

Biostatisztika és informatika alapjai

2. előadás: Leíró statisztika

2016. szeptember 15.

Veres Dániel

Tatisztika? Ammeg mi?

(Békásmegyeri aluljáró „átlagos” „lakója”)

A **statisztika** a véletlen tömegjelenségek leírója.



- Adatgyűjtés
 - Adatok rendszerezése, áttekintése
 - Adatok elemzése
 - Következtetések levonása
- Leíró statisztika
- Következtető statisztika
(induktív statisztika)

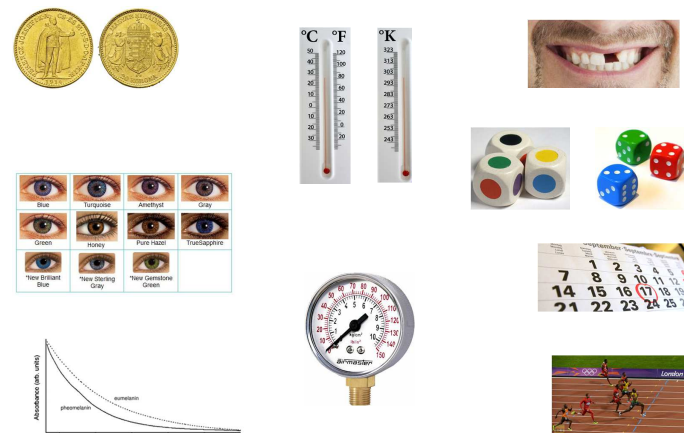
Tatisztika? Ammeg mi?



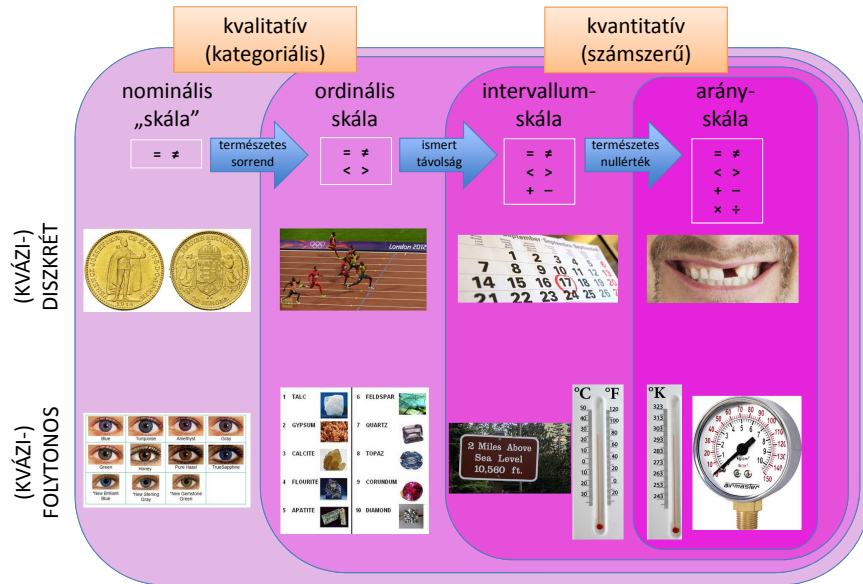
- Adatgyűjtés
 - Adatok rendszerezése, áttekintése
 - Adatok elemzése
 - Következtetések levonása
- Leíró statisztika
- Következtető statisztika
(induktív statisztika)

