OSNOVE STATISTIKE

UVOD

• **DEFINICIJA: Statistika** je grana matematike koja obuhvaća sakupljanje, analizu, interpretaciju i prezentaciju podataka te izradu predviđanja koja se temelje na tim podacima.

• Smatra se granom matematike, veliku važnost u korištenju statistike imaju i **planiranje i provođenje pokusa**, odnosno **skupljanje podataka** koji će se analizirati (negativni primjer: *Hawthorne study*), ali i **interpretacija** dobivenih rezultata (lažna korelacija)!

• Navodno su prve statističke metode korištene čak u 5 stoljeću p.n.e.

• Najstariji zapisi o korištenju statistike potječu iz 9. stoljeća (arapski znanstvenik Al-Kindi u svrhu izučavanja kodiranih poruka).

• U 14 stoljeću nastaju zapisi *Nuova Cronica* (povijest Firenze) sadrže niz statističkih podataka o populaciji, edukaciji i sl.

• Matematički razvoj ide usporedno s razvojem teorije vjerojatnosti

UVOD

• Pojam statistika je prvobitno izveden iz latinskog izraza *statisticum collegium* (vijeće država) te talijanske riječi ***statista*** (državnik ili političar).

• Njemačka riječ *Statistik* uvedena od Gottfrieda Achenwalla (1749 god.) je originalno značila analizu podataka o državi.

• Značenje sakupljanja i analize podataka statistika je dobila početkom 19. stoljeća, a riječ je u engleski jezik uveo Sir John Sinclair.

• Statistiku dijelimo na **deskriptivnu** i **induktivnu** te **matematičku** i **egzaktnu**.

PODJELA STATISTIKE

• **Deskriptivna statistika** (engl. *descriptive statistics)* bavi se organizacijom sakupljenih podataka te njihovim **sažetim opisom** s pomoću numeričkih i grafičkih prikaza.

• **Induktivna statistika** (engl. *inferential statistics*) bavi se izvođenjem zaključaka o populaciji na temelju svojstava uzorka.

• **Matematička statistika** je proučavanje statistike s matematičke točke gledišta korištenje teorije vjerojatnosti, matematičke analize i linearne algebre.

• **Egzaktna statistika** je grana statistike koja daje točne rezultate za pripadne statističke testove.

• poddiscipline statistike korištene u **prirodnim znanostima**: biostatistika, kemometrika, *data mining*, …

OSNOVNI POJMOVI

• **Populacija** (engl. *population*) je skup realnih ili hipotetskih objekata koji nas zanima.

• Populacija može imati **konačno** ili **beskonačno** mnogo objekata. Npr. populacija svih studenata kemije druge godine Preddiplomskog studija kemije ili populacija eksperimentalnih mjerenja koja bi sadržavala sve rezultate koji bi mogli biti opaženi ako se mjerenja provedu beskonačno mnogo puta pod istim uvjetima.

• **Varijabla** (engl. *variable*) je neko svojstvo svakog člana populacije (kontinuirane i diskretne varijable; dimenzionalnost).

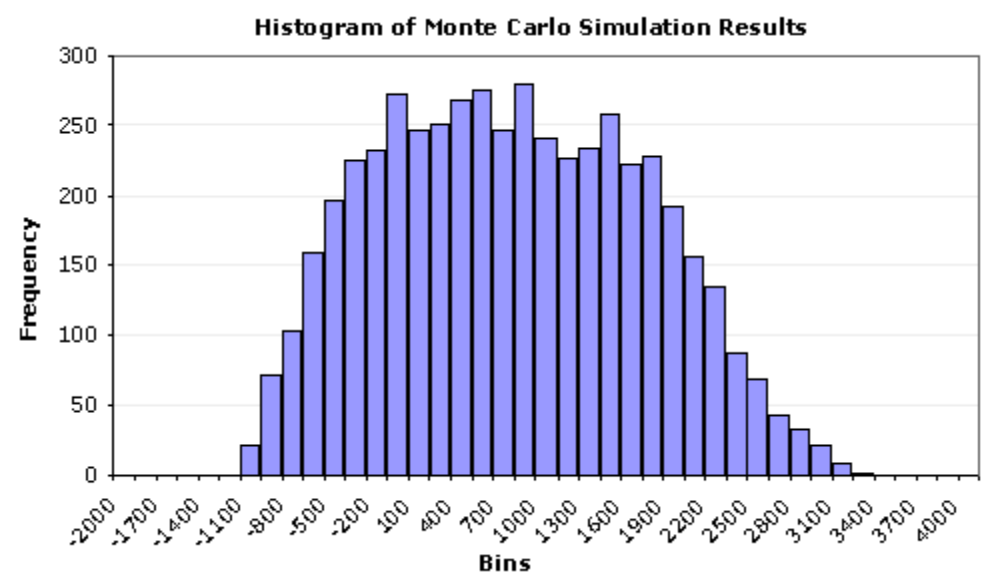
• **Uzorak** (engl. *sample*) je skup opaženih rezultata.

