Elementy statystyki opisowej

Agnieszka Goroncy

UNIWERSYTET
MIKOŁAJA KOPERNIKA
W TORUNIU
Wydział Matematyki i Informatyki

Statystyki pozycyjne

Niech $X_1, \ldots, X_n$ będzie próbą losową.
Statystyki pozycyjne uzyskuje się poprzez uporządkowanie zmiennych X_i, $i = 1, \ldots, n$ w kolejności niemalejącej:

$$X_1, \ldots, X_n \Longrightarrow X_{1:n} \leqslant \ldots \leqslant X_{n:n},$$

gdzie $X_{i:n}$ jest ***i*-tą statystyką pozycyjną**, $i = 1, \ldots, n$.

Przykład: Wagi pięciu 14-letnich dziewcząt wynoszą (w kg):
40, 52, 48, 60, 52.

Statystyki pozycyjne

Niech $X_1, \ldots, X_n$ będzie próbą losową.
Statystyki pozycyjne uzyskuje się poprzez uporządkowanie zmiennych X_i, $i = 1, \ldots, n$ w kolejności niemalejącej:

$$X_1, \ldots, X_n \Longrightarrow X_{1:n} \leqslant \ldots \leqslant X_{n:n},$$

gdzie $X_{i:n}$ jest i-**tą statystyką pozycyjną**, $i = 1, \ldots, n$.

Przykład: Wagi pięciu 14-letnich dziewcząt wynoszą (w kg): 40, 52, 48, 60, 52.

Mamy więc $n = 5$ oraz:

$x_{1:5}$	$x_{2:5}$	$x_{3:5}$	$x_{4:5}$	$x_{5:5}$
40	48	52	52	60

Tabelaryczna prezentacja rozkładu zmiennej

Plik `babyboom.ods` zawiera informacje dotyczące 44 dzieci urodzonych w ciągu jednego 24-godzinnego okresu w szpitalu w Brisbane, Australia (na podstawie danych *babyboom* z pakietu *Using R*, środowisko *R*):

- godzina na zegarze
- płeć
- waga w gramach
- minuty urodzenia po północy

Tabelaryczna prezentacja rozkładu zmiennej

Plik `babyboom.ods` zawiera informacje dotyczące 44 dzieci urodzonych w ciągu jednego 24-godzinnego okresu w szpitalu w Brisbane, Australia (na podstawie danych *babyboom* z pakietu *Using R*, środowisko *R*):

- godzina na zegarze
- płeć
- waga w gramach
- minuty urodzenia po północy

Szereg

- szczegółowy

Tabelaryczna prezentacja rozkładu zmiennej

Plik `babyboom.ods` zawiera informacje dotyczące 44 dzieci urodzonych w ciągu jednego 24-godzinnego okresu w szpitalu w Brisbane, Australia (na podstawie danych *babyboom* z pakietu *Using R*, środowisko *R*):

- godzina na zegarze
- płeć
- waga w gramach
- minuty urodzenia po północy

Szereg

- szczegółowy
  zmienna *godzina* przyjmuje następujące wartości:

Tabelaryczna prezentacja rozkładu zmiennej

Plik `babyboom.ods` zawiera informacje dotyczące 44 dzieci urodzonych w ciągu jednego 24-godzinnego okresu w szpitalu w Brisbane, Australia (na podstawie danych *babyboom* z pakietu *Using R*, środowisko *R*):

- godzina na zegarze
- płeć
- waga w gramach
- minuty urodzenia po północy

Szereg

- szczegółowy
  zmienna *godzina* przyjmuje następujące wartości:
  $5, 104, 118, 155, 257, 405, 407, \ldots, 2327, 2355$

Tabelaryczna prezentacja rozkładu zmiennej

Plik `babyboom.ods` zawiera informacje dotyczące 44 dzieci urodzonych w ciągu jednego 24-godzinnego okresu w szpitalu w Brisbane, Australia (na podstawie danych *babyboom* z pakietu *Using R*, środowisko *R*):

- godzina na zegarze
- płeć
- waga w gramach
- minuty urodzenia po północy

Szereg

- szczegółowy
  zmienna *godzina* przyjmuje następujące wartości:
  $5, 104, 118, 155, 257, 405, 407, \ldots, 2327, 2355$

Szeregi, c.d.

