Adam Wosatko
Magdalena Jakubek
Piotr Pluciński

Matematyka stosowana i metody numeryczne

Konspekt z wykładu

14 Podstawy statystyki

14.1 Wstęp

Statystyka – nauka o metodach badań właściwości populacji (zbiorowości), które można wyrazić za pomocą liczb lub w sposób opisowy.

Głównym zadaniem statystyki jest wykrycie prawidłowości występujących w zjawiskach powiązanych z daną populacją.

Zakres zastosowań:

- sposób gromadzenia, prezentacji, analizy, przetwarzania i interpretacji danych,
- zrozumienie i klasyfikacja zachodzących procesów (zjawisk),
- przewidywanie przyszłych procesów (zjawisk) i podejmowanie decyzji.

Klasyfikacja danych:

- **Dane pierwotne** wcześniej nie istniały i są zbierane specjalnie dla potrzeb prowadzonego badania. Zaletą jest ich aktualność i dopasowanie do celu analizy, a wadę stanowią koszty (również rozumiane jako czas) pozyskania.
- **Dane wtórne** istniały wcześniej i są przetworzone na potrzeby nowego, innego badania. Zaletą jest niski koszt ich pozyskania, a wadą może być ich niedopasowanie oraz anachroniczność.
- **Dane wewnętrzne** pochodzą z organizacji prowadzącej badanie, również na zlecenie.
- **Dane zewnętrzne** uzyskuje się ze źródeł zewnętrznych, np. krajowych (publikowanych przez GUS) lub zagranicznych.

Obserwacja statystyczna jest to gromadzenie informacji o właściwościach poszczególnych jednostek zbiorowości, czyli o cechach zmiennych.
Wyniki obserwacji statystycznej można przedstawić w formie:

- szeregów statystycznych (szereg szczegółowy – uporządkowany),
- tablic statystycznych (prostych, złożonych),
- wykresów statystycznych (punktowych, obrazkowych, powierzchniowych – histogramów, liniowych, mapowych złożonych).

Badanie statystyczne jest to zespół czynności zmierzających do uzyskania za pomocą metod statystycznych informacji charakteryzujących badaną zbiorowość statystyczną.

Badanie statystyczne może być:

- pełne – obejmuje całą populację (zbiorowość),
- częściowe – dotyczy wybranej części populacji, nazywanej także próbą statystyczą, np. wybierani są tylko studenci.

Zbiorowość statystyczna (populacja) jest to zbiór **jednostek statystycznych**, które nie są identyczne, ale stanowią jedną logiczną całość. Zbiorowość może składać się z osób, rzeczy lub zdarzeń. Zbiorowość statystyczna musi być precyzyjnie określana pod względem rzeczowym, przestrzennym i czasowym.

Cechy statystyczne (zmienne) są to właściwości, którymi poszczególne jednostki statystyczne różnią się między sobą, przyjmując odmienne **warianty cechy**.

Cecha statystyczna stanowi przedmiot badania statystycznego. Wyróżnia się cechy:

- mierzalne – wyrażone za pomocą liczb:
  - skokowe – skończone, przeliczalne liczby wartości w pewnym przedziale, np. liczby pracowników małych przedsiębiorstw,
  - ciągłe – nieprzeliczalnie wiele wartości, np. objętości zużytej wody w mieszkaniach osiedla,
  - quasi-ciągłe (prawie ciągłe) – cechy skokowe, ale ze względu na duże liczby wartości potraktowane są jak ciągłe, np. ceny tony stali na przestrzeni pięćdziesięciolecia,
- niemierzalne (jakościowe, opisowe) – wyrażone w sposób opisowy (warianty, kategorie), np. wykształcenie mieszkańców kraju (cecha wielodzielna), płeć (cecha dwudzielna).

Rozkład cechy statystycznej stanowi uporządkowany zbiór wartości cechy z przyporządkowanymi im liczebnościami czyli liczbą jednostek przyjmujących daną wartość cechy. Wyróżnia się:

- **szereg punktowy** dla cechy mierzalnej skokowej,
- **szereg przedziałowy** dla cechy mierzalnej ciągłej.

Rozkład cechy oznacza przyporządkowanie liczby obserwacji odpowiednim wartościom cechy zmiennej. Opis struktury zbiorowości statystycznej określony jest przez parametry (dla populacji) lub statystyki (dla próby).

