Biostatisztika és informatika alapjai

2. előadás: Leíró statisztika

2018. szeptember 20.

Veres Dániel

Ez az előadás a leíró statisztika alapvető fogalmairól szól.
Az előadás első részében a statisztikai tevékenység egy lehetséges csoportosítását írom le. Utána egyszerűsített definíciót adok a változókra és kimenetelükre, majd a kimeneteleket vizsgálva – mérési skálák szerint - csoportosítom őket.
Ezt követi az előadás magja, amelyben a leíró statisztika elemeit ismertetem a mérési skálák fonala mentén először egy, majd több változó esetében.
Az utolsó néhány dián bemutatom a percentilis görbéket, illetve azok használatát.
Végül az utolsó dián néhány szót ejtek az adatgyűjtés és adatrögzítés néhány olyan kérdéséről, amelyek tapasztalataim szerint sokszor problémát okoznak, bár ezek nehézségek könnyen orvosolhatóak lennének.

Tatisztika? Ammeg mi?

(Békásmegyeri aluljáró „átlagos” „lakója”)

Ebben az előadásban alapvető statisztikai fogalmakat írunk le – így remélhetőleg elkerülve a fenti kínos kérdést...
A címben nem véletlenül szerepelnek az idézőjelek – a lakója szociális értelemben, de vajon mit jelenthet, hogy átlagos?

Tatisztika? Ammeg mi?

(Békásmegyeri aluljáró „átlagos” „lakója”)

A **statisztika** a véletlen tömegjelenségek leírója.

Bár számos definíció létezik a statisztikára, én mégis egy újabbat adok: *a statisztika a véletlen tömegjelenségek leírója*.
Véletlen (azaz *egyénre vonatkozóan előre meg nem határozható*) tömegjelenségeket, tehát a *számos mérhető vagy megfigyelhető tulajdonságot (testmagasság, életkor, pólószín) több „dolgon” jellemez.*

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 néhány 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.

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.

Számos módon csoportosíthatjuk a változókat, én egy gyakorlati szempontból hasznos csoportosítást mutatok itt be, amely a *változó kimeneteleinek tulajdonságain* alapul (ismétlés az előző óráról).
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 mérésszinthierarchia 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

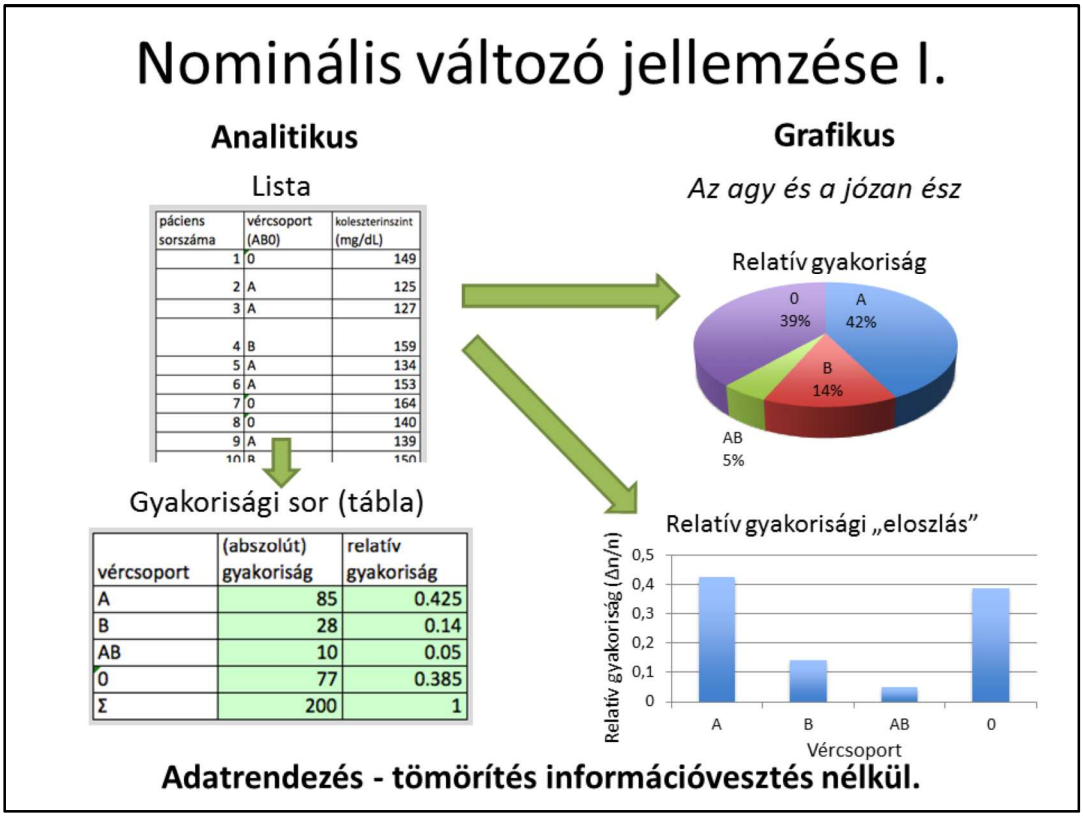
sorszámmal jelölik, amit észben kell tartani, hiszen sorszámokon nem végezhetők el a szokásos matematikai műveletek. Az *ordinális skála kategóriái közötti eltérés vagy távolság nem egyenlő vagy nem tudjuk megállapítani*.