- rozdzielczy

Szeregi, c.d.

- rozdzielczy
  - punktowy

Szeregi, c.d.

- rozdzielczy
  - punktowy
    zmienną *płeć* można pogrupować następująco:

Szeregi, c.d.

- rozdzielczy
  - punktowy
    zmienną *płeć* można pogrupować następująco:

płeć	dziewczynka	chłopiec
ilość (n_i, i=1,2)	18	26

Szeregi, c.d.

- rozdzielczy
  - punktowy
    zmienną *płeć* można pogrupować następująco:

płeć	dziewczynka	chłopiec
ilość (n_i, i=1,2)	18	26

  - przedziałowy

Szeregi, c.d.

- rozdzielczy
  - punktowy
    zmienną *płeć* można pogrupować następująco:

płeć	dziewczynka	chłopiec
ilość (n_i, i=1,2)	18	26

  - przedziałowy
    zmienną *waga* można pogrupować następująco:

Szeregi, c.d.

- rozdzielczy
  - punktowy
    zmienną *płeć* można pogrupować następująco:

płeć	dziewczynka	chłopiec
ilość (n_i, i=1,2)	18	26

  - przedziałowy
    zmienną *waga* można pogrupować następująco:

waga	(1744.5, 2147.5]	(2147.5, 2550.5]	(2550.5, 2953.5]
ilość (n_i,i=1,...,3)	2	3	4

waga	(2953.5, 3356.5]	(3356.5, 3759,5]	(3759.5, 4162.5]
ilość (n_i,i=4,...,6)	10	19	6

Szeregi, c.d.

- rozdzielczy
  - punktowy
    zmienną *płeć* można pogrupować następująco:

płeć	dziewczynka	chłopiec
ilość (n_i, i=1,2)	18	26

  - przedziałowy
    zmienną *waga* można pogrupować następująco:

waga	(1744.5, 2147.5]	(2147.5, 2550.5]	(2550.5, 2953.5]
ilość (n_i,i=1,...,3)	2	3	4

waga	(2953.5, 3356.5]	(3356.5, 3759,5]	(3759.5, 4162.5]
ilość (n_i,i=4,...,6)	10	19	6

liczba klas: k$\simeq \sqrt{n} = 6$, $n = n_1 + \ldots + n_6 = 44$, min=1745, max=4162, długości klas=403, środki przedziałów: $x_1^0 = 1946, \ldots, x_6^0 = 3961$

Oznaczenia

n - ilość obserwacji,
k - ilość klas w szeregu rozdzielczym, dla szeregu przedziałowego liczymy np. ze wzoru $k \simeq \sqrt{n}$,
α - dokładność, z jaką podawane są wartości obserwacji,
x_i^- - dolna granica i-tego przedziału, np. $x_1^- = x_{1:n} - \frac{\alpha}{2}$,
b - długość przedziału w szeregu rozdzielczym przedziałowym, np. $b = \frac{x_{n:n} - x_1^-}{k}$,
n_i - liczebność i-tej klasy, $i = 1, \ldots, k$,
x_i^0 - środek przedziału i-tej klasy (w szeregu przedziałowym), $i = 1, \ldots, k$.

Podstawowe statystyki opisowe

Miary:

Podstawowe statystyki opisowe

Miary:

- położenia (tendencji centralnej)
  - średnia arytmetyczna
  - mediana
  - kwantyle
  - dominanta (moda)

Podstawowe statystyki opisowe

Miary:

- położenia (tendencji centralnej)
  - średnia arytmetyczna
  - mediana
  - kwantyle
  - dominanta (moda)
- rozproszenia (zróżnicowania)
  - wariancja
  - odchylenie standardowe
  - odchylenie przeciętne od średniej
  - odchylenie przeciętne od mediany
  - rozstęp
  - współczynnik zmienności
  - współczynnik nierównomierności

Podstawowe statystyki opisowe

Miary:

