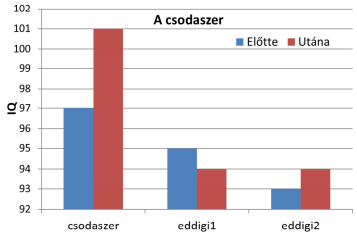
Biostatisztika I.
fogorvostan hallgatóknak

1. előadás:
Biostatisztika I.
2020. Szeptember 7.
Veres Dániel

Biostatisztika – nade miért?

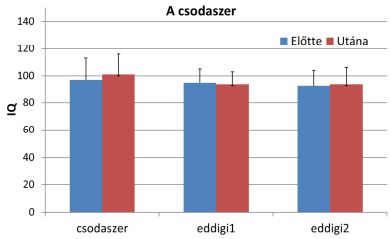
- „Azért, hogy el tudjuk dönteni, elhiggyünk-e valamit, amit olvasunk, vagy hogy észrevegyük, hol van benne a hiba, vagyis hogy ne dőljünk be olyan könnyen a statisztikai bűvészkedéseknek, műtermékeknek és tévedéseknek."



Miért tanuljunk statisztikát – úgy általában, illetve hallgatói pályafutásunkhoz.
1. Statisztikai bűvészkedés pl. ábrával

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."



Az ábra nulla pontjának, illetve az eredmények változatosságának feltüntetésével jobban érzékeljük a valós helyzetet.

+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

További érdekességek elérhetőek:
Korreláció – ok-okozat:
https://www.fastcompany.com/3030529/hilarious-graphs-prove-that-correlation-isnt-causation
Többszörös összehasonlítás: 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." (pl. pozitív eredmény egy nagy érzékenységű teszttel influenza és AIDS esetében)
- „Azért, hogy jobban meg tudjuk ítélni, mi mennyit ér, miért mennyit érdemes kockáztatni." (pl. adott kezelés kockázata)

Miért tanuljunk statisztikát – úgy általában.

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." (pl. pozitív eredmény egy nagy érzékenységű teszttel influenza és AIDS esetében)
- „Azért, hogy jobban meg tudjuk ítélni, mi mennyit ér, miért mennyit érdemes kockáztatni." (pl. adott kezelés kockázata)
- „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)

- „Azért, hogy együtt tudjunk dolgozni egy statisztikussal"

Miért tanuljunk statisztikát – úgy általában, illetve hallgatói pályafutásunkhoz.
A fenti idezétek a Reiczigek J., Harnos. A., Solymosi N.: Biostatisztika nem statisztikusoknak könyvből származik. Javaslom mindenkinek a könyvet.

A statisztika kulcsszavai

VÁLTOZÉKONYSÁG

Sztohasztikus

Véletlen

A világunk egyik érdekessége, hogy változatos: az emberek (páciensek), jelenségek bár hasonlóak, de sok tulajdonságukban kicsit mások: VÁLTOZÉKONYAK. Ezt a változékonyságot meg kell figyelnünk, mert egy adott pillanatban meglévő ismereteink alapján nem tudjuk egyértelműen meghatározni megfigyelés nélkül az adott tulajdonságot, annak értékét, jövőbeli viselkedését – azaz ezen tulajdonságok sztochasztikusak. Ezen túlmenően általában nem tudunk megfigyelni minden lehetséges pácienst, esetet – ezeknek csak egy részhalmazát látjuk, amely véletlenszerű, hogy mely elemeket tartalmazza.

Tatisztika? Ammeg mi?

(Békásmegyeri aluljáró „átlagos" „lakója")

Ebben az előadásban alapvető statisztikai fogalmakat írunk le – így remélhetőleg tudunk majd válaszolni egyszerűen a fenti kérdésre…

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)

Bár számos definíció létezik a statisztikára, én mégis egy újabbat adok: *a statisztika a véletlen tömegjelenségek leírója*.
A statisztika, azaz véletlen (azaz ahogyan tanultuk korábban *egyénre vonatkozóan előre meg nem határozható*) tömegjelenségek – tehát a *számos mérhető vagy megfigyelhető tulajdonságok* – jellemzéséhez a következő tevékenységek tartoznak: *adatgyűjtés, adatok rendszerezése, áttekintése, adatok elemzése és a következtetések levonása*. Az első kettő a *leíró statisztika* tárgykörébe, míg az utóbbiak a *következtető statisztikához* (más néven induktív statisztika) tartoznak. Megjegyzendő azonban, hogy ezen tevékenységek között a határvonal nem éles. Kiemelném még azt is, hogy a leíró statisztika mindig a következtető statisztika alapja: mind a megfelelő adatgyűjtés, mind a megfelelő rendszerezés és áttekintés elengedhetetlen az adatok elemzéséhez és helyes következtetések levonásához.

Tatisztika? Ammeg mi?

- Adatgyűjtés
- **Adatok rendszerezése, áttekintése**

Leíró statisztika

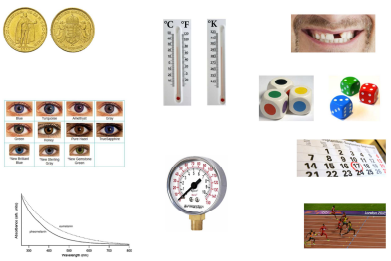
- Adatok elemzése
- Következtetések levonása

Következtető statisztika (induktív statisztika)

Az adatgyűjtés néhány lényeges momentumára még visszatérünk később, illetve néhányra a későbbi előadásban is; most az adatok rendszerezését vesszük górcső alá. Az adatok rendezése, áttekintése segít az adathalmaz jelentéssel bíró leírásában, összefoglalásában – a helyes adatrendezés kiemelhet számunkra lényeges mintázatokat, lehetséges összefüggéseket, érdekességeket, továbbá ötletet adhat a további elemzésekhez is.

