STATYSTYKA OPISOWA

Dr Alina Gleska

Instytut Matematyki WE PP

28 września 2018

1. Wprowadzenie
2. Pojęcia podstawowe
3. Szeregi rozdzielcze

Zwykle wyróżnia się dwa podstawowe działy statystyki:

- statystyka opisowa (jedno- i wielowymiarowa),
- statystyka matematyczna.

STATYSTYKA OPISOWA zajmuje się:

- zasadami programowania badań statystycznych,
- metodami obserwacji statystycznej - zbierania danych statystycznych,
- sposobami opracowania i prezentacji danych statystycznych,
- syntetycznym opisem zebranego materiału statystycznego za pomocą statystyk (miar opisowych).

Metody statystyki opisowej można podzielić na:

- metody statystyki jednowymiarowej służące do analizy zjawisk prostych,
- metody statystyki wielowymiarowej umożliwiające analizę zjawisk złożonych.

STATYSTYKA MATEMATYCZNA zajmuje się metodami wnioskowania o całej zbiorowości na podstawie zbadania pewnej jej części zwanej próbą, którą wybiera się w sposób losowy. Podstawę statystyki matematycznej stanowi rachunek prawdopodobieństwa, który pozwala ocenić dokładność i wiarygodność wniosków dotyczących całej badanej zbiorowości.

Dyscyplinę tę tworzą m. in. takie działy, jak:

- teoria estymacji,
- weryfikacja hipotez statystycznych,
- metoda reprezentacyjna,
- statystyczna analiza wielowymiarowa w przypadku podejścia stochastycznego,
- planowanie eksperymentu,
- statystyczna kontrola jakości.

ETAPY badania statystycznego:

- pozyskiwanie danych,
- opracowanie danych,
- przetwarzanie danych,
- interpretacja wyników.

ZBIOROWOŚĆ STATYSTYCZNA (tzw. populacja) - zbiór dowolnych elementów objętych badaniem statystycznym, które noszą nazwę JEDNOSTEK STATYSTYCZNYCH. Wyróżniamy dwa rodzaje zbiorowości statystycznej:

- skończone,
- nieskończone.

PRÓBA - część zbiorowości statystycznej, która może być wybrana w sposób losowy lub przez dobór celowy.
CECHA - właściwość występująca u wszystkich jednostek zbiorowości, która pozwala odróżniać jednostki zbiorowości. Zbiór cech charakteryzujących jednostki nazywamy PRZESTRZENIĄ BADAŃ STATYSTYCZNYCH.

Klasyfikacja cech:

- ilościowe (mierzalne) i jakościowe,
- ciągłe, quasi-ciągłe i dyskretne (skokowe),
- proste i złożone,
- stymulanty, destymulanty i nominanty.

Forma przedstawienia wartości obejmuje WSKAŹNIKI STRUKTURY, NATĘŻENIA i DYNAMIKI.

SKALE POMIAROWE:

- nominalna,
- porządkowa,
- przedziałowa (klasyczny przykład to skala Likerta),
- ilorazowa.

W wyniku badania 100 sklepów detalicznych w powiecie poznańskim w 2005 r uzyskano następujące dane:

Lp	Badana cecha	Warianty cechy
1	Rodzaj sklepu	0, 5, 2, …, 3
2	Liczba pracujących w sklepie	20, 2, 8, …, 5
3	Liczba ludności na 1 sklep	200, 100, 70, …, 148
4	Wykształcenie kierownika sklepu	3, 1, 0, …, 2
5	Powierzchnia sprzedażowa (m^2)	300, 20, 150, …, 100
6	Metoda sprzedaży	2, 1, 1, …, 2

gdzie:

- rodzaje sklepów: 0 – ogólnospożywczy, 1 – owocowo-warzywny, 2 – mięsny, 3 – rybny, 4 – piekarniczy, 5 – z napojami alkoholowymi;
- poziom wykształcenia: 0 – podstawowe, 1 – średnie ogólne, 2 – średnie handlowe, 3 – wyższe;
- metody sprzedaży: 1 – sklepy z obsługą, 2 – sklepy samoobsługowe.