- położenia (tendencji centralnej)
  - średnia arytmetyczna
  - mediana
  - kwantyle
  - dominanta (moda)
- rozproszenia (zróżnicowania)
  - wariancja
  - odchylenie standardowe
  - odchylenie przeciętne od średniej
  - odchylenie przeciętne od mediany
  - rozstęp
  - współczynnik zmienności
  - współczynnik nierównomierności
- asymetrii: współczynnik skośności

Podstawowe statystyki opisowe

Miary:

- położenia (tendencji centralnej)
  - średnia arytmetyczna
  - mediana
  - kwantyle
  - dominanta (moda)
- rozproszenia (zróżnicowania)
  - wariancja
  - odchylenie standardowe
  - odchylenie przeciętne od średniej
  - odchylenie przeciętne od mediany
  - rozstęp
  - współczynnik zmienności
  - współczynnik nierównomierności
- asymetrii: współczynnik skośności
- koncentracji: kurtoza, eksces

Miary położenia

Średnia arytmetyczna:

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$ - szereg szczegółowy,

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$$ - szereg punktowy,

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{k} x_i^0 n_i$$ - szereg przedziałowy.

waga	(1744.5, 2147.5]	(2147.5, 2550.5]	(2550.5, 2953.5]
środki klas (x_i^0)	1946	2349	2752
ilość (n_i)	2	3	4

waga	(2953.5, 3356.5]	(3356.5, 3759,5]	(3759.5, 4162.5]
środki klas (x_i^0)	3155	3558	3961
ilość (n_i)	10	19	6

Miary położenia

Średnia arytmetyczna:

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$ - szereg szczegółowy,

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$$ - szereg punktowy,

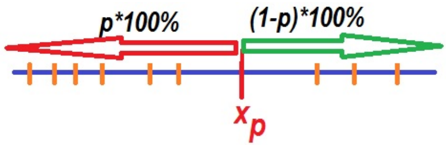
$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{k} x_i^0 n_i$$ - szereg przedziałowy.

waga	(1744.5, 2147.5]	(2147.5, 2550.5]	(2550.5, 2953.5]
środki klas (x_i^0)	1946	2349	2752
ilość (n_i)	2	3	4

waga	(2953.5, 3356.5]	(3356.5, 3759,5]	(3759.5, 4162.5]
środki klas (x_i^0)	3155	3558	3961
ilość (n_i)	10	19	6

$$\bar{x}_{44} = \frac{1}{44}(2 \cdot 1946 + \ldots + 6 \cdot 3961) \simeq 3292,4$$

Miary położenia: kwantyle rzędu $p \in (0,1)$



Kwantyle rzędu $\frac{i}{p}$, $i = 1, \ldots, p-1$:
Szereg szczegółowy i punktowy:

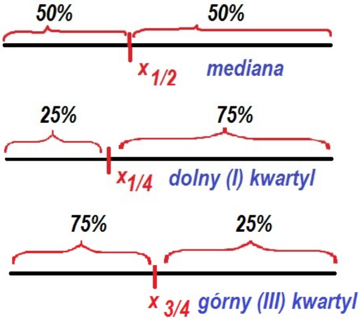
$$Q_{\frac{i}{p}} = \left(k + 1 - (n+1)\frac{i}{p}\right) x_{k:n} + \left((n+1)\frac{i}{p} - k\right) x_{k+1:n},$$

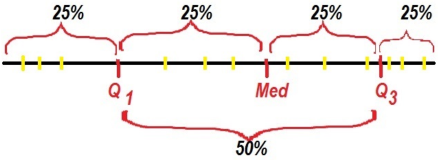
gdzie $k = \left[(n+1)\frac{i}{p}\right]$.
Szereg przedziałowy:

$$Q_{\frac{i}{p}} = x_q^- + \frac{b_q}{n_q}\left(\frac{i}{p}n - \sum_{i=1}^{q-1} n_i\right),$$

gdzie: q - numer klasy, w której znajduje się kwantyl, zaś b_q to jej długość.