Változók, kimenetek

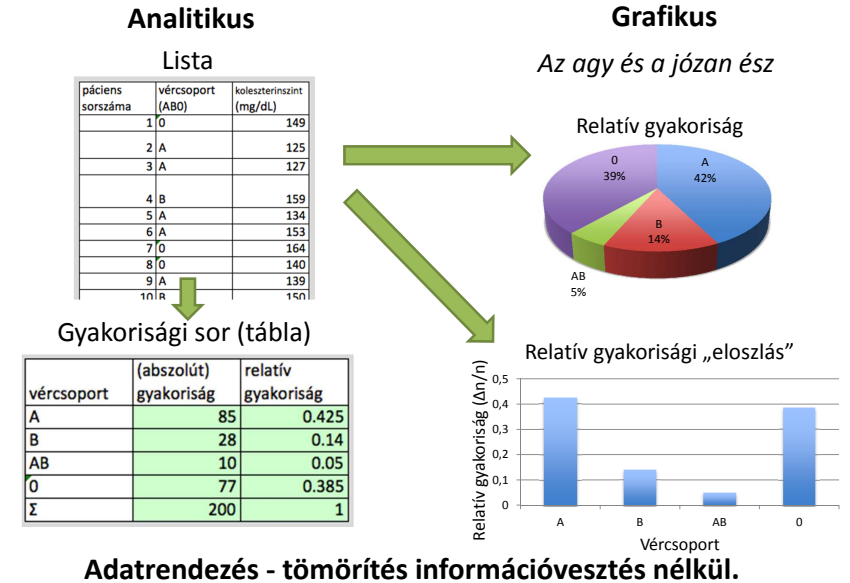
Amit meg tudunk mérni vagy meg tudunk figyelni.



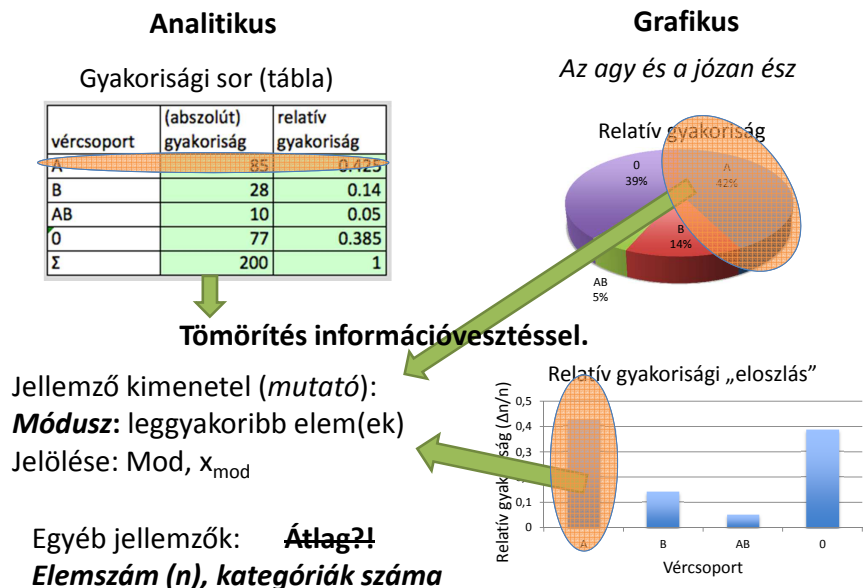
Változók típusai, mérési skálák



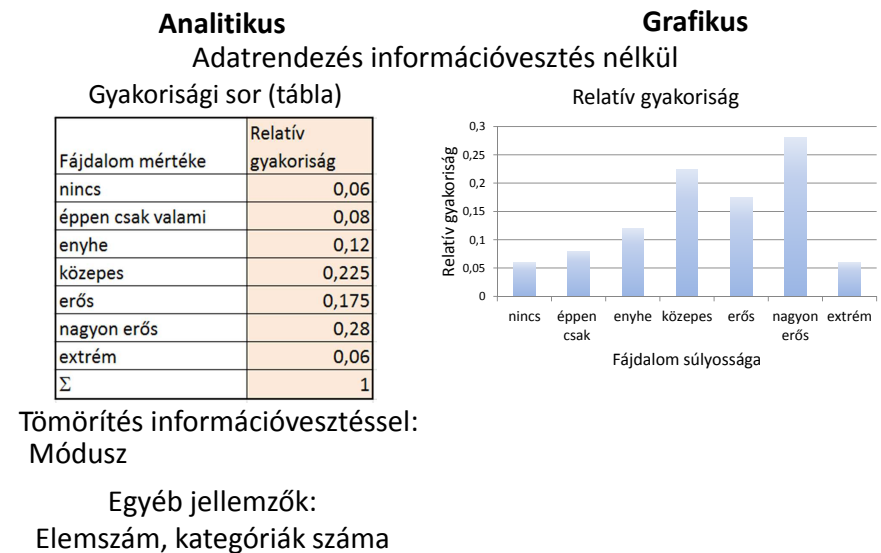
Nominális változó jellemzése I.



