

Elementarna statystyka

Podstawowa analiza danych

(Examining distributions)

Alexander Bendikov

6 kwietnia 2016

- Zbiór danych zawiera informacje o pewnej grupie *przypadków* lub *obserwacji*: ludzi, rzeczy itp. Dla każdego przypadku dane zawierają wartość jednej bądź większej ilości *zmiennych*: wysokości, płci itp.

- Zbiór danych zawiera informacje o pewnej grupie *przypadków* lub *obserwacji*: ludzi, rzeczy itp. Dla każdego przypadku dane zawierają wartość jednej bądź większej ilości *zmiennych*: wysokości, płci itp.
- *Zmienne kategoryjne* zaliczają każdy przypadek do pewnej *kategorii*: mężczyzna/kobieta, status społeczny itp.

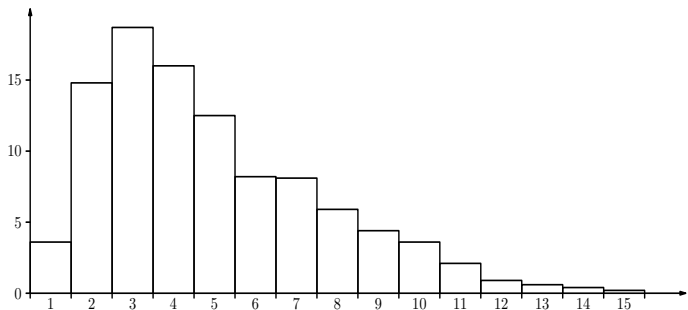
- Zbiór danych zawiera informacje o pewnej grupie *przypadków* lub *obserwacji*: ludzi, rzeczy itp. Dla każdego przypadku dane zawierają wartość jednej bądź większej ilości *zmiennych*: wysokości, płci itp.
- *Zmienne kategoryjne* zaliczają każdy przypadek do pewnej *kategorii*: mężczyzna/kobieta, status społeczny itp.
- *Zmienne ilościowe* podają pewną charakterystykę numeryczną, wysokość, wagę, pensję itp.

- Zbiór danych zawiera informacje o pewnej grupie *przypadków* lub *obserwacji*: ludzi, rzeczy itp. Dla każdego przypadku dane zawierają wartość jednej bądź większej ilości *zmiennych*: wysokości, płci itp.
- *Zmienne kategoryjne* zaliczają każdy przypadek do pewnej *kategorii*: mężczyzna/kobieta, status społeczny itp.
- *Zmienne ilościowe* podają pewną charakterystykę numeryczną, wysokość, wagę, pensję itp.
- *Badawcza analiza danych* stosuje wykresy i parametry liczbowe aby opisać zmienne oraz zależności pomiędzy zmiennymi w zbiorze danych.

- Zbiór danych zawiera informacje o pewnej grupie *przypadków* lub *obserwacji*: ludzi, rzeczy itp. Dla każdego przypadku dane zawierają wartość jednej bądź większej ilości *zmiennych*: wysokości, płci itp.
- *Zmienne kategoryjne* zaliczają każdy przypadek do pewnej *kategorii*: mężczyzna/kobieta, status społeczny itp.
- *Zmienne ilościowe* podają pewną charakterystykę numeryczną, wysokość, wagę, pensję itp.
- *Badawcza analiza danych* stosuje wykresy i parametry liczbowe aby opisać zmienne oraz zależności pomiędzy zmiennymi w zbiorze danych.
- *Rozkład* zmiennej ilościowej opisuje wartości zmiennej i ich częstotliwości.

Zmienne ilościowe: histogramy

1. Rozkładamy zakres wartości zmiennej na przedziały równej długości
2. Zliczamy ilość przypadków z wartościami w poszczególnych przedziałach
3. Rysujemy prostokąty: podstawa każdego pokrywa kolejny przedział, wysokość to zliczona ilość przypadków w danym przedziale.



Rysunek: Histogram zmiennej ilościowej „Długość słów”

Długość	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Odsetek	3,6	14,8	18,7	16,0	12,5	8,2	8,1	5,9	4,4	3,6	2,1	0,9	0,6	0,4	0,2

Tablica: Długość słów w czasopiśmie *Popular Science*

Opis rozkładu przy pomocy parametrów

- $X = x_1, x_2, \dots, x_n$ (uporządkowane rosnąco)
- Średnia $\bar{x} = \frac{1}{n} (x_1 + x_2 + \dots, x_n)$
- Mediana $M = \begin{cases} x_{(n+1)/2} & n \text{ nieparzysta} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & n \text{ parzysta.} \end{cases}$

Opis rozkładu przy pomocy parametrów

- $X = x_1, x_2, \dots, x_n$ (uporządkowane rosnąco)
- Średnia $\bar{x} = \frac{1}{n} (x_1 + x_2 + \dots, x_n)$
- Mediana $M = \begin{cases} x_{(n+1)/2} & n \text{ nieparzysta} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & n \text{ parzysta.} \end{cases}$

Uwagi: 1) Dla $n \gg 1$, $\bar{x} \approx E(X)$ z prawa wielkich liczb,

Opis rozkładu przy pomocy parametrów

- $X = x_1, x_2, \dots, x_n$ (uporządkowane rosnąco)
- Średnia $\bar{x} = \frac{1}{n} (x_1 + x_2 + \dots, x_n)$
- Mediana $M = \begin{cases} x_{(n+1)/2} & n \text{ nieparzysta} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & n \text{ parzysta.} \end{cases}$

Uwagi: 1) Dla $n \gg 1$, $\bar{x} \approx E(X)$ z prawa wielkich liczb,
2) Dla niewielkich n średnia $\bar{x}$ istotnie zależy od obserwacji nietypowych (outlier), natomiast mediana M nie zależy.