Parametry określające charakterystyki liczbowe:

X	– badana cecha zmienna,
x_i	– wartość cechy i-tej jednostki w szeregu punktowym,
$\hat{x}_i$	– środki i-tych przedziałów w szeregu przedziałowym,
k	– liczba przedziałów,
N	– liczebność populacji,
n_i	– liczebność i-tego przedziału: $\sum_i n_i = N$,
$f_i = \frac{n_i}{N}$	– częstość (frakcje): $\sum_i f_i = 1$,
$W_i = \frac{n_i}{N} \cdot 100\%$	– wskaźnik struktury,
$N(x_i) = \sum_{j:x_j \leqslant x_i} n_j$	– liczebności skumulowane,
$F(x_i) = \sum_{j:x_j \leqslant x_i} f_j$	– częstości skumulowane – dystrybuanta.

14.2 Miary opisu populacji

Moment zwykły rzędu r dla szeregu punktowego i przedziałowego:

$$m_r = \frac{1}{N}\sum_{i=1}^{k}(x_i)^r n_i \qquad m_r = \frac{1}{N}\sum_{i=1}^{k}(\hat{x}_i)^r n_i$$

Moment centralny rzędu r szeregu punktowego i przedziałowego:

$$\mu_r = \frac{1}{N}\sum_{i=1}^{k}(x_i - \mu)^r n_i \qquad \mu_r = \frac{1}{N}\sum_{i=1}^{k}(\hat{x}_i - \mu)^r n_i$$

Średnia arytmetyczna jest to suma wartości cechy mierzalnej jednostek populacji podzielona przez liczbę tych jednostek:

- dla szeregu punktowego:
$$\mu = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$
- dla szeregu przedziałowego:
$$\mu = \bar{x} = \frac{1}{N}\sum_{i=1}^{k} \hat{x}_i\, n_i$$

Średnia arytmetyczna nazywana jest również wartością przeciętną, wartością średnią, wartością oczekiwaną, nadzieją matematyczną. Inne oznaczenia: $E(X), m, m_1$.

Średnia arytmetyczna stanowi moment zwykły rzędu 1 (m_1).

Średnia harmoniczna to odwrotność średniej arytmetycznej, obliczonej z odwrotności wartości cechy:

- dla szeregu punktowego:
$$\mu_H = \bar{x}_H = N\frac{1}{\sum\limits_{i=1}^{N}\frac{1}{x_i}}$$
- dla szeregu przedziałowego:
$$\mu_H = \bar{x}_H = N\frac{1}{\sum\limits_{i=1}^{N}\frac{n_i}{\hat{x}_i}}$$

Wariancja (dyspersja) jest to średnia arytmetyczna kwadratów odchyleń wartości jednostek cechy od wartości średniej dla populacji:

- dla szeregu punktowego:
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$
- dla szeregu przedziałowego:
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}(\hat{x}_i - \mu)^2 n_i$$

Wariancja stanowi moment drugi centralny.

Odchylenie standardowe informuje o przeciętnym odchyleniu wartości cechy od średniej arytmetycznej i liczy się go jako pierwiastek kwadratowy z wariancji:
$$\sigma = \sqrt{\sigma^2}$$

Asymetria informuje o tym czy i jakie wartości cechy przeważają w rozkładzie populacji:

- dla szeregu punktowego:
$$\mu_3 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3, \qquad \gamma_3 = \frac{\mu_3}{\sigma^3}$$
- dla szeregu przedziałowego:
$$\mu_3 = \frac{1}{N}\sum_{i=1}^{k}(\hat{x}_i - \mu)^3 n_i, \qquad \gamma_3 = \frac{\mu_3}{\sigma^3}$$

- $\mu_3 < 0$ lub $-2 \leq \gamma_3 < 0$ – rozkład jest lewostronnie asymetryczny,
- $\mu_3 = 0$ lub $\gamma_3 = 0$ – rozkład jest symetryczny,
- $\mu_3 > 0$ lub $0 < \gamma_3 \leq 2$ – rozkład jest prawostronnie asymetryczny.

Asymetria stanowi moment trzeci centralny.

Miara skupienia i odpowiadający jej współczynnik skupienia czyli **kurtoza**:

- dla szeregu punktowego:

$$\mu_4 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4, \qquad \gamma_4 = \frac{\mu_4}{(\sigma^2)^2}$$

- dla szeregu przedziałowego:

$$\mu_4 = \frac{1}{N}\sum_{i=1}^{k}(\hat{x}_i - \mu)^4 n_i, \qquad \gamma_4 = \frac{\mu_4}{(\sigma^2)^2}$$

Miara skupienia stanowi moment czwarty centralny.

Kurtoza zwana także koncentracją oznacza stopień skupienia się rozkładu wartości cechy wokół średniej arytmetycznej.