Nominális változó jellemzése II.



Ordinális változó jellemzése I.



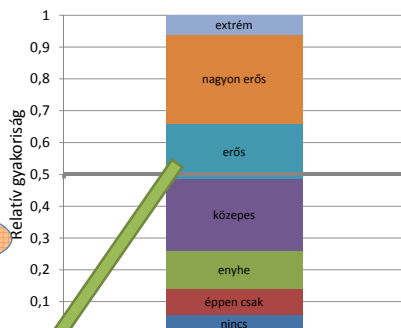
Ordinális változó jellemzése II.

Analitikus

Gyakorisági sor (tábla)

Fájdalom mértéke	Kumulatív relatív gyakoriág
nincs	0,06
éppen csak	0,14
enyhe	0,26
közepes	0,485
erős	0,66
nagyon erős	0,94
extrém	1

Grafikus



Új Jellemző (információvesztéssel):

Medián: „középső” elem(ek)

Jelölése: Me , Med , x_{med}

Kvantitatív (számszerű) változó jellemzése I.

Analitikus

Gyakorisági sor (tábla)

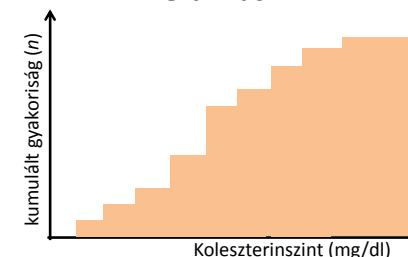
osztályok	osztályok felső (zárt) határa	(abszolút) gyakoriság (GYAKORISÁG)	(abszolút) gyakoriság (DARABT)
$x \leq 100$	100	0	0
$100 < x \leq 110$	110	0	0
$110 < x \leq 120$	120	2	2
$120 < x \leq 130$	130	5	5
$130 < x \leq 140$	140	22	22
$140 < x \leq 150$	150	31	31
$150 < x \leq 160$	160	48	48
$160 < x \leq 170$	170	40	40
$170 < x \leq 180$	180	22	22

Adatrendezés információvesztéssel járhat.

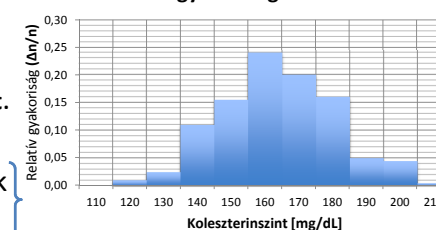
Osztályszélesség meghatározása:

- szakmai és esztétikai szempontok
- statisztikai szempontok alapján

Grafikus



Relatív gyakorisági eloszlás

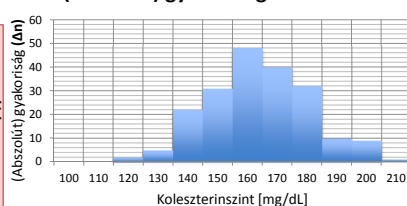


Az adatok szemléltetése I.B

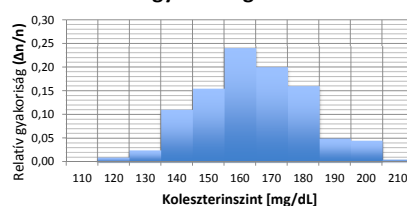
A gyakoriságok eloszlások típusai *kvantitatív* változó esetén

normális az adatok számára (n)

