Biostatisztika és informatika alapjai

2. előadás: Leíró statisztika

2017. szeptember 21.

Veres Dániel

Tatisztika? Ammeg mi?

(Békásmegyeri aluljáró „átlagos” „lakója”)

A **statisztika** a véletlen tömegjelenségek leírója.

- Adatgyűjtés
- Adatok rendszerezése, áttekintése

Leíró statisztika

- Adatok elemzése
- Következtetések levonása

Következtető statisztika
(induktív statisztika)

Tatisztika? Ammeg mi?

- Adatgyűjtés
- **Adatok rendszerezése, áttekintése**

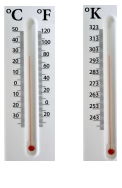
Leíró statisztika

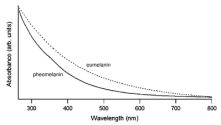
- Adatok elemzése
- Következtetések levonása

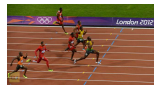
Következtető statisztika
(induktív statisztika)

Változók, kimenetelek

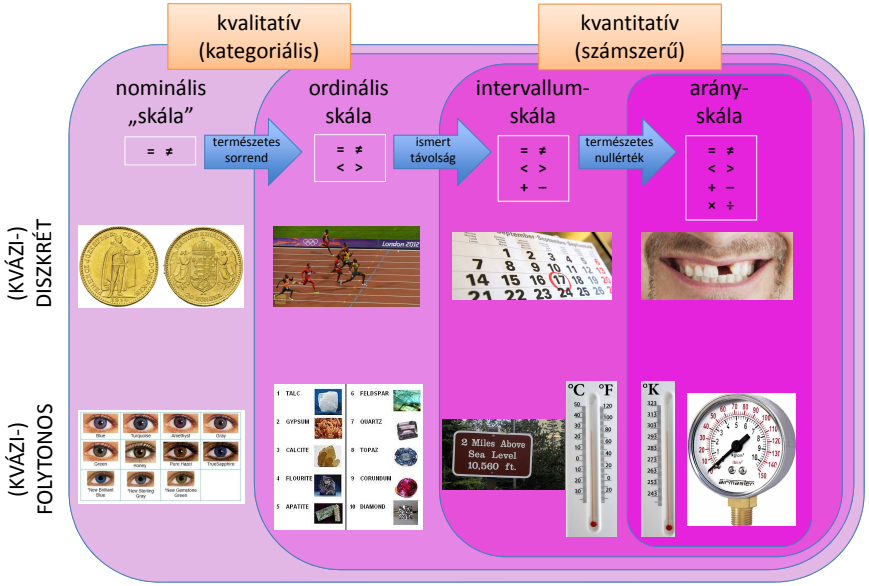
Amit meg tudunk mérni vagy meg tudunk figyelni.







Változók típusai, mérési skálák



Nominális változó jellemzése I.

Analitikus

Lista

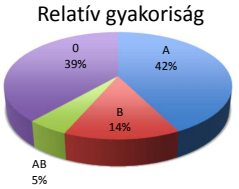
páciens sorszáma	vércsoport (ABO)	koleszterinszint (mg/dL)
1	0	149
2	A	125
3	A	127
4	B	159
5	A	134
6	A	153
7	0	164
8	0	140
9	A	139
10	B	150

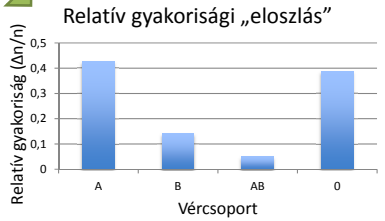
Gyakorisági sor (tábla)

vércsoport	(abszolút) gyakoriság	relatív gyakoriság
A	85	0.425
B	28	0.14
AB	10	0.05
0	77	0.385
Σ	200	1

Grafikus

Az agy és a józan ész





Adatrendezés - tömörítés információvesztés nélkül.

Nominális változó jellemzése II.

Analitikus

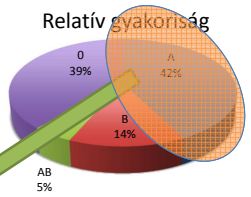
Gyakorisági sor (tábla)

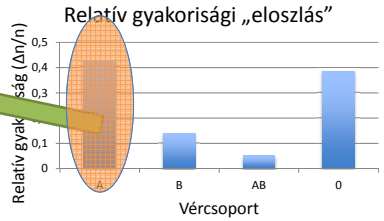
vércsoport	(abszolút) gyakoriság	relatív gyakoriság
A	85	0.425
B	28	0.14
AB	10	0.05
0	77	0.385
Σ	200	1

Tömörítés információvesztéssel.