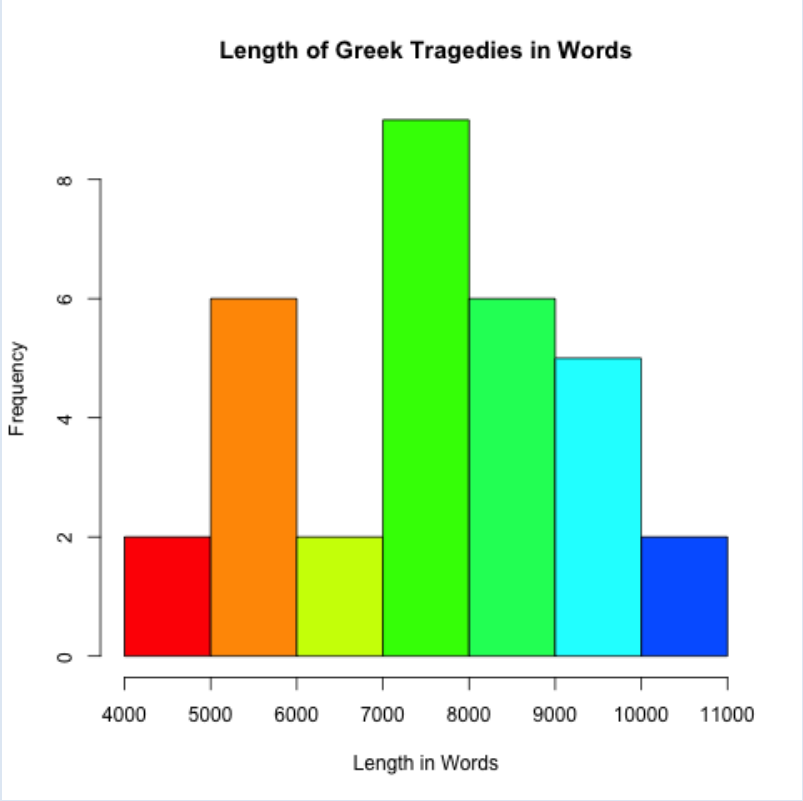
• **Cenzus** (lat. *census*, u doba antičke Rimske republike popis svih odraslih muškaraca sposobnih za vojnu službu) je poznavanje podataka o svim objektima populacije. Cenzus rijetko postoji jer je prikupljanje svih podataka najčešće ili nemoguće ili preskupo.