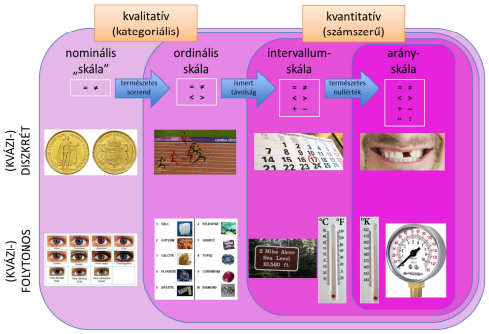
Változók, kimenetelek

Amit meg tudunk mérni vagy meg tudunk figyelni.



A statisztikában az adatok különböző *változó*khoz tartoznak.
A változó egyszerűsített definíciója: olyan „jelenség”, amit meg tudunk mérni, vagy meg tudunk figyelni. Például egy érme feldobásának eredménye, a szem vagy hajszín, hőmérséklet, vérnyomás...
A változó adott körülmények között, adott esetben mérhető, megfigyelhető „eredményét”, „értékét” a változó *kimenetel*ének nevezzük.
Például az érmefeldobáskor kapható kimenetelek a fej és az írás; a vérnyomás kimenetele lehet alacsony, közepes, magas – de megadhatjuk konkrét számmal is: 75 Hgmm; 125 Hgmm; 160 Hgmm
Ez utóbbi példa alapján is látható, hogy nagyon lényeges, hogy a *változót az adott kimeneteleivel együtt* értelmezzük, használjuk.

Változók típusai, mérési skálák



Számos módon csoportosíthatjuk a változókat, én egy gyakorlati szempontból hasznos csoportosítást mutatok itt be, amely a *változó kimeneteleinek tulajdonságain* alapul. Első lépésben 2 nagyobb csoportot különíthetünk el: *kvalitatív (kategoriális)* és *kvantitatív (számszerű)* változók csoportját (ahogyan ez az előző előadáson is szerepelt). A változók további besorolása az úgynevezett *mérési skálák*on történik.

A legprimitívebb skála, a *nominális* vagy névleges skála, mely a mérészinthierarchia alján áll. Ilyen lehet például maga a névadás, vagy a vércsoport, a hajszín, szemszín, állampolgárság stb. A skála létrehozása úgy történik, hogy kategóriákat hozunk létre, a kategóriák egyszerű névadással azonosíthatók. Az egyes megfigyelések során megállapítható, hogy *két elem azonos vagy nem azonos*. A kategóriák között nincs természetes sorrend, de praktikus okokból kialakíthatnak (ABC-rend, sorszámmal való jelölés), amiket a szokásoknak megfelelően használnak,

hogy később könnyebb legyen az összehasonlítás. Azonban ezeknek a *sorrendeknek semmiféle mennyiségi jelentése* nincs. Emiatt a skála megnevezés is kissé megtévesztő, helyesebb inkább rendszert használni, ami nem enged természetes sorrendiségre következtetni. A kategóriák elhatárolása lehet könnyebb (magától értetődő, pl. fej, írás) vagy nehezebb (mesterséges, pl. szemszín).

Az *ordinális* skála szintén kategóriákat jelöl ki, azonban ezek között már *természetes, jelentéssel bíró sorrend* van, ilyen például az iskolai osztályzat, a betegségek, sérülések súlyossága vagy a Mohs-skála. Az ordinális skálán tehát nem csak azonosságot tudunk megállapítani, hanem *kisebb/nagyobb relációt* is. A skálaelemeket rendszerint sorszámmal jelölik, amit észben kell tartani, hiszen sorszámokon nem végezhetők el a szokásos matematikai műveletek. Az *ordinális skála kategóriái közötti eltérés vagy távolság nem egyenlő vagy nem tudjuk megállapítani*.

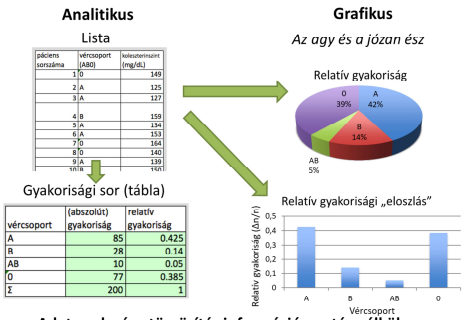
Az *intervallumskála* annyiban fejlettebb az ordinális skálánál, hogy ismert a felvehető *értékek közötti távolság*, vagyis már nem csak a sorrend, hanem a különbség és az összeg *is értelmezhető*. A mindennapi életből ismertpéldák pl. az évszám, az adott nap, a Celsius-fokban mért hőmérséklet vagy a tengerszinthez viszonyított magasság. A példákból látható az intervallumskálák egy további közös tulajdonsága: a nullapont kijelölése egyezmény alapján történik.

