2. STATISTIKE Z ENO SPREMENLJIVKO

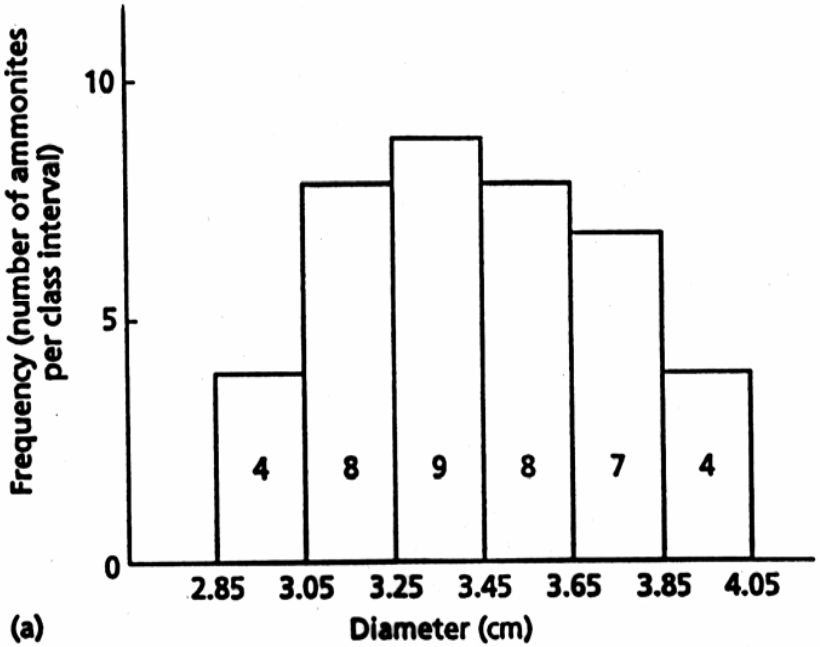
- Prvi pregled nad podatki
- Podatki s sorazmerno in intervalno lestvico
- Numerični in grafični opis

2.1. GRAFIČNI IN NUMERIČNI OPIS

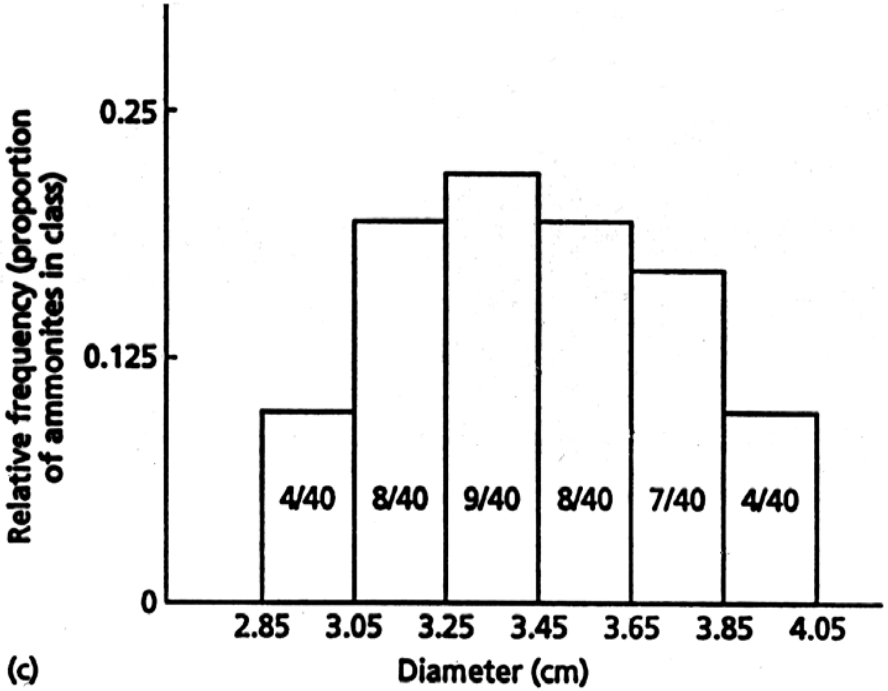
- 2.1.1. Frekvenčna porazdelitev populacije in histogram
- 2.1.2. Lastnosti in parametri frekvenčne porazdelitve populacije
- 2.1.3. Statistike frekvenčne porazdelitve vzorca

2.1.1. Frekvenčna porazdelitev in histogram populacije

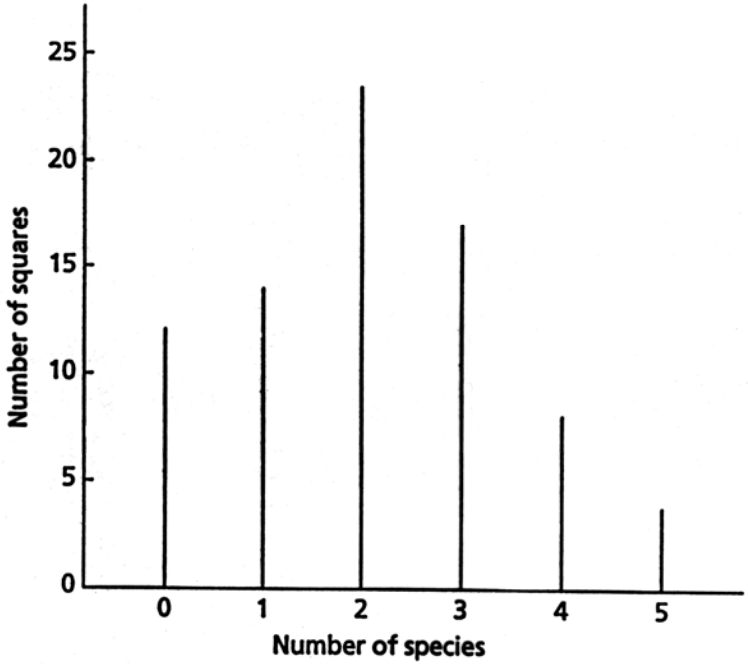
- Enodimenzionalna razvrstitev točk na premici
- Sprememba gostote točk vzdolž premice $\Rightarrow$ razredi $\Rightarrow$ frekvenčna porazdelitev $\Rightarrow$ histogram
- Nezvezne spremenljivke – črtovni graf
- Kumulativna frekvenčna porazdelitev

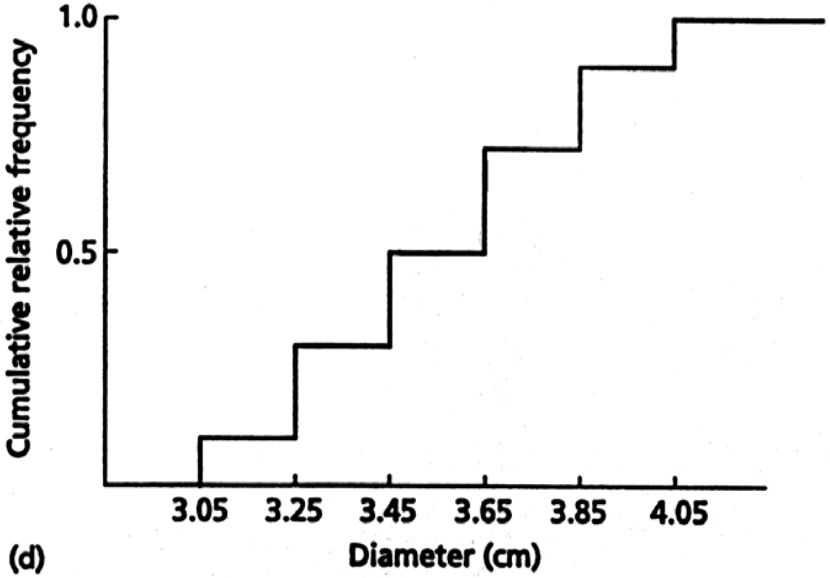

Frequency (number of ammonites per class interval)
10
5
0
4
8
9
8
7
4
2.85
3.05
3.25
3.45
3.65
3.85
4.05
Diameter (cm)

(a)


Relative frequency (proportion of ammonites in class)
0.25
0.125
0
4/40
8/40
9/40
8/40
7/40
4/40
2.85
3.05
3.25
3.45
3.65
3.85
4.05
Diameter (cm)

(c)

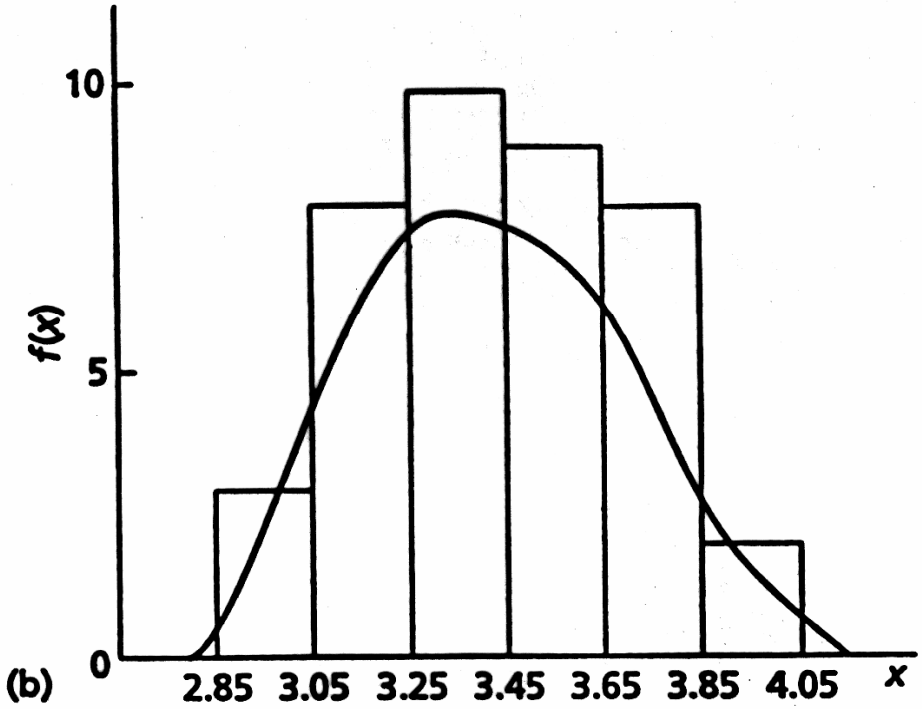

Number of squares
25
20
15
10
5
0
0
1
2
3
4
5
Number of species


Cumulative relative frequency
1.0
0.5
3.05
3.25
3.45
3.65
3.85
4.05
Diameter (cm)

(d)