OSNOVNI POJMOVI

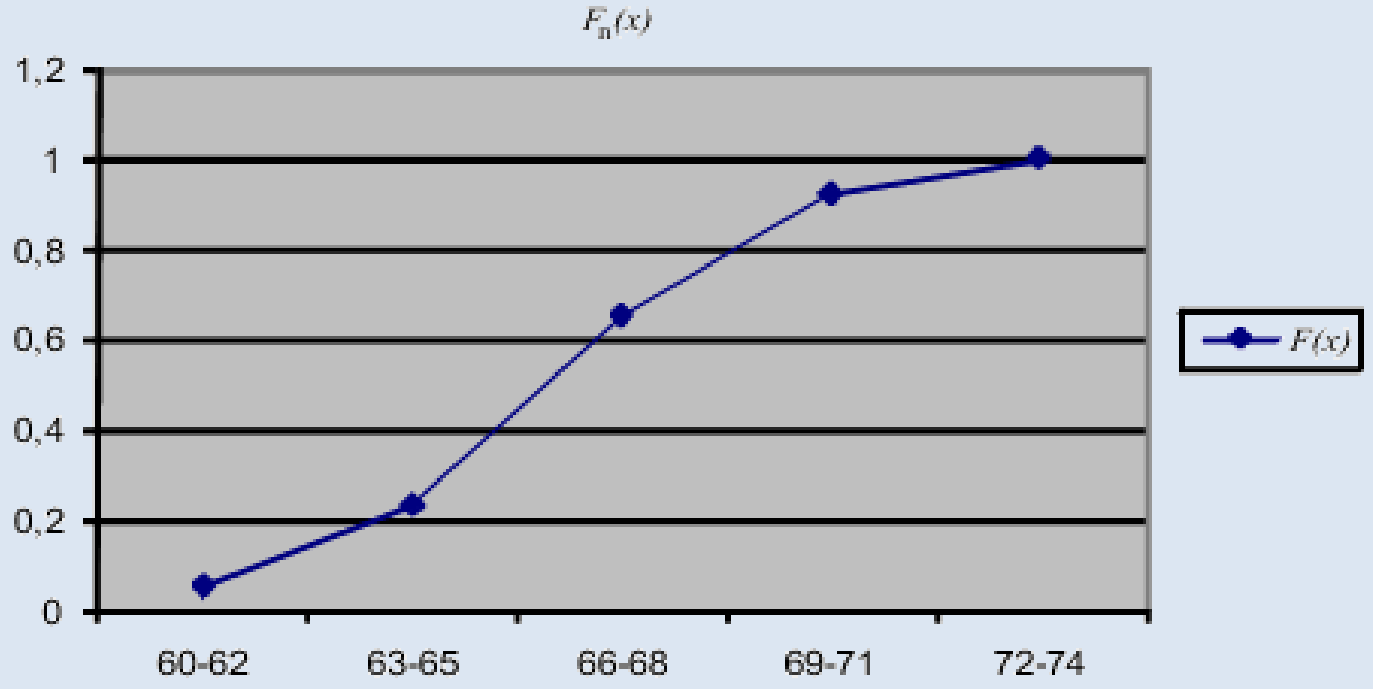
- Promatramo li empirijske podatke, često će se među njima javiti mjerenja jednakih vrijednosti
- **Frekvencija** (f_i) nam kaže koliko se puta vrijednost neke varijable javila u uzorku ili populaciji
- **relativna frekvencija -** f_i **/N**
- **kumulativne frekvencije**
- **kumulativne relativne frekvencije**
- distribucija frekvencija: poligon frekvencija i histogram, “Pareto chart”, …

histogram

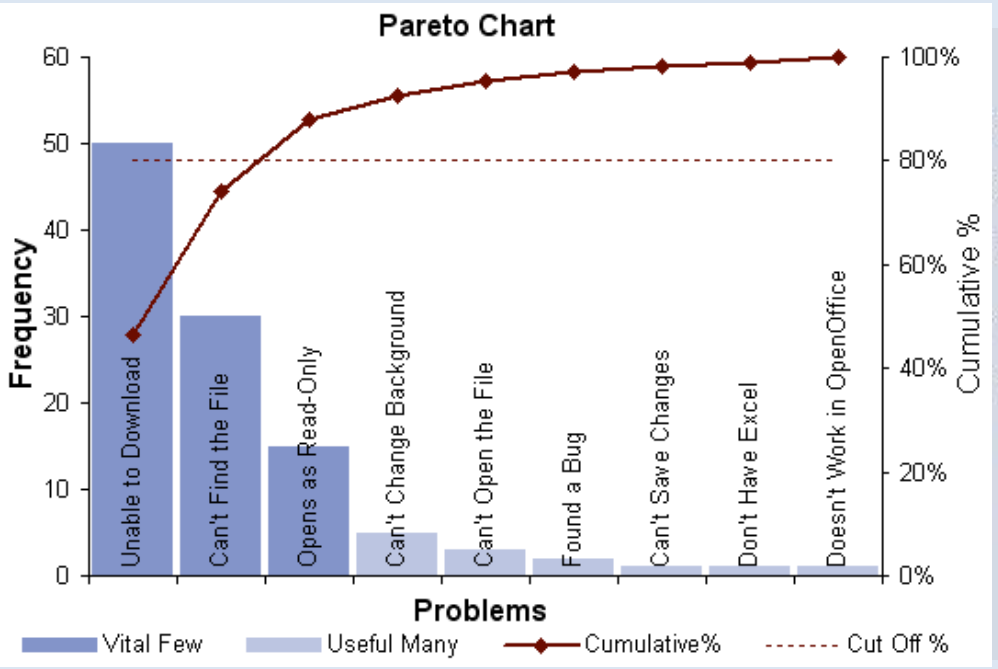

Histogram of Monte Carlo Simulation Results
Frequency
Bins
300
250
200
150
100
50
0
-2000
-1700
-1400
-1100
-800
-500
-200
100
400
700
1000
1300
1600
1900
2200
2500
2800
3100
3400
3700
4000