Grafikus

Az agy és a józan ész





Jellemző kimenetel (*mutató*):
Módusz: leggyakoribb elem(ek)
Jelölése: Mod, x_{mod}

Egyéb jellemzők: ~~**Átlag?!**~~
Elemszám (n), kategóriák száma

Ordinális változó jellemzése I.

Analitikus

Adatrendezés információvesztés nélkül

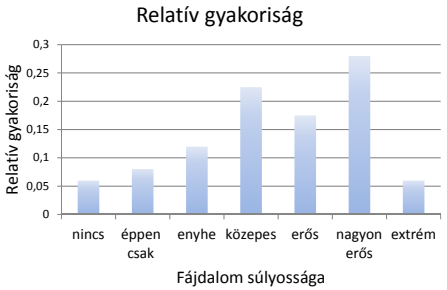
Gyakorisági sor (tábla)

Fájdalom mértéke	Relatív gyakoriság
nincs	0,06
éppen csak valami	0,08
enyhe	0,12
közepes	0,225
erős	0,175
nagyon erős	0,28
extrém	0,06
Σ	1

Tömörítés információvesztéssel:
Módusz

Egyéb jellemzők:
Elemszám, kategóriák száma

Grafikus



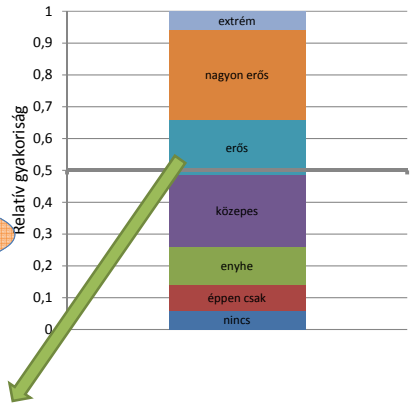
Ordinális változó jellemzése II.

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

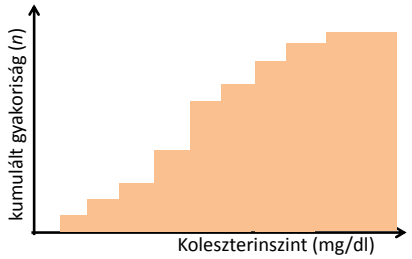
Gyakorisági sor (tábla)

gyakorisági eloszlások (differenciáldiszkriminációs függvénye			
osztályok	osztályok felső (zárt) határa	(abszolút) gyakoriság (GYAKORISÁG)	(abszolút gyakorisá (DARABT
x ≤ 100	100	0	
100 < x ≤ 110	110	0	
110 < x ≤ 120	120	2	
120 < x ≤ 130	130	5	
130 < x ≤ 140	140	22	
140 < x ≤ 150	150	31	
150 < x ≤ 160	160	48	
160 < x ≤ 170	170	40	

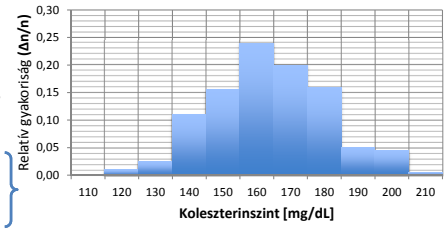
Adatrendezés **információvesztéssel** jár**hat.**

Osztályszélesség meghatározása:
- szakmai és esztétikai szempontok
- statisztikai szempontok alapján

Grafikus



Relatív gyakorisági eloszlás



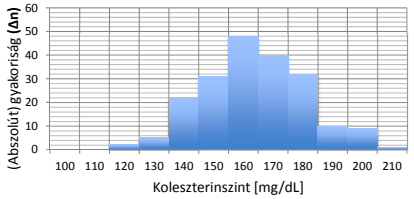
Az adatok szemléltetése I.B

*A gyakoriságok eloszlások típusai **kvantitatív** változó esetén*

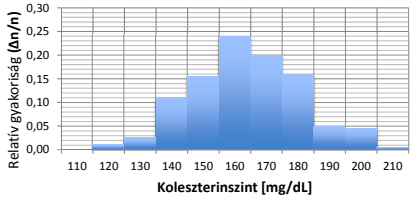
normálás az adatok számára (*n*)

normálás az osztályszélességre (Δx)