Ezen konvencionális nullaérték helyett *természetes nullaértéken* alapulnak az *arányskálák*: az arányosságot maga a természetes nullapont létezése teszi lehetővé. Az ilyen skálákon így már az arányossághoz kapcsolódó műveletek, az *osztás és a szorzás is értelmezhető*.

Mindegyik skálaszinten elkülöníthetők többé-kevésbé *diszkrét és folytonos* változók. Nominális változókat tekintve például az érmefeldobásnál egyértelműek a diszkrét kategóriák, míg a szemszín esetén eléggé homályos az egyes kategóriák határa, illetve a kategóriák száma, igazából csak rajtunk múlik, hogy mennyire finom felosztást hozunk létre. A gyakorlatban diszkrétnek szoktuk tekinteni a változót, ha kimeneteleinek száma kisebb, mint 20; folytonosnak, ha kimenetelek száma legalább 20 a mintában.

A statisztika szempontjából lényeges – amint ezt később látni fogjuk – hogy hány, illetve milyen változótípus jelenik meg az adathalmazban.

Nominális változó jellemzése I.



páciens sorszáma	vércsoport (AB0)	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

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

Adatrendezés - tömörítés információvesztés nélkül.

Kezdjük az adatrendezés leírását egy nominális változó jellemzésével. Példánkban az AB0 vércsoport szerepel, mint nominális változó.
Minden statisztikai jellemzésnél alapvetően kétféle lehetőségünk van: *analitikus* és *grafikus* elemzés. Az analitikus leírásban alapvetően számokat, illetve csoportazonosítókat használunk, amíg a grafikus leírásban ábrákat. Soha ne feledjük a grafikus ábrázolást. Az emberi agy ezt jól fel tudja dolgozni – a „józan paraszti ész” sokszor elengedhetetlen, hogy lássuk a számok mögötti értelmet, hibákat és ebben az ábrázolás sokat segít.
Az adatgyűjtést követően egy *lista* áll rendelkezésünkre. Ebből a listából állíthatjuk elő a *gyakorisági sort* (megszámláljuk az adott kiemenetelek számát, vagy ennek arányát fejezzük ki a mintában – abszolút, illetve relatív gyakoriságokat képezve), illetve *gyakorisági eloszlást mutató ábrákat*.
Nominális változók esetében a lista tömörítése a gyakorisági sorral, illetve a gyakorisági ábrákkal *nem okoz információvesztést* – azaz a változót leíró eredeti adathalmaz visszaállítható (amennyiben ez a változó érdekes csak számunkra, az nem, hogy kihez tartozik az adott vércsoport, illetve milyen adatai vannak még az adott páciensnek).

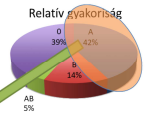
Nominális változó jellemzése II.

Analitikus **Grafikus**

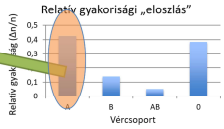
Gyakorisági sor (tábla)

Az agy és a józan ész

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.



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

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

További elemzésekhez, összehasonlításokhoz túl sok az „információ” – valamilyen jellemző eredményt (mutatót) kell megadnunk.
Ebben az esetben ez a *módusz*, azaz a *legnagyobb gyakoriságú elem(ek)*. Ennek a megoldásnak azonban hátránya is van: csak a módusz ismeretében nem állítható vissza az eredeti adathalmaz – *azaz információt vesztettünk*.
Ismét felhívnám a figyelmet az ábrákra. A grafikus ábrázolásból első pillantásra látszik a módusz, de az is, hogy az A és 0 gyakoriságai között kicsi a különbség – a példa esetében a módusz „nem annyira jó”, viszont nincs más lehetőségünk. (Az átlag itt nem értelmezhető – mit is jelenthetne az AB,AB0A például?)
A minta leírásához kapcsolódó egyéb jellemzők a minta elemszáma és a kategóriák száma – józan ésszel belátható, hogy ezek a paraméterek is lényegesek további elemzéseinkhez (túl kevés elemszám, túl sok kategória megnehezíti az információ feldolgozását).

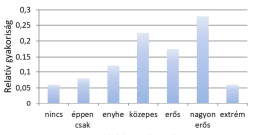
Ordinális változó jellemzése I.

Analitikus **Grafikus**

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

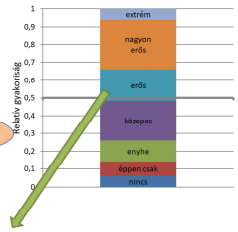
Ordinális változóra hozott példánk legyen a fájdalom mértékének súlyossági szubjektív skálája.
Az ilyen típusú változónál már van értelme egy másfajta eloszlásnak is: a kumulatív eloszlásnak. A fájdalom skála esetében a kumulatív eloszlás megadja egy adott fájdalomnál nem nagyobb fájdalomérzet gyakoriságát – összegezzük az adott értéknél kisebbekeket.
Ordinális változók esetében a nominális változónál ismertetett megoldások mind alkalmazhatóak a jellemzésre. De tudunk-e adni további jellemzőt, kihasználva a sorbarendezhetőséget?

Ordinális változó jellemzése II.

