

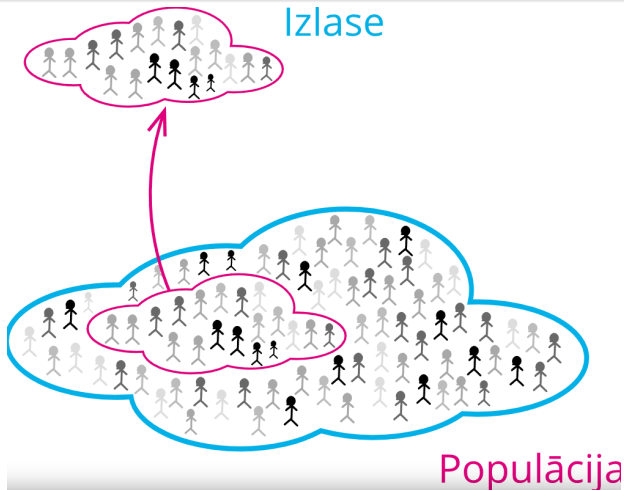
Robustas metodes: kāpēc tās ir svarīgas un nozīmīgas?

Emīls Siliņš

06.12.2025

Mazā Matemātikas Universitāte

Izlases un Populācijas



Attēls 1: Izlase un Populācija

Izlases vidējā vērtība

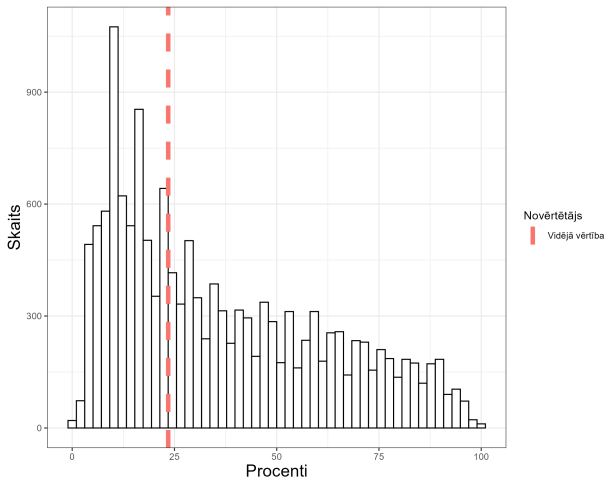
Kas ir vidējā vērtība?

Vidējā vērtība apraksta sagaidāmo novērojumu vērtību.

Neformāla definīcija: Ja $X_1, X_2, X_3, \dots, X_n$ ir mūsu novērojumi (kopā n), tad **izlases vidējo vērtību** varam definēt kā

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} (X_1 + X_2 + \dots + X_n).$$

Eksāmenu rezultātu datu piemērs



Attēls 2: 2021. gada 12. klases matemātikas eksāmena rezultāti.

Dispersija (1)

Kas ir dispersija?

Dispersija apraksta datu vidējo kvadrātisko novirzi no vidējās vērtības (jeb datu izkliedi)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left((X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \right. \\ \left. + (X_3 - \bar{X})^2 + \dots + (X_n - \bar{X})^2 \right)$$

Arī vidējajai vērtībai mēs varam aprēķināt dispersiju, pēc formulas: s^2/n .

Dispersija (2)

Dispersijas aprēķina piemērs

Pieņemsim, ka ir doti sekojoši novērojumi :

x	1	2	7	4	6
---	---	---	---	---	---

Vidējā vērtība: $\bar{X}_n = \bar{X}_5 = \frac{1}{5}(1 + 2 + 7 + 4 + 6) = \frac{20}{5} = 4$