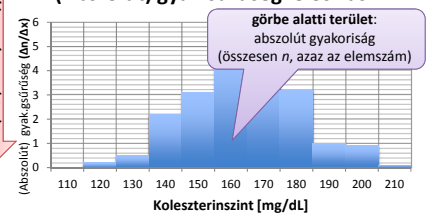
(Abszolút) gyakorisági eloszlás



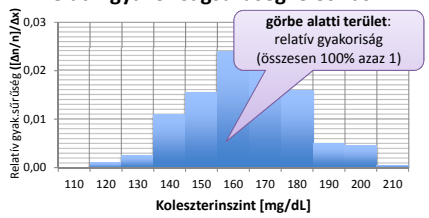
Relatív gyakorisági eloszlás



(Abszolút) gyak.sűrűségi eloszlás

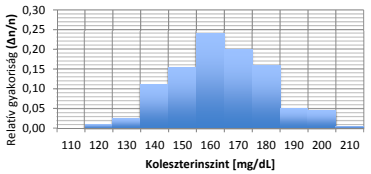


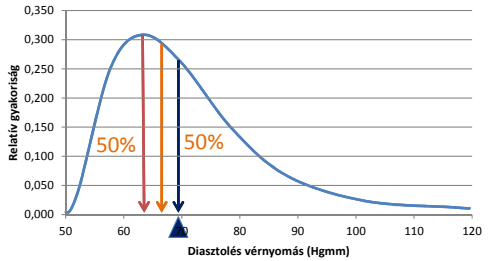
Relatív gyakoriságsűrűségi eloszlás



Kvantitatív változó jellemzése II.

Relatív gyakorisági eloszlás



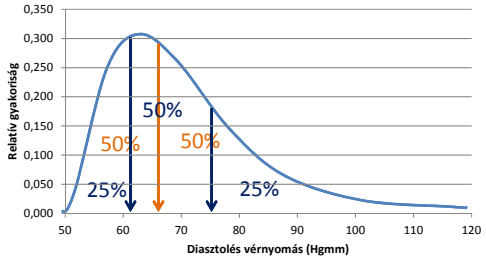


Jellemzők – ***középértékek*** (speciális ***helyparaméterek***):
- Módusz(ok): leggyakoribb elem(ek) ?
- Medián: „középső” elem(ek?)
- ***Átlag*** (számtani közép): „súlypont” ,érzékeny a „kiszóró” adatokra ?!
  Jelölése: $x_{átl}$, $\overline{x}$

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

Képletek: képlettárban

Kvantilisek I.



Egyéb helyparaméterek:

- Medián: 50-50% (Q_2)
- ***Kvartilisek*** : 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*p és amelynél nagyobb adatok darabszáma legfeljebb n*(1 – p), ahol p 0 és 1 közötti szám

Kitérő I.

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

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

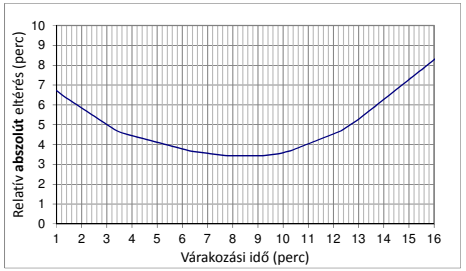
Medián, kvantilisek elméletben és gyakorlatban eltérhetnek.
Átlag érzékeny a kiszóró adatokra, de kvantilisek nem érzékenyek.
Módusz?

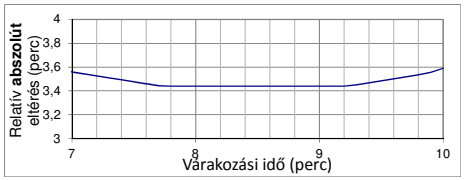
Kitérő II.

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

Minimális, ha:

$$x^* = Medián$$

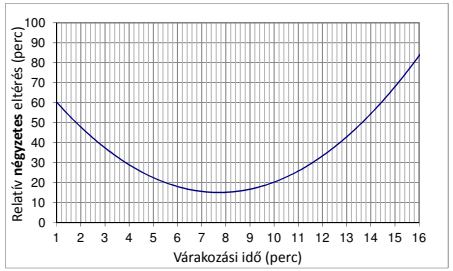




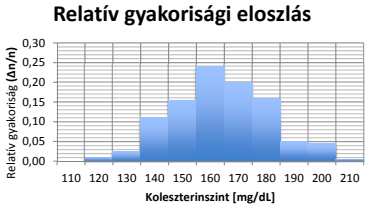
$$\frac{1}{n}\sum(x_i - x^*)^2$$

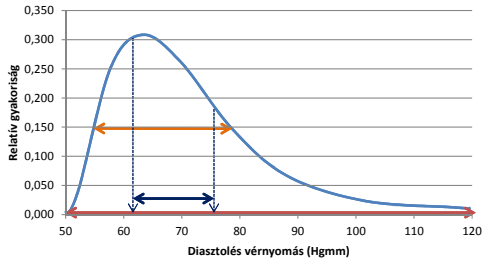
Minimális, ha:

$$x^* = Átlag$$



Kvantitatív változó jellemzése III.