Analitikus **Grafikus**

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



Új Jellemző (információvesztéssel):
Medián: „középső” elem(ek)
Jelölése: Me, Med, x_{med}

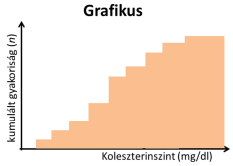
A sorbarendezhetőséggel egy új jellemzőt is találhatunk, a *mediánt*, amely megmutatja egy sorba rendezett adatsorban a *„középső” elemet(ek)*, „középső pontot(ok)” az ábrán. Ezt azt jelenti tehát, hogy az adatok 50%-a „alatta”, míg 50%-a „felette” helyezkedik el a kumulált gyakoriági sorban. Jelen esetben a minta mediánja az erős fájdalom. A ()-es többes számra, illetve az idézőjelekre még később visszatérünk, de hasonlóan arra is, hogy a medián, mint felező érték, mintájára nem használhatnánk-e negyedelő, ötödölő... értékeket is

Kvantitatív (számszerű) változó jellemzése I.

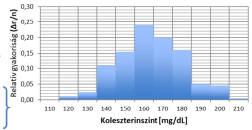
Analitikus **Grafikus**

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	



Relatív gyakorisági eloszlás



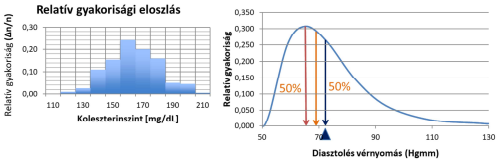
Adatrendezés **információvesztéssel** jár**hat.**

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

A következő diákon a *kvantitatív (számszerű)* változók (de egyszerre csak egy) leírását tekintjük át.
Ebben az esetben a veszteségmentes tömörítés grafikusan, kumulatív gyakorisági függvény ábrázolásával lehetséges.
Ennél a mérési skálánál, ha a változó folytonos (mint a legtöbb esetben), a minta gyakorisági függvényeinek létrehozásához mesterségesen *osztályokat* (intervallumokat, vagy az excel terminológiájával élve bineket) kell meghatároznunk. Az így végzett adatrendezés egyrészt *információvesztéssel* jár, másrészt pedig felmerül a kérdés, hogy hogyan alkossuk meg az osztályokat?
Az *osztályszélesség meghatározására* alapvetően két megoldásunk van. Az egyik *statisztikai szempontok* alapján határozza meg azt, például az osztályszélesség = (maximális-minimális érték)/(elemszám négyzetgyöke). A másik a *szakmai illetve „szépészeti” szempontok* alapján határozza meg az osztályszélességet. Ebben az esetben kevésbé tudunk egzakt megoldást mondani, de néhány

példát említenék. Például nincs értelme kisebb osztályszélességet használnunk, mint a legkisebb mérhető különbség. Továbbá érdemes egészszámokat használnunk az osztályhatároknál, ha a mérendő értékeink is csak egész számok lehetnek. „Szépészeti” szempontból pedig elmondható, hogy osztályhatároknak a „kerek számokat” szeretjük, például 0;5;10;15... vagy 10;20;30...
Összességében elmondhatjuk, hogy az osztályszélesség meghatározására bár két szempontunk lehet, de (amennyiben lehetséges) mindkettőt figyelembe kell vennünk. Javaslom, hogy először statisztikai szempont szerint határozzuk meg az osztályszélességet, majd kerekítsük ezt felfelé a szakmai, esztétikai szempontoknak megfelelően.

Kvantitatív változó jellemzése II.



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

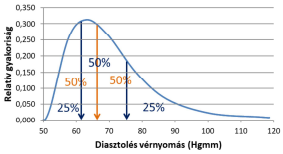
Előny: tömörítés, ***kevés adatból is számíthatóak***

Képletek: képlettárban

A kvantitatív változók leírására a nominális és ordinális változóknál ismertetett megoldaások mind alkalmazhatóak, valamint újabb lehetőségeink is vannak. Hogy jobban megérthessük a különböző jellemzők „jelentését” egy végtelen kicsi osztályszélességgel létrehozott (megfelelően nagyszámú) eloszlás grafikonján mutatnám be ezeket. A grafikon a 4 éves gyermekek diasztolés vérnyomását mutatja.
Az eloszlás „közepét” valamilyen módon jellemző jellemzőket *középértékeknek* nevezzük (ezek *speciális helyparaméterek*). Ezek a következők.
A *módusz(ok)*, azaz *leggyakoribb elem(ek)*, amely a legnagyobb gyakoriság(ok)hoz tartozik az ábrán, tehát a *grafikon csúcsá(ai)nak* értékére mutat.
A *medián(ok)* a görbe alatti területet 50-50%-os arányban osztja (felezi).
Az *átlag* a *görbe súlypontja*, azaz ha egy lapból kivágnám a görbét, akkor azt az átlag értékénél lehetne alátámasztani a kiegyensúlyozáshoz, mint egy libikókát.

Az ábráról leolvasható, hogy egy nem szimmetrikus (ferde – erre még később visszatérünk) eloszlás esetében a medián és az átlag – ebben a sorrendben –, az eloszlás „farka” felé tolódik. Ezen jellemzők előnye az eloszlásgörbékkel szemben, hogy kevés adatból is meghatározhatóak.