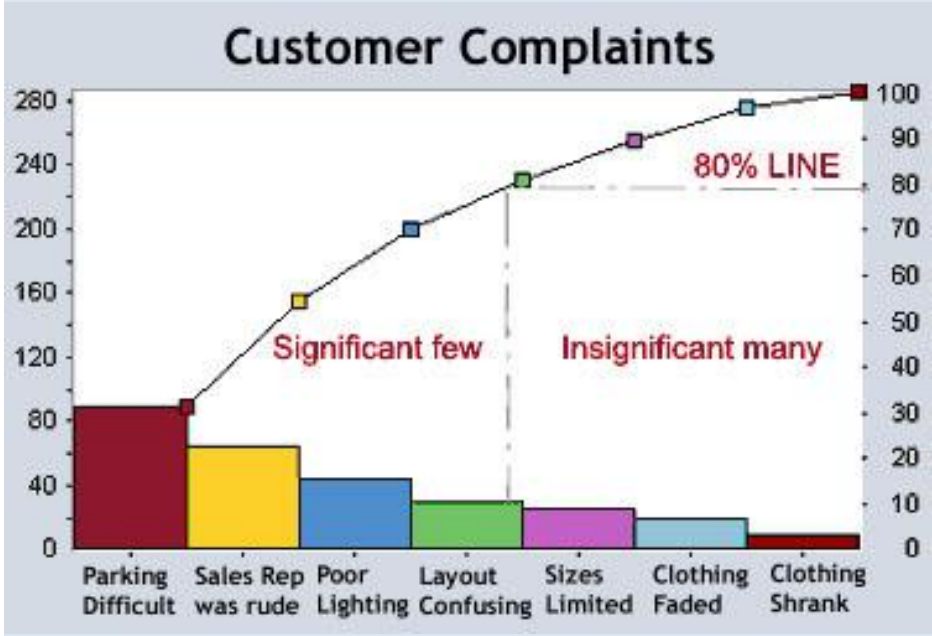

Length of Greek Tragedies in Words
Frequency
Length in Words
0
2
4
6
8
4000
5000
6000
7000
8000
9000
10000
11000

