Biostatisztika I.
fogorvostan hallgatóknak

1. előadás:

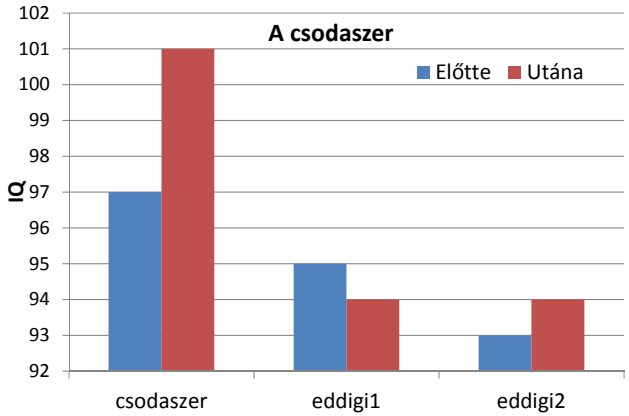
Biostatisztika I.

2019. Szeptember 9.

Veres Dániel

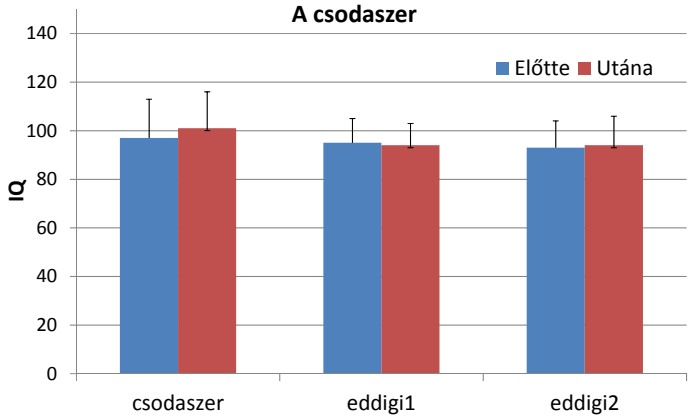
Biostatisztika – nade miért?

- „Azért, hogy el tudjuk dönteni, elhiggyünk-e valamit, amit olvasunk, vagy hogy észrevegyük, hol van benne a hiba, vagyis hogy ne dőljünk be olyan könnyen a statisztikai bűvészkedéseknek, műtermékeknek és tévedéseknek.”



Biostatisztika – nade miért?

- „Azért, hogy el tudjuk dönteni, elhiggyünk-e valamit, amit olvasunk, vagy hogy észrevegyük, hol van benne a hiba, vagyis hogy ne dőljünk be olyan könnyen a statisztikai bűvészkedéseknek, műtermékeknek és tévedéseknek.”



+Példák

- ok-okozat?

(Ananászfogyasztás – daganatos betegség, testmagasság – alvászavar)

☺ pl: http://www.fastcodesign.com/3030529/infographic-of-the-day/hilarious-graphs-prove-that-correlation-isnt-causation

☺pl: A csoki segít a lefogyásban
https://io9.gizmodo.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800

Biostatisztika – nade miért?

- „Azért, hogy el tudjuk dönteni, elhiggyünk-e valamit, amit olvasunk, vagy hogy észrevegyük, hol van benne a hiba, vagyis hogy ne dőljünk be olyan könnyen a statisztikai bűvészkedéseknek, műtermékeknek és tévedéseknek.” (*ld. excel file csodaszer, továbbiak később…)
- „Azért, hogy jobban meg tudjuk ítélni, szerencsénk volt-e vagy pechünk – vagy éppen egyik sem.”
- „Azért, hogy jobban meg tudjuk ítélni, mi mennyit ér, miért mennyit érdemes kockáztatni.”

Biostatisztika – nade miért?