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

További helyparamétereket is meghatározhatunk a medián mintájára. Így például negyedelő pontokat, amelyek a görbe alatti területet negyedekre (például 25-75% arányban) osztják. Ezeket nevezzük *kvartilis*eknek. (A quartus latinul negyediket, quarta pars pedig negyedet jelent.) Pontosabban *alsó kvartilis*(ek)nek, vagy első kvartilisnek (1/4=0,25, Q1) azt a számot hívjuk, amely a görbe alatti területet 25-75%-ban osztja. *Felső kvartilisnek* (3. kvartilisnek, ¾=75, Q3), pedig amely 75-25% arányban oszt. Ehhez hasonlóan lehet definiálni a mediánt, mint 2. kvartilist.
Általánosításként használhatunk bármilyen osztópontot. Ezeket hívjuk kvantiliseknek. *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.

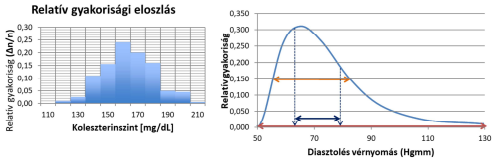
Nap sorszáma	Várakozási idő		
1	1,77	medián:	8,475
2	3,2	alsó kvartilis	3,59
3	3,44	átlag	7,723333
4	3,64		
5	6,33		
6	7,72		
7	9,23		
8	9,87		
9	10,31		
10	12,29		
11	12,3		
12	12,98		

Nap sorszáma	Várakozási idő		
1	1,77	medián:	8,475
2	3,2	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?

Ezen a dián megpróbálok rámutatni a korábban ?-lel, többes számmal, „”-lel jelölt néhány kérdésre.
A példában azt tüntettem fel, hogy az egyes napokon mennyi időt kellett várni a buszra. Ez a várakozási idő lesz a változó, amit vizsgálunk. Az ábrán a mért értékek nagyság szerint sorba vannak rendezve. A minta elemszáma 12.
Vizsgáljuk meg először, hogy miért is használtam többes számot a medián, a kvartilisek, illetve kvantilisek esetében. Az előző dián adott definíciónak megfelelően a medián keresése: p=0,5, így 12*0,5=6 adat kisebb, illetve ugyanennyi nagyobb a mediánnál. A definíció szerint ennek az állításnak minden 7,72 és 9,23 közötti szám megfelel, így ezek *mind mediánnak tekinthetőek*. Ugyanígy a definíciónak megfelelően minden 3,44 és 3,64 közötti szám az adathalmaz alsó kvartilise. A gyakorlatban (például excelben) azonban láthatjuk, hogy csak egy szám van megadva. Ezt különböző módokon számíthatják. A leggyakoribb megoldás (ahogyan excel is számítja), hogy az adott p-kvantilis „határszámait” 1-p, illetve p arányban (tehát fordítva, mint a kvanitilis érték) vesszük figyelembe. Például az alsó kvartilis (25-75%-os osztópont) elméletileg 3,44 és 3,65 közötti minden szám, tehát a „határszámok” a 3,44 és 3,64. Vegyük ezek különbségét, amely 3,64-3,44=0,2, majd adjuk hozzá az alsó értékhez ezen különbség 0,75-szeresét (0,15) és megkapjuk a gyakorlatban számított 3,59-es értéket.
Másodszor vizsgáljuk meg, hogy hogyan változik a medián, illetve az átlag, ha kiszoró („nagyon eltérő”) adatunk van (a kiszoró adatot majd később definiáljuk). Példánkban a legnagyobb elemet 12,98 helyett vegyük 30-nak. Jól látható, hogy amíg a medián változatlan maradt, addig az átlag erősen megváltozott. *Ezért mondjuk, hogy az átlag érzékeny, amíg a medián érzéketlen a kiszoró adatokra.*
Végül vizsgáljuk meg, hogy mi is az adathalmaz módusa? Azt mondhatjuk, hogy nincsen, vagy mindegyik elem az – ez azonban így nem bír jelentéssel. Tehát numerikus (és kisebb elemszámú) minta esetében gyakran nincs értelme meghatározni móduszt. Ebben az esetben legfeljebb gyakorisági eloszlás alapján (ha van elég adat és ezért van értelme elkészíteni) van értelme egy tartományt, mint móduszt számítani.

Kvantitatív változó jellemzése III.



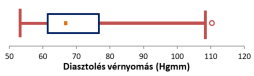
Jellemzők – szóródási paraméterek:

- ***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

A jellemzők egy másik részét képezik a *szóródási paraméterek*, amelyek a minta *változékonyság*át, az *eloszlásgörbe szélességét* mutatják. Ezek a jellemzők a következőek.
Terjedelem, amely a maximális és minimális érték különbsége.
Variancia (szórásnégyzet), amely az átlagtól vett átlagos négyzetes eltérés. Ha minta leírására használjuk, akkor a Bessel korrigált formál, míg ha populáció, mint minta leírására használjuk, akkor a korrigálatlan formát használjuk.
A *szórás* a variancia négyzetgyöke.
Az *interkvartilis távolság* a felső és alsó kvartilisértékek különbsége.
Amíg a terjedelem, a variancia és a szórás érzékeny a kiszóró adatokra, addig az interkvartilis terjedelem nem.