kumulativne relativne frekvencije

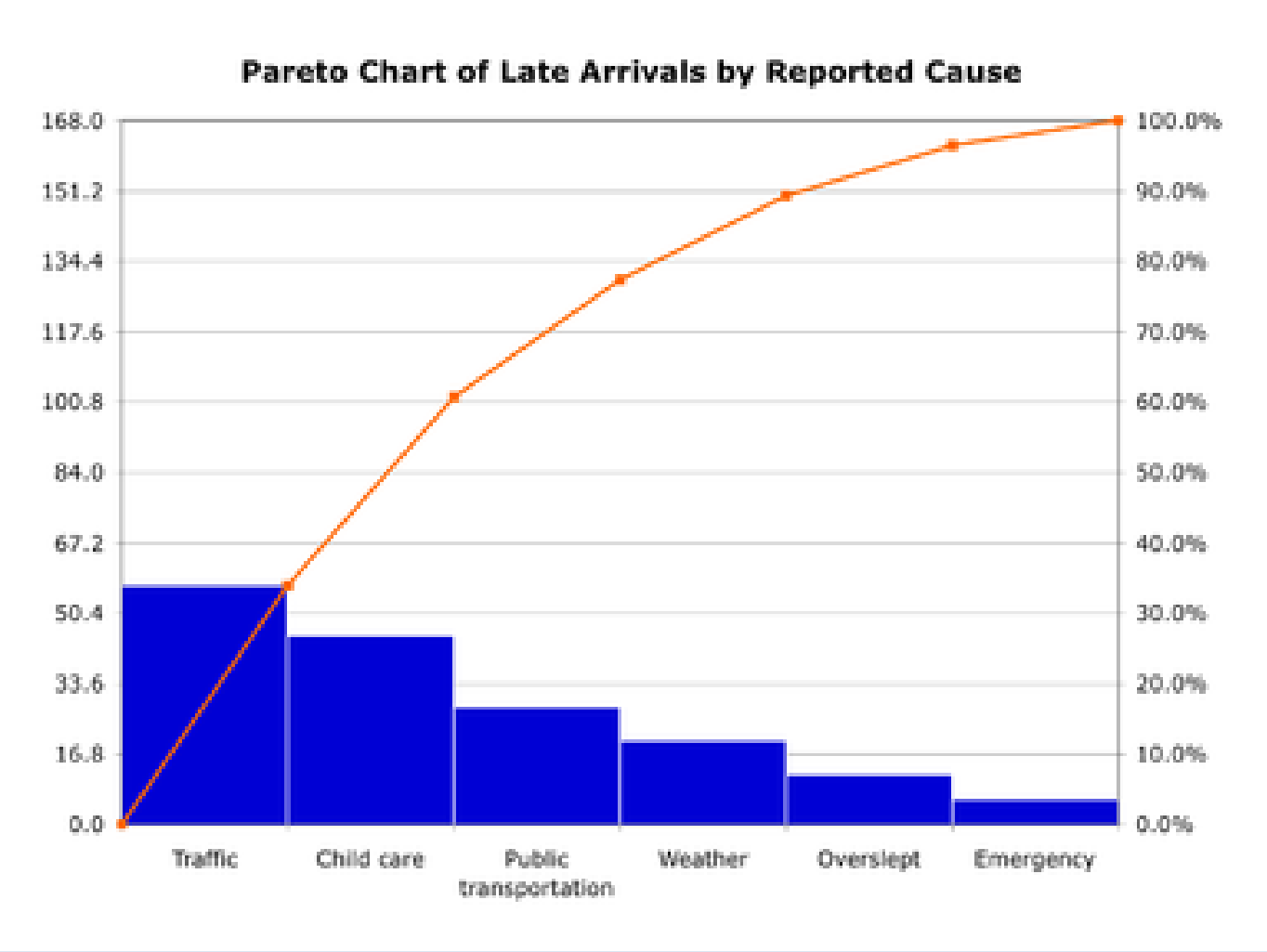


“Pareto chart”





“Pareto chart”



MJERE SREDIŠNJE (CENTRALNE) TENDENCIJE

1. Srednja vrijednost
2. Medijan
3. Mod

SREDNJA VRIJEDNOST

Srednja vrijednost ili **aritmetička sredina** (engl. *mean* ili *arithmetic mean*) predstavlja sumu svih podataka podijeljenu s ukupnim brojem podataka.

Srednja vrijednost uzorka je definirana sa

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

gdje ***n*** predstavlja ukupan broj podataka u uzorku.

Srednja vrijednost populacije *μ* je definirana sa

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

gdje ***N*** predstavlja ukupan broj podataka u populaciji.

- izvod – srednja vrijednost je ona oko koje je suma odstupanja nula

PRIMJER

Neka sljedeći podaci predstavljaju rezultate nekog mjerenja

41 11 29 7 37 1013 17 1009 5 23 31 13 2 19 3

Izračunajte srednju vrijednost.

Rješenje:
Srednja vrijednost tih podataka iznosi 150, 7.
(Empirijsko je pravilo da se srednja vrijednost izrazi s jednim decimalnim mjestom više u odnosu na originalne podatke.)

Ponovite prethodni račun bez vrijednosti 1013 za koju smo naknadno ustvrdili da je nepouzdana

Rješenje: 89,1

Na srednju vrijednost znatno utječu veliki ili mali granični podaci!

MEDIJAN

• **Medijan** (engl. *median*) je vrijednost središnjeg podatka koja podatke poredane po veličini dijeli u dva jednako brojna dijela. Ako je broj podataka neparan medijan je **vrijednost središnjeg podatka**, a ako je broj podataka paran medijan predstavlja **srednju vrijednost dva središnja podatka**.

• također razlikujemo medijan uzorka i medijan populacije

• To znači da u sortiranom nizu podataka 50 % elemenata ima vrijednost manju ili jednaku medijanu te da 50 % elemenata ima vrijednost veću ili jednaku medijanu.