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.



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*, 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.

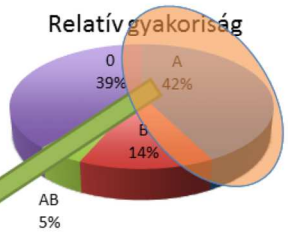
Analitikus

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

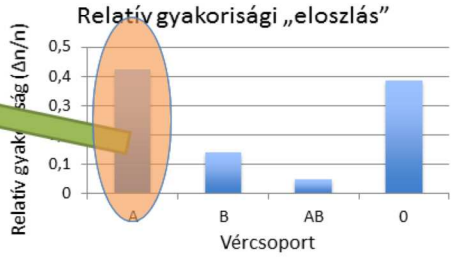
Grafikus

Az agy és a józan ész



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).

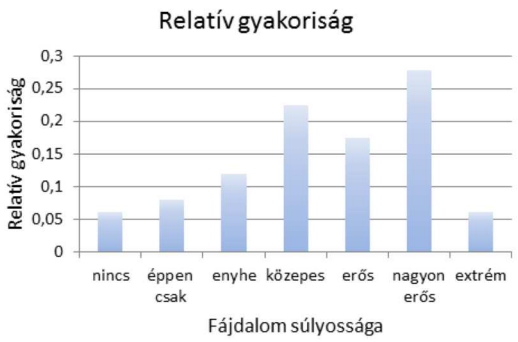
Ordinális változó jellemzése I.

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.
Emlékeztetőül: az ilyen típusú változónál már van értelme a kumulatív eloszlásnak is. A fájdalom skála esetében a kumulatív eloszlás megadja egy adott fájdalomnál nem nagyobb fájdalomérzet gyakoriságát.
Ordinális változók esetében a nominális változónál ismertetett megoldások mind alkalmazhatóak a jellemzésre. De tudunk-e adni további jellemzőt, kihasználva a sorbarendezhetőséget?

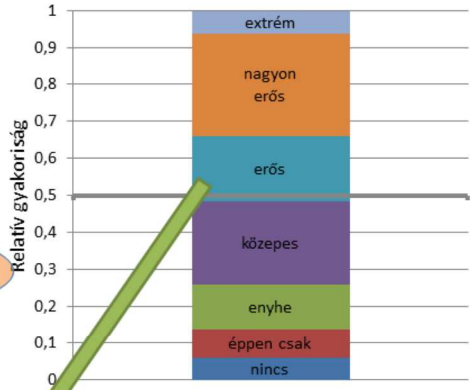
Ordinális változó jellemzése II.

