STATYSTYKA dla ZOM II
dr inż Krzysztof Bryś
Wykład 2

Statystyka - pojęcia wstępne

populacja - cały zbiór badanych przedmiotów lub wartości.

próba - skończony podzbiór populacji podlegający badaniu.

próba losowa - próba losowana (najczęściej) zgodnie z rozkładem równomiernym, tzn. wylosowanie każdej próby jest jednakowo prawdopodobne.

cechy: mierzalne, niemierzalne

badana cecha = zmienna losowa X

Poszukiwany: rozkład cechy w populacji = rozkład zmiennej losowej X

próba n-elementowa = ciąg n niezależnych zmiennych losowych $(X_1, \ldots, X_n)$ o jednakowym rozkładzie (takim jak poszukiwany rozkład zmiennej losowej X).

Etapy badania statystycznego

1) Przygotowanie (formatowanie) badania (określenie celu, rodzaju, potrzebnych parametrów wejściowych badania).

2) Przeprowadzenie badania (wylosowanie próby i określenie wartości badanych cech w próbie).

3) Zebranie uzyskanych podczas badania danych.

4) Opis i wnioskowanie statystyczne (obliczenie parametrów, estymacja, weryfikacja hipotez).

5) Przedstawienie wyników.

Szeregi statystyczne

1) **Szereg wyliczający uporządkowany:** $(x_1, x_2, \ldots, x_n)$
przy czym $x_1 \leq x_2 \leq \ldots \leq x_n$.

2) **Szereg rozdzielczy punktowy:** $(x_1, x_2, \ldots, x_k)$, $(n_1, n_2, \ldots, n_k)$,
gdzie $x_1 < x_2 < \ldots < x_k$ oraz dla każdego $i = 1, 2, \ldots, k$: n_i-liczba realizacji (obserwacji) wartości x_i, $\sum_{i=1}^{k} n_i = n$.

3) **Szereg rozdzielczy przedziałowy:** $(y_0; y_1 >, (y_1; y_2 >, \ldots, (y_{k-1}; y_k)$, $(n_1, n_2, \ldots, n_k)$,
gdzie $y_0 < y_1 < y_2 < \ldots < y_{k-1} < y_k$ oraz dla każdego $i = 1, 2, \ldots, k$: n_i-liczba realizacji (obserwacji) wartości należącej do przedziału $(y_{i-1}; y_i)$, $\sum_{i=1}^{k} n_i = n$.
Wszystkie wartości należące do przedziału $(y_{i-1}; y_i >$, $i = 1, 2, \ldots, k$ utożsamia się z jego środkiem x_i.
Reguły wyznaczania liczby przedziałów (klas): $k \approx \sqrt{n}$, $k \leq 5 \log n$.

Parametry empiryczne

Miary położenia rozkładu

1) **Średnia z próby** $\overline{x}$
- dla szeregu wyliczającego:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- dla szeregu rozdzielczego:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} n_i \cdot x_i$$

2) **Dominanta (moda, wartość modalna)** D = punkt, w którym funkcja prawdopodobieństwa osiąga największą wartość
- dla szeregu wyliczającego: najczęściej występująca wartość,
- dla szeregu rozdzielczego punktowego: punkt, dla którego liczebność (częstość) osiąga największą

wartość, - dla szeregu rozdzielczego przedziałowego (wzór interpolacyjny):

$$D = x_{0d} + \frac{n_d - n_{d-1}}{(n_d - n_{d-1}) + (n_d - n_{d+1})} \cdot h_d,$$

gdzie
x_{0d} - początek przedziału zawierającego dominantę (przedziału o najwiekszej liczebności),
h_d - szerokość przedziału zawierającego dominantę (przedziału o najwiekszej liczebności),
n_d - liczebność przedziału zawierającego dominantę (najwieksza liczebność),
n_{d-1} - liczebność przedziału poprzedzającego przedział zawierający dominantę,
n_{d+1} - liczebność przedziału następnego po przedziale zawierającym dominantę.

3) **Dystrybuanta empiryczna (częstość skumulowana** $F_n(x)$
- dla szeregu wyliczającego:

$$F_n(x) = \frac{1}{n}|\{i : x_i < x, i = 1, \ldots, n\}|$$

- dla szeregu rozdzielczego:

$$F_n(x) = \sum_{i:x_i<x} \frac{n_i}{n}$$

4) **Kwantyl empiryczny rzędu** p $x_{p,n}$:
(punkt w którym dystrybuanta empiryczna po raz pierwszy osiąga wartość niemniejszą niż p)
- dla szeregu wyliczającego:

$$x_{p,n} = x_{\lceil np \rceil}$$

- dla szeregu rozdzielczego punktowego:

$$x_{p,n} = x_q \text{ gdzie } q = \min\{r : p \leq \sum_{i=1}^{r} \frac{n_i}{n}\}$$

- dla szeregu rozdzielczego przedziałowego (wzór interpolacyjny):

$$x_{p,n} = x_{0p} + (np - \sum_{x_i<x_{0p}} n_i) \cdot \frac{h_p}{n_p},$$

gdzie
x_{0p} - początek przedziału zawierającego $x_{p,n}$ (przedziału w którym dystrybuanta empiryczna po raz pierwszy osiąga wartość niemniejszą niż p),
h_p -szerokość przedziału zawierającego $x_{p,n}$,
n_p -liczebność przedziału zawierającego $x_{p,n}$,
$\sum_{x_i<x_{0p}} n_i$ - liczebność skumulowana dla przedziału poprzedzającego przedział zawierający $x_{p,n}$ (suma liczebności przedziałów poprzedzających)
Mediana: Me = kwantyl rzędu $\frac{1}{2}$
Kwartyl dolny: Q_1 = kwantyl rzędu $\frac{1}{4}$
Kwartyl górny: Q_3 = kwantyl rzędu $\frac{3}{4}$.

Miary rozproszenia rozkładu

5) **Wariancja z próby** s^2
- dla szeregu wyliczającego:

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

- dla szeregu rozdzielczego:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot (x_i - \overline{x})^2$$

6) **Odchylenie standardowe z próby** $s = \sqrt{s^2}$.

7) **Współczynnik zmienności** $V = \frac{s}{\overline{x}} \cdot 100\%$.

8) **Rozstęp** $R =$ różnica między największą i najmniejszą wartością w próbie.

9) **Współczynnik asymetrii** A_s:

- dla szeregu wyliczającego:

$$A_s = \frac{1}{s^3} \cdot (\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^3)$$

- dla szeregu rozdzielczego:

$$A_s = \frac{1}{s^3} \cdot (\frac{1}{n} \sum_{i=1}^{k} n_i \cdot (x_i - \overline{x})^3)$$

10) **Kurtoza (współczynnik skupienia)** A_s:

- dla szeregu wyliczającego:

$$K = \frac{1}{s^4} \cdot (\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^4)$$

- dla szeregu rozdzielczego:

$$K = \frac{1}{s^4} \cdot (\frac{1}{n} \sum_{i=1}^{k} n_i \cdot (x_i - \overline{x})^4)$$

11) **Współczynnik skośności** A_1:

$$A_1 = \frac{\overline{x} - D}{s}$$