PRIMJER

Neka sljedeći podaci predstavljaju rezultate nekog mjerenja:

41 11 29 7 37 1013 17 1009 5 23 31 13 2 19 3

Pronađite medijan.

Rješenje:
Podatke je prvo potrebno sortirati po veličini

2 3 5 7 11 13 17 19 23 29 31 37 41 1009 1013

Kako je broj podataka neparan, medijan je vrijednost središnjeg podatka - 19.

Ponovite prethodni račun bez vrijednosti 1013 za koju smo naknadno ustvrdili da je nepouzdana

Ako je broj podataka paran, medijan će biti srednja vrijednost dva

središnja podatka

2 3 5 7 11 13 17 19 23 29 31 37 41 1009

U ovom slučaju medijan iznosi 18,0.

Na medijan znatno manje utječu veliki ili mali granični podaci nego što je to slučaj kod srednje vrijednosti!

MOD

Mod (engl. *mode*) je vrijednost podatka koji se najčešće ponavlja.

PRIMJER

Neka sljedeći podaci predstavljaju rezultate nekog mjerenja

11 11 3 7 13 11 17 11 5 23 13 13 2 19 11

Pronađite mod.

Rješenje:

Podatke je korisno poredati sortirati po veličini (ali nije nužno)

2 3 5 7 **11 11 11 11 11** 13 13 13 17 19 23

Mod je 11.

PRIMJER

Neka sljedeći podaci predstavljaju rezultate nekog mjerenja

2 3 5 7 11 13 17 19 23 29 31 37 41 1009 1013

Pronađite mod.

Rješenje:

Moda nema (nije pravilno reći da je mod 0 !).

Neka sljedeći podaci predstavljaju rezultate nekog mjerenja

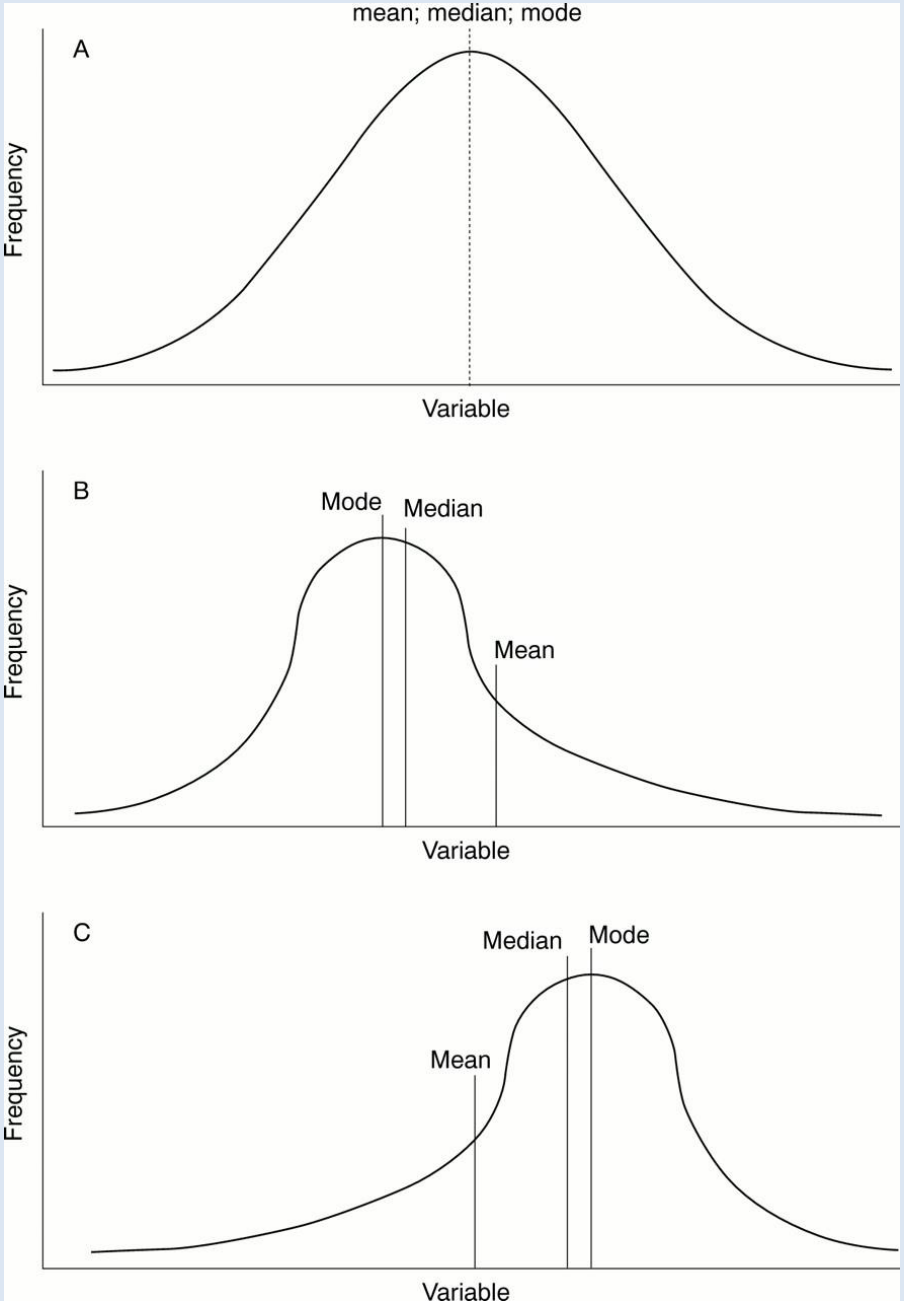
2 3 5 7 11 11 11 13 17 17 17 19 23 29 31