- „Azért, hogy el tudjuk dönteni, elhiggyünk-e valamit, amit olvasunk, vagy hogy észrevegyük, hol van benne a hiba, vagyis hogy ne dőljünk be olyan könnyen a statisztikai bűvészkedéseknek, műtermékeknek és tévedéseknek.” (*ld. excel file csodaszer, továbbiak később…)
- „Azért, hogy jobban meg tudjuk ítélni, szerencsénk volt-e vagy pechünk – vagy éppen egyik sem.”
- „Azért, hogy jobban meg tudjuk ítélni, mi mennyit ér, miért mennyit érdemes kockáztatni.”
- „Azért, hogy saját vizsgálataink tervezését, illetve kiértékelését ügyesebben el tudjuk végezni.” (diplomamunka…)
- „Érdekes, váratlan eredményt kaptam? Most felfedeztem valamit, vagy csak a véletlen játéka, amit látok?”
- „Azért, hogy eredményeinket érthetőbben és hatásosabban, a lényeget kiemelve tudjuk közölni.”
- „Azért, hogy pontosan értsük a szakirodalmat.”

(Reiczigel J. – Harnos A. – Solymosi N.: Biostatisztika nem statisztikusoknak)

A statisztika kulcsszavai

VÁLTOZÉKONYSÁG

Sztohasztikus

Véletlen

Tatisztika? Ammeg mi?

(Békásmegyeri aluljáró „átlagos” „lakója”)

Tatisztika? Ammeg mi?

(Békásmegyeri aluljáró „átlagos” „lakója”)

A **statisztika** a véletlen tömegjelenségek leírója.

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**

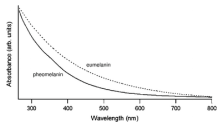
Leíró statisztika

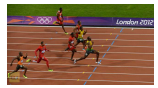
- Adatok elemzése
- Következtetések levonása

Következtető statisztika
(induktív statisztika)

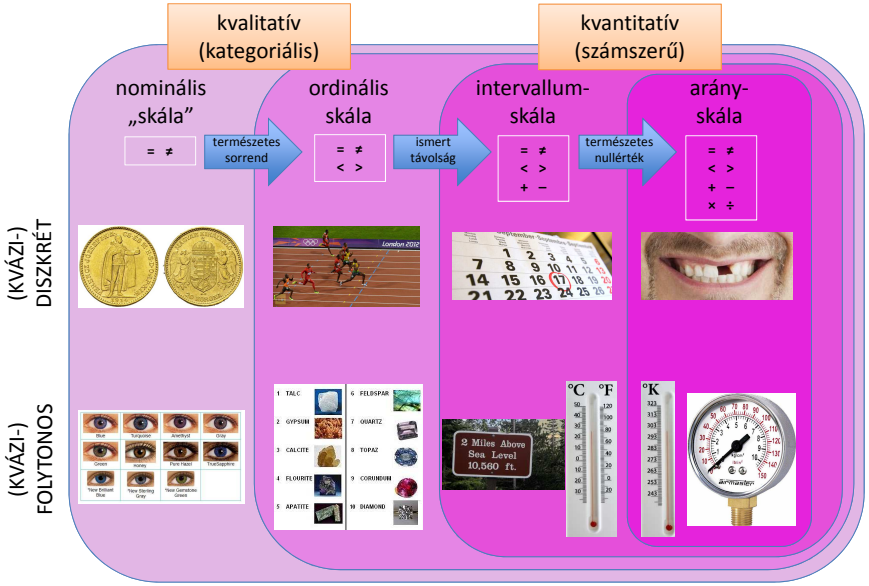
Változók, kimenetelek

Amit meg tudunk mérni vagy meg tudunk figyelni.





Változók típusai, mérési skálák



Nominális változó jellemzése I.

Analitikus

Lista

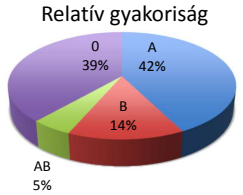
páciens sorszáma	vércsoport (ABO)	koleszterinszint (mg/dL)
1	0	149
2	A	125
3	A	127
4	B	159
5	A	134
6	A	153
7	0	164
8	0	140
9	A	139
10	B	150

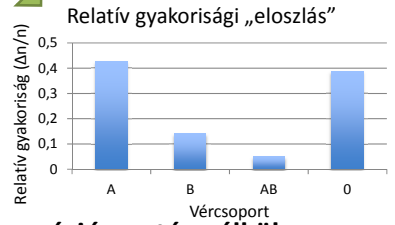
Gyakorisági sor (tábla)

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

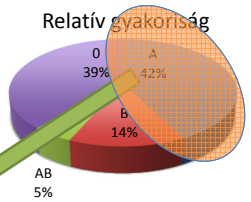
Gyakorisági sor (tábla)