- $\gamma_4 < 3$ – rozkład jest spłaszczony,
- $\gamma_4 = 3$ – rozkład jest normalny,
- $\gamma_4 > 3$ – rozkład jest wysmukły.

Dominanta $D(x)$ zwana także **modą** albo **wartością modalną** jest to wartość, która najczęściej występuje (jest najbardziej liczna) w populacji. W szeregu punktowym wyznaczenie dominanty sprowadza się do jej odpowiedniego wskazania. W szeregu przedziałowym wskazujemy przedział dominanty $D(x)$ i wyznaczamy przybliżoną wartość na podstawie interpolacji liniowej.

Kwantyle to cała grupa miar, które są pozycyjnymi miarami położenia. Dzielą one uporządkowaną (według rosnących wartości cechy) populację na liczebnie równe części, które nazywa się grupami kwantylowymi. Wśród kwantyli wyróżnia się:

- mediana – dzieli populację na 2 części,
- kwartyle – dzielą populację na 4 części,
- kwintyle – dzielą populację na 5 części,
- decyle – dzielą populację na 10 części,
- centyle (percentyle) – dzielą populację na 100 części.

Najczęściej wykorzystywany kwantyl to **mediana** $M(x)$, która dzieli populację na dwie równe części, a więc jest drugim kwartylem populacji. Mediana jest wyznaczana następująco:

- dla szeregu punktowego:
  - populacja o liczebności nieparzystej: $M(x) = x_{\frac{N+1}{2}}$,
  - populacja o liczebności parzystej: $M(x) = (x_{\frac{N}{2}} + x_{\frac{N+2}{2}})/2$,
- dla szeregu przedziałowego:

$$M(x) = x_{pm} + (\frac{N}{2} + \sum_{i=1}^{m-1} n_i)\frac{h_m}{n_m}$$

gdzie: m – numer przedziału mediany, x_{pm} – dolna granica przedziału mediany, $N/2$ – pozycja mediany, $\sum_{i=1}^{m-1} n_i$ – liczebność wszystkich przedziałów poprzedzających przedział mediany, h_m – rozpiętość przedziału mediany, n_m – liczebność przedziału mediany.

14.3 Elementy probabilistyki

Wnioskowanie statystyczne to uogólnienie wyników uzyskanych z próby losowej na całą populację generalną przy założeniu, że dobór próby podlega pewnym regułom obiektywnym. Sprowadza się ono do analizy zdarzeń losowych.

Zdarzenie losowe to wynik doświadczenia, które może być wielokrotnie powtarzane i nie da się przewidzieć jego wyniku, a możliwość zajścia tego zdarzenia określone jest przez jego **prawdopodobieństwo**.

Przykład: Jednokrotny rzut kostką.

Zdarzenie:

- liczba oczek wynosi 3 – może być zrealizowane tylko na 1 sposób → **zdarzenie elementarne**,
- liczba oczek > 4 – może być zrealizowane tylko na 2 sposoby,
- liczba oczek < 1 – nie może być zrealizowane → **zdarzenie niemożliwe**,
- liczba oczek jest parzysta lub nieparzysta – zdarzenie jest zawsze zrealizowane → **zdarzenie pewne**.

Zbiór zdarzeń elementarnych to taki zbiór zdarzeń, które się wzajemnie wykluczają oraz wyczerpują wszystkie możliwości czyli w każdym możliwym przypadku przynajmniej jedno z nich musi zachodzić). Zbiór ten stanowi **przestrzeń zdarzeń elementarnych** Ω. Zdarzenie losowe stanowi dowolny podzbiór zdarzeń elementarnych.

Dla jednokrotnego rzutu kostką zbiór ten wynosi 6 zdarzeń elementarnych, a dla dwukrotnego rzutu kostką (2 rzuty po kolei!) – 36 zdarzeń elementarnych.

Jeżeli przestrzeń zdarzeń elementarnych Ω jest skończona i składa się z n elementów, natomiast zdarzenie A składa się z m zdarzeń elementarnych (jednakowo możliwych) to **prawdopodobieństwo zdarzenia** A:

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

Zmienna losowa X – odpowiednik pojęcia cechy statystycznej i jednoznaczna funkcja, która przyporządkowuje wartości liczbowe x zdarzeniom elementarnym (wynikom doświadczenia losowego) ω, a więc przekształca przestrzeń zdarzeń elementarnych Ω w przestrzeń liczb rzeczywistych $\mathbb{R}$.

Dzięki wprowadzeniu pojęcia zmiennej losowej X, prawdopodobieństwo zdarzenia $P(A)$ można zastąpić prawdopodobieństwem $P(x)$ przyjęcia przez zmienną losową określonej wartości x ze zbioru X. Zmienna losowa i rozkład jej prawdopodobieństwa pozwalają zapisać pewne procesy (zjawiska) w postaci modeli matematycznych.