Określ populację i jednostkę statystyczną, rodzaje cech oraz skale pomiarowe, w jakich dokonano pomiaru cech.

Populacja – sklepy detaliczne w powiecie poznańskim

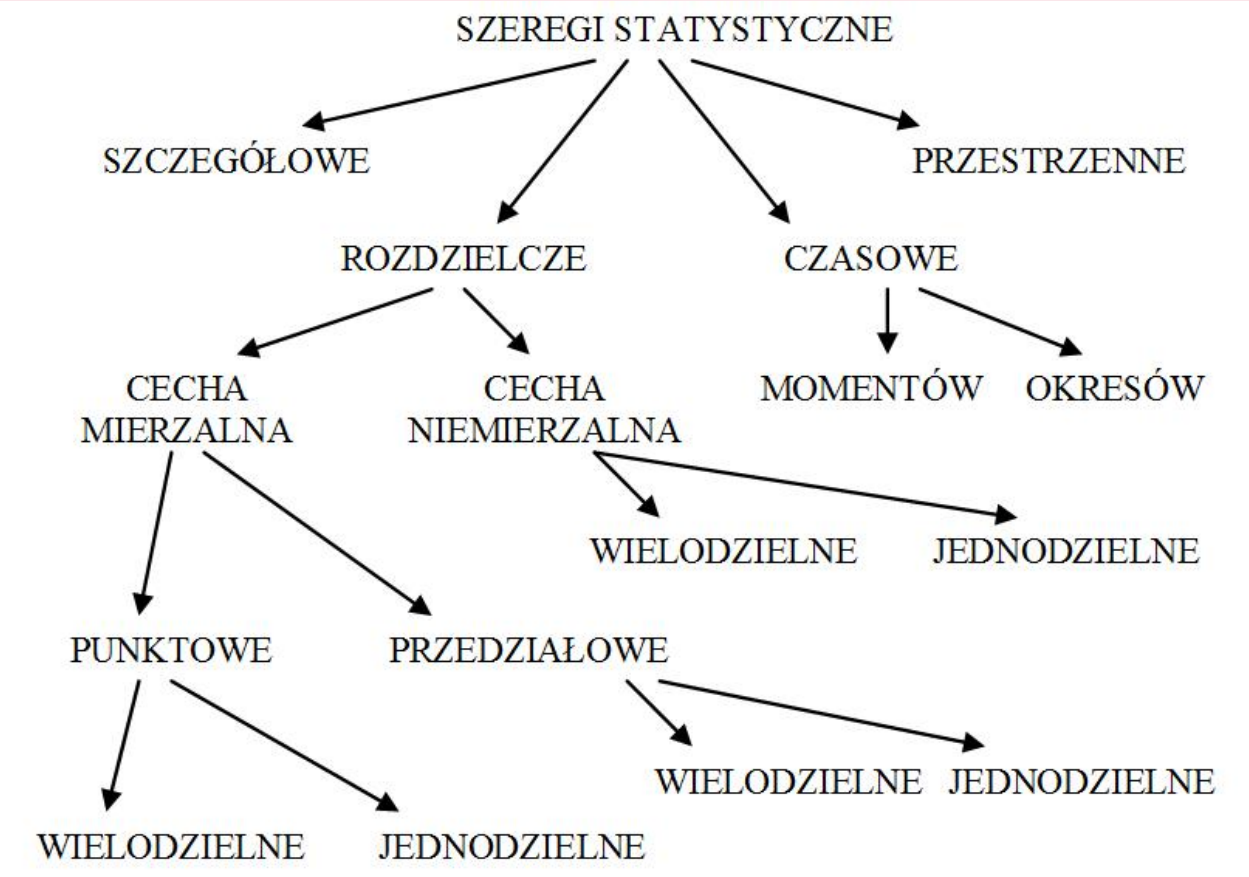
Jednostka statystyczna – sklep detaliczny