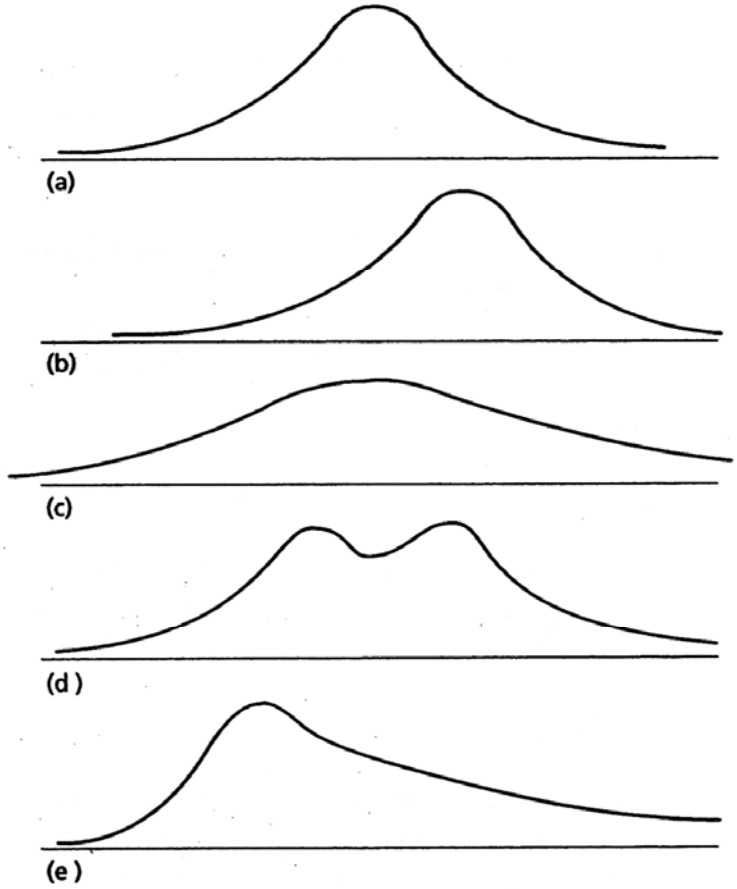
2.1.2. Lastnosti in parametri frekvenčne porazdelitve

- Frekvenčna gostota *f(x)* – krivulja prirejena obrisu histograma



2.1.2. Lastnosti in parametri frekvenčne porazdelitve

- Položaj – “povprečna” lega populacije na merilni lestvici = srednja vrednost
- Razpršenost – obseg, v katerem so podatki razporejeni vzdolž merilne lestvice (okrog srednje vrednosti)



2.1.2.1. Mere položaja

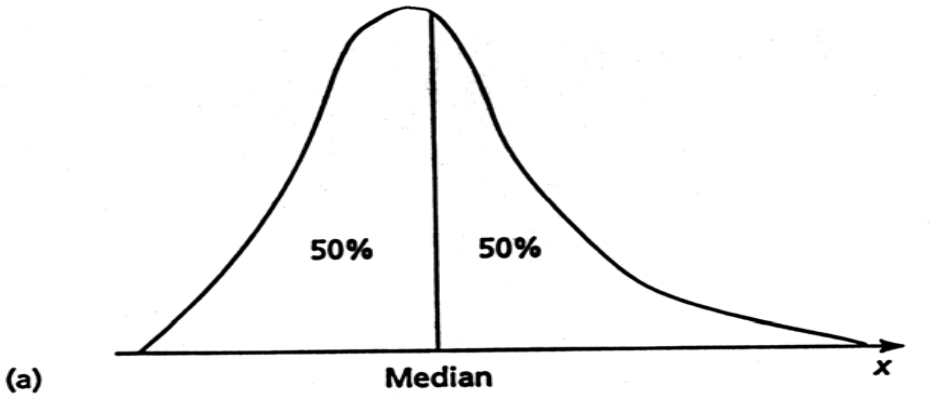
- Aritmetična sredina populacije

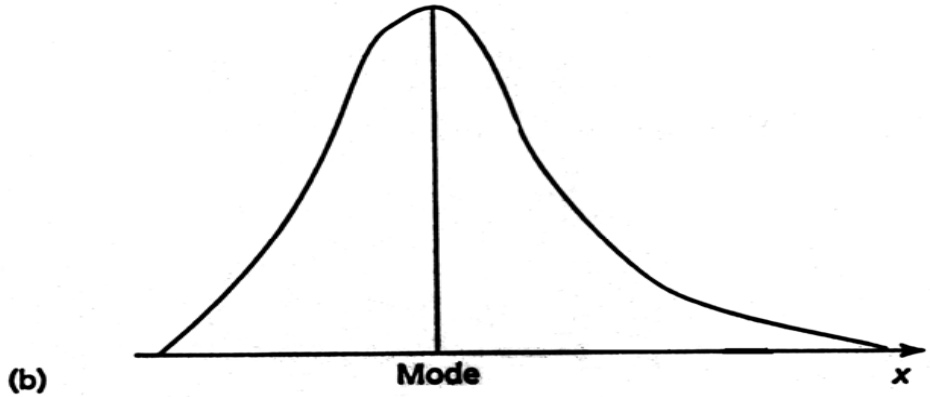
$$\mu = (X_1 + X_2 + \ldots + X_N)/N = {}^{1}/_{N} \sum_{i=1}^{N} X_i$$

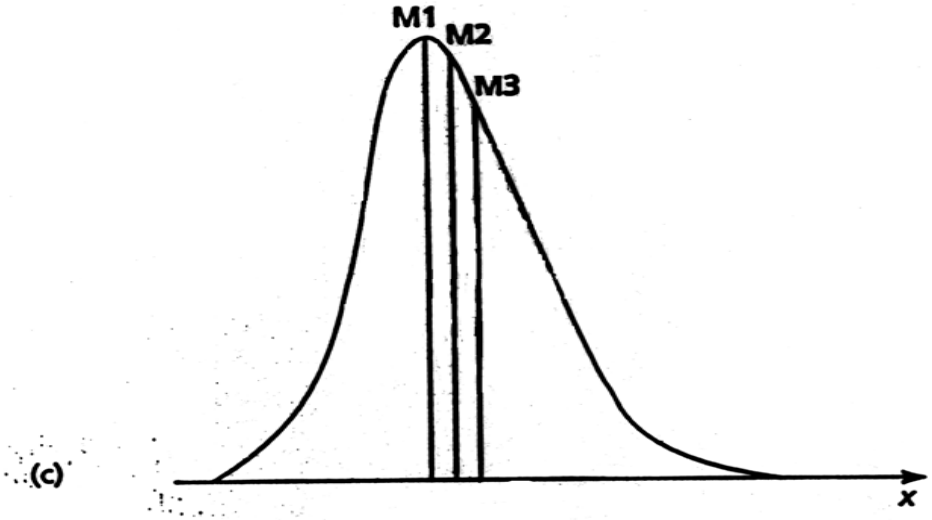
- Geometrijska sredina populacije – antilogaritmirana srednja vrednost logaritmiranih vrednosti

$$\mu_G = (X_1 \cdot X_2 \cdot \ldots \cdot X_N)^{1/N}$$

- Mediana – vrednost na sredini niza podatkov, razvrščenih v naraščajočem vrstnem redu
- Modus – vrednost, ki se najpogosteje pojavlja = vrh krivulje frekvenčne gostote


50%
50%
(a)
Median
x


(b)
Mode
x

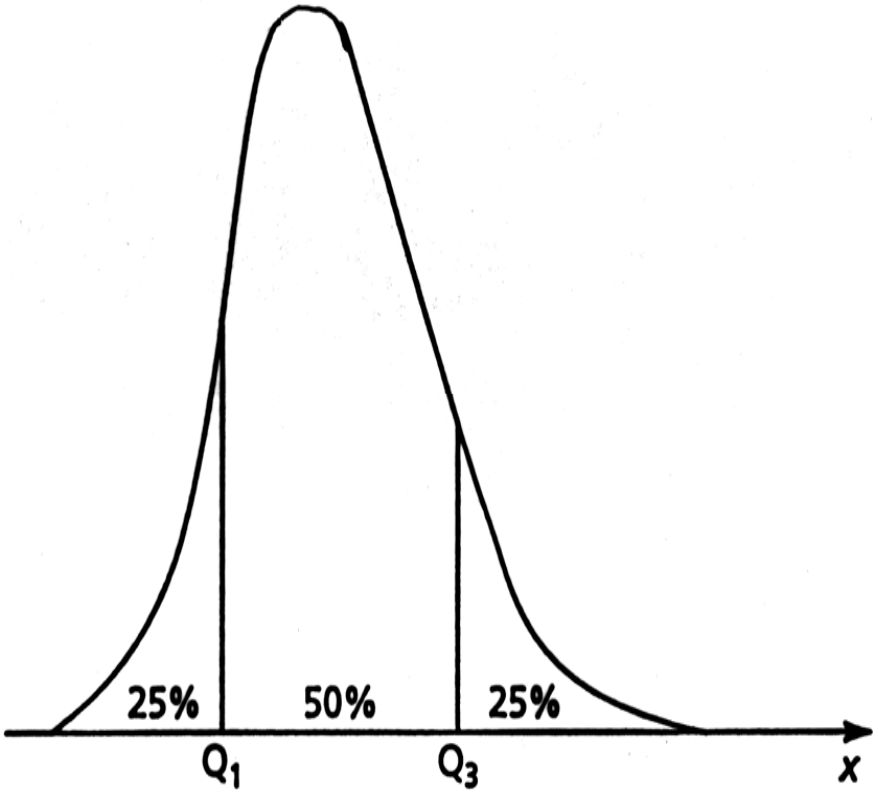

M1
M2
M3
(c)
x

2.1.2.1. Mere razpršenosti

- Razpon (range) – razlika med najnižjo in najvišjo vrednostjo populacije

- Odstopanje znotraj kvartilov (IQD)