Zmienną losową:

- określamy dla argumentu w postaci zdarzenia elementarnego ω,
- zapisujemy bez argumentu jako X, zamiast $X(\omega)$,
- oznaczamy tylko dużą literą czyli $X, Y, \ldots$.

Zmienna losowa może być typu:

- **skokowego** (dyskretnego) – przyjmuje tylko przeliczalny zbiór wartości (czyli zmienia się skokowo): x_i, $P(x_i)$,
- **ciągłego** – przyjmuje dowolne wartości z przedziału zmienności (czyli zmienia się w sposób ciągły):

$$x, P(x_0 < X < x_0 + \Delta x), \quad P(x_i) = \frac{n(x_i)}{N} \quad \text{dla} \quad N \longrightarrow \infty$$

gdzie: $n(x_i)$ – liczba zdarzeń, którym przypisana jest zmienna x_i, N – liczba wszystkich zdarzeń.

14.4 Rozkłady zmiennej losowej

Funkcja rozkładu prawdopodobieństwa (inne nazwy: rozkład prawdopodobieństwa, funkcja prawdopodobieństwa) odnosi się tylko do przypadku zmiennych dyskretnych.

Jeżeli $x_1, x_2, \cdots, x_k, \cdots$ jest skończonym lub przeliczalnym zbiorem wartości zmiennej losowej X to funkcja

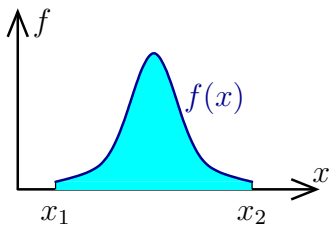
$$P(X = x_i) = P(\{\omega : X(\omega) = x_i\}) = p(x_i) = p_i > 0$$

przyporządkowuje temu zbiorowi odpowiednie prawdopodobieństwa $p_1, p_2, \cdots, p_k, \cdots$, gdzie:

$$\sum_i p_i = 1 \quad \text{dla} \;\; i \in N$$

Funkcja gęstości prawdopodobieństwa – prawdopodobieństwo, że zmienna losowa trafi do przedziału x_1, x_2 jest równe polu pod krzywą gęstości między punktami x_1, x_2:

$$\int_{x_1}^{x_2} f(x)dx \equiv P(x_1 \leq X \leq x_2)$$



Właściwości funkcji gęstości prawdopodobieństwa:

- $f(x) \geq 0$,
- $f(x)$ jest unormowana tj. $\int_{-\infty}^{+\infty} f(x)dx = 1$.

Dystrybuanta F(x) stanowi prawdopodobieństwo tego, że zmienna losowa X przyjmie wartość mniejszą od x.

Dystrybuanta zmiennej losowej typu skokowego:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} p_i$$

Dystrybuanta zmiennej losowej typu ciągłego:

$$F(x) \equiv P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

- $0 \leq F(x) \leq 1$ i $F(x)$ jest niemalejącą funkcją,
- $P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1)$, $x_1, x_2 \in \mathbb{R}$,
- $f(x) = \dfrac{\mathrm{d}F(x)}{\mathrm{d}x} \geq 1$,
- $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$ i $F(+\infty) = \lim_{x \to +\infty} F(x) = 1$,
- $\displaystyle\int_{-\infty}^{+\infty} f(t)dt = F(+\infty) - F(-\infty) = 1.$

14.4.1 Rozkład dwumianowy (Bernoulliego)

Zmienna losowa skokowa X przyjmuje wartości, które są liczbami całkowitymi nieujemnymi $k \in \{0, 1, 2, \ldots, n-1, n\}$.

Funkcja rozkładu prawdopodobieństwa jest określona następująco:

$$P_n(k) = \binom{n}{k} p^k \, q^{n-k}, \quad k = 0, 1, \cdots, n,$$

gdzie: $p \in (0, 1)$ jest ustalone, $q = 1 - p$, $n \in \mathbb{N}$. Otrzymujemy prawdopodobieństwo uzyskania k sukcesów w n próbach.