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***

A sodrófadiagram (más nevén box plot, vagy whisker plot) az adatok nagyon látványos grafikus leírását adja.
A következő részekből áll a sodrófadiagram. (A dián az ábrán használt jelölt paramétereket dőlt betűkkel jelöltem.)
A sodrófa szeme, amely általában a medián, ritkábban az átlag értéke. Szimmetrikus eloszlás esetén használhatjuk az átlagot (de a medián ekkor is jó), míg asszimetrikus eloszlás, vagy kiszóró pontokat tartalmazó adathalmaz esetében mindig a mediánt használjuk.
A sodrófa testeként (a box) általában az interkvartilis távolságot (1,5*IQR) adjuk meg, de a szórást, standard hibát (lásd későbbi előadáson) is feltüntethetjük némely esetekben. Ha az átlagot használtuk, mint a sodrófa szemét, akkor a szórást, vagy a standard hibát (ez utóbbit, ha kevés adatunk van) tüntessük fel. A medián mellett az interkvartilis távolságot szoktuk megjeleníteni.
A sodrófa száraként, ha az adathalmaz nem tartalmaz kiszóró értékeket, akkor a minimum és maximum értékeket használjuk. Egyébként a szórás 2-szeresét, illetve az

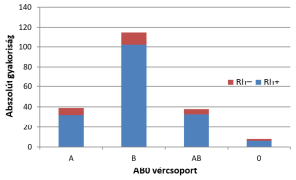
interkvartilis távolság 1,5-szeresét használjuk az átlag, illetve a medián mellett. Kiszóró adatnak az interkvartilis távolság 1,5-szeresén túlnyúló adatokat szoktuk tekinteni.
Amint az látható a sodrófadiagram elemei sokfélék lehetnek, én csak egy általánosan elfogadott javaslatot írtam le – ezt a javaslatot azonban *tudni kell*. A többféle megjelenítés miatt az is lényeges, hogy *mindig tüntessük fel, hogy mit használtunk a sodrófa elemeiként.*

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

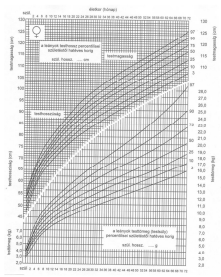
Grafikus: ***mozaik ábra***



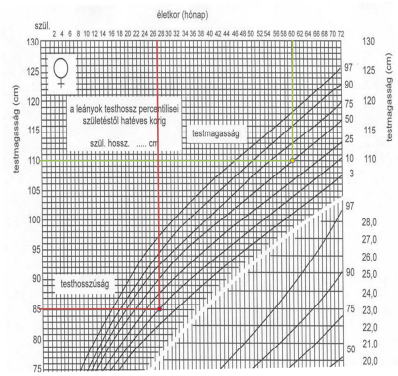
Több változó együttes leírása igen bonyolult. A következőkben csak néhányat emelek ki ezek közül. Több kvalitatív változó analitikus jellemzésére a *kontingencia táblázatokat* szoktuk használni. Grafikus megjelenítésre pedig kiválóan alkalmasak a *mozaik ábrák*. Ismét rámutatnék arra, hogy mennyivel egyszerűbb az ábrák értelmezése, mint a táblázaté, ha egyszerűen csak ránézünk.

Több kvantitatív változó jellemzése

Grafikus: ***percentilis ábrák***
Percentilis: %-ban kifejezett kvantilis



Több kvantitatív változó jellemzésére általában pontdiagramokat használunk.
Azonban az orvosi gyakorlatban (főleg a gyerekgyógyászatban) ehhez a jellemzéshez gyakran használunk úgynevezett percentilis görbéket. Egy ilyet tüntettem fel a dián is.



A görbék értelmezését az előadás során megbeszéltük. (A piros pont azt mutatja, hogy a 27 hónapos leányok 10%-ának testmagassága 85 cm alatt van.)

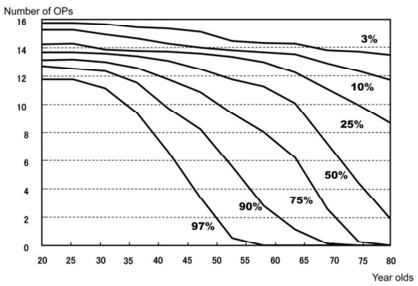


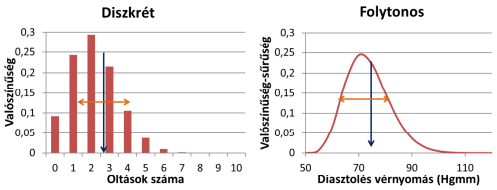
Fig. 1 Percentile curves of occluding pairs in males (n = 1,535)

Fogorvosi példa.

ismeretében bármely értékhez tartozó valószínűség meghatározható.

[Akit érdekel: Vizsgáljuk meg a diszkrét esetben kapott várható érték definíciót. Egy kis utánagondolással E felírható $E=(\Sigma(\text{abs.gyak.}_i * x_i))/n$, azaz az egyforma elemeket annyiszor adom össze, ahány van belőle (ez az $\text{abs.gyak.}_i * x_i$), és ezt minden elemnél megteszem – tehát összességében minden elemet összeadok, majd az így kapott értéket osztom az elemszámmal. Vegyük észre, hogy ez nem más, mint az átlag definíciója végtelen elemszám esetében.
Folytonos változó esetében ugyanezt az összegzést végzem el, csak végtelenül kicsi osztályszélességgel (ez az integrálás).]