$$IQD = Q_3 - Q_1$$



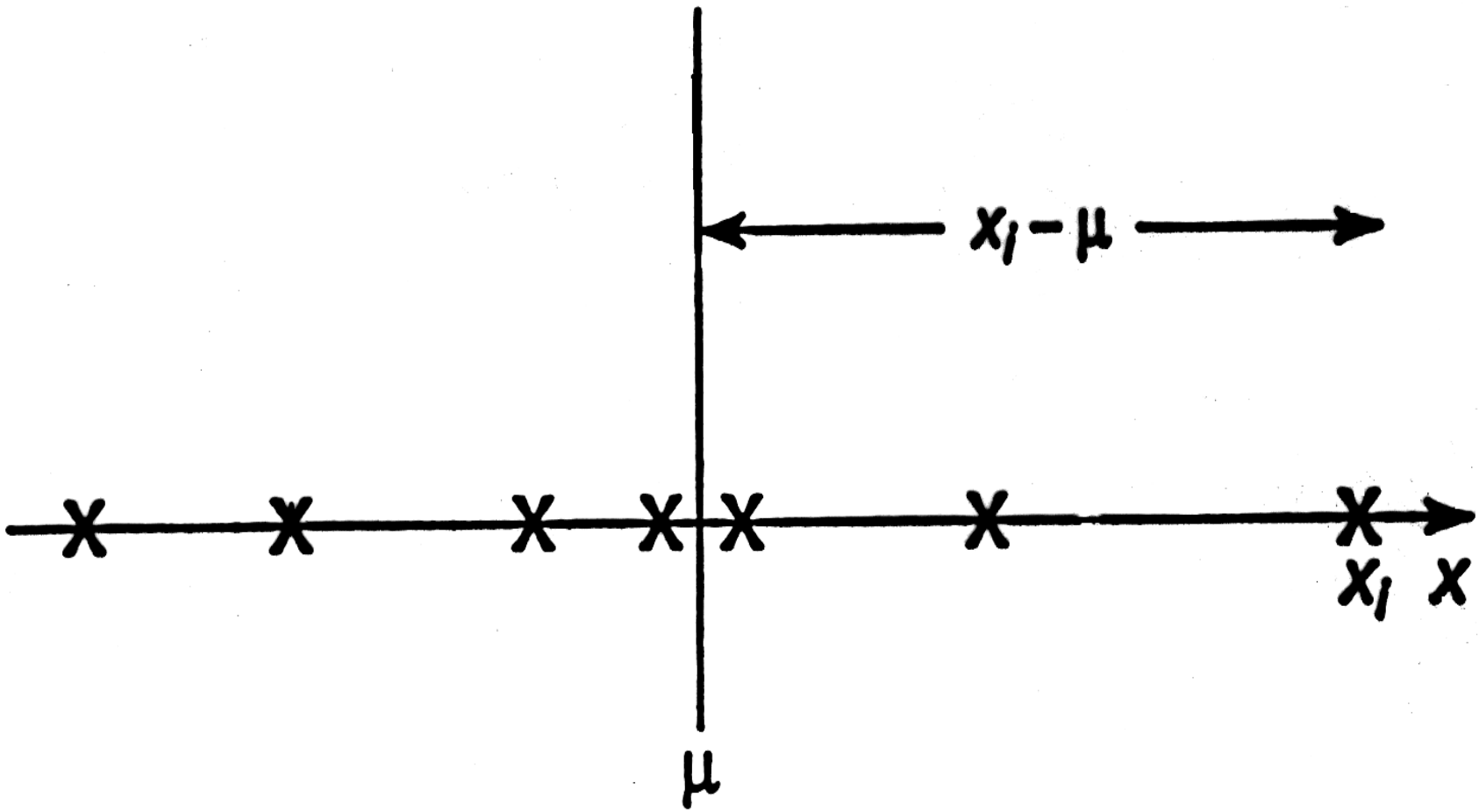
2.1.2.1. Mere razpršenosti

- Spremenljivost (varianca)

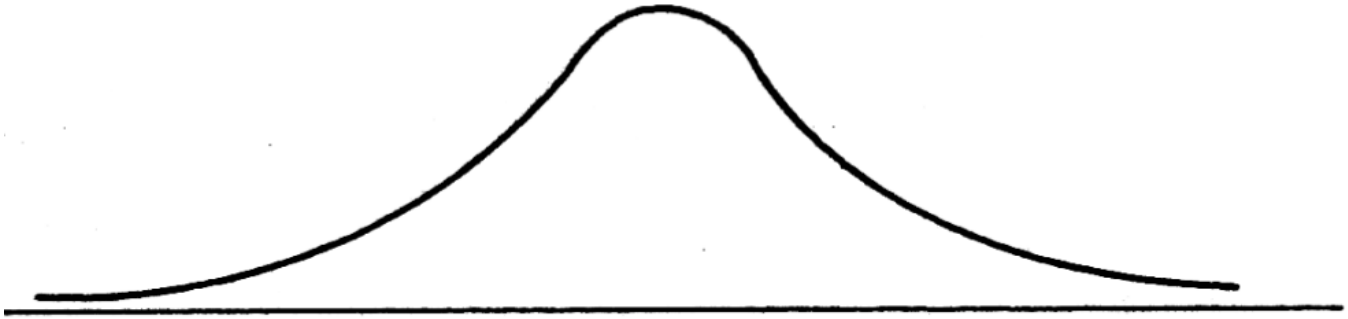
$$\sigma^2 = ((X_1 - \mu)^2 + \ldots + (X_N - \mu)^2)/N = {}^1/_N \sum_{i=1}^{N} (X_i - \mu)^2$$

- Enote so kvadrirane enote enot meritev; zato navadno uporabimo kvadratni koren variance, ki ima enake enote kot meritev – standardni odklon (deviacija)

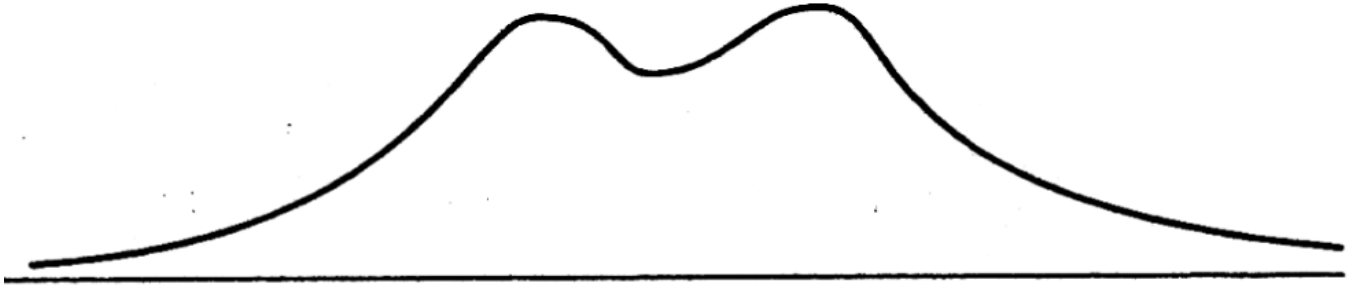
$$\sigma = \sqrt{\sigma^2}$$


$x_i - \mu$
x_i
x
μ

2.1.2.3. Opis oblike



- Število modusov
  - Unimodalna
  - Bimodalna

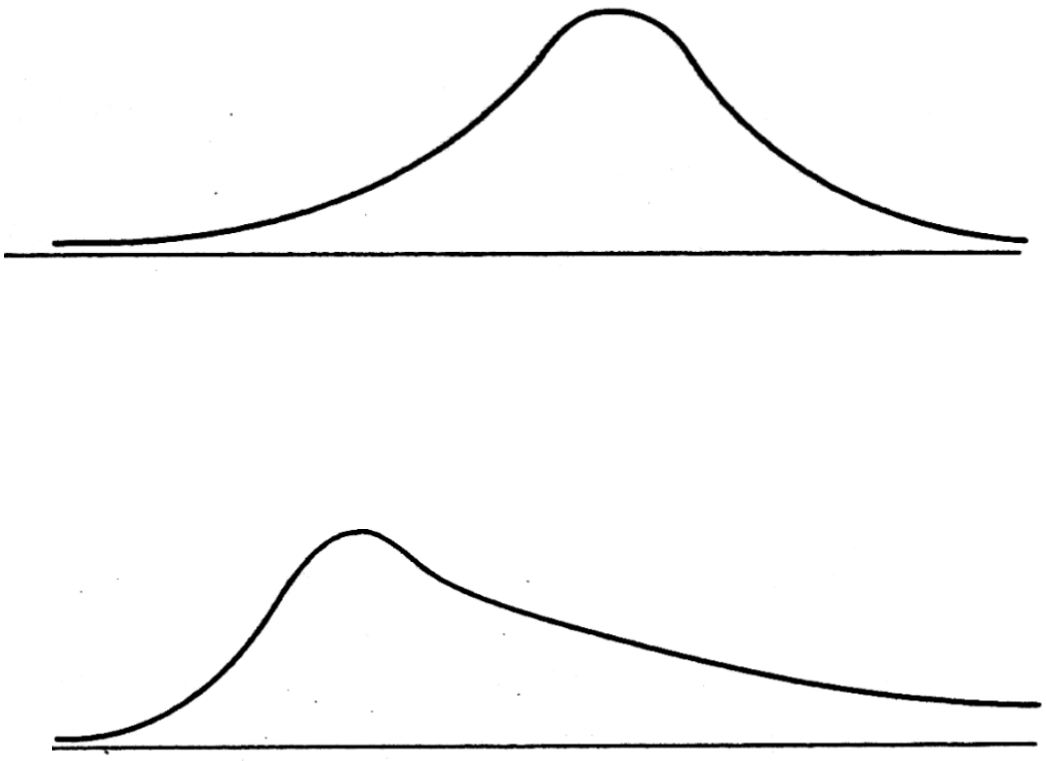


2.1.2.3. Opis oblike

- Asimetričnost (skewness)
  - Pozitivno (desno) premaknjena

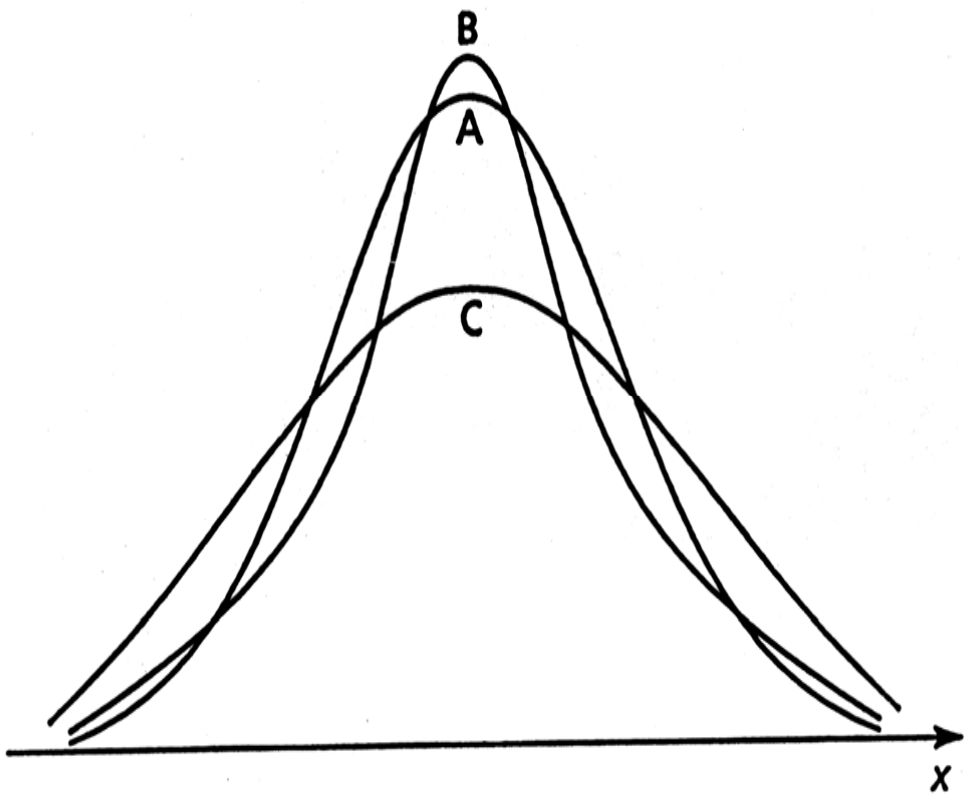
  $\mu > Me$
  - Negativno (levo) premaknjena