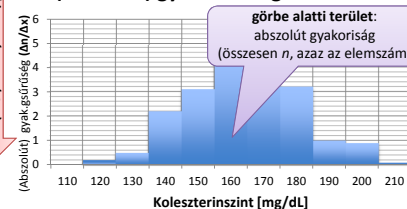
(Abszolút) gyakorisági eloszlás



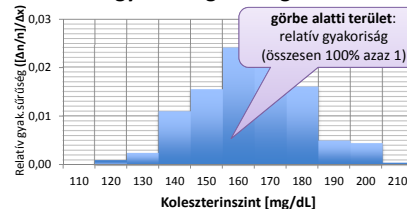
Relatív gyakorisági eloszlás



(Abszolút) gyak.sűrűségi eloszlás

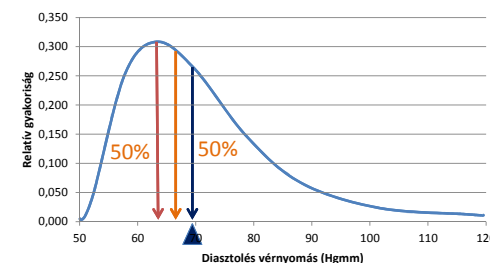
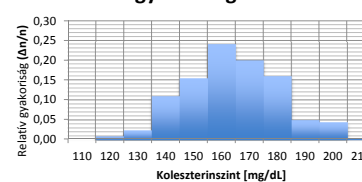


Relatív gyakoriságsűrűségi eloszlás



Kvantitatív változó jellemzése II.

Relatív gyakorisági eloszlás



Jellemzők – **középértékek** (speciális **helyparaméterek**):

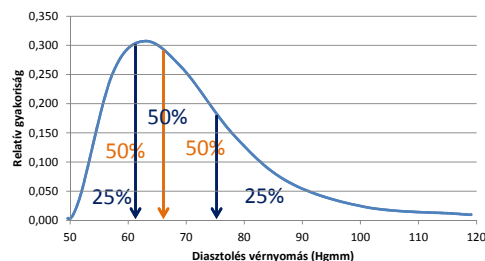
- **Módusz(ok)**: leggyakoribb elem(ek) ?
- **Medián**: „középső” elem(ek) ?
- **Átlag** (számtani közép): „súlypont”, érzékeny a „kiszóró” adatokra ?!

Jelölése: $x_{\text{átl}}$, \bar{x}

Előny: tömörítés, **kevés adathból is számíthatóak**

Képletek: képlettárban

Kvantilisok I.



Egyéb helyparaméterek:

- **Medián:** 50-50% (Q_2)
- **Kvantilisok:** alsó kvartilis (Q_1): 25-75%; felső kvartilis (Q_3): 75-25%

Általánosan

p-kvantilis(ek): az adatrendszer p-kvantilisének nevezzük azt a számot, amelynél kisebb adatok darabszáma legfeljebb $n \cdot p$ és amelynél nagyobb adatok darabszáma legfeljebb $n \cdot (1 - p)$, ahol p 0 és 1 közötti szám

Kitérő I.

Nap sorszáma	Várakozási idő		Nap sorszáma	Várakozási idő	
1	1,27	medián: 8,475	1	1,27	medián: 8,475
2	3,3	alsó kvartilis 3,59	2	3,3	alsó kvartilis 3,59
3	3,44	átlag 7,723333	3	3,44	átlag 9,141667
4	3,64		4	3,64	
5	6,33		5	6,33	
6	7,72		6	7,72	
7	9,23		7	9,23	
8	9,87		8	9,87	
9	10,31		9	10,31	
10	12,29		10	12,29	
11	12,3		11	12,3	
12	12,98		12	30	

Medián, kvantilisok elméletben és gyakorlatban eltérhetnek.