Elméleti eloszlások



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

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

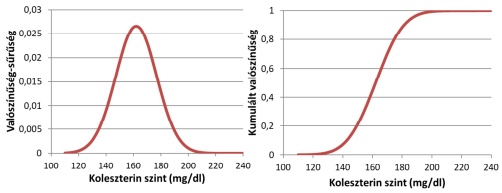
- Elméleti szórásnégyzet (Var, D^2, σ^2) (szóródási paraméter)

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

Vizsgáljuk meg először, hogy az elméleti eloszlásoknak melyek ezek a jellemzői, paraméterei.
Két alapvető jellemzőt használunk - a leíró statisztikában tanultakhoz hasonlóan: egy *közép értéket* és egy *szóródási paramétert*.
Kétféle eloszlást mutatok be az ábrán: egy folytonos (Hány Hgmm a diasztolés vérnyomása az embereknek) és egy diszkrét változó (hány oltás szükséges az előző év alapján) eloszlását.

A középérték jellemzésére a *várható értéket* (E, M vagy μ jelöléssel), a szóródás jellemzésére az *elméleti szórást* (Var, D^2 vagy σ^2 jelöléssel) használjuk. A várható érték az eloszlás „közepét” mutatja (ahogyan a minta esetén a módusz, a medián és az átlag – tehát ezek használatosak a becslésére). Az elméleti variancia az elméleti eloszlás „szélességét” mutatja (ahogyan a minta varianciája, illetve kvantilistávolságai). Ez a két jellemző egyértelműen leírja az általunk használt speciális eloszlásokat – azaz ezek

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: $\sim \mu \pm 2*\sigma$

A *normál (Gauss) eloszlásnak* egy lényeges különlegessége van: *ez az orvosi és fogorvosi gyakorlatban a leggyakoribb eloszlás.*
Az ábrán az eloszlás sűrűségfüggvényét és kumulatív eloszlásgörbéjét egyaránt feltüntettem, ugyanis ez nagyon fontos eloszlás számunkra. Mint látható, az előzőekkel ellentétben a normál eloszlás egy szimmetrikus eloszlás.

Látható, hogy valóban sok a változó, amely normál eloszlást követ: a koleszterinszint, a vércukorszint, a legtöbb enzimszint, a testmagasság, a BMI, a disztolés vérnyomás.... De vajon miért van ez?

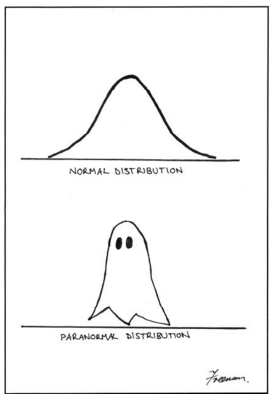
Normál eloszlás II.

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.

A normál eloszlás gyakoriságának okára a *centrális határeloszlás tétele* utal. Ez kimondja, hogy ha sok független valószínűségi változót összegzünk, akkor elég általános feltételek teljesülése esetén az összeg normális eloszlású valószínűségi változó lesz.
Az emberi test különböző jellemzői (mérhető értékei, változói) általában nagy sok más változó együtteséből alakulnak ki. Például az emberi testmagasság függ az apai és anyai génektől, a táplálkozástól, az életviteltől...

Másik megfogalmazása a centrális határeloszlás tételének a mintavétel során kapható átlagra vonatkozik: ha többször veszünk n elemű mintát egy sokaságból, mintából, akkor (ha n elég nagy), a minták átlagainak eloszlása normál eloszlás lesz, és ennek szórásnégyzete az eredeti szórásnégyzet n-ed része lesz. Ez egy lényeges tétel a statisztikában, ajánlott megjegyezni.
Erre példát a következő előadásban is láthatunk.



:)

Tatisztika? Ammeg mi?

- Adatgyűjtés
- Adatok rendszerezése, áttekintése

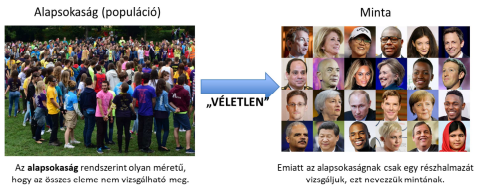
} Leíró statisztika

- Adatok elemzése
- Következtetések levonása

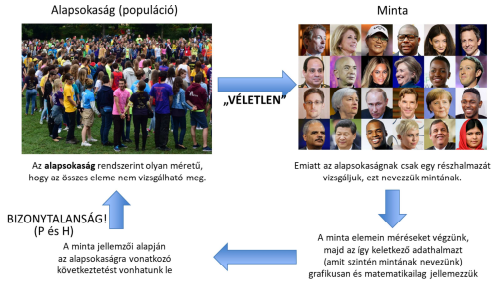
} **Következtető statisztika (induktív statisztika)**

A továbbiakban a következtető statisztikára térünk rá.

Alapsokaság és minta



Alapsokaság és minta



Először tisztázzuk az *alapsokaság* és a *minta* fogalmát.

Mint említettük, a statisztika végtelen tömegjelenségek leírója. Ez azt jelenti, hogy a jelenségek vizsgálata során *sok*, akár *végtelensok* mérést is végezhetünk. Ezen elméletileg lehetséges összes mérés kimeneteleinek, eredményeinek összefoglaló halmazát nevezzük *alapsokaságnak* (*populációnak*). Elméletben a statisztikai változó teljes megismeréséhez ezt a végtelensok mérést el kellene végeznünk, de erre nyilvánvalóan nincs lehetőségünk.