  $\mu < Me$



2.1.2.3. Opis oblike

- Stopnja pogostnosti (kurtosis)
  - Sploščenost
  - Zašiljenost



2.1.3. Statistike frekvenčne porazdelitve vzorca

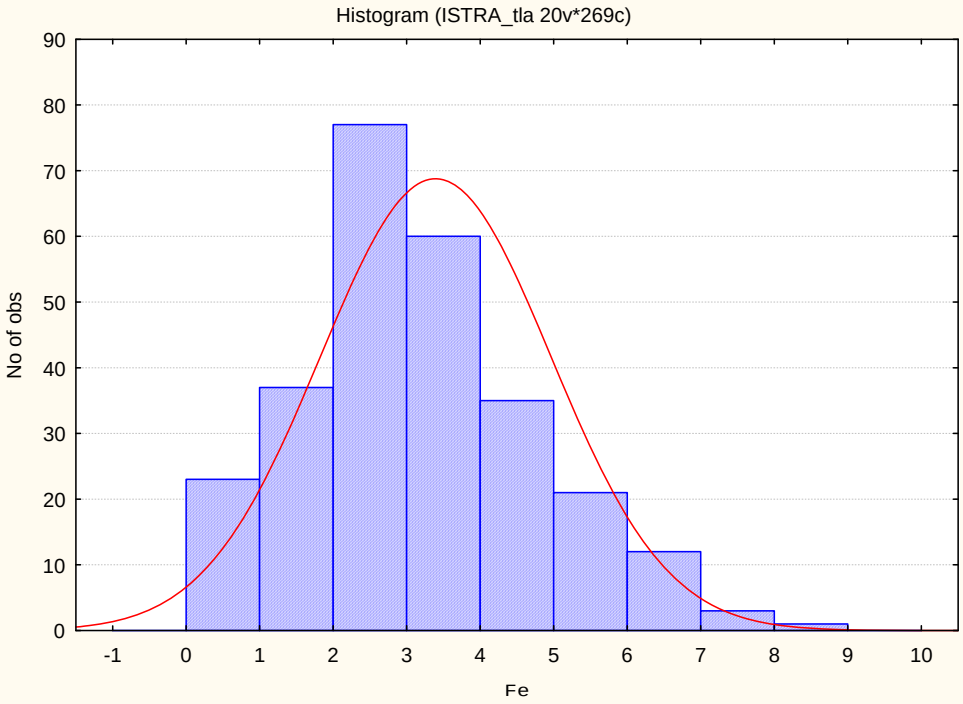
- Vedno razpolagamo z meritvami vzorca, iz katerih z izračunom statistik ocenimo parametre (značilnosti) populacije.
- Različni vzorci iste populacije dajo različne rezultate – nikoli ne vemo, katere vrednosti parametrov so “prave”.

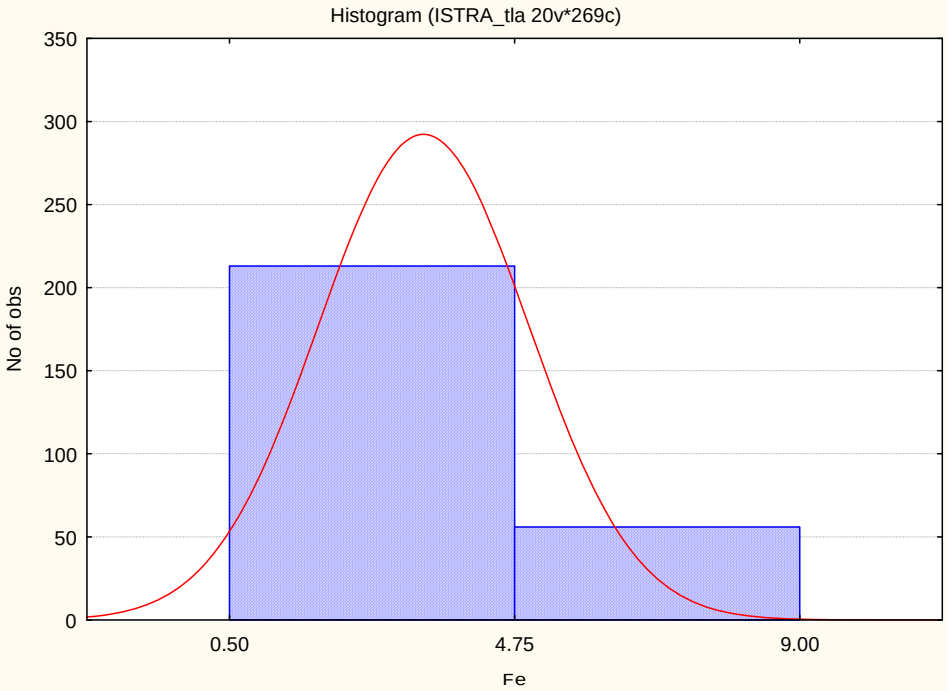
2.1.3.1. Histogram vzorca

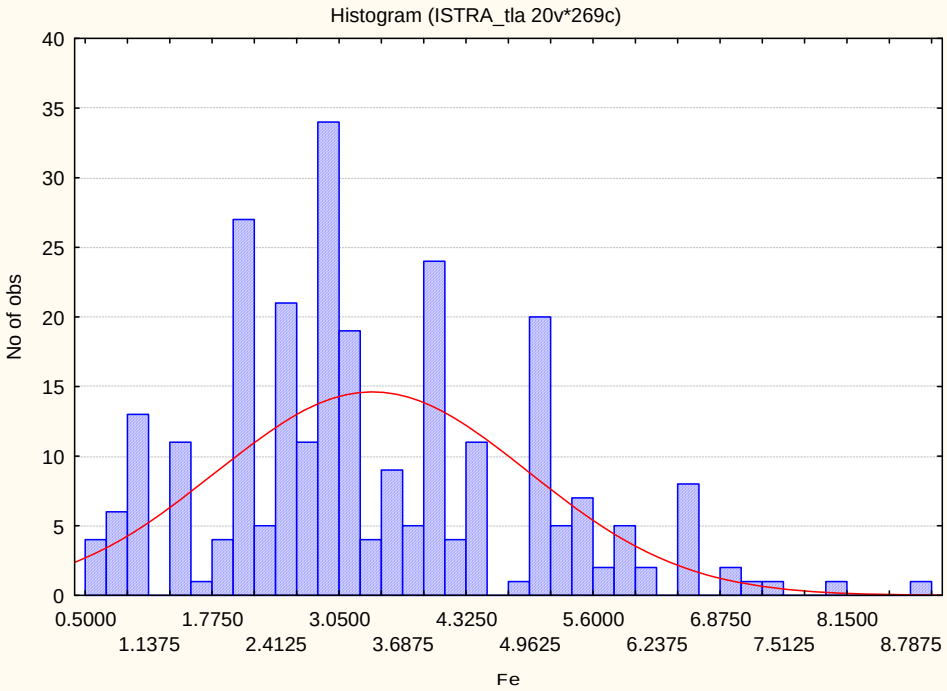
- Ugotavljanje pravilnosti porazdelitve
- Odločitev o nadaljnih postopkih
- Neposredna (geološka) interpretacija

2.1.3.1. Histogram vzorca

- Pravila izdelave histograma
  - Število razredov
    - Dovolj veliko, da postene oblika porazdelitve vidna
    - Dovolj majhno, da ni praznih razredov
    - Kvadratni koren števila opazovanj - $\sqrt{n}$
  - Meje razredov
    - Vključene morajo biti vse opazovane vrednosti
    - Nobeno opazovanje naj ne sovpada z mejo $\Rightarrow$ imajo eno decimalno mesto več kot meritve
    - Enake širine $\Rightarrow$ višine in površine stolpcev so sorazmerne frekvencam razredov – razpon podatkov delimo s številom razredov
  - Frekvence razredov ugotovimo z uvrščanjem (štetjem) opazovanj v posamezen razred


Histogram (ISTRA_tla 20v*269c)
No of obs
Fe
90
80
70
60
50
40
30
20
10
0
-1
0
1
2
3
4
5
6
7
8
9
10


Histogram (ISTRA_tla 20v*269c)
No of obs
Fe
350
300
250
200
150
100
50
0
0.50
4.75
9.00


Histogram (ISTRA_tla 20v*269c)
No of obs
Fe
40
35
30
25
20
15
10
5
0
0.5000
1.1375
1.7750
2.4125
3.0500
3.6875
4.3250
4.9625
5.6000
6.2375
6.8750
7.5125
8.1500
8.7875

2.1.3.2. Mere položaja vzorca

- Aritmetična sredina ≈ povprečje
- Geometrijska sredina
- Mediana
- Modus

2.1.3.2. Mere položaja vzorca Aritmetična sredina

- Običajna mera opisovanja vzorcev, čeprav nanjo vplivajo skrajne vrednosti.