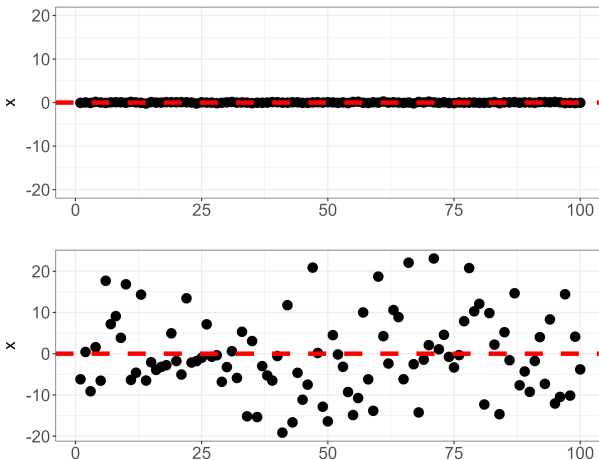
Dispersija:

$$s^2 = \frac{1}{5-1} \left((1-4)^2 + (2-4)^2 + (7-4)^2 + (4-4)^2 + (6-4)^2 \right) =$$

$$= \frac{1}{4}(9 + 4 + 9 + 4) = \frac{26}{4} = \frac{13}{2}$$

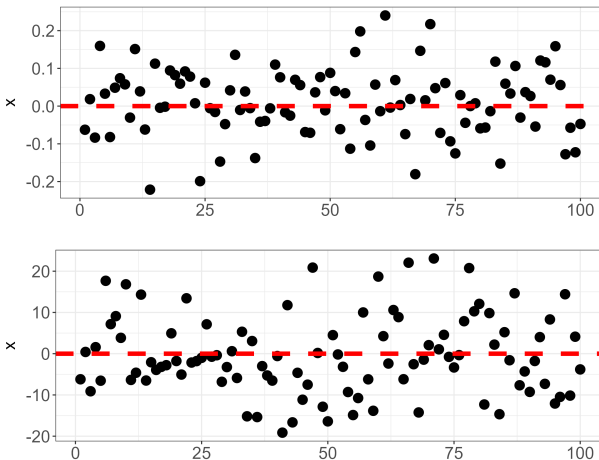
Vidējās vērtības dispersija: $s^2/n = s^2/5 = 13/10 = 1.3$

Lielas un mazas dispersijas piemērs (1)



Attēls 3: Mazas un lielas dispersijas piemērs

Lielas un mazas dispersijas piemērs (2)



Attēls 4: Mazas un lielas dispersijas piemērs

Kas ir izlecēji un kā tos noteikt?

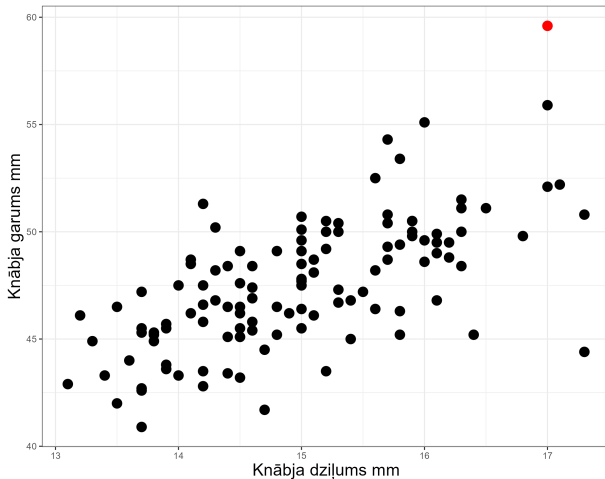
Izlecēji

Izlecējiem nav formālas definīcijas. Varam teikt, ka tie ir tādi novērojumi, kuri ir **būtiski lielāki** vai **mazāki** par pārējiem novērojumiem.

Piemērs

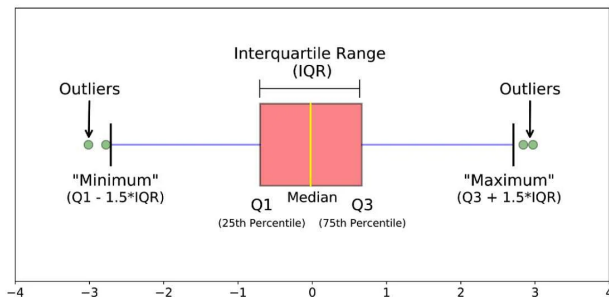
Ja gandrīz visi skolēni eksāmenā saņem $< 50\%$, taču daži saņem $> 90\%$. Tad šos skolēnus varētu uzskatīt par izlecējiem.