Analitikus

Gyakorisági sor (tábla)

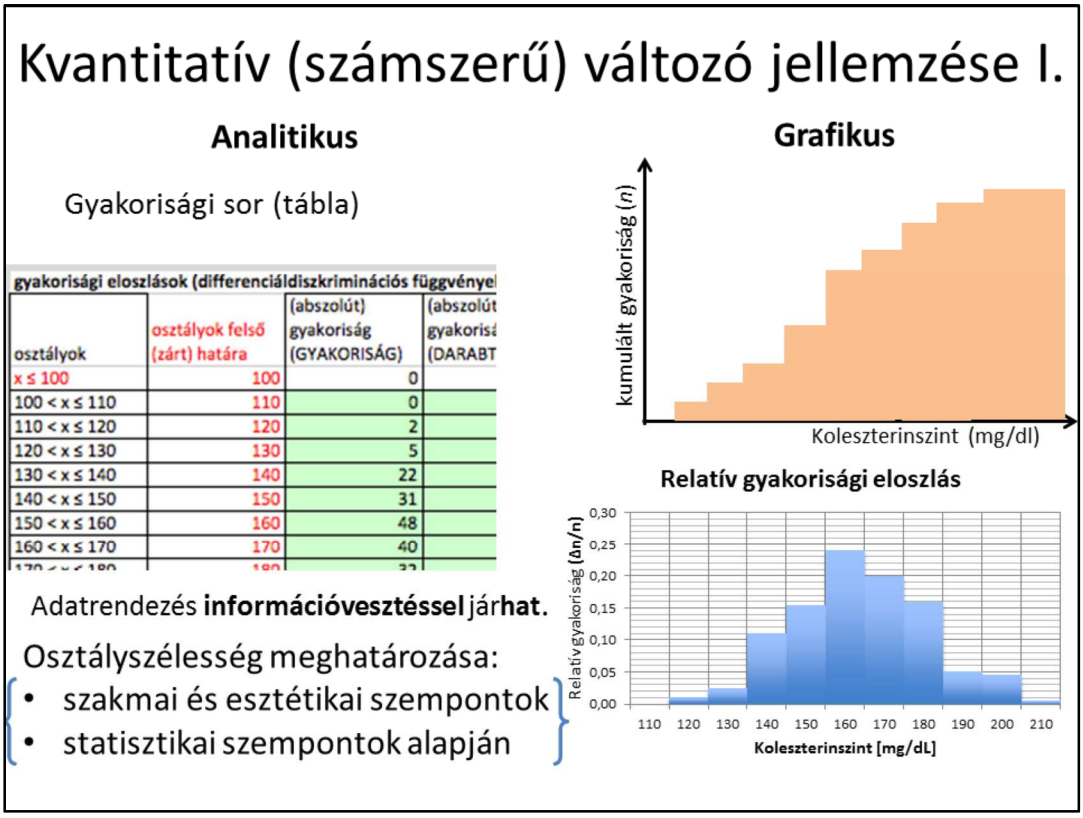
Fájdalom mértéke	Kumulatív relatív gyakoriág
nincs	0,06
éppen csak	0,14
enyhe	0,26
közepes	0,485
erős	0,66
nagyon erős	0,94
extrém	1