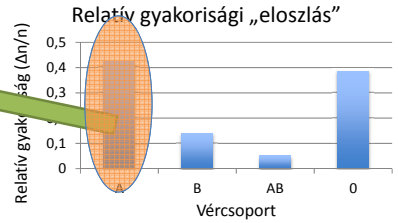
vércsoport	(abszolút) gyakoriság	relatív gyakoriság
A	85	0.425
B	28	0.14
AB	10	0.05
0	77	0.385
Σ	200	1

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

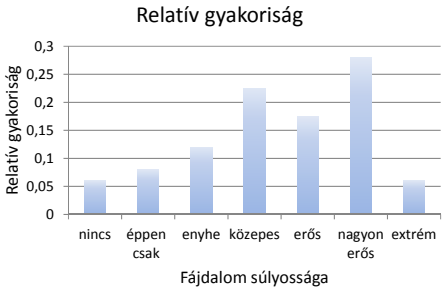
Gyakorisági sor (tábla)

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



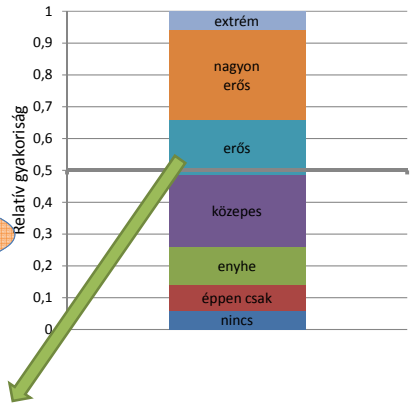
Ordinális változó jellemzése II.

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

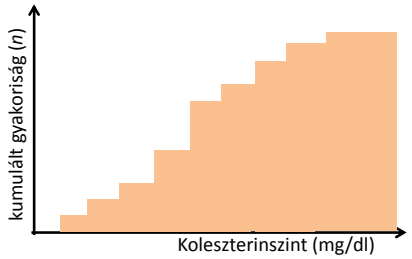
Gyakorisági sor (tábla)

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	

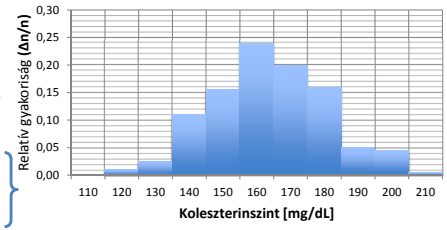
Adatrendezés **információvesztéssel** jár**hat.**

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

Grafikus

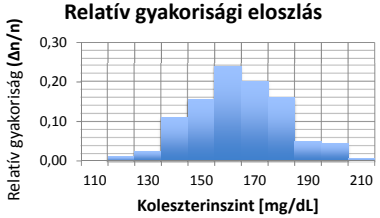


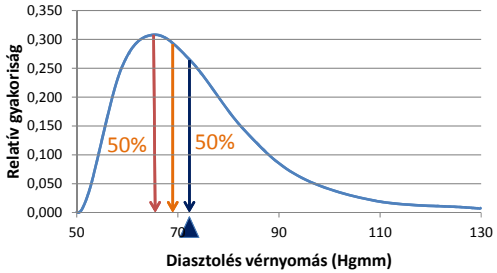
Relatív gyakorisági eloszlás



Kvantitatív változó jellemzése II.

Relatív gyakorisági eloszlás



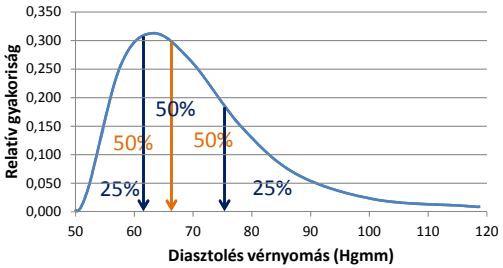


Jellemzők – ***középértékek*** (speciális ***helyparaméterek***):
- Módusz(ok): leggyakoribb elem(ek) ?
- Medián: „középső” elem(ek?)
- ***Átlag*** (számtani közép): „súlypont” ,érzékeny a „kiszóró” adatokra ?!
  Jelölése: $x_{átl}$, $\overline{x}$