Cecha	Rodzaj	Skala
Rodzaj sklepu	jakościowa (politomiczna), zmienna, prosta	nominalna
Liczba pracujących w sklepie	ilościowa, zmienna, prosta	ilorazowa
Liczba ludności na 1 sklep	ilościowa, zmienna, prosta, wskaźnik struktury	ilorazowa
Wykształcenie kierownika sklepu	jakościowa (politomiczna), zmienna, prosta	porządkowa
Powierzchnia sprzedażowa (m^2)	ilościowa, zmienna, prosta	ilorazowa
Metoda sprzedaży	jakościowa (dychotomiczna), zmienna, prosta	nominalna

KLASYFIKACJA SZEREGÓW STATYSTYCZNYCH WG SOBCZYKA 1998

Szeregi statystyczne:

- szczegółowe (wyliczające)
- rozdzielcze (strukturalne)
  - cecha mierzalna
    - punktowe (jednodzielne i wielodzielne)
    - przedziałowe (jednodzielne i wielodzielne)
  - cecha niemierzalna
    - jednodzielne
    - wielodzielne
- czasowe (dynamiczne)
  - momentów
  - okresów
- przestrzenne (geograficzne)



Rysunek: Klasyfikacja szeregów

KONSTRUKCJA SZEREGU ROZDZIELCZEGO DLA CECHY CIĄGŁEJ:

1) wyszukujemy z n-elementowej zbiorowości wartości cechy $x_1, x_2, \ldots, x_n$ jednostki o wartości minimalnej x_{min} oraz maksymalnej x_{max},
2) określamy obszar zmienności szeregu poprzez rozstęp $R = x_{max} - x_{min}$,
3) ustalamy liczbę przedziałów klasowych k za pomocą jednego ze wzorów: $k \approx \sqrt{n}$, $k \approx \frac{3}{4}\sqrt{n}$, $k \leq 5\log(n)$, $k \approx 1 + 3,322\log(n)$ (H. A. Sturges 1926),
4) wyznaczamy długość przedziału klasowego d ze wzoru: $d \approx R/k$ (zaokrąglenie w górę),

5) przyjmujemy za dolną granicę pierwszego przedziału klasowego x_{min} lub arbitralnie ustalamy wartość tej granicy tak, aby jednostka o najmniejszej wartości była zakwalifikowana do pierwszego przedziału klasowego, a o największej wartości - do ostatniego przedziału klasowego,
6) konstruujemy k przedziałów klasowych o długości d, a następnie zliczamy jednostki o wartościach należących do odpowiednich przedziałów klasowych.

ZASADY BUDOWANIA SZEREGÓW ROZDZIELCZYCH:

- rzadko stosujemy mniej niż 6 i więcej niż 15 przedziałów klasowych,
- zawsze ustalamy takie klasy, które obejmują wszystkie dane,
- klasy powinny być rozłączne,
- przedziały klasowe powinny być niepuste, czyli powinny zawierać co najmniej jedną jednostkę,
- ustalamy przedziały klasowe w miarę możliwości o jednakowej długości, a w przypadku występowania obserwacji odstających stosujemy otwarty górny lub dolny przedział,
- jednostki zaliczone do tej samej klasy nie powinny być - ze względu na wartości badanej cechy - zbyt mocno zróżnicowane.

Przyjmujemy zasadę, że lewy koniec przedziału klasowego jest domknięty, a prawy otwarty. Umownie można zapisać to tak: 20-49,9; 50-79,9; 80-109,9 itd. Z kolei przedziały prawostronnie domknięte dla cechy ciągłej zapisujemy tak: 20,1-50; 50,1-80; 80,1-110 itd.

Symbol x_i^0 oznacza środek i-tego przedziału klasowego: $x_i^0 = (x_i + x_{i+1})/2$ dla $i = 1, 2, \ldots, k$.