Pronađite mod.

Rješenje:

Modovi su 11 i 17. Ovakav skup podataka je bimodalan.

MJERE SREDIŠNJE (CENTRALNE) TENDENCIJE - ZAKLJUČAK



MJERE VARIJABILNOSTI

(rasipanja vrijednosti oko središnje tendencije)

Raspon podatka (engl. *data range*) je razlika između maksimalne i minimalne vrijednosti podataka

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

VARIJANCIJA

Varijancija uzorka (engl. *sample variance*) je suma kvadrata odstupanja svih podataka od njihove srednje vrijednosti podijeljene s *n − 1*

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

gdje *n* predstavlja ukupan broj podataka u uzorku.

Varijancija populacije (engl. population variance) je vrijednost sume kvadrata odstupanja svih podataka od njihove srednje vrijednosti podijeljene s *N*

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

gdje *N* predstavlja ukupan broj podataka u populaciji.

- izvod

STANDARDNA DEVIJACIJA

Standardna devijacija uzorka (engl. *sample standard deviation*) je pozitivna vrijednost drugog korijena varijancije uzorka

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

gdje *n* predstavlja ukupan broj podataka u uzorku.

Standardna devijacija populacije (engl. *population standard deviation*) je pozitivna vrijednost drugog korijena varijancije populacije

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

gdje *N* predstavlja ukupan broj podataka u populaciji.

KOEFICIJENT VARIJACIJE

Koeficijent varijacije uzorka (engl. *sample* coefficient of variation) je omjer vrijednosti standardne devijacije uzorka sa srednjom vrijednosti podataka u uzorku

$$\mathrm{CVar} = \frac{s}{\bar{x}}$$

Koeficijent varijacije populacije (engl. *population coefficient of variation*) je omjer vrijednosti standardne devijacije populacije sa srednjom vrijednosti podataka u populaciji

$$\mathrm{CVar} = \frac{\sigma}{\mu}$$

- normalizirana mjera disperzije – pogodnija je za uspoređivanje od standardne devijacije, jer je standardna devijacija ovisna o srednjoj vrijednosti, CVar je bezdimenzijska veličina, može biti izražena u postocima, često se koristi u kromatografiji, problem nastupa kada je srednja vrijednost oko nule.

KOEFICIJENT VARIJACIJE

Primjer:
Mjerili smo visinu učenica u nekom razredu i dobili smo srednju vrijednost od 140,91cm i standardnu devijaciju 10,34 cm.

Zatim smo mjerili visinu učenika u istom razredu i dobili smo srednju vrijednost 152,36 cm, sa standardnom devijacijom s=7,25 cm.