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

Primer:

I. 1, 2, 2, 3 $\bar{x} = 2$

II. 1, 2, 2, 30 $\bar{x} = 8{,}75$

2.1.3.2. Mere položaja vzorca Geometrijska sredina

- Uporabna pri pozitivno premaknjenih podatkih, kjer je aritmetična sredina nereprezentativno visoka.

$$\overline{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

Primer:

I. 1, 2, 2, 3 $\overline{x}_G = 1{,}86$

II. 1, 2, 2, 30 $\overline{x}_G = 3{,}31$

2.1.3.2. Mere položaja vzorca Mediana

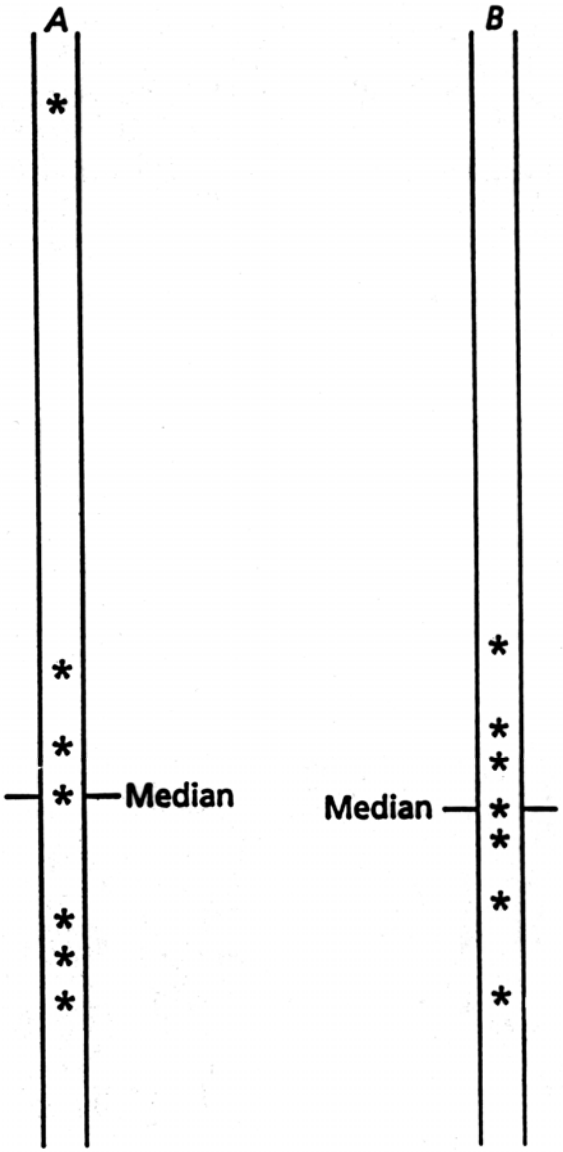
- Vrednost (ali aritmetična sredina dveh) na sredini niza podatkov, razvrščenih v naraščajočem (ali padajočem) vrstnem redu.
- Nanjo ne vplivajo izredno visoke ali nizke vrednosti. Uporabna je pri podatkih s sorazmerno, razvrstitveno (ordinalno) in intervalno lestvico. Je mera srednje vrednosti pri neparametrični statistiki.

2.1.3.2. Mere položaja vzorca
Mediana

Primer:

I. 1, 2, 2, 3 Me = 2

II. 1, 2, 2, 30 Me = 2



2.1.3.2. Mere položaja vzorca Modus

- Je najpogostejša vrednost (ali več vrednosti) v nizu podatkov – najvišja točka krivulje frekvenčne porazdelitve.
- Pri majhnem številu opazovanj modusa ni, ali pa je več navideznih.

Primer:

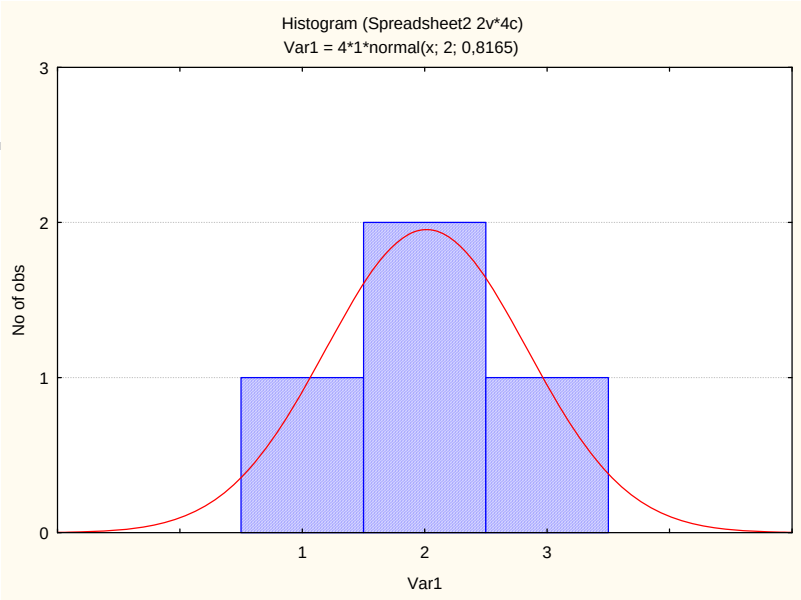
I. 1, 2, 2, 3 Mo = 2

II. 1, 2, 2, 30 Mo = 2

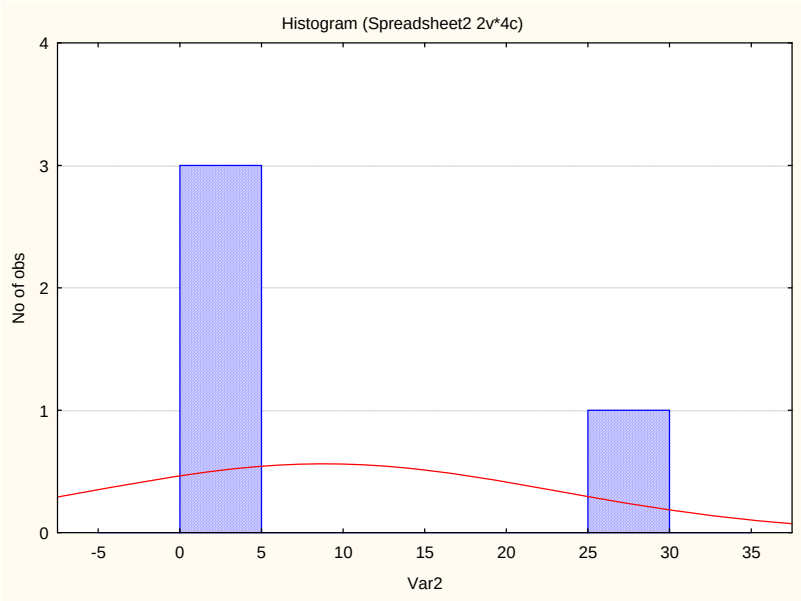
2.1.3.2. Mere položaja vzorca
Modus

Primer:

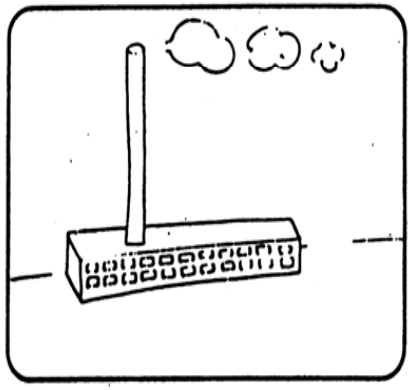
I. 1, 2, 2, 3 Mo = 2



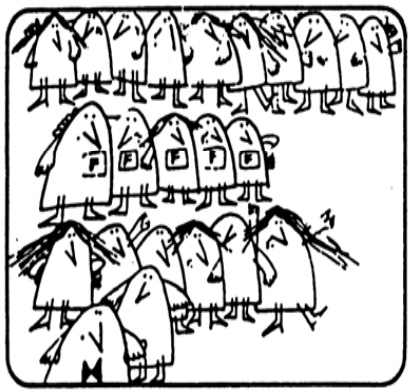
II. 1, 2, 2, 30 Mo = 2



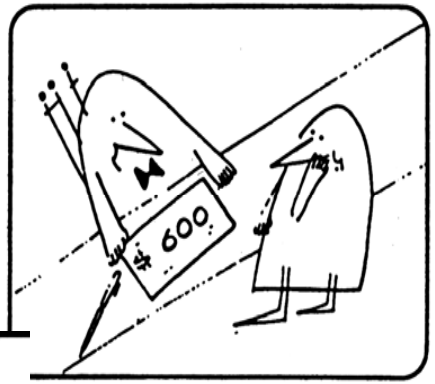
Varljivi »povpreček«



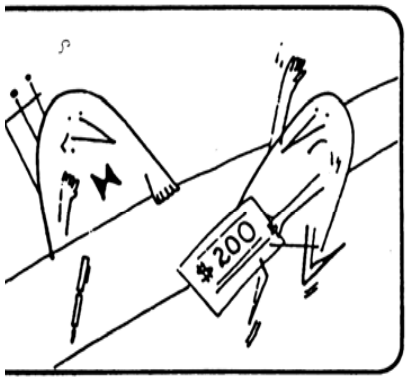
Družina Gismo ima majhno tovarno, v kateri izdelujejo mehanske nadizume.



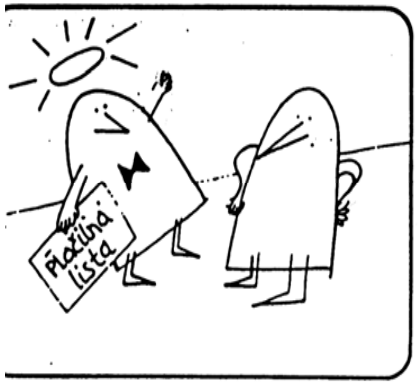
Tovarno vodijo Gismo, njegov brat in šest sorodnikov, v njej pa dela pet preddelavcev in deset delavcev. Posel gre dobro in tovarna potrebuje novega delavca.



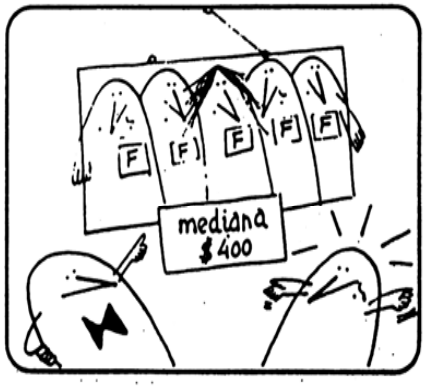
Gismo se glede službe pogovarja s Samom.
Gismo: Tukaj so zelo dobre plače – povprečna je 600 dolarjev na teden. Med poskusno dobo boste dobivali po 150 dolarjev na teden, toda to se bo kmalu zvišalo.



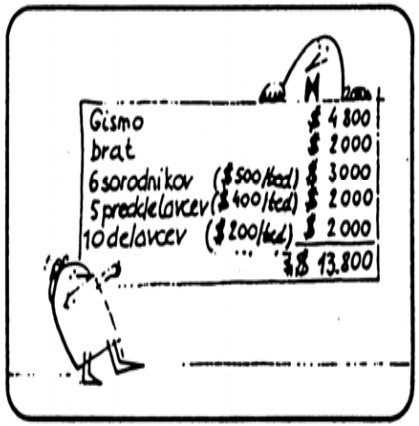
Potem ko je Samo nekaj dni delal, je prišel k šefu.
Samo: Osleparili ste me! Preveril sem plače drugih delavcev in ugotovil, da niti eden ne dobiva več kot 200 dolarjev na teden. Kako je potem lahko povprečna plača 600 dolarjev na teden?