Liczebność n_i oznacza liczbę jednostek zaliczonych do i-tego przedziału, a liczebność skumulowana n_i^{skum} określa sumy częściowe liczebności dla i-tego przedziału i jest wyrażona wzorem:

$$n_i^{skum} = \sum_{j=1}^{i} n_j$$

dla $i = 1, 2, \ldots, k$.

Przez ω_i określa się częstość wyrażoną ilorazem liczebności i-tego przedziału i liczebności zbiorowości $n = \sum_{i=1}^{k} n_i$, czyli $\omega_i = \frac{n_i}{n}$ i $\sum_{i=1}^{k} \omega_i = 1$. Częstość wyrażona w procentach określona jest wzorem $p_i = (\frac{n_i}{n})100\%$.

Częstość skumulowaną wyznacza się według wzoru $\omega_i^{skum} = \sum_{j=1}^{i} \omega_j$ (dla udziału) lub $p_i^{skum} = \sum_{j=1}^{i} p_j$ (dla udziału procentowego), gdzie $i = 1, 2, \ldots, k$. Wówczas otrzymujemy $\omega_k^{skum} = 1$ i $p_k^{skum} = 100\%$.

Gdy przedziały klasowe mają różne długości, wprowadza się pojęcie przeciętnej gęstości liczebności oraz przeciętnej gęstości częstości. Przeciętna gęstość liczebności to $g_i = n_i/d_i$, gdzie n_i oznacza liczebność i-tego przedziału, a $d_i = x_{i+1} - x_i$ jego długość ($i = 1, 2, \ldots, k$). Oznacza ona, ile jednostek zbiorowości danego przedziału klasowego przypada przeciętnie na jednostkę miary analizowanej cechy ilościowej. Z kolei przeciętna gęstość częstości $\tilde{g}_i$ jest określona ilorazem częstości ω_i i długości d_i i-tego przedziału klasowego: $\tilde{g}_i = \omega_i/d_i$ i określa, jaka częstość przypada na jednostkę miary cechy ilościowej.

BŁĘDY POPEŁNIANE

PRZY TWORZENIU SZEREGU ROZDZIELCZEGO

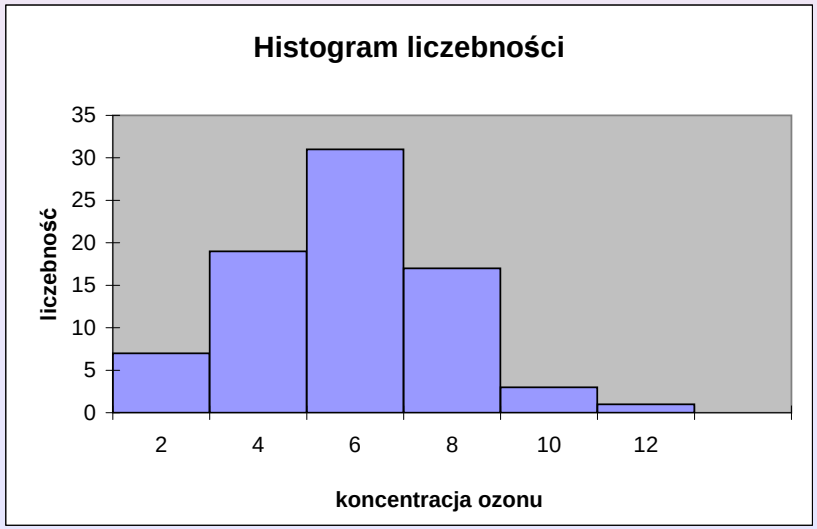
1) Jeżeli zbiorowość jest niejednorodna i występuje duża koncentracja jednostek w jednej klasie, szereg z równymi przedziałami nie daje zadowalających rezultatów (przykład).

2) Wnioskowanie porównawcze dotyczące zbiorowości o różnej liczności (przykład).