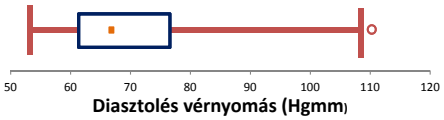


Jellemzők – szóródási paraméterek:

- ***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)



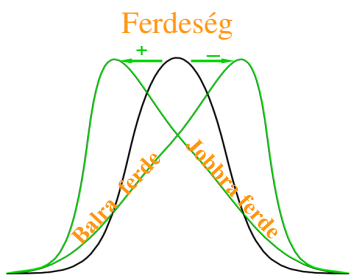
Sodrófa szeme: átlag, illetve *medián*
Sodrófa teste: átlagtól mért szórás, illetve *interkvartilis távolság*
Sodrófa szára: minimum és maximum értékek, 0,5-ös és 0,95-ös kvantilisek, szórás 2-szerese, *IQR 1,5-szerese*...
sodrófa szárán túl: ***kiszóró pont***

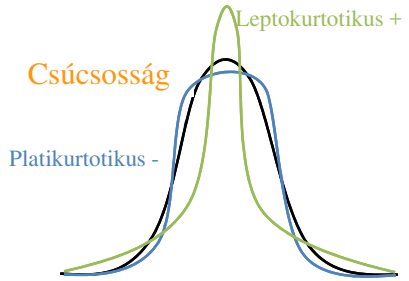
Kvantitatív változó jellemzése V.

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



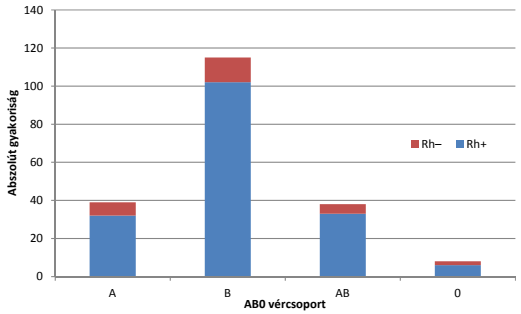


Több kvalitatív változó jellemzése

Analitikus: ***kontingencia*** táblázat

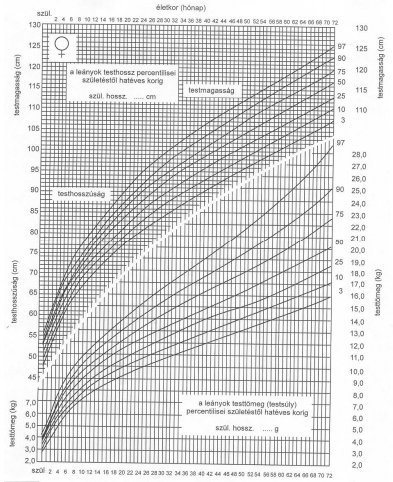
	A	B	AB	0	Σ
Rh+	32	102	33	6	173
Rh–	7	13	5	2	27
Σ	39	115	38	8	200

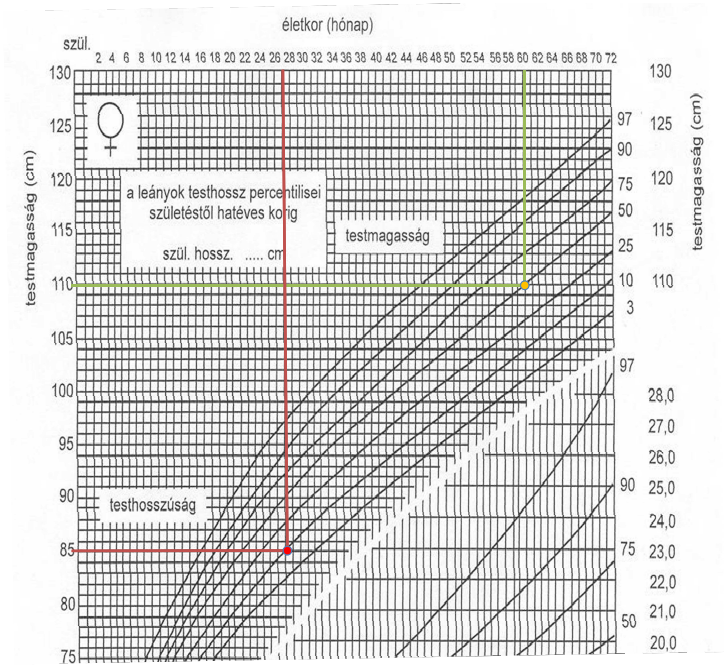
Grafikus: ***mozaik ábra***



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 módsz a terjedelem, az interkvartilis terjedelem 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.
- Definiált az alsó kvartilist. Mi a szemléletes jelentése?
- Mi a különbség a második kvartilis és a medián között?
- Micsoda, illetve hogyan számolandó az alsó kvartilis 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?