych statystycznych są wartość średnia i mediana. **Wartość średnia srytmetyczna**. Wartością średnią arytmetyczną danych n liczb $a_1, a_2, ..., a_n$ nazywamy liczbę

$$\text{Srednia arytmetyczna } = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

Wartość 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 nazywamy następującą sumę iloczynów

$$\text{Srednia arytmetyczna wazona } = \rho_1 a_1 + \rho_2 a_2 + \cdots + \rho_n a_n$$

Istotnie, 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 środku danych. Mianowicie, w pierwszej kolejności sortujemy dane porządkując je od najmiejszej do największej lub od największej do najmieszej. Wtedy liczba, która leży w równej odległości od początku i od końca uporządkowanych danych nazywa się mmedianą. 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
o]wnym

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

Zauważamy, że liczba 6 jest odległa od począdku 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 począdku 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 $\frac{4+5}{2}=4.5$. Odpowiedź: medianą danych (ii) jest liczba 4.5

1.2.1 Korelacja danych statystycznych

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.

Defimnicja 1.1 *Korelacją danych statystycznych*

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

nazywamy następujący iloraz:

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

Dane statystyczne piszemy również w ich unormowanej formie. Mianowicie, niech

$$\begin{aligned}\hat{a}=\{\hat{a}_1,\hat{a}_2,...,\hat{a}_n\}&=\frac{\{a_1,a_2,...,a_n\}}{\sqrt{a_1^2+a_2^2+\cdots+a_n^2}},\\ \hat{b}=\{\hat{b}_1,\hat{b}_2,...,\hat{b}_n\}&=\frac{\{b_1,b_2,...,b_n\}}{\sqrt{b_1^2+b_2^2+\cdots+b_n^2}},\end{aligned}\tag{1.1}$$

gdzie

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

Zauważamy, że dane statystyczne (**??**) 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 korelacja pomiędzy danymi a i b oraz korelacja pomiędzy danymi unormowanymi $\hat{a}$ i $\hat{b}$ jest ta sama i określana jak następuje:

Defimnicja 1.2 *Korelacją danych statystycznych*

$$\hat{a} = \{\hat{a}_1, \hat{a}_2, ..., \hat{a}_n\}, \quad \hat{b} = \{\hat{b}_1, \hat{b}_2, ..., \hat{b}_n\}$$

nazywamy sumę następujących iloczynów:

$$Cor(a, b) = Cor(\hat{a}\ \hat{b}) = \hat{a}_1\ \hat{b}_1 + \hat{a}_2\ \hat{b}_2 + \cdots + \hat{a}_n\ \hat{b}_n,$$

Przykład 1.3 *Oblicz korelację pomiędzy danymi*

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

Rozwiązanie. Podstawiając do wzoru dane

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

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

obliczamy współczynnik korelacji

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

1.3 Wariancja i odchylenie standardowe

Wariancja σ^2 danych statystycznych

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

związana jest z ich średnią arytmetyczną

$$s = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

następującym wzorem:

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

Czytamy sigma.
Odchylenie standardowe σ jest pierwiastkiem kwadratowym z wariancji

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

Przykład 1.4 *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

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

następnie obliczamy wariancję

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

oraz odchylenie standardowe

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

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

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

następnie obliczamy wariancję

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

oraz chylenie standardowe

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