Pomiary koncentracji ozonu

3,5	1,7	3,1	4,5	3,0	6,1
6,8	1,1	5,8	4,2	6,0	8,1
2,4	7,5	4,7	5,4	1,4	2,0
6,8	5,8	5,7	6,5	2,8	6,2
5,5	3,4	6,0	7,4	2,5	5,6
6,2	3,1	4,4	5,5	3,7	4,0
5,7	4,4	4,7	5,8	3,3	3,4
9,4	6,6	4,7	1,6	9,4	11,7
6,8	5,4	5,6	1,4	5,3	6,6
6,6	5,6	6,0	5,9	3,5	4,1
1,4	5,3	5,8	3,7	4,1	6,6
2,5	5,1	7,6	4,4	6,7	3,7
3,0	6,2	3,8	4,7	3,9	7,6

Końce przedziałów klasowych		*liczebność*
0	2	7
2	4	19
4	6	31
6	8	17
8	10	3
10	12	1



PREZENTACJA GRAFICZNA SZEREGÓW ROZDZIELCZYCH

HISTOGRAMEM nazywamy wykres słupkowy oparty na układzie współrzędnych, składający się z przylegających do siebie prostokątów, których długości podstaw są proporcjonalne do rozpiętości przedziałów klasowych, a wysokości - do ich liczebności (częstości). W przypadku prezentacji graficznej szeregów rozdzielczych punktowych otrzymuje się wykres odcinków pionowych kreślonych od punktu będącego wartością cechy i o długości proporcjonalnej do liczebności (częstości) poszczególnych wartości cechy skokowej.

PRAWO HISTOGRAMU - suma pól powierzchni prostokątów tworzących histogram musi być równa:

- liczbie n, czyli liczbie obserwacji badanej zmiennej, gdy wartości zmiennej w rozkładzie waży się częstościami,
- 1,0 lub równoważnie 100, gdy wartości zmiennej w rozkładzie waży się częstościami lub procentowymi częstościami względnymi, odpowiednio.

WIELOBOK LICZEBNOŚCI (CZĘSTOŚCI) jest linią łamaną łączącą środki górnych krawędzi prostokątów. KRZYWĄ LICZEBNOŚCI (CZĘSTOŚCI) tworzy połączenie dużej liczby punktów o współrzędnych: środki przedziałów klasowych i odpowiadające im liczebności (częstości) otrzymanych poprzez zmniejszanie rozpiętości przedziałów klasowych, a tym samym zwiększanie liczby przedziałów.

Rezultaty obserwacji statystycznej, która polega na ustaleniu wartości cech ilościowych lub wariantów cech jakościowych u wszystkich jednostek analizowanej zbiorowości można też przedstawić w postaci TABLICY STATYSTYCZNEJ. Tablice statystyczne składają się z jednego lub kilku szeregów statystycznych. Gdy tablica zawiera jeden szereg, określa się ją jako jednodzielną. Tablice zawierające wiele szeregów statystycznych zalicza się do wielodzielnych. Tablica statystyczna może zawierać kombinację szeregów rozdzielczych dla cechy ilościowej i jakościowej, szeregów czasowych lub przestrzennych.

Macierz danych (dla cech ilościowych)

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix} \begin{matrix} \leftarrow 1 \text{ jednostka statystyczna} \\ \leftarrow 2 \text{ jednostka statystyczna} \\ \\ \leftarrow n-\text{ta jednostka statystyczna} \end{matrix}$$

$\uparrow$ Cecha 1 $\uparrow$ Cecha 2

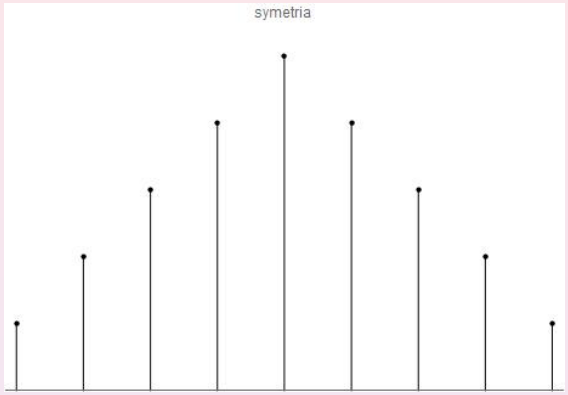
x_{ij} - wartość j-tej cechy ilościowej dla i-tej jednostki statystycznej

