

Grupowanie i histogramy

SW¹

¹Wybrane przykłady z prezentacji A. Mensfelt

Analiza statystyczna

- ▶ Uogólnianie informacji zawartej w danych
- ▶ Wnioskowanie *próba* → *populacja*

Populacja

Zbiór wyników wszystkich pomiarów istotnych dla danego problemu (*uniwersum*)

Próba

Podzbiór wyników pomiarów wybrany z populacji.

Zazwyczaj wybór **losowy** (każda próba z taką samą szansą) →
(prosta) **próba losowa**

Przykładowe dane

ID	Płeć	Wiek	Wzrost	Temperatura [C]	Glukoza [mg/dl]
P2	K	50	średni	37.7	90
P4	K	25	wysoki	36.6	101
P10	M	49	niski	35.7	72
P21	M	32	średni	38.1	120
...					
P33	K	51	średni	36.7	110

Skale pomiarowe

- ▶ ilościowe / liczbowe

- ▶ dyskretne
- ▶ ciągłe

- ▶ jakościowe

- ▶ nominalne
- ▶ porządkowe

1. nominalne ($=, \neq$)
2. porządkowe ($<, >$)
3. interwałowe ($+, -$)
4. ilorazowe ($\times, /$)

Szeregi rozdzielcze i histogramy

Szereg rozdzielczy

Podsumowanie danych pokazujące liczbę wystąpień (liczbość, częstość) poszczególnych wartości lub *klas* (*przedziałów*)

- ▶ Szereg punktowy (skala nominalna)
- ▶ Szereg przedziałowy (skala porządkowa i lepsze)

Histogram

Graficzna reprezentacja szeregu rozdzielczego

Punktowy szereg rozdzielczy

- ▶ Wyniki rzutu kostką:

3 2 2 1 6 4 1 3 1 1 6 5 3 1 5 4 4 2 6 3 6 2 5 3 5 1 3 5 1 1

- ▶ Rozkład wyników:

Wynik	Liczność	Częstość
1	8	$\frac{8}{30}$
2	4	$\frac{4}{30}$
3	6	$\frac{6}{30}$
4	3	$\frac{3}{30}$
5	5	$\frac{5}{30}$
6	4	$\frac{4}{30}$

Przedziałowy szereg rozdzielczy

► Wiek pacjentów:

32 31 57 70 72 66 34 47 42 53 52 32 50 47 31 63 55 57 46 47
39 63 70 30 30 25 58 32 48 59

► Rozkład wieku:

Przedział	Obserwacje	Liczność	Częstość
(20, 30]	25, 30, 30	3	$\frac{3}{30}$
(30, 40]	31, 31, 32, 32, 32, 34, 39	7	$\frac{7}{30}$
(40, 50]	42, 46, 47, 47, 47, 48, 50	7	$\frac{7}{30}$
(50, 60]	52, 53, 55, 57, 57, 58, 59	7	$\frac{7}{30}$
(60, 70]	63, 63, 66, 70, 70	5	$\frac{5}{30}$
(70, 80]	71	1	$\frac{1}{30}$

Konstrukcja szeregu rozdzielczego

1. Wybór liczby przedziałów (klas) k
2. Obliczenie szerokości przedziału h
3. Ustalenie początku pierwszego przedziału
4. Utworzenie przedziałów
5. Zliczenie obserwacji należących do poszczególnych przedziałów

Wybór liczby przedziałów

n – liczba obserwacji

- ▶ $k \approx \sqrt{n}$
- ▶ $k \approx 1 + 3.22 \cdot \log(n)$
- ▶ $k < 5 \cdot \log(n)$

Obliczenie szerokości przedziału

k – liczba przedziałów

R – rozstęp ($R = x_{max} - x_{min}$)

$$h \approx \frac{R}{k}$$

IQR – odstęp międzykwartylowy ($IQR = Q3 - Q1$)

$$h_0 = 2.64 \cdot IQR \cdot n^{-1/3}$$

$$k \approx \frac{R}{h_0}$$