Grafikus

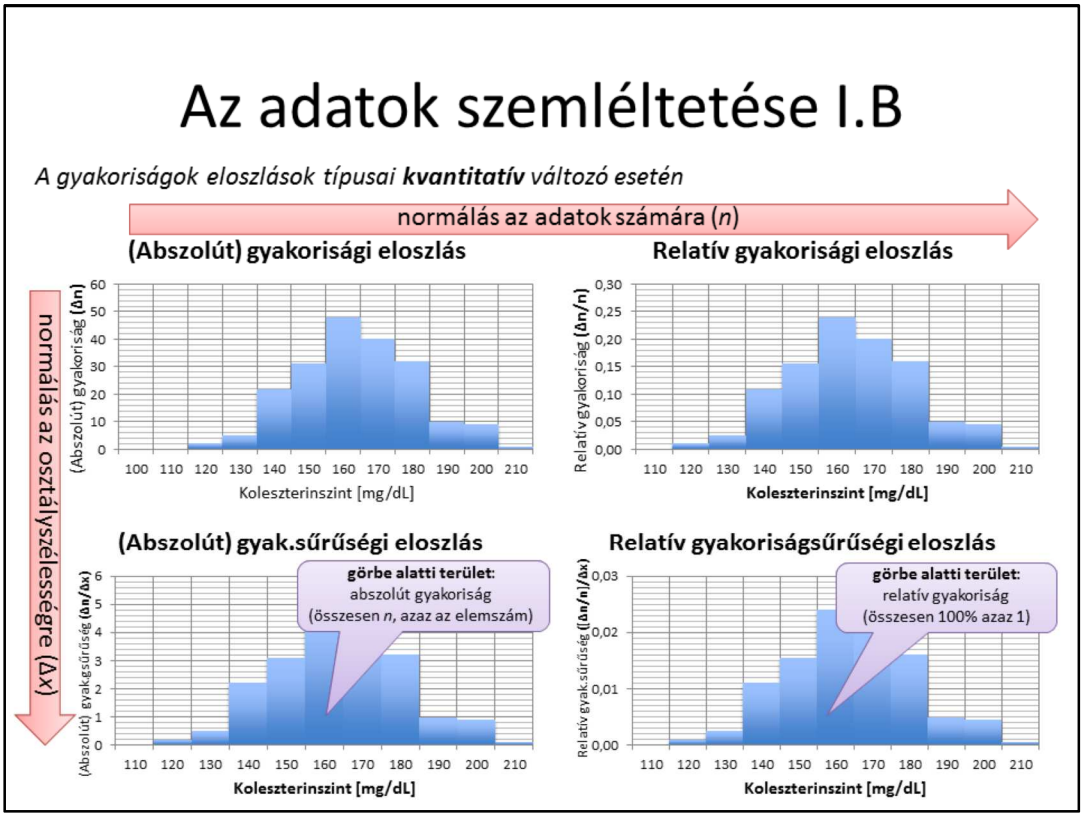


Új Jellemző (információvesztéssel):
Medián: „középső” elem(ek)
Jelölése: Me, Med, x_{med}

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 gyakorisági 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



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.



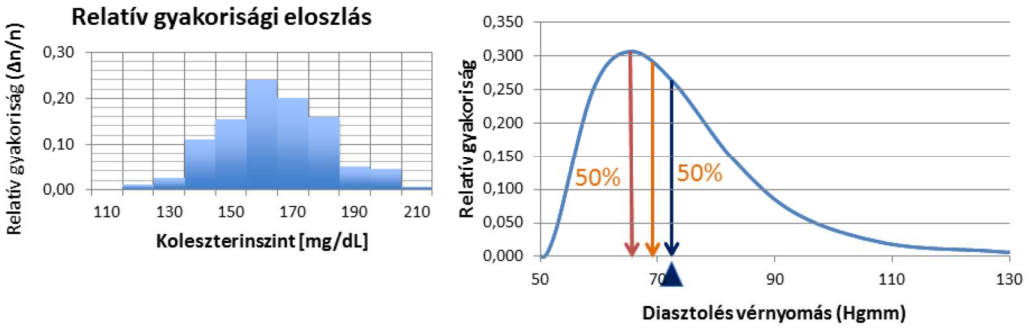
Kvantitatív változók esetében többféle gyakorisági eloszlást definiálhatunk, a különböző összehasonlíthatóság kedvéért, erről itt egy rövid összefoglalót adok.

Az y tengelyen felvehetjük az abszolút gyakoriságot, ekkor az oszlopok magasságáról közvetlenül leolvasható az adott osztályba tartozó elemek száma. Ha a relatív gyakoriságot ábrázoljuk, akkor az oszlopok magassága az adott osztályba tartozó elemek mintán belüli hányadának felel meg.

Ha az abszolút gyakoriságot normáljuk az osztályszélességre, akkor megkapjuk a(z abszolút) gyakoriságsűrűséget. Ezt felvéve az y tengelyen a oszlopok (téglalapok) területe fog megfelelni az adott osztály elemszámának. A teljes Hasonlóan, a relatívgyakoriságsűrűség

Megjegyzés: a sűrűség kifejezés általános jelentése, hogy valamilyen egységre (egységnyi térfogatra, felületre vagy szakaszra) vonatkoztatunk. Háromdimenziós példát említve a hagyományos “sűrűség” (tömegsűrűség) az egységnyi térfogatra vonatkoztatott tömeget adja meg. Két dimenzióban használt mennyiség a felületi sűrűség (például papírt szokták vele jellemezni), amely az egységnyi felületre vonatkoztatott tömeg (pl. egy négyzetméter papírlap tömege). A gyakoriságsűrűségnél lényegében a Δx szakaszhosszra vonatkoztatjuk a gyakoriságot («tömegét»), vagyis ez egy egydimenziós sűrűségnek tekinthető.