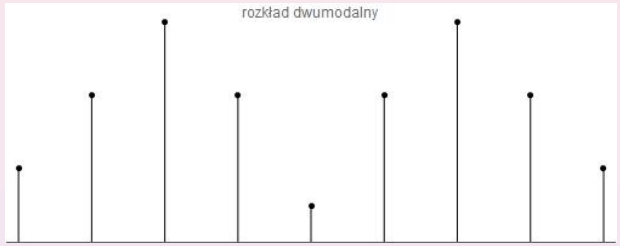
Macierz indykatywna (dla cech jakościowych):

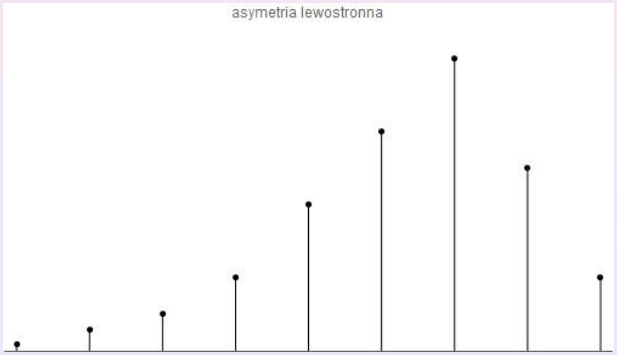
Cechy jakościowe		1		2			…	m	
Warianty cechy		1	2	3	4	5	…	$p-1$	p
Jednostki statystyczne	1	0	1	0	1	0	…	0	1
	2	1	0	1	0	0	…	1	0
	…	…	…	…	…	…	…	…	…
	n	1	0	0	0	1	…	0	1

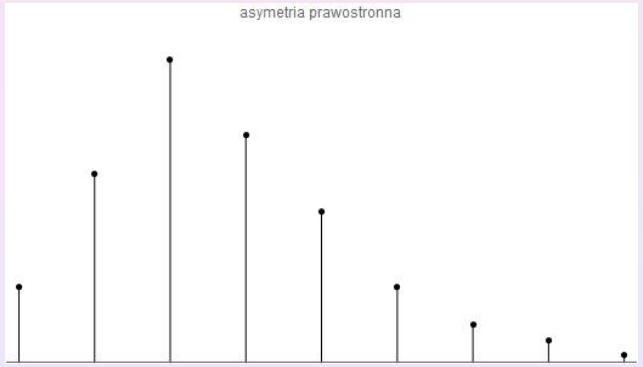
m – liczba cech
n – liczba jednostek statystycznych

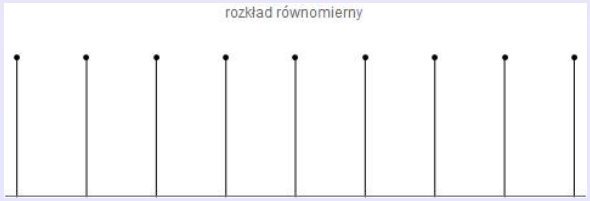
Rozkłady

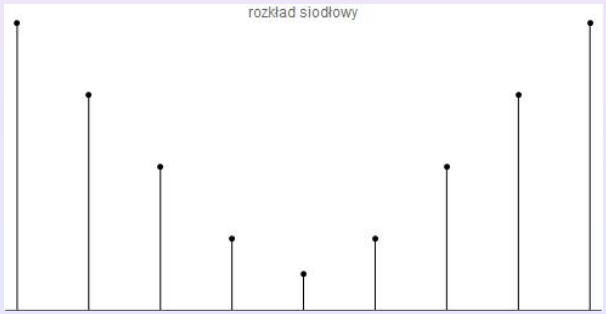

symetria


rozkład dwumodalny


asymetria lewostronna


asymetria prawostronna


rozkład równomierny


rozkład siodłowy