Dystrybuanta:

$$F(x) = \sum_{k<x} P_n(k) = \sum_{k<x}^{n} \binom{n}{k} p^k q^{n-k}$$

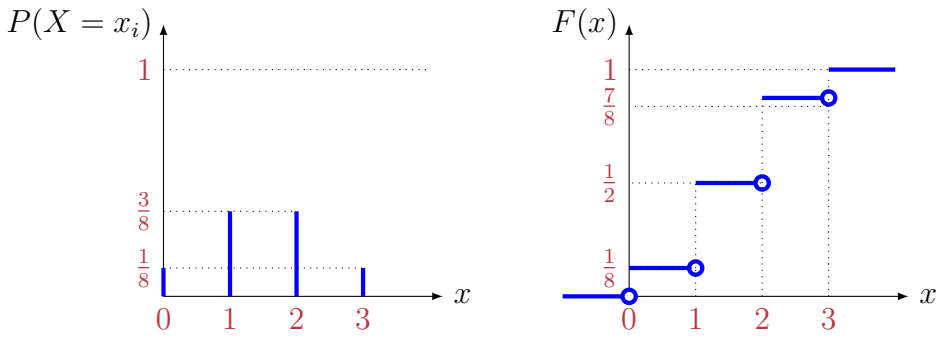
Przykład:

Doświadczenie polega na **trzykrotnym** rzucie monetą. Zmienną losową X jest liczba wyrzuconych orłów i może przyjmować wartości $k = 0, 1, 2, 3$.

Zbiór zdarzeń elementarnych	Liczba orłów k	Prawdopodobieństwo $P(X = k)$
(R,R,R)	0	1/8
(R,R,O), (R,O,R), (O,R,R)	1	3/8
(R,O,O), (O,R,O), (O,O,R)	2	3/8
(O,O,O)	3	1/8

Dystrybuanta:

$$F(x) = \begin{cases} 0 & \text{dla } x < 0 \\ \frac{1}{8} & \text{dla } 0 \le x < 1 \\ \frac{4}{8} & \text{dla } 1 \le x < 2 \\ \frac{7}{8} & \text{dla } 2 \le x < 3 \\ 1 & \text{dla } x \ge 3 \end{cases}$$



Funkcja prawdopodobieństwa i dystrybuanta.

14.4.2 Rozkład normalny (Gaussa)

Najważniejsze rozkłady zmiennej losowej typu ciągłego:

- Rozkład jednostajny
- **Rozkład normalny (Gaussa)**
- Rozkład log-normalny (logarytmiczno normalny)
- Rozkład Gumbela
- Rozkład Weibulla

Gęstość rozkładu normalnego (rozkładu Gaussa) wyznacza funkcja:

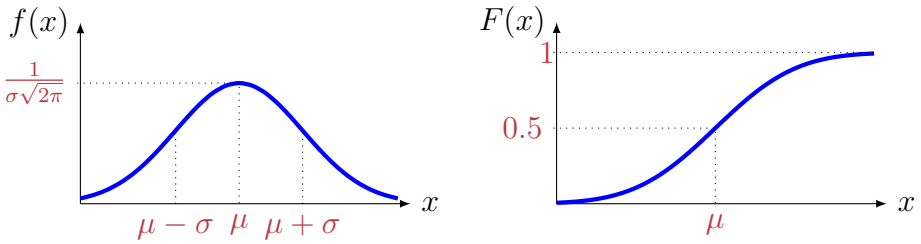
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

gdzie $\mu, \sigma \in \mathbb{R},\ \sigma > 0$ są ustalonymi parametrami, które określają odpowiednio położenie środka (przesunięcie krzywej) i szerokość (smukłość) krzywej Gaussa. Wartość funkcji gęstości jest niezerowa dla dowolnego x.

Jeżeli zmienna losowa X ma rozkład normalny wówczas mówimy, że jest normalną zmienną losową. Rozkład Gaussa oznaczamy $N(\mu, \sigma)$.

Dystrybuanta $F(x)$ dla rozkładu normalnego:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2} dt.$$



Funkcja gęstości prawdopodobieństwa i dystrybuanta.

W praktyce można założyć, że dane statystyczne mają rozkład normalny, mimo, że przedział zmiennych losowych dla funkcji gęstości $f(x)$ nie jest ograniczony. W rzeczywistości dane cechy są ograniczone, ale rozkład normalny daje wystarczająco dobre przybliżenie do oszacowania **przedziału ufności**:

- **68%** wartości cechy leży w odległości σ od wartości średniej μ,
- **95.5%** wartości cechy leży w odległości $2\,\sigma$ od wartości średniej μ,
- **99.7%** wartości cechy leży w odległości $3\,\sigma$ od wartości średniej μ.

Ostatnia właściwość jest nazywana **regułą trzech sigm**.

Przedział ufności z prawdopodobieństwem **95%** (np. potrzebny do wyznaczenia wytrzymałości gwarantowanej betonu na ściskanie) zawiera się w $[\mu - 1.95996\,\sigma, \mu + 1.95996\,\sigma]$.