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 ?!

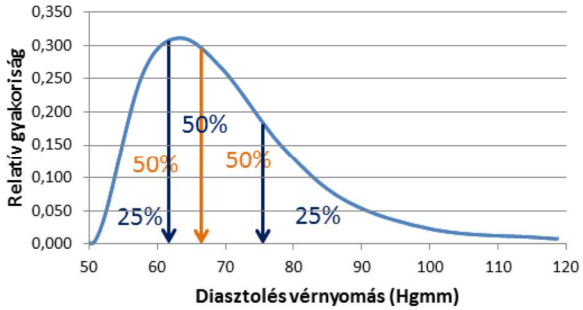
  Jelölése: $x_{átl}$, $\overline{x}$

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.
A ?-ek és többesszámok jelentésére még a későbbiekben visszatérek.

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.

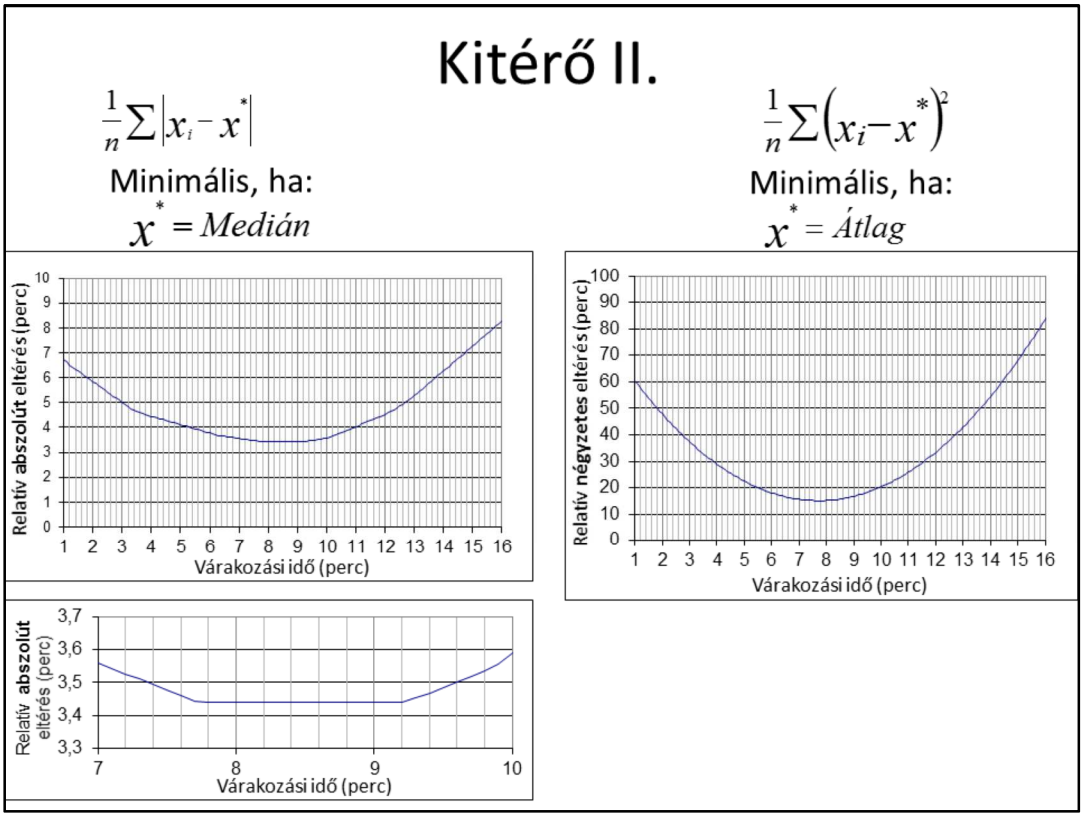
Nap sorszáma	Várakozási idő		
1	1,27	medián:	8,475
2	3,3	alsó kvartilis	3,59
3	3,44	átlag	7,723333
4	3,64		
5	6,33		
6	7,72		
7	9,23		
8	9,87		
9	10,31		
10	12,29		
11	12,3		
12	12,98		

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?