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

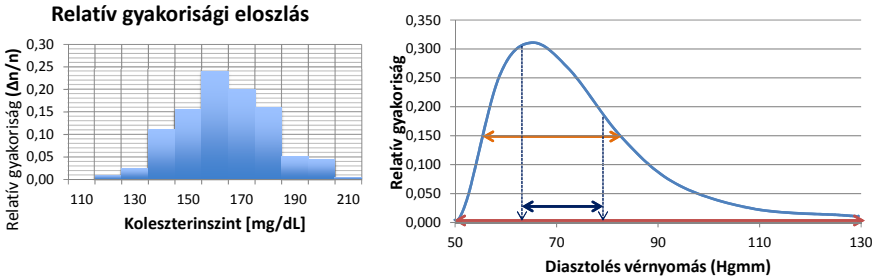
Kitérő I.

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?

Kvantitatív változó jellemzése III.

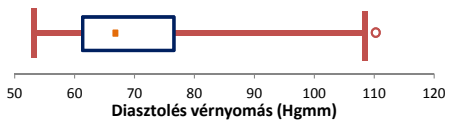


Jellemzők – szóródási paraméterek:

- ***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)



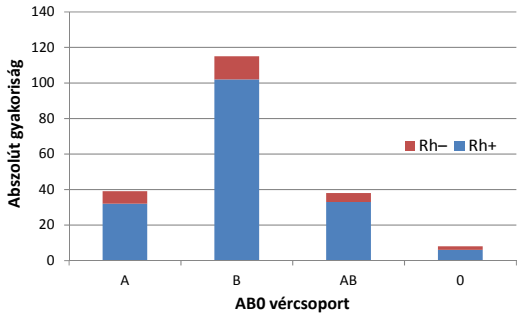
Sodrófa szeme: átlag, illetve *medián*
Sodrófa teste: átlagtól mért szórás, illetve *interkvartilis távolság*
Sodrófa szára: minimum és maximum értékek, 0,5-ös és 0,95-ös kvantilisek, szórás 2-szerese, *IQR 1,5-szerese*...
sodrófa szárán túl: ***kiszóró pont***