Pingvīnu knābju datu piemērs



Attēls 5: Džentū pingvīnu knābju dziļums pret garumu *mm*.

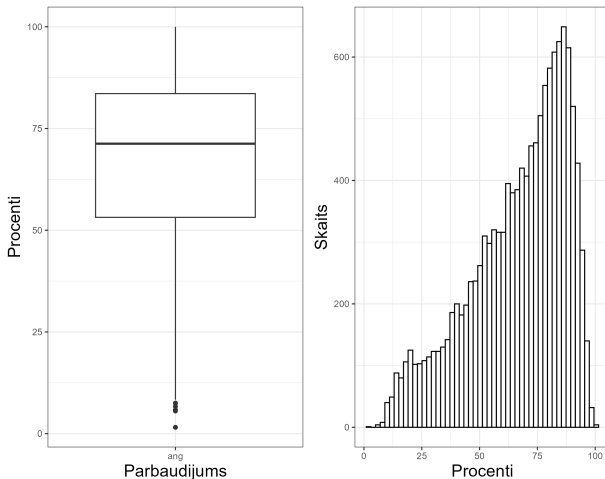
Kastu grafiki



Attēls 6: Kastu grafika konstruēšana.

Kastu grafiku bieži iesaka izmantot, lai noteiktu izlecējus. Taču tas nav tam **domāts**. Ar kastu grafika palīdzību varam noteikt izlecējus **tikai**, kad dati atbilst Gausa sadalījumam.

Kastu grafiks uz eksāmena datiem



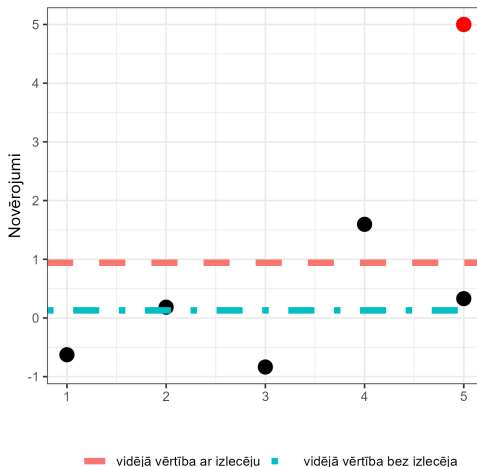
Attēls 7: 2021. gada 12. klases angļu valodas eksāmena rezultāti.

Izlecēju noteikšana

Izlecēju noteikšana praksē

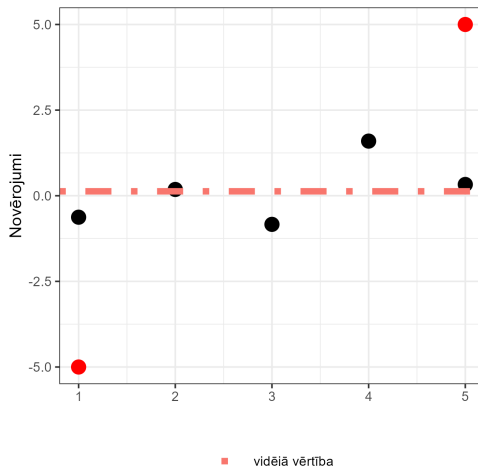
- Izlecēju noteikšana praksē ir ļoti sarežģīta un ne vienmēr tas ir iespējams.
- Bieži ir nepieciešams praktisks pamatojums, lai šādus novērojumus atmestu.
- Labāk izmantot metodes, kuras ir mazāk jūtīgas pret izlecējiem.

Problēmas ar vidējo vērtību, kad datos ir izlecēji (1)



Attēls 8: Vidējā vērtība ar vienu izlecēju.

Problēmas ar vidējo vērtību, kad datos ir izlecēji (2)



Attēls 9: Vidējā vērtība ar diviem izlecējiem.