Gismo: No, Samo, ne razburjajte se. Povprečna plača je 600 dolarjev in to vam bom dokazal.

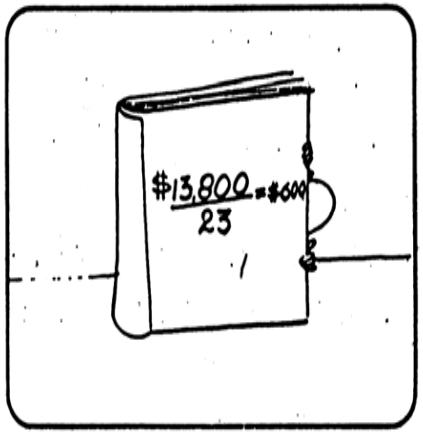


Gismo: To pa ni res. Samo vi me niste razumeli. Lahko bi naštel plače po vrsti in vam povedal, da je srednja plača 400 dolarjev, toda to ni povprečna; to je *mediana*.

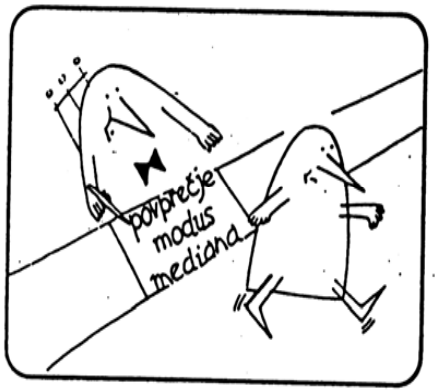


Tu je poročilo o tem, kar vsak teden izplačamo. Jaz dobim 4800 dolarjev, moj brat 2000 dolarjev, vsak od mojih šestih sorodnikov po 500 dolarjev, vsak preddelavec 400 dolarjev in vsak delavec po 200 dolarjev. To znese 13 800 dolarjev na teden za 23 ljudi. Ali ne?

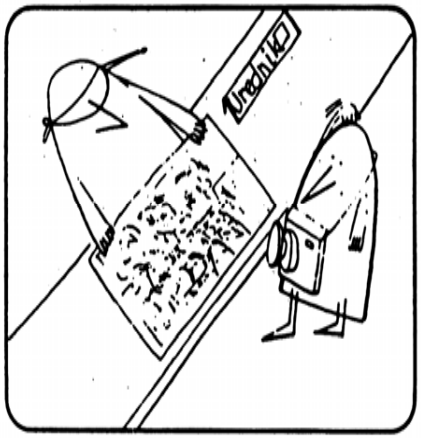
Samo: Kaj pa je potem plača 200 dolarjev?
Gismo: To imenujemo *modus*. To je plača, ki jo dobi *večina* ljudi.



Samo: Dobro, dobro, prav imate. Povprečna plača je res 600 dolarjev na teden, a vi ste me *vseeno* osleparili.



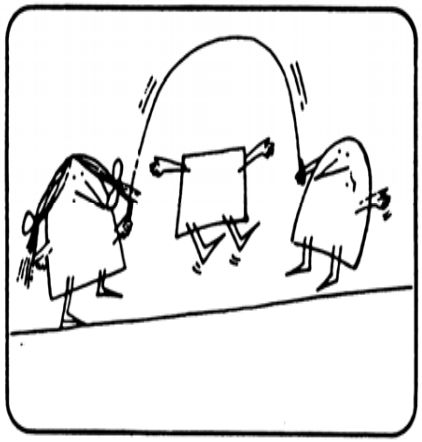
Gismo: Vaša nesreča, dragi moj, je v tem, ker ne poznate razlike med povprečjem, mediano in modusom.
Samo: No, zdaj jo poznam. Zato si kar koga drugega poiščite za to službo.



Urednik je bil s sliko zadovoljen. **Urednik:** Imenitno sliko si naredil, Bregar. Zdaj pa imam zate drug posel. Naredi mi sliko povprečno velike družine iz tega mesta.

Mati leta

Ob koncu leta je Samova žena dobila od župana nagrado kot mati leta.



Bregar pa tega ni mogel narediti. Zakaj? Ker niti ena družina v mestu ni bila povprečna! Izračunali so namreč, da je povprečno število otrok $2\frac{1}{2}$!

Lokalni časopis je priobčil sliko Sama, njegove žene in njunih 13 otrok.

1. Če hoče urednik sliko »tipične« družine – v smislu modusa – ali fotograf vedno lahko najde tako družino,

2. Ali je možno, da je več kot en modus? Ali je lahko vsaka od dveh družin, od katerih ima ena dva otroka, druga pa tri, primer modusa?

3. Ali bi urednik vedno lahko našel družino – mediano, če bi želel njeno sliko?

(Da, tipični primer očitno obstaja.)

(Da, če je v mestu 1476 družin z dvema otrokoma in 1476 družin s tremi otroki, vse druge družine v mestu pa imajo več ali manj otrok; potem bi mesto imelo dva tipa družin, od katerih bi bil vsak pravi modus.)

(Običajno bi lahko, ne pa vedno. Če v mestu živi sodo število družin in se osrednji dve družini razlikujeta po številu otrok, potem mediana ni nujno celo število.)

2.1.3.3. Mere razpršenosti vzorca

- Uporaba je za:
  - Neposredna geološka interpretacija
  - Določanje števila podatkov, potrebnih za zanesljivo oceno parametrov
- Vrste mer razpršenosti:
  - Razpon (range)
  - Odstopanje znotraj kvartilov (IQD)
  - Mere odstopanja od povprečja – varianca (s^2) in standardni odklon (s)

2.1.3.3. Mere razpršenosti vzorca Razpon

- Razlika med najvišjo (max.) in najnižjo (min.) vrednostjo vzorca.

$$razpon = x_{max} - x_{min}$$

- Zelo zanesljiva mera pri majhnem številu opazovanj; pri velikem lahko nanj vplivajo izstopajoče vrednosti.

Primer:

I. 1, 2, 2, 3 razpon = 3 – 1 = 2
II. 1, 2, 2, 30 razpon = 30 – 1 = 29

2.1.3.3. Mere razpršenosti vzorca Odstopanje znotraj kvartilov

- IQD (Intra quartile deviation) je območje med 25% in 75% vrednosti (razpona).

  $IQD = Q_3 - Q_1$

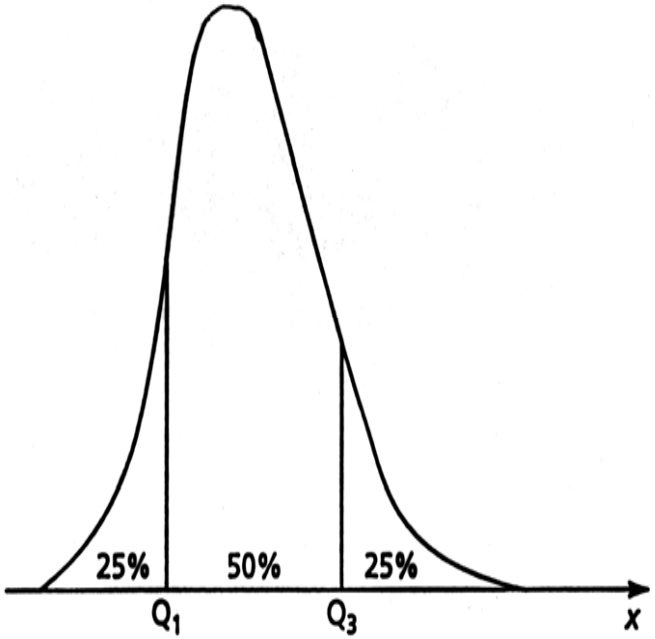
- Folkovi parametri so različica, ki se uporablja v sedimentologiji.

2.1.3.3. Mere razpršenosti vzorca
Odstopanje znotraj kvartilov

Primer:

I. 1, 2, 2, 3 IQD = 1

II. 1, 2, 2, 30 IQD = 14,5



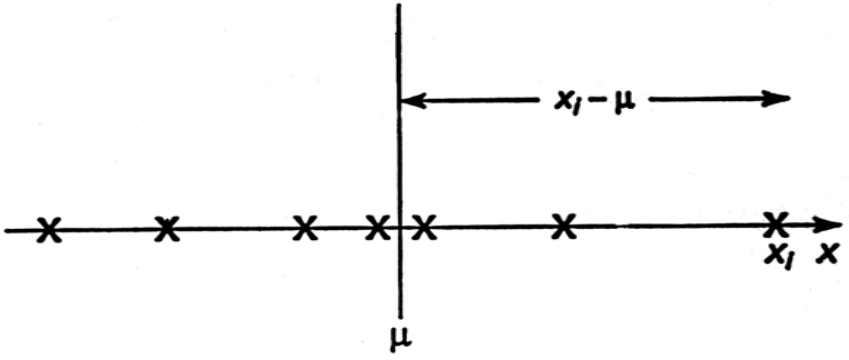
$$Q_1 = \frac{razpon \cdot 25}{100} = \frac{(3-1)\cdot 25}{100} = 0{,}5$$

$$Q_3 = \frac{razpon \cdot 75}{100} = \frac{(3-1)\cdot 75}{100} = 1{,}5$$

2.1.3.3. Mere razpršenosti vzorca Odstopanje od povprečja

- Temeljijo na sipanju podatkov okrog srednje vrednosti vzorca.
- Ker se negativna in pozitivna odstopanja izničijo, za izračun srednje vrednosti odstopanj uporabimo kvadrirane vrednosti (kvadrate) odstopanj. Popravljena vsota kvadratov (CSS) je:

$$CSS = \sum_{i=1}^{n} (x_i - \bar{x})^2$$



2.1.3.3. Mere razpršenosti vzorca Varianca vzorca

- Na velikost CSS vpliva število opazovanj v vzorcu (n), zato oceno uporabimo varianco (s^2). Ta za razliko od σ^2 ni asimetrična, zato nekoliko podceni parameter populacije. Problem rešimo tako, da namesto z n, CSS delimo z (n-1).