Átlag érzékeny a kiszóró adatokra, de kvantilisok nem érzékenyek.

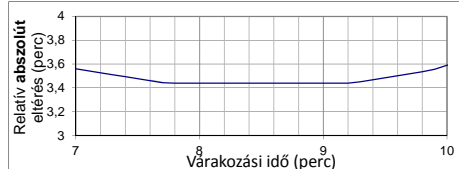
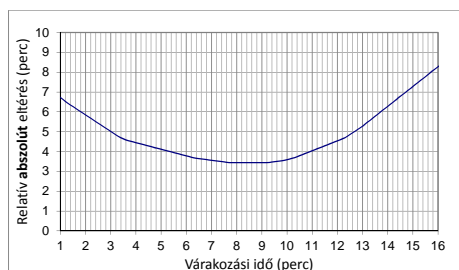
Módusz?

Kitérő II.

$$\frac{1}{n} \sum |x_i - x^*|$$

Minimális, ha:

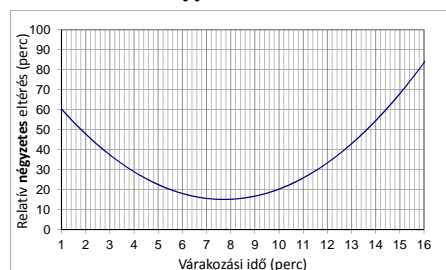
$$x^* = \text{Medián}$$



$$\frac{1}{n} \sum (x_i - x^*)^2$$

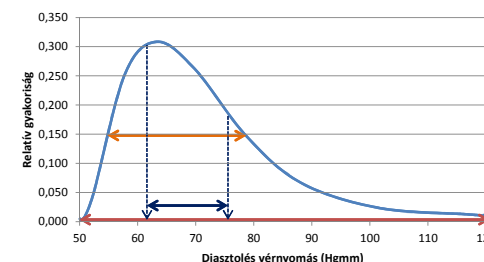
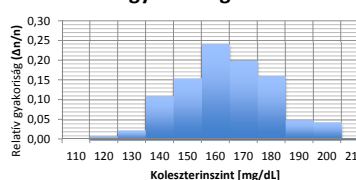
Minimális, ha:

$$x^* = \text{Átlag}$$



Kvantitatív változó jellemzése III.

Relatív gyakorisági eloszlás

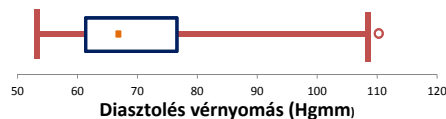


Jellemzők – szóródási paraméterek:

- **Terjedelem:** **maximális** érték és **minimális** érték különbsége
- **Variancia (szórásnégyzet, s^2):** átlagtól vett átlagos négyzetes eltérés (korrigált - minta, korrigálatlan - sokaság)
- **Szórás (s):** variancia négyzetgyöke – eloszlásgörbe „szélessége”
- **Interkvartilis távolság (IQR):** felső és alsó kvartilis értékek különbsége, előnye: nem érzékeny a „kiszóró” pontokra

Kvantitatív változó jellemzése IV.

Box plot – (sodrófadiagram)



Sodrófa szeme: átlag, illetve *medián*

Sodrófa teste: átlagtól mért szórás, illetve *interkvartilis távolság*

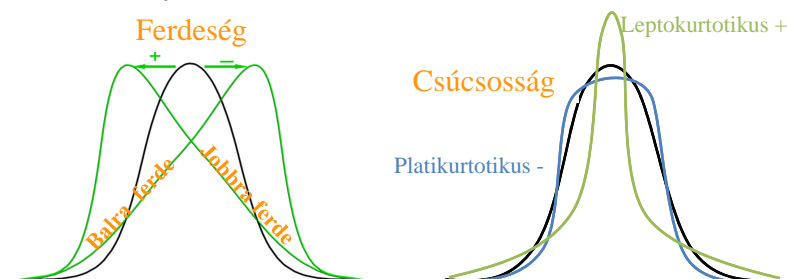
Sodrófa szára: minimum és maximum értékek, 0,5-ös és 0,95-ös kvantilisok, szórás 2-szerese, *IQR 1,5-szerese...*

sodrófa szárán túl: **kiszóró pont**

Kvantitatív változó jellemzése V.