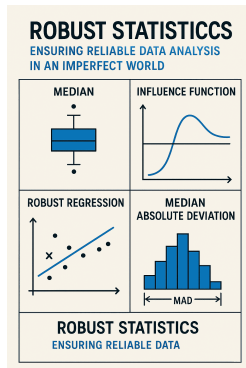
Problēmas ar vidējo vērtību, kad datos ir izlecēji (3)

Lielums	Vērtība
Vidējā vērtība	0.09
Izlases dispersija (ar izlecējiem)	8.95
Izlases dispersija (bez izlecējiem)	0.92

Tabula 1: Aprakstošās statistikas

Kas ir robustas metodes?

Par **robustām** metodēm sauc tādas metodes, kas ir mazāk jūtīgas pret izlecējiem salīdzinājumā ar vidējo vērtību.



Kvantiles

Populācijas kvantiles

Lielumu Q sauc par p -to kvantili, ja tam izpildās

$$P(X \leq Q) = p,$$

kur $0 \leq p \leq 1$. Piemēram, mediāna ir 0.5-tā kvantile (Q_2), tātad tai izpildās $P(X \leq Q_2) = 0.5$.

Izlases kvantiles

Ar $X_{(i)}$ apzīmēsim augošā secībā sakārtotus novērojumus, t.i. $X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \dots \leq X_{(n)}$. Tad izlases p -tā kvantile ir $n \cdot p$ -tais novērojums, jeb $X_{(n \cdot p)}$.

Mediāna

Mediāna ir **visrobustākais** lielums!

Izsoles mediānas aprēķināšana

Izsoles (izmērā n) mediānu var definēt vairākos veidos, piemēram:

- pirmā pieeja
 - ja n ir nepāra skaitlis, tad no sakārtotās izsoles $X_{(k)}$ novērojums būs mediāna, kur $k = (n + 1)/2$,
 - ja n ir pāra skaitlis, tad mediānu varam ņemt kā divu tuvāko novērojumu vidējo, t.i.

$$\frac{X_{(n/2)} + X_{((n+2)/2)}}{2},$$

- otrā pieeja ir noapaļot lielumu $n \cdot 0.5$ uz **augšu** vai **leju** un ņemt šo novērojumu kā mediānu.

Visas šīs pieejas ir **ekvivalentas**!

Mediānas aprēķina piemērs

Piemērs

Pieņemsim, ka ir doti sekojoši novērojumi :

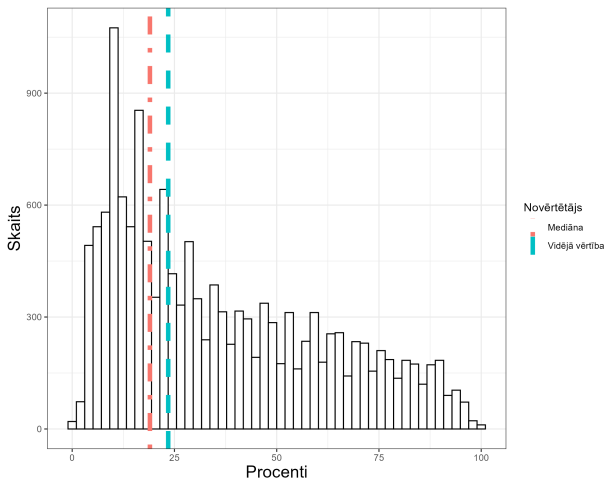
x	1	2	7	4	6
---	---	---	---	---	---

Tos sakārtojot augošā secībā iegūstam:

x	1	2	4	6	7
---	---	---	---	---	---

- Pēc pirmās pieejas mediāna ir $X_{(3)} = 4$ novērojums, jo $(5 + 1)/2 = 3$.
- Pēc otrās pieejas $5 \cdot 0.5 = 2.5$. Apaļojot uz augšu mediāna ir tas pats **trešais** novērojums $X_{(3)} = 4$, bet apaļojot uz leju tā būs **otrais** novērojums $X_{(2)} = 2$.

Mediāna eksāmena rezultātu datiem



Attēls 10: 2021. gada 12. klases matemātikas eksāmena rezultāti.

Nošķeltā vidējā vērtība

Nošķeltās vidējās vērtības mērķis ir atņemt noteiktu procentu novērojumu ar **mazākajām** un **lielākajām** vērtībām.