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

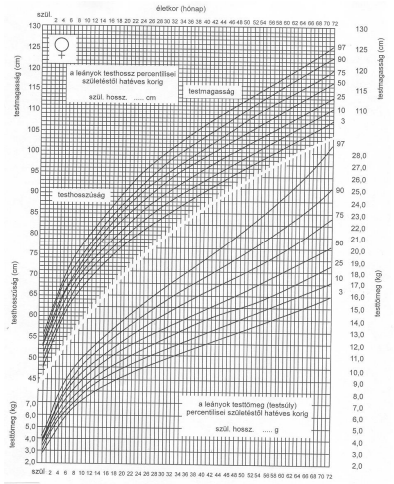
Grafikus: ***mozaik ábra***

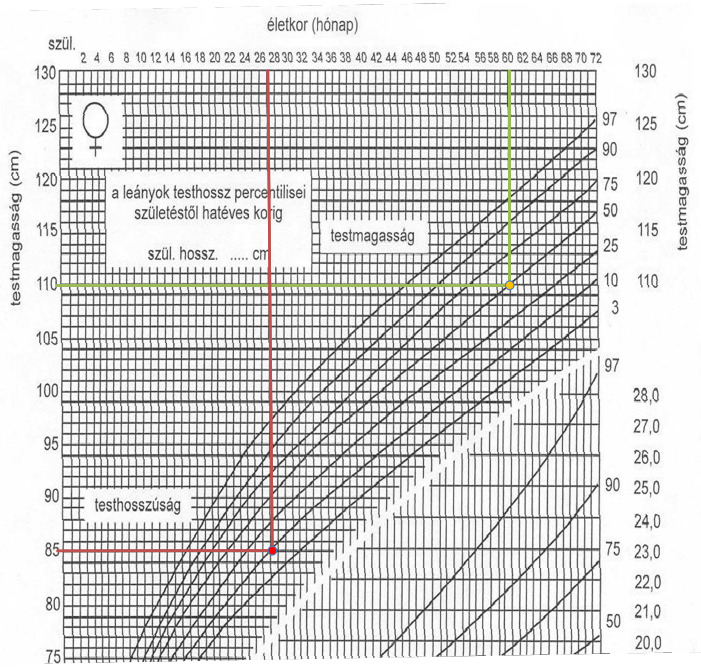


Több kvantitatív változó jellemzése

Grafikus: ***percentilis ábrák***

Percentilis: %-ban kifejezett kvantilis





Yoshino K *et al.*

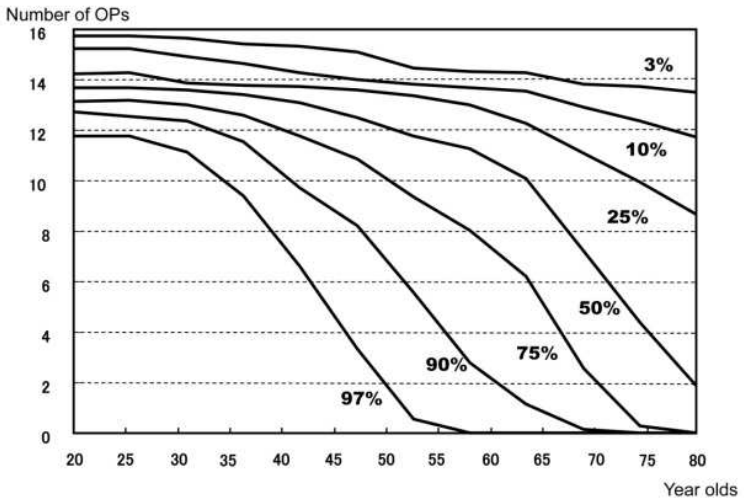
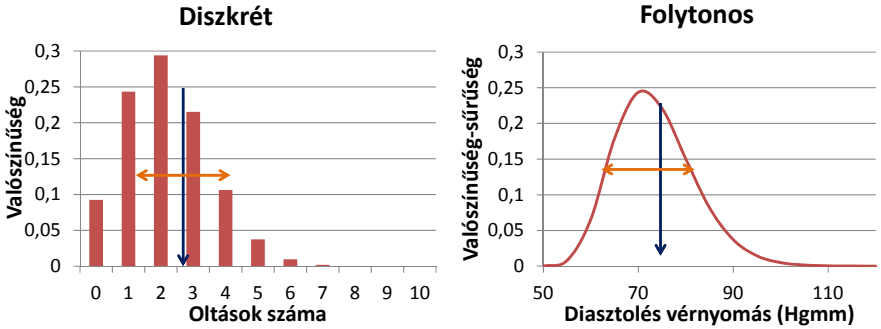


Fig. 1 Percentile curves of occluding pairs in males (n = 1,535)

Elméleti eloszlások



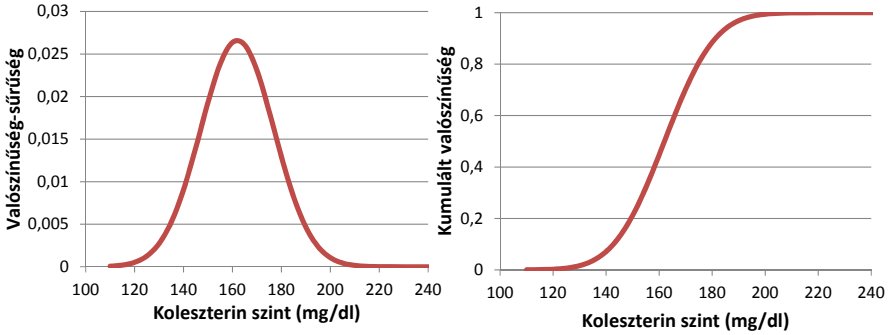
- ***Várható érték (E, M, μ) (hely paraméter)***

$$E(\xi) = \sum_{i=1}^{m} p_i \cdot x_i \qquad E(\xi) = \int_{-\infty}^{\infty} p_i \cdot x_i$$

- Elméleti szórásnégyzet (Var, D^2, σ^2) (szóródási paraméter)

$$Var(\xi) = E\left[\left(\xi - E(\xi)\right)^2\right]$$

Normál (Gauss) eloszlás I.