Opis rozkładu przy pomocy parametrów

- $X = x_1, x_2, \dots, x_n$ (uporządkowane rosnąco)
- Średnia $\bar{x} = \frac{1}{n} (x_1 + x_2 + \dots, x_n)$
- Mediana $M = \begin{cases} x_{(n+1)/2} & n \text{ nieparzysta} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & n \text{ parzysta.} \end{cases}$

Uwagi: 1) Dla $n \gg 1$, $\bar{x} \approx E(X)$ z prawa wielkich liczb,
2) Dla niewielkich n średnia $\bar{x}$ istotnie zależy od obserwacji nietypowych (outlier), natomiast mediana M nie zależy.

- Kwartyle Q_1 i Q_3 :

$$Q_1 = M(x_1, \dots, x_k; k = \lfloor \frac{n+1}{2} \rfloor),$$

$$Q_3 = M(x_m, \dots, x_n; m = \lceil \frac{n+1}{2} \rceil).$$

Przykłady:

• 9 9 22 32 33 39 40 42 49 52 52 70

 ↑ ↑ ↑

$Q_1 = 27$ $M = 39,5$ $Q_3 = 50,5$

Przykłady:

- 9 9 22 32 33 39 40 42 49 52 52 70
 $Q_1 = 27$ $M = 39,5$ $Q_3 = 50,5$
- 22 25 34 35 41 41 46 46 46 47 49 54 54 55 60
 Q_1 M Q_3

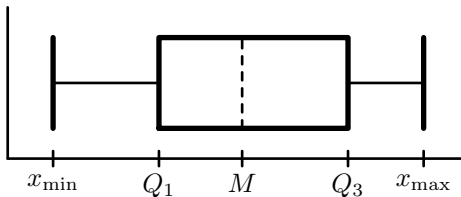
Przykłady:

- 9 9 22 32 33 39 40 42 49 52 52 70
 $Q_1 = 27$ $M = 39,5$ $Q_3 = 50,5$
- 22 25 34 35 41 41 46 46 46 47 49 54 54 55 60
 Q_1 M Q_3
- 5 - liczbowe podsumowanie: $x_{\min} - Q_1 - M - Q_3 - x_{\max}$.

Przykłady:

- 9 9 22 32 33 39 40 42 49 52 52 70
 $Q_1 = 27$ $M = 39,5$ $Q_3 = 50,5$
- 22 25 34 35 41 41 46 46 46 47 49 54 54 55 60
 Q_1 M Q_3
- 5 - liczbowe podsumowanie: $x_{\min} - Q_1 - M - Q_3 - x_{\max}$.

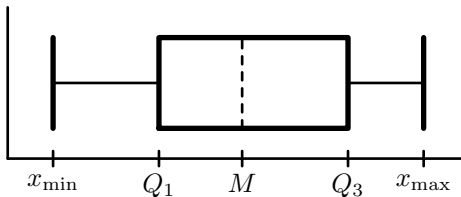
- Wykres pudełkowy:



Przykłady:

- 9 9 22 32 33 39 40 42 49 52 52 70
 $Q_1 = 27$ $M = 39,5$ $Q_3 = 50,5$
- 22 25 34 35 41 41 46 46 46 47 49 54 54 55 60
 Q_1 M Q_3
- 5 - liczbowe podsumowanie: $x_{\min} - Q_1 - M - Q_3 - x_{\max}$.

- Wykres pudełkowy:



- Rozrzut: odchylenie standardowe próbki s :

$$X : x_1, x_2, \dots, x_n,$$
$$s^2 = \frac{1}{n-1} ((x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2) = \frac{1}{n-1} \sum x_i^2 - \frac{n}{n-1} \bar{x}^2.$$

Własności:

- $E(\bar{X}) = E(X)$,
- $E(s^2) = \text{Var}(X)$.

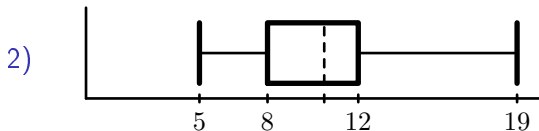
Własności:

- $E(\bar{x}) = E(X)$,
- $E(s^2) = \text{Var}(X)$.

Przykład: $X : 5 \quad 7 \quad 8 \quad 9 \quad 10 \quad 11 \quad 12 \quad 12 \quad 15 \quad 19$

$\uparrow \quad \uparrow \quad \uparrow$
 $Q_1 \quad M \quad Q_3$

1) $\min = 5, \max = 19, M = \frac{1}{2}(10 + 11) = 10,5, Q_1 = 8, Q_3 = 12$



3) $\bar{x} = \frac{1}{10}(5 + 7 + \dots + 19) = 10,8 > M, s^2 = 16,4, s = 4,06$.