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

2.1.3.3. Mere razpršenosti vzorca Standardni odklon vzorca

- Enote variance so kvadrirane, zato pogosto uporabimo kvadratni koren variance – standardni odklon (s), kjer so enote enake meritvam in ga je zato enostavneje interpretirati.

$$s = \sqrt{s^2}$$

2.1.3.3. Mere razpršenosti vzorca Odstopanje od povprečja

Primer:

I. 1, 2, 2, 3

$$s^2 = \frac{(1-2)^2 + (2-2)^2 + (2-2)^2 + (3-2)^2}{4-1} = 0{,}816 \qquad s = \sqrt{0{,}816} = 0{,}903$$

II. 1, 2, 2, 30

$s^2 = 14{,}174$ $s = 3{,}765$

2.1.3.4 Opis in razlaga oblike histograma

- Momenti

  Za n vrednosti $x_1, x_2, \ldots, x_n$ je r-ti moment definiran kot povprečje r-tih potenc odklonov od $\bar{x}$.

$$m_r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^r}{n}$$

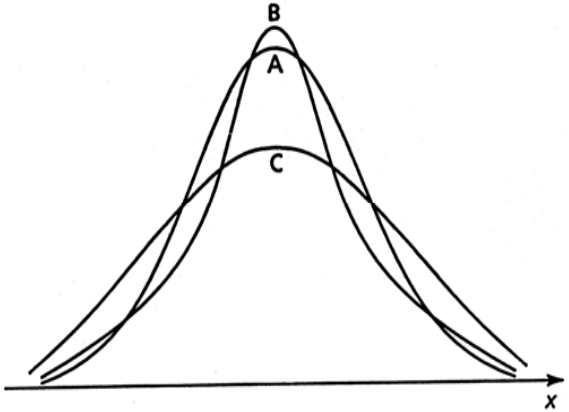
$$m_2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

2.1.3.4 Opis in razlaga oblike histograma

- Poševnost (A, $\sqrt{b_1}$) je mera asimetričnosti (skewness). Ko je krivulja simetrična je A = 0.

$$A = \sqrt{b_1} = \frac{m_3}{\sqrt{m_2^3}}$$

- Sploščenost (E, b_2) je mera zašiljenosti (leptokurtična E>3) ali sploščenosti (platikurtična E<3). Normalna krivulja je mezokurtična (E=3).



$$E = b_2 = \frac{m_4}{m_2^2} - 3 = \left(\frac{m_4}{m_2^2}\right)^2$$

2.1.3.4 Opis in razlaga oblike histograma

- Normalna porazdelitev

  je posebna unimodalna, simetrična porazdelitev, izpeljana iz verjetnostne teorije, na osnovi naključnega odstopanja okrog prepostavljene srednje vrednosti; tako stanje je v naravi pogosto.
- Mešanje (prekrivanje)

  Porazdelitev razstavimo v več normalnih porazdelitev z različnimi srednjimi vrednostmi in/ali standardnimi odkloni. Postopek je upravičen pri jasno razvidni bimodalnosti; v primeru nekaterih bimodalnosti, zašiljenosti in sploščenosti ga lahko napačno uporabimo.

2.1.3.4 Opis in razlaga oblike histograma

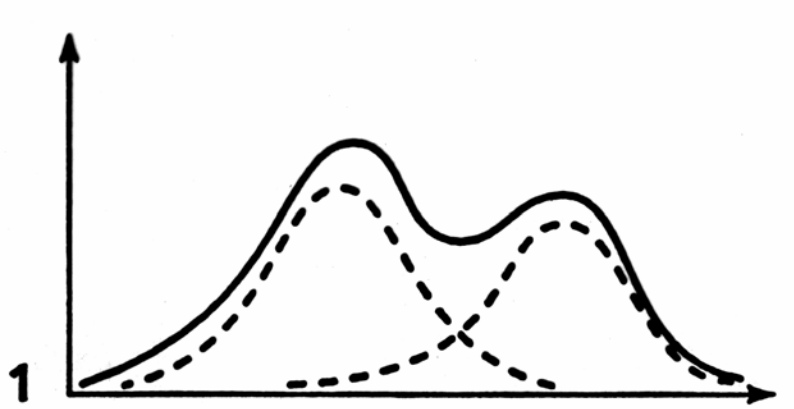
- Outlier

  je podatek z nenormalno (visoko) vrednostjo. Je posledica napake v postopku ali kakega drugega lažnega vpliva. Kadar so pristni, jih razlagamo kot predstavnike manjše populacije z izjemnimi vrednostmi.
- Zaprtost podatkov

  Spreminjanje količine glavne komponente povzroča pzitivno asimetričnost podrejene spremenljivke in obratno. Uporabimo logaritmirane podatke.
- Potence

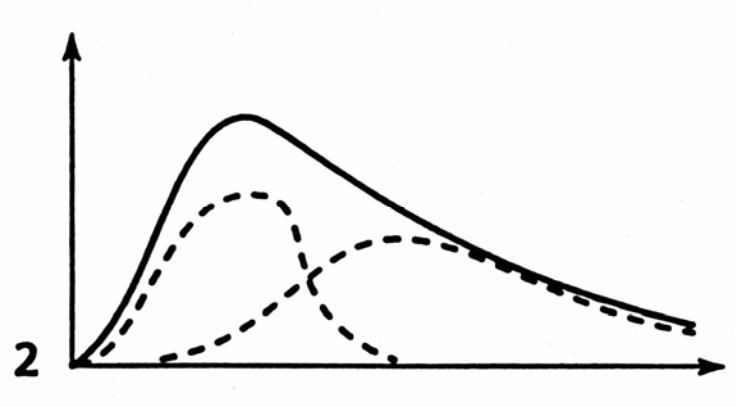
  Nekatere procese v naravi najboljše opišemo s funkcijami, ki vključujejo potence. Vsako potenciranje normalno porazdeljene spremenljivke povzroči asimetričnost.

2.1.3.4 Opis in razlaga oblike histograma



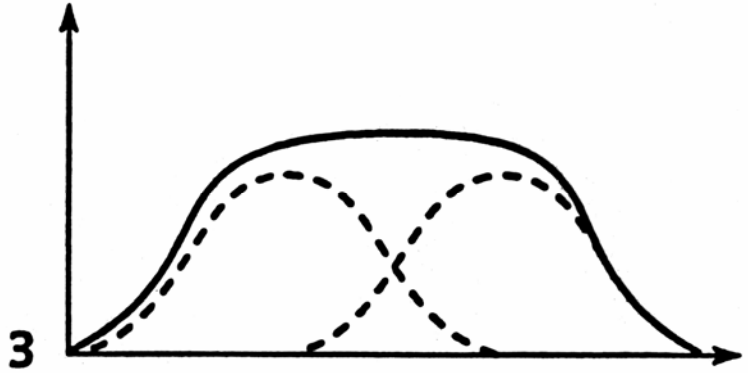
- Bimodalnost:

  mešanje dveh podpopulacij z različnima srednjima vrednostima.



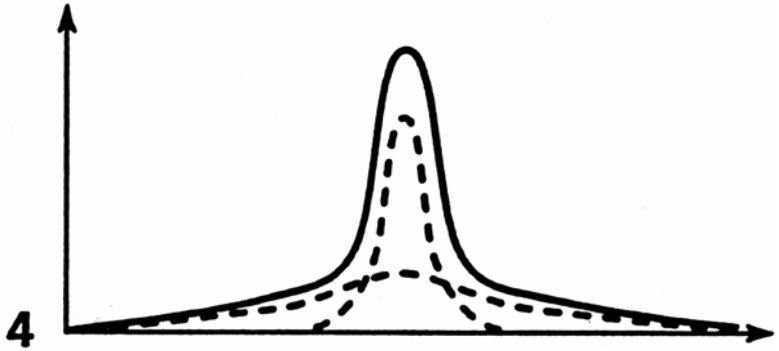
- Pozitivna asimetričnost: mešanje dveh podpopulacij z različnima srednjima vrednostima in standardnima odklonoma.

2.1.3.4 Opis in razlaga oblike histograma



- Sploščenost:

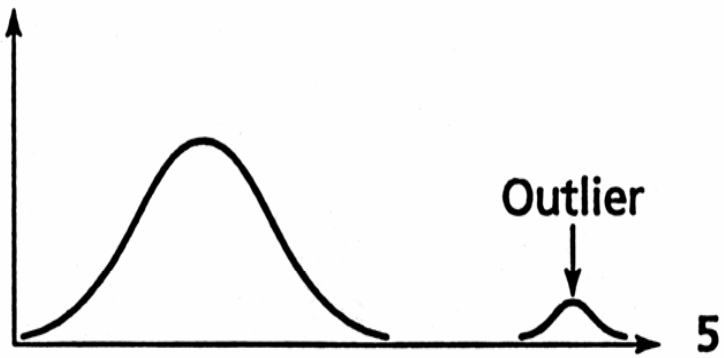
  mešanje dveh podpopulacij z različnima standardnima deviacijama.

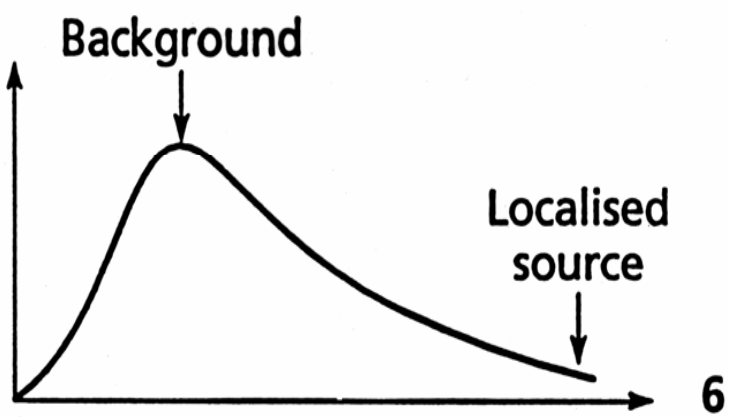


- Zašiljenost:

  mešanje dveh podpopulacij z različnima standardnima deviacijama.