Koleszterinszin, vércukorszint....
Testmagasság, BMI
Diasztolés vérnyomás felnőtteknél
.....

$$E(\xi) = \mu$$

$$Var(\xi) = \sigma^2$$

$$P = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Normál (referencia) tartomány: 95% az adatoknak itt: $\sim \mu \pm 2 * \sigma$

Normál eloszlás II.

Centrális határeloszlás tétele (változókra): ha sok független valószínűségi változót összegzünk, akkor elég általános feltételek teljesülése esetén az összeg normális eloszlású valószínűségi változó lesz.

Centrális határeloszlás tétele (mintavételi átlagokra): ha egy adathalmazból *n* elemű mintákat veszünk, akkor elég általános feltételek teljesülése esetén a minták átlagai normál eloszlásúak lesznek, és az eloszlás varianciája az eredeti eloszlás varianciájának *n*-ed része lesz.

Tatisztika? Ammeg mi?

- Adatgyűjtés
- Adatok rendszerezése, áttekintése

→ Leíró statisztika

- Adatok elemzése
- Következtetések levonása

→ **Következtető statisztika (induktív statisztika)**

Alapsokaság és minta

Alapsokaság (populáció)

Az **alapsokaság** rendszerint olyan méretű, hogy az összes eleme nem vizsgálható meg.

Alapsokaság és minta

Alapsokaság (populáció)

„VÉLETLEN”

Minta

Az **alapsokaság** rendszerint olyan méretű, hogy az összes eleme nem vizsgálható meg.

Emiatt az alapsokaságnak csak egy részhalmazát vizsgáljuk, ezt nevezzük mintának.

Alapsokaság és minta

Alapsokaság (populáció)

„VÉLETLEN”

Minta

Az **alapsokaság** rendszerint olyan méretű, hogy az összes eleme nem vizsgálható meg.

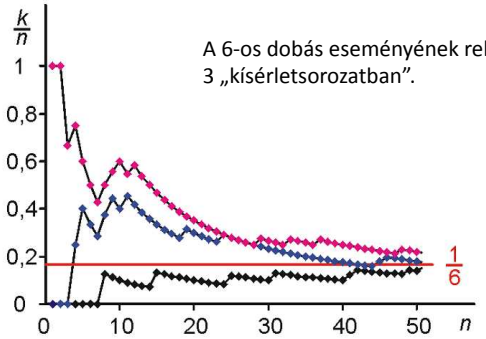
Emiatt az alapsokaságnak csak egy részhalmazát vizsgáljuk, ezt nevezzük mintának.

BIZONYTALANSÁG! (P és H)

A minta jellemzői alapján az alapsokaságra vonatkozó következtetést vonhatunk le

A minta elemein méréseket végzünk, majd az így keletkező adathalmazt (amit szintén mintának nevezünk) grafikusan és matematikailag jellemezzük

Valószínűség I.

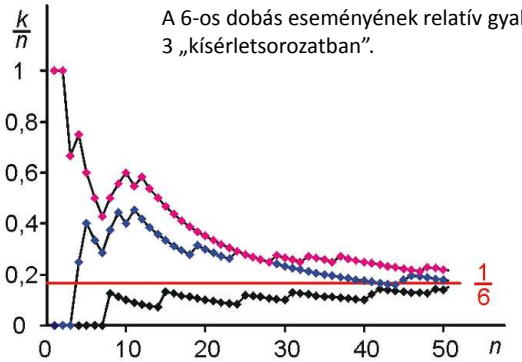


A 6-os dobás eseményének relatív gyakoriságai 3 „kísérletsorozatban”.

$$\text{Rel.gyak.} = \frac{N_{"kedvező"}}{N_{összes}}$$

Azt tapasztaljuk, hogy a **relatív gyakoriságok** ilyen sorozatai – bár ingadozásokat mindig mutatnak – a „kísérletsorozat” hosszának növekedtével egyre inkább **stabilizálódnak valamilyen érték körül.** Továbbá ez az érték az aktuális „kísérletsorozattól” függetlenül lényegében ugyanakkora.

Valószínűség, mint mennyiség?



A 6-os dobás eseményének relatív gyakoriságai 3 „kísérletsorozatban”.

$$P = \text{Rel.gyak.} = \frac{N_{"kedvező"}}{N_{összes}}$$

$$N_{összes} \to \infty$$

A **nagy számok *(**relatív gyakoriságokra vonatkozó**)* tapasztalati törvénye:** a relatív gyakoriság értéke egy végtelen sorozatban egy adott értékhez tart.
Az adott **eseményhez** hozzárendelhetjük ezt az **értéket**: 6 dobáshoz az 1/6-ot.
Ezt az értéket nevezzük az **esemény valószínűségének**.