Egyéb paraméterek:

- **momentum:**
a k. momentum: $\Sigma(x_i)^k / n$
- **centrális momentum:**
a k. centrális momentum: $\Sigma(x_i - \mu)^k / n$
- **ferdeség,**
- **csúcsosság** } az eloszlásgörbe alakját mutatják

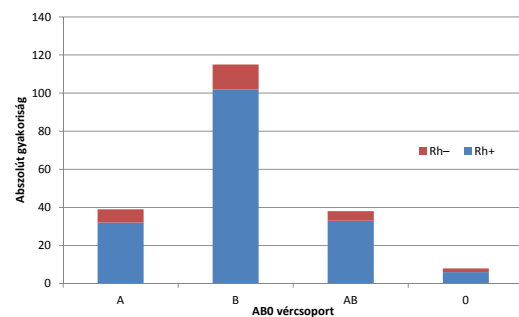


Több kvalitatív változó jellemzése

Analitikus: **kontingencia** táblázat

	A	B	AB	O	Σ
Rh+	32	102	33	6	173
Rh-	7	13	5	2	27
Σ	39	115	38	8	200

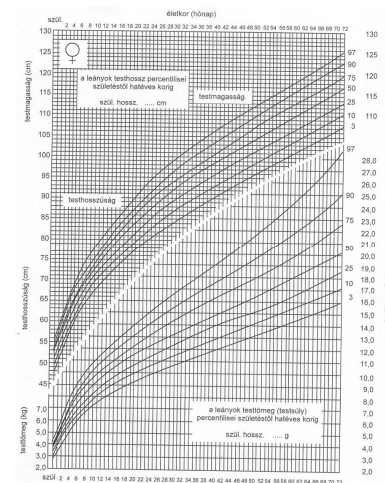
Grafikus: **mozaik ábra**

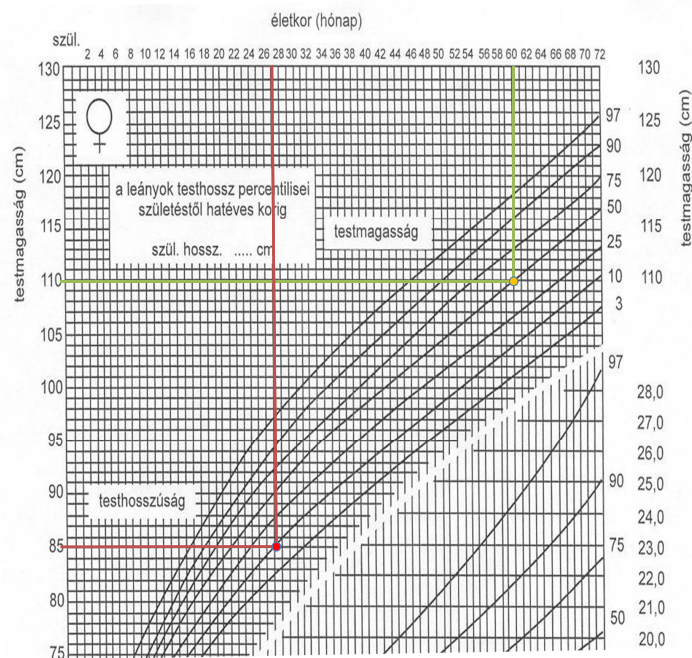


Több kvantitatív változó jellemzése

Grafikus: **percentilis ábrák**

Percentilis: %-ban kifejezett kvantilis





Adatgyűjtés

Adatgyűjtésnek **célja** van és nem változói!
Változó legyen a lehető legmagasabb skálájú.