Definīcija

Pieņemsim, ka mums ir sakārtota izlase $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$, tad par nošķelto vidējo vērtību sauc lielumu

$$\bar{X}_{\alpha\beta} = \frac{1}{[n(1-\alpha-\beta)]} \sum_{i=[n\alpha]+1}^{[n(1-\beta)]} X_{(i)},$$

kur $[\cdot]$ apzīmē noapaļošanu uz leju un $0 \leq \alpha < 1/2$, $0 \leq \beta < 1/2$ ir šķelšanas proporcijas, jeb cik % no **lielākajiem** / **mazākajiem** novērojumiem atmetīsim.

Redzams, ka tā kā nošķeltā vidējā vērtība ignorē kādu daļu no izlecējiem, tad tā ir **robustāka** par parasto vidējo vērtību.

Nošķeltās vidējās vērtības aprēķināšana

Pieņemsim, ka mums ir izlase ar $n = 10$ novērojumiem un vēlamies aprēķināt nošķelto vidējo vērtību ar **šķelšanas parametriem** $\alpha = 0.1$ un $\beta = 0.1$. Tas nozīmē, ka atmetīsim (nošķelsim) 10% lielākos un mazākos novērojumus.

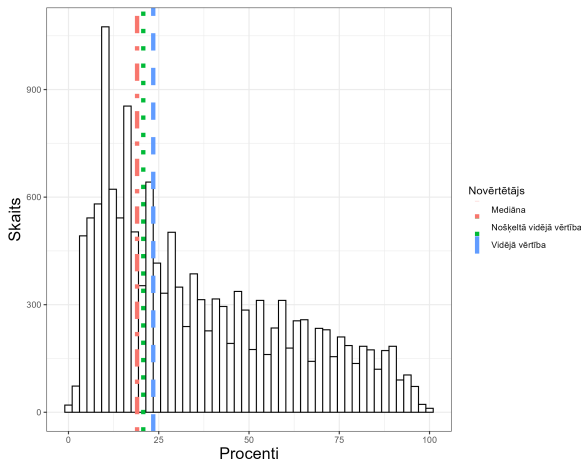
Aprēķinam lielumus :

- $1 - \alpha - \beta = 1 - 0.2 = 0.8$,
- $\lfloor n(1 - \alpha - \beta) \rfloor = \lfloor 10 \cdot 0.8 \rfloor = 8$,
- $\lfloor n\alpha \rfloor = \lfloor 10 \cdot 0.1 \rfloor = 1$,
- $\lfloor n(1 - \beta) \rfloor = \lfloor 10 \cdot 0.9 \rfloor = 9$.

Līdz ar ko nošķeltā vidējā vērtība izlasei izmērā 10 ar šķelšanas parametriem $\alpha = \beta = 0.1$ ir aprēķināma pēc formulas

$$\bar{X}_{\alpha\beta} = \frac{1}{8} \sum_{i=2}^9 X_{(i)} = \frac{1}{8} (X_{(2)} + X_{(3)} + \dots + X_{(9)}).$$

Nošķeltā vidējā vērtība eksāmena rezultātu datiem



Attēls 11: 2021. gada 12. klases matemātikas eksāmena rezultāti.
Nošķeltā vidējā vērtība aprēķināt ar $\alpha = \beta = 0.2$

Svērtā vidējā vērtība

Kas tā tāda ir?

Svērtā vidējā vērtība tiek aprēķināta katram novērojumam X_i piešķirot kādu svaru a_i .

Definīcija. Pieņemsim, ka mums ir sakārtoti novērojumi $X_{(i)}$, tad par svērto vidējo vērtību sauc lielumu

$$\bar{X}_S = \sum_{i=1}^n a_i X_{(i)},$$

kur $a_i \geq 0$ un $\sum_{i=1}^n a_i = 1$.

Patiesībā visi šodien apskatītie lielumi : **vidējā vērtībā**, **kvantiles** un **nošķeltā vidējā vērtība** arī ir svērtās vidējās vērtības.

Paldies par uzmanību!
Jautājumi?