Ez a törvény *tapasztalati* törvény, tehát logikai úton nem bizonyítható.

Események valószínűségei I.

Események valószínűségének alaptörvényei ***(Kolmogorov-axiómák)***:
1. **0 ≤ P(A) ≤ 1**

2. **P(*biztos*) = 1** *(a páciens előbb vagy utóbb meghal)*
P(*lehetetlen*) = 0 *(a páciens teljesen egészséges*)*

Események valószínűségei I.

Események valószínűségének alaptörvényei ***(Kolmogorov-axiómák)***:
1. **0 ≤ P(A) ≤ 1**

2. **P(*biztos*) = 1** *(a páciens előbb vagy utóbb meghal)*
P(*lehetetlen*) = 0 *(a páciens teljesen egészséges*)*

*3. Egymást kölcsönösen kizáró eseményekre: (*P(AésB)=0)
P(AvagyB)=P(A)+P(B)
*(annak a valószínűsége, hogy páciensünk terhes **vagy** férfi)*

Ezekből levezethető:
+4. *Független eseményekre*: **P(AésB)=P(A)*P(B)**
*(annak a valószínűsége, hogy az első páciensünk férfi **és** a második nő)*

Események valószínűségei II.

Feltételes valószínűség számítása
általános forma 2 eseményre: **P(A|B)=P(AésB)/P(B)**

Események valószínűségei IIa.

Feltételes valószínűség számítása
általános forma 2 eseményre: **P(A|B)=P(AésB)/P(B)**

Különleges esetek:
I. Független eseményekre:
annak a valószínűsége, hogy a második páciensünk férfi,
***HA** az első nő volt*
P(A|B)=P(AésB)/P(B)
P(A|B)=P(A)*P(B)/P(B)
P(A|B)=P(A)

annak a valószínűsége, hogy a második páciensünk férfi,
***HA** az első nő volt = annak a valószínűsége, hogy a második páciensünk férfi*

Események valószínűségei IIb.

II. A esemény részhalmaza B eseménynek:
annak a valószínűsége, hogy a páciensünknek influenza fertőzése van
HA ismert, hogy fertőzése vírusos eredetű

P(A|B)=P(AésB)/P(B)
P(A|B)=P(A)/P(B)

Számolási példa:
Annak a valószínsége, hogy páciensünknek vírusos fertőzése van:
P(B)=8%
Annak a valószínűsége, hogy páciensünknek influenza fertőzése van:
P(A)=2%
annak a valószínűsége, hogy a páciensünknek influenza fertőzése van
HA ismert, hogy fertőzése vírusos eredetű:
P(A|B) = 2% / 8 % = 25%.

Valószínűségszámítás.....

Permutációk
Variációk
Kombinációk

Na mire is lehet jó....

Influenzaszezont megelőzően a rendelőnkben az adott napra 4 oltóanyag áll rendelkezésre. Az előző években átlagosan 2989 páciensből 402 személyt kellett beoltanunk. Az előző év alapján mekkora a valószínűsége, hogy a rendelkezésre álló 4 oltóanyag elegendő lesz és el is fogy, ha 25 embert várunk aznapra?

$$P=\binom{n}{k}\cdot(p)^{k}\cdot(1-p)^{(n-k)}=\binom{25}{4}\cdot\left(\frac{402}{2989}\right)^{4}\cdot\left(1-\frac{402}{2989}\right)^{(25-4)}\approx 0{,}2$$

Na mire is lehet jó....

Mekkora a valószínűsége annak, hogy páciensünk 3.45 mmol/l-es (normál tartományon kívüli) K^+ szintje még „egészséges”?

Hány szülés várható az esti ügyeletben, ha az éves statisztika 1000 szülést mutat éjfél és 8:00 között?

Az évfolyamból várhatóan hányan lesznek alkalmasak egy csípőprotézis elvégzésére (tömegük alapján)?

Vajon hat-e az adott gyógyszer?...