Adatgyűjtésnek lehetőségei:

- **ismert – meg kell kérdezni**
- **nem ismert – meg kell figyelni, vagy meg kell mérni**

Adatrögzítés

- olyan formában, amely könnyen rendszerezhető, alakítható - excel
- **változókat külön-külön**
- **kódolás egyértelmű** legyen (változó minősége, eltérő kategóriák)

Ellenőrző kérdések#1

- Add meg a statisztikai tevékenységek csoportosítását.
- Milyen két nagy csoportra oszthatóak a statisztikai megoldások?
- Mik tartoznak a leíró statisztika tárgykörébe?
- Mik tartoznak a következtető statisztika tárgykörébe?
- Milyen két módon rendezhetünk, és jellemezhetünk egy változót?
- Mely esetekben történik a változó jellemzése adatvesztéssel, illetve adatvesztés nélkül?
- Milyen jellemzőket használhatunk nominális változó leírására?
- Milyen jellemzőket használhatunk ordinális változó leírására?
- Milyen jellemzőket használhatunk ordinális változó leírására?
- Definiáld a móduszt.
- Hogyan jelöljük a móduszt?
- Definiáld a mediánt.
- Hogyan jelöljük a mediánt?
- Hogyan határozható meg az osztályszélesség egy számszerű változónál?
- Mik tartoznak a középértékek közé?
- Hogyan számítandó az osztályszélesség statisztikai szempontból?
- Mi az átlag, a medián, a módusz terjedeleme, az interkvartilis terjedeleme és a szórás szemléletes jelentése?
- Hogyan határozható meg egy minta átlaga?
- Hogyan jelöljük az átlagot?
- Melyik középérték érzékeny a kiszóró értékekre?
- Mi a középértékek előnye az eloszlásgörbével szemben?
- Melyek a helyparaméterek?
- Definiáld általánosan a kvantiliseket.
- Definiáld az alsó kvantilis. Mi a szemléletes jelentése?
- Mi a különbség a második kvantilis és a medián között?
- Micsoda, illetve hogyan számolandó az alsó kvantilis az elméletben és a gyakorlatban.
- Milyen értékre lesz az átlagos abszolút eltérés minimális?
- Milyen értékre lesz az átlagos négyzetes eltérés minimális?
- Melyek a szóródási paraméterek?
- Definiáld a varianciát.
- Definiáld a szórást.
- Mit jelent a ferdeség?
- Mit jelent a csúcsosság?
- Definiáld az interkvartilis távolságot.
- Hogyan jelöljük (rövidítjük) az interkvartilis távolságot?

Ellenőrző kérdések#2

- Mi az a sodrófadiagram?
- Milyen részei vannak a sodrófadiagramnak?
- Mi lehet a sodrófadiagram szeme?
- Mit használhatunk a sodrófadiagram testének?
- Mit használhatunk a sodrófadiagram szárának?
- Mit használjunk a sodrófadiagram részeinek, ha nem szimmetrikus eloszlásunk van kiszóró pontokkal?
- Mit használjunk a sodrófadiagram részeinek, ha szimmetrikus eloszlásunk van kiszóró pontok nélkül?
- Mit használjunk a sodrófa testének, ha a sodrófa szeme a medián?
- Mit jelent a részátlag?
- Hogyan szokták definiálni a kiszóró pontokat?
- Hogyan számítandók a momentumok és a centrális momentumok?
- Mekkora az értéke az első centrális momentumnak?
- Mekkora az értéke az első momentumnak?
- Mivel egyenlő a második centrális momentum?
- Mit jelent a percentilis?
- Mit tudunk leolvasni a percentilis görbékről?