Kwartyle


50%
50%
x 1/2
mediana
25%
75%
x1/4
dolny (I) kwartyl
75%
25%
x 3/4
górny (III) kwartyl


25%
25%
25%
25%
Q 1
Med
Q3
50%

Drugi kwartyl: mediana

$$\text{Med}=\begin{cases} x_{\frac{n+1}{2}:n} & n\text{-nieparzyste} \\ \frac{x_{n/2:n}+x_{n/2+1:n}}{2} & n-\text{parzyste} \end{cases}$$ - szereg szczegółowy,

$$\text{Med}=\begin{cases} x_{\frac{n+1}{2}:n} & n\text{-nieparzyste} \\ \simeq x_{n/2:n} & n-\text{parzyste} \end{cases}$$ - szereg punktowy,

$$\text{Med}=x_m^- + \frac{b_m}{n_m}\left(\frac{n}{2}-\sum_{i=1}^{m-1} n_i\right)$$ - szereg przedziałowy,

gdzie: m - numer klasy, w której znajduje się mediana, zaś b_m to jej długość.

Kwantyle - przykłady

Percentyle

Przykład: Siatki centylowe

Miary położenia

Dominanta (moda) W przypadku szeregu szczegółowego lub punktowego $D = x_i$, gdzie x_i jest najczęstszym wariantem badanej cechy (odpowiada mu największa liczebność). W przypadku szeregu przedziałowego jest to albo środek najliczniejszej klasy, gdy liczności klas sąsiednich są identyczne, albo

$$D = x_d^- + \frac{n_d - n_{d-1}}{(n_d - n_{d-1}) + (n_d - n_{d+1})} b_d,$$

gdy liczności sąsiednich klas są różne, gdzie d jest klasą najliczniej reprezentowaną, zaś b_d jej długoscią.
Jeżeli w szeregu rozdzielczym najliczniejsze są klasy skrajne, to szereg ten nazywamy antymodalnym (typu U lub typu J).
Przykład: Jaka jest dominanta wagi noworodków na podstawie danych z pliku `babyboom.ods`?

Miary rozproszenia: wariancja, odchylenie standardowe

Wariancja

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ - szereg szczegółowy,

$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i(x_i - \bar{x}_n)^2$ - szereg punktowy,

$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i(x_i^0 - \bar{x}_n)^2$ - szereg przedziałowy.

Odchylenie standardowe: $s = \sqrt{s^2}$.

Przykład:

waga	(1744.5, 2147.5]	(2147.5, 2550.5]	(2550.5, 2953.5]
środki klas (x_i^0)	1946	2349	2752
ilość (n_i)	2	3	4

waga	(2953.5, 3356.5]	(3356.5, 3759,5]	(3759.5, 4162.5]
środki klas (x_i^0)	3155	3558	3961
ilość (n_i)	10	19	6

Miary rozproszenia: wariancja, odchylenie standardowe

Wariancja

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ - szereg szczegółowy,

$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i(x_i - \bar{x}_n)^2$ - szereg punktowy,

$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i(x_i^0 - \bar{x}_n)^2$ - szereg przedziałowy.

Odchylenie standardowe: $s = \sqrt{s^2}$.

Przykład:

waga	(1744.5, 2147.5]	(2147.5, 2550.5]	(2550.5, 2953.5]
środki klas (x_i^0)	1946	2349	2752
ilość (n_i)	2	3	4

waga	(2953.5, 3356.5]	(3356.5, 3759,5]	(3759.5, 4162.5]
środki klas (x_i^0)	3155	3558	3961
ilość (n_i)	10	19	6

$s = \sqrt{\frac{1}{44}(2 \cdot (1946 - 3292,4)^2 + \ldots + 6 \cdot (3961 - 3292,4)^2)} \simeq 521$

Miary rozproszenia

Odchylenie przeciętne od średniej:

$d_1 = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}_n|$ - szereg szczegółowy,

$d_1 = \frac{1}{n}\sum_{i=1}^{k}n_i|x_i - \bar{x}_n|$ - szereg punktowy,

$d_1 = \frac{1}{n}\sum_{i=1}^{k}n_i|x_i^0 - \bar{x}_n|$ - szereg przedziałowy.