Az influenza/AIDS teszt pozitív – milyen valószínűséggel vagyok tényleg beteg?

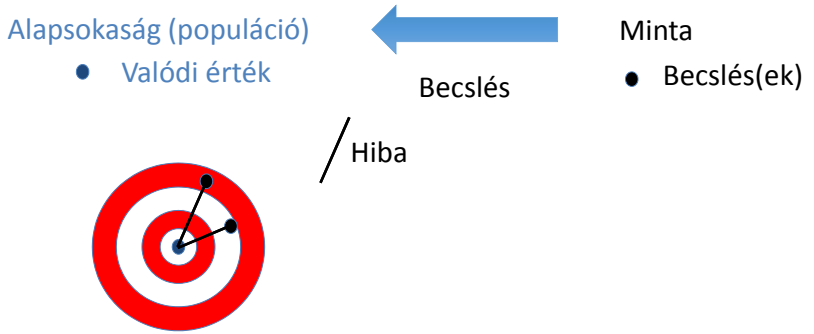
..... Hogyan számoljunk? Ismerjük a „képletet”? milyen „egyenletet”, táblázatot, excel függvényt... válasszunk, mikor melyiket?

Az emberi gondolkodás...

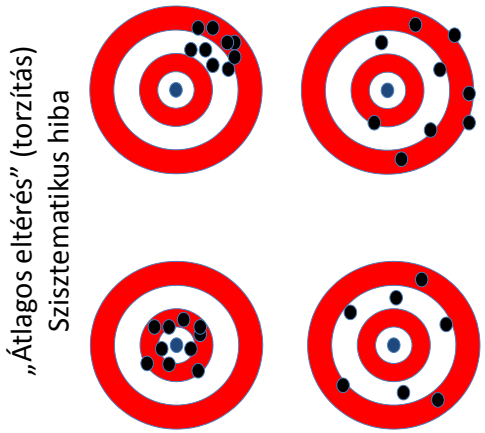
Linda tehetséges, független, filozófia szakot végzett 31 éves nő. Nagyon érzékeny a társadalmi igazságtalanságokra. Diákként részt vett az antinukleáris demonstrációkban. Sorszámozza meg az alábbi állításokat aszerint, hogy mennyire tartja valószínűnek (1-es sorszám a legvalószínűbb):

a) Linda tanító egy általános iskolában,
b) Linda könyvesboltban dolgozik, és jóga tanfolyamra jár,
c) Linda a nőszavazók ligájának tagja,
d) Linda bankpénztáros,
e) Linda biztosítási ügynök,
f) Linda bankpénztáros és feminista.

Hiba



Hiba – 2 dimenziója



Jó becslés, ha:
Torzítatlan:
az ingadozásának „közepe” (várható értéke) a valódi érték
Hatásos:
az ingadozás lehető legkisebb
(Konzisztens:
az elemszám növelésével csökken az ingadozás)

Ellenőrző kérdések#1

- 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.
- Definiáld a mediánt.
- Mik tartoznak a középértékek közé?
- Mi az átlag, a medián, a módsz a terjedelem, az interkvartilis terjedelem és a szórás szemléletes jelentése?
- Hogyan határozható meg egy minta átlaga?
- Melyik középérték érzékeny a kiszóró értékekre?
- Mi a középértékek előnye az eloszlásgörbével szemben?
- Melyek a helyparaméterek?
- Melyek a szóródási paraméterek?
- Definiáld a varianciát.
- Definiáld a szórást.
- Definiáld az interkvartilis távolságot.
- Mi az a sodrófadiagram?
- Milyen részei vannak a sodrófadiagramnak?
- Mit tudunk leolvasni a percentilis görbékről?
- Definiáld a valószínűséget a nagy számok törvénye alapján.
- Ismertesd a nagy számok törvényét.
- Hogyan bizonyítható a nagy számok törvénye?
- Mik a Kolmogorov axiómák?
- Mit tudsz A és B esemény viszonyáról, ha P(AvagyB)=P(A)+P(B) igaz?
- Mit tudsz A és B esemény viszonyáról , ha P(AB)=P(A)*P(B) igaz?
- Miről szól a centrális határeloszlás tétele?
- Mik a jó becslés tulajdonságai?
- Mit jelent a torzítatlan becslés?