Usporedbom koeficijenata varijance:

CV_A= 10,34/140.91=0,07

CV_B=7,25/152,36=0,05

Vidimo da je varijabilnost u visini učenica 1,4 puta veća od varijabilnosti u visini učenika:
CV_A/ CV_B=0,07/0,05=1,4

Kvantili

•***kvantili*** su vrijednosti statističkog obilježja koje statistički niz dijele na q jednakih dijelova

• kvantili se dijele s obzirom na broj intervala na koji dijele niz podataka:

- kvantil drugog reda dijeli niz podataka na dva jednaka dijela (*medijan?*)
- kvantili trećeg reda su **tercili** i dijele niz podataka na tri dijela
- kvantili četvrtog reda su **kvartili** i dijele niz podataka na četiri dijela
- **kvintili …**
- najčešće se koriste kvartili, a zatim decili i percentili

• broj kvantila je uvijek **q-1**, odnosno za kvantil drugog reda imamo jedan kvantil , za kvantil trećeg reda (**tercil**) imamo **2** kvantila, za kvantil četvrtog reda **(kvartil)** imamo **3** kvantila, …

• sa statističke točke gledišta, **k**-ti kvantil **q** predstavlja onu vrijednost **x** za koju možemo reći da je **vjerojatnost** da nasumična varijabla bude **manja** od **x** iznosi najviše **k/q**. Također, vjerojatnost da nasumična varijabla bude veća od **x** iznosi najviše **(q-k/q)** ili **1-(k/q).**

Kvantili

• ***kvartili*** su vrijednosti statističkog obilježja koje statistički niz dijele na 4 jednaka dijela, mogu se podijeliti na donji i gornji kvartil

• ***donji kvartil*** dijeli statistički niz na dva dijela u omjeru 1:3, odnosno preciznije 25 % elemenata statističkog skupa ima vrijednost manju ili jednaku donjem kvartilu,

• ***gornji kvartil*** dijeli statistički niz na dva dijela u omjeru 3:1, odnosno preciznije 75 % elemenata statističkog skupa ima vrijednost manju ili jednaku gornjem kvartilu

• ***srednji kvartil*** *je* često medijan i dijeli statistički niz u dva jednaka dijela 1:1**,** odnosno 50 % elemenata statističkog skupa ima vrijednost manju ili jednaku srednjem kvartilu, a 50 % elemenata statističkog skupa ima vrijednost veću od srednjeg kvartila

• **interkvartilna razlika** - razlika između donjeg i gornjeg kvartila - predstavlja raspon unutar kojeg se nalazi središnjih 50 % statističkog niza – vrlo često se koristi.

Kvantili

PRIMJER

Odredite prvi i treći kvartil za slijedeći niz podataka:
3, 6, 7, 8, 8, 10, 13, 15, 16, 20

Rješenje
- prvi kvartil – 10*(1/4) = 2,75 – zaokružimo na 3, dakle treći element u slijedu predstavlja prvi kvartil – **7**
3, 6, **7**, 8, 8, 10, 13, 15, 16, 20

-treći kvartil – 10*(3/4)=7,5 – zaokružimo na 8, dakle **15**
3, 6, 7, 8, 8, 10, 13, **15**, 16, 20

Kvantili

PRIMJER

Mjerili smo visinu učenica u nekom razredu i dobili smo slijedeće vrijednosti u cm:
140, 141, 138, 140, 122, 160, 154, 132, 148, 135, 140
Odredite prvi i treći kvartil.

Rješenje:
Najprije vrijednosti poredamo po veličini:
122, 132, 135, 138, 140, 140, 140, 141, 148, 154, 160

Imamo ukupno 11 vrijednosti pa je prvi kvartil ¼*11= 2,75, zaokružimo na prvi veći broj, u ovom slučaju 3, znači prvi kvartil predstavlja treća vrijednost u nizu, a to je 135 cm.
To znači da 25% učenica ima visinu jednaku ili manju 135 cm. Ili 75 % učenica je više od 135 cm.

Treći kvartil: ¾*11= 8,25, znači treći kvartil predstavlja deveta vrijednost u nizu - 148 cm.
To znači da 75% učenica ima visinu jednaku ili manju od 148 cm. Ili 25 % učenica je više od 148 cm.

Kvantili

Primjer:
Odredite prvi i treći kvartil za slijedeći niz podataka:
28 23 59 25 23 20 31 48 32

Rješenje:
Poredamo ih po redu:
20 23 23 25 28 31 32 48 59

Imamo 9 podataka.
Prvi kvartil je ¼ * 9 = 2,25, znači treći podatak je prvi kvartil – 23.
Treći kvartil je ¾ * 9 = 6,75, znači treći kvartil je sedmi podatak – 32.