2.1.3.4 Opis in razlaga oblike histograma





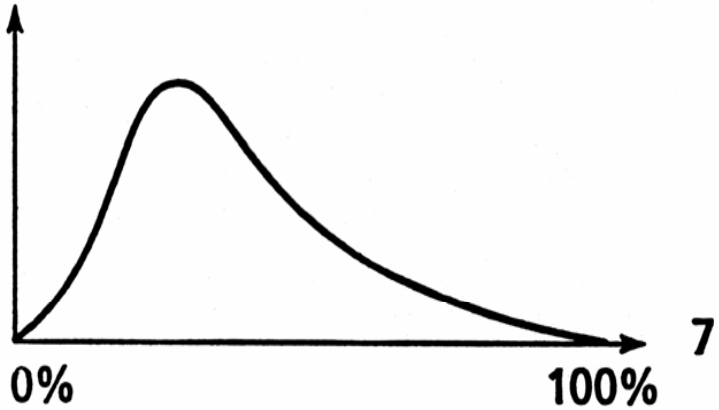
- Outlier:

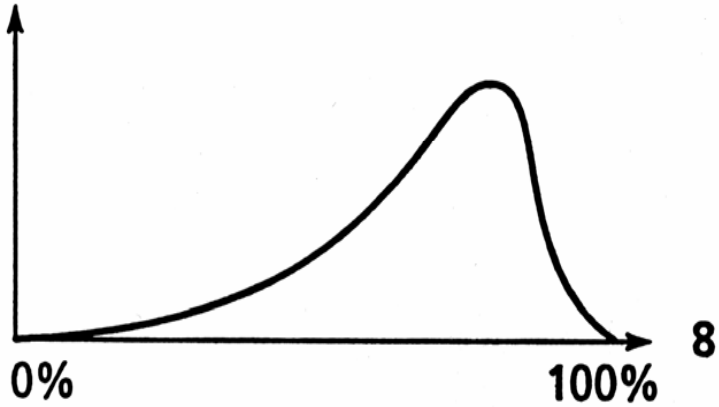
  zelo majhna podpopulacija z drugačno srednjo vrednostjo ali napačna oz. anomalna meritev.

- Pozitivna asimetričnost:

  V geokemiji redek, lokalni pojav visoke koncentracije, morda blizu izvora prvine (orudenje, onesnaženje, vulkan...)

2.1.3.4 Opis in razlaga oblike histograma





- Pozitivna asimetričnost:

  pri zaprtih setih podatkov se ta odraža pri podrejenih komponentah, zaradi razredčenja z glavno prvino.

- Negativna asimetričnost:

  pri zaprtih setih podatkov je to obraten učinek zgoraj opisanega, ki se pojavi pri glavni komponenti.

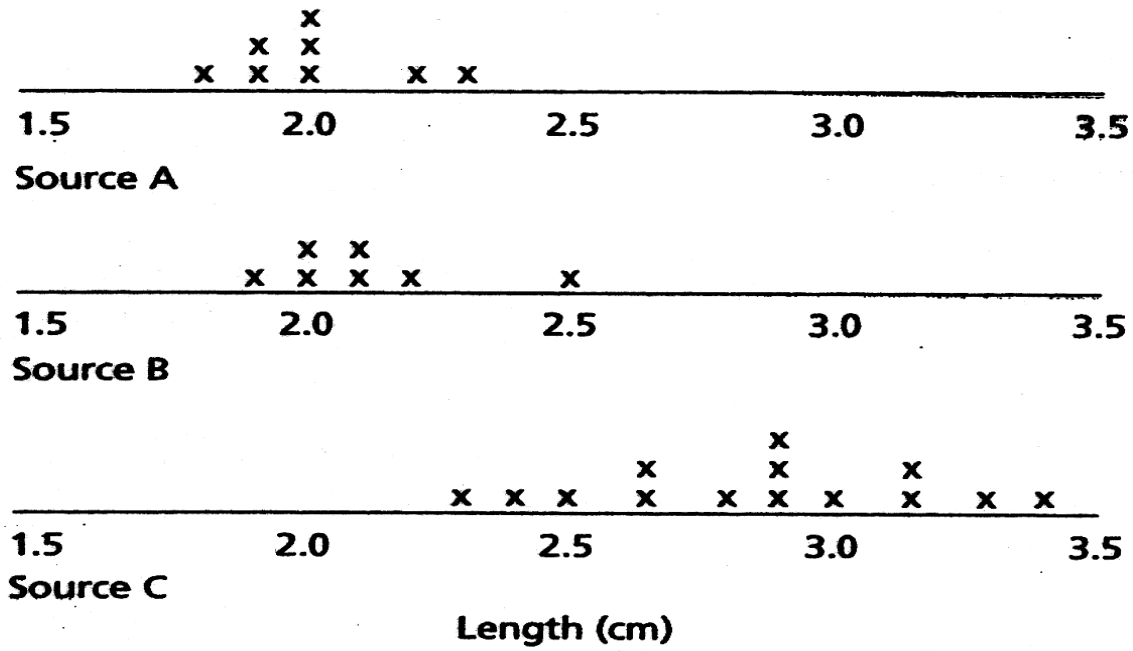
2.2. GRAFIČNO PREISKOVANJE PODATKOV

- Podatki vključujejo napačne ali zavajujoče vrednosti
- Narava spremenljivosti je očitna
- Ni potrebna dodatna analiza – (geološko) zanimivi učinki so vidni že na prvi pogled; testiramo le njihovo statistično značilnost
- Podatki nakazujejo (geološko) hipotezo ali model, ki ju testiramo z novimi, neodvisnimi podatki.

2.2. GRAFIČNO PREISKOVANJE PODATKOV

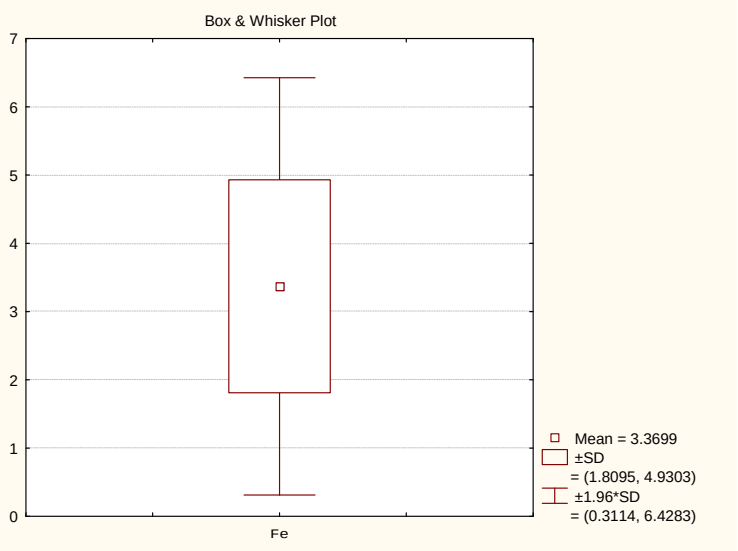
- Točkovni graf

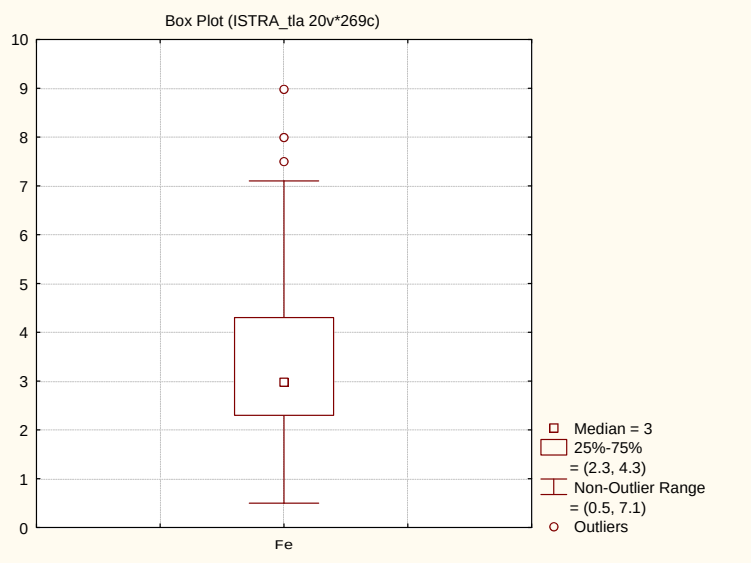
  Podatki so prikazani kot točke na premici; ponovljene vrednosti se kopičijo ena vrh druge.



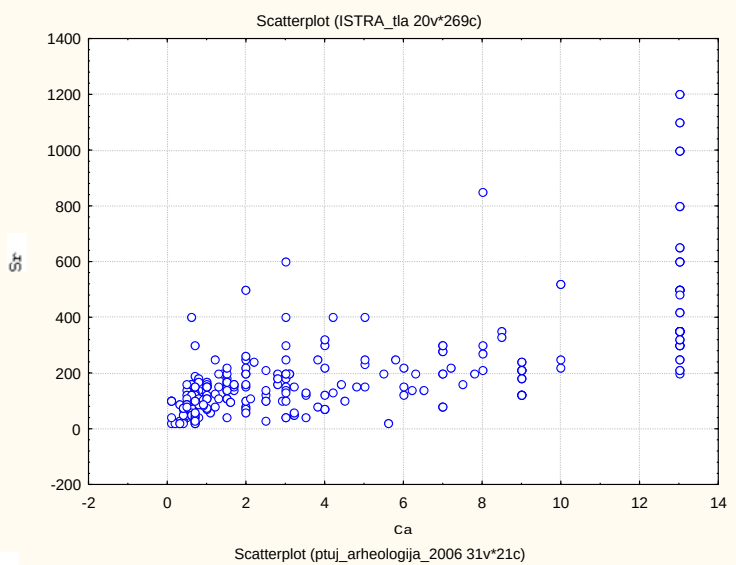
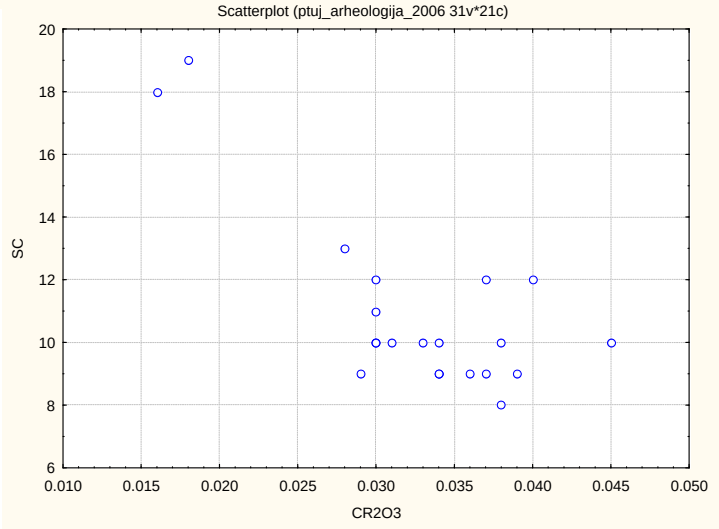
2.2. GRAFIČNO PREISKOVANJE PODATKOV

- Škatla – črta (box and whisker) graf
  - Uporaben za primerjavo nizov podatkov ter odkrivanje “sumljivih” vrednosti.
  - Glavnina podatkov je v škatli, “repi” porazdelitve pa so črte, outlierji točke.





2.2. GRAFIČNO PREISKOVANJE PODATKOV

- Razsevni (scatter) diagrami

  Na vsaki nosi je nanešena ena spremenljivka. Pokažejo odnos (povezavo) med spremenljivkama in izstopajoče vrednosti.