Odchylenie przeciętne od mediany:

$d_2 = \frac{1}{n}\sum_{i=1}^{n}|x_i - \mathrm{Med}|$ - szereg szczegółowy,

$d_2 = \frac{1}{n}\sum_{i=1}^{k}n_i|x_i - \mathrm{Med}|$ - szereg punktowy,

$d_2 = \frac{1}{n}\sum_{i=1}^{k}n_i|x_i^0 - \mathrm{Med}|$ - szereg przedziałowy.

Rozstęp: $R = x_{n:n} - x_{1:n}$

Współczynnik zmienności wyraża jaką część średniej arytmetycznej stanowi odchylenie standardowe:

$$\gamma = \frac{s}{\bar{x}_n} 100\%.$$

Im mniejsze γ, tym mniejsze jest zróżnicowanie zmiennej.
Współczynnik nierównomierności:

$$H = \frac{d_1}{\bar{x}_n} 100\%.$$

Momenty

- **Moment zwykły rzędu** k:

$$m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \quad k \in \mathbb{N}.$$

- **Moment centralny rzędu** k:

$$M_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}_n)^k, \quad k \in \mathbb{N}.$$

- **Moment absolutny (zwykły) rzędu** k:

$$a_k = \frac{1}{n}\sum_{i=1}^{n} |x_i|^k, \quad k \in \mathbb{N}.$$

- **Absolutny moment centralny rzędu** k:

$$A_k = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}_n|^k, \quad k \in \mathbb{N}.$$

Miara asymetrii

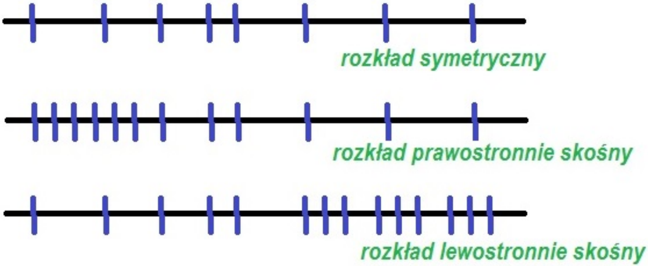
Współczynnik asymetrii (skośności):

$$sk = c \cdot \frac{M_3}{s^3},$$

gdzie $c = \frac{n^2}{(n-1)(n-2)}$ lub $c = (\frac{n}{n-1})^{3/2}$.
Jeżeli

- $sk > 0$, to rozkład jest **prawostronnie skośny** (ma dłuższy prawy „ogon", dane skupiają się "na lewo") - często mediana jest mniejsza niż średnia,
- $sk < 0$, to rozkład jest **lewostronnie skośny** (ma dłuższy lewy „ogon", dane skupiają się "na prawo") - często mediana jest wyższa niż średnia,
- $sk = 0$, to rozkład jest **symetryczny**.

Skośność rozkładu



Miara koncentracji (skupienia)

Kurtoza

$$K = \frac{n^2[n(n+1)M_4 - 3(n-1)M_2^2]}{(n-1)(n-2)(n-3)s^4}.$$

Im większa jest kurtoza, tym większa jest koncentracja, tzn. tym bardziej wartości zmiennej koncentrują się wokół średniej. Jeżeli

- $K < 0$ - dane są mało skoncentrowane wokół średniej, rozkład jest bardziej spłaszczony od standardowego normalnego,
- $K > 0$ - dane są bardzo skoncentrowane wokół średniej, rozkład jest bardziej wysmukły od standardowego normalnego,
- $K = 0$ - dane są "normalnie" skoncentrowane wokół średniej.

Jeżeli kurtozę liczymy ze wzoru

$$K_1 = \left(\frac{n}{n-1}\right)^2 \cdot \frac{M_4}{s^4},$$

to porównujemy jej wartość nie z 0, ale z 3 (dla rozkładu standardowego normalnego $K_1 = 3$).

Współczynnik spłaszczenia - eksces:

$$ek = \frac{M_4}{s^4} - 3.$$

Literatura

- A. Maksimowicz-Ajchel, „*Wstęp do statystyki. Metody opisu statystycznego*", WUW, 2007.
- M. Sobczyk, „*Statystyka opisowa*", C.H.Beck, 2010.