Ezért az alapsokaságnak csak egy részhalmazát vizsgáljuk, amit *mintának* nevezünk. A minta tehát az alapsokaság részhalmaza, amelynek alapja a *véletlen* kiválasztás.

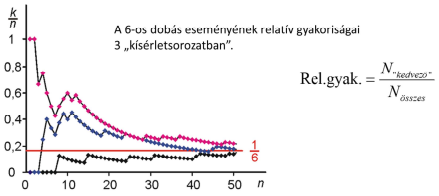
A létrehozott mintán méréseket végzünk, a keletkező mérési eredmények (kimenetelek) halmazát szintén mintának nevezzük. (Magyarán: kevésbé precíz fogalmazással pl. az évfolyam mint alapsokaság egyik csoportjának tagjait tekinthetjük az évfolyamból vett mintának; precízebben fogalmazva az évfolyam hallgatóinak vércsoportadatai jelenthetnek alapsokaságot, míg az egyik csoport tagjainak vércsoportadatai egy lehetséges mintát.)

A mintát jellemezhetjük grafikusan és számszerűen, *ahogyan azt megtanultuk az előzőekben*. A minta így megállapított tulajdonságait extrapolálhatjuk, azaz kiterjeszthetjük az alapsokaságra. Az előbbi példánál maradva: amilyen arányban a mintában előfordulnak az egyes vércsoportok, kb. olyasmit várunk el az egész alapsokaságtól is. Mivel a minta összeállítása véletlenszerűen történik, nem biztos, hogy tökéletesen reprezentálja az alapsokaságot, a különböző értékek alapsokaságon belüli előfordulási arányát. Így a minta alapján levont következtetésekhez *mindig társul valamekkora bizonytalanság*.

Milyen mennyiség ez a bizonytalanság? Hogyan definiálhatjuk?

2 mennyiséget szoktunk használni erre: valószínűség (P) és hiba (H)

Valószínűség I.



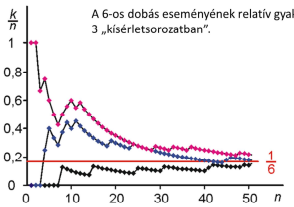
Azt tapasztaljuk, hogy a **relatív gyakoriságok** ilyen sorozatai – bár ingadozásokat mindig mutatnak – a „kísérletsorozat" hosszának növekedtével egyre inkább **stabilizálódnak valamilyen érték körül.** Továbbá ez az érték az aktuális „kísérletsorozattól" függetlenül lényegében ugyanakkora.

Ebben a kísérletben egy dobókockát dobunk 50-szer és közben rögzítjük a 6-os dobások relatív gyakoriságát. Ezt a kísérletet 3-szor végezzük el. Ezeknek az eredményei láthatók a dián. A 6-os dobási eredményt (kimenetelt) nevezzek a kedvező kimenetelnek – eseménynek.
Azt tapasztalhatjuk, hogy a *relatív gyakoriság* („kedvező"/összes dobások száma) – bár folyamatosan ingadozva –, de *egy adott értékhez tart*, továbbá ez a stabilizálódó érték az aktuális „*kísérletsorozattól" függetlenül* lényegében ugyanakkora.

Ebben a kísérletben egy dobókockát dobunk 50-szer és közben rögzítjük a 6-os dobások relatív gyakoriságát. Ezt a kísérletet 3-szor végezzük el. Ezeknek az eredményei láthatók a dián.
Azt tapasztalhatjuk, hogy a *relatív gyakoriság* („kedvező"/összes dobások száma) – bár folyamatosan ingadozva –, de *egy adott értékhez tart*, továbbá ez a stabilizálódó érték az aktuális „*kísérletsorozattól"*

függetlenül lényegében ugyanakkora.

Valószínűség, mint mennyiség?



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

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

$$N_{\text{összes}} \to \infty$$

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.

Ezt a megfigyelést nevezzük a *nagy számok (relatív gyakoriságokra vonatkozó) tapasztalati törvényének: :* a relatív gyakoriság értéke egy végtelen sorozatban egy adott értékhez tart.
Rendeljük hozzá a „kedvező” eseményünkhöz a „stabilizálódó” értéket, mint mennyiséget: jelen esetben a 6-os dobáshoz az 1/6-ot.
Ezt az értéket nevezzük az *esemény valószínűségének*. (Egy adott helyzetben mért változó kimenetelének valószínűsége.)
Tehát azt is mondhatjuk, hogy egy esemény *relatív gyakorisága megegyezik az esemény valószínűségével, ha az ismétlések száma* (kísérletsorozat hossza) *végtelen*.
Ez a törvény *tapasztalati törvény*, tehát logikai úton nem bizonyítható.

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.

OMHV
(Oktatók Hallgatói Véleményezése)

link:
http://report.semmelweis.hu/linkreport.php?qr=E2RE1NG37E2VY1Q8

Pin kód: VQ5

Az egyetemi előadásokról a report.semmelweis.hu oldalon tudok visszajelzést küldeni.
köszönjük.

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?
- Miről szól a centrális határeloszlás tétele?

A kérdéseket önellenőrzésnek szánjuk. A kérdések megválaszolhatók az előadáson elhangzottak, a gyakorlatvezetővel folytatott konzultációk, illetve saját utánaolvasás segítségével.