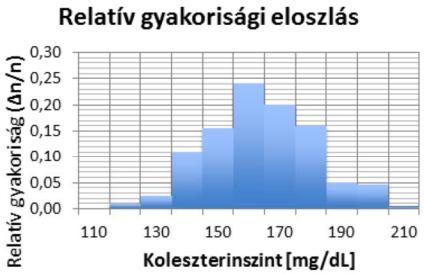
Ezen a dián megpróbálok rámutatni a korábban ?-lel, többes számmal, „”-lel jelölt néhány kérdésre.
A példában azt tüntettem fel, hogy az egyes napokon mennyi időt kellett várni a buszra. Ez a várakozási idő lesz a változó, amit vizsgálunk. Az ábrán a mért értékek nagyság szerint sorba vannak rendezve. 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,

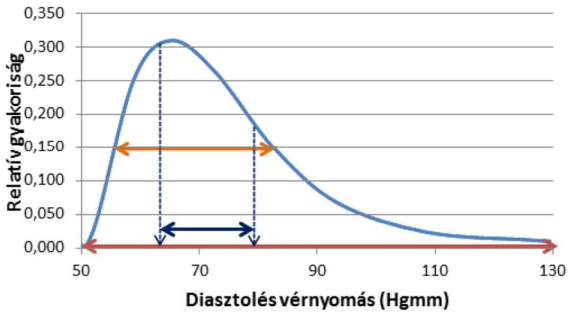
vagy mindegyik elem az – ez azonban így nem bír jelentéssel. Tehát numerikus (és kisebb elemszámú) minta esetében gyakran nincs értelme meghatározni móduszt. Ebben az esetben legfeljebb gyakorisági eloszlás alapján (ha van elég adat és ezért van értelme elkészíteni) van értelme egy tartományt, mint móduszt számítani.



Egy kis matek.
Egy adathalmazban az adatok *átlagos abszolút eltérése* egy adott értéktől a *mediánra* lesz minimális.
Egy adathalmazban az adatok *átlagos négyzetes eltérése* egy adott értéktől az *átlagra* lesz minimális.
Ezt kipróbálhatjuk excelben is az I. kitérőben használt adatsorral. Ekkor is hasonló eredményre jutunk.
Felhívnám a figyelmet arra is, hogy az abszolút eltérésnél 7,72 és 9,23 között ugyanakkora (és minimális) értéket kaptunk. Ez is mutatja, hogy elméletben miért is tekintünk minden számot ezek között mediánnak.

Kvantitatív változó jellemzése III.





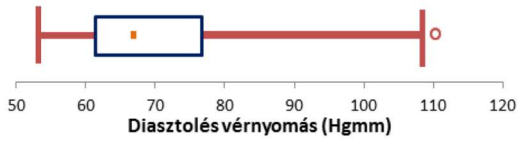
Jellemzők – szóródási paraméterek:

- ***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***

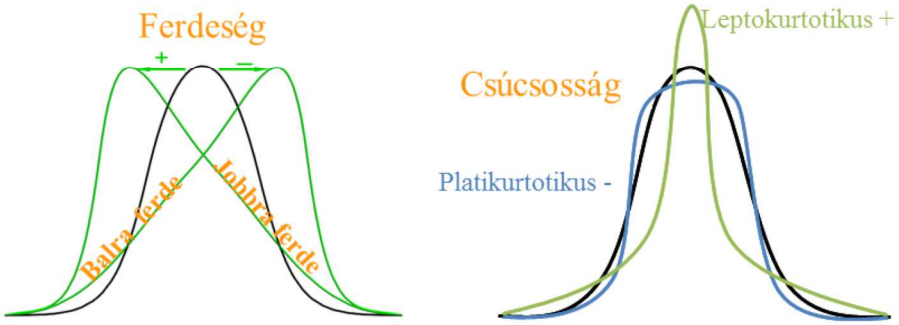
A sodrófadiagram (más nevén box plot, vagy whisker plot) az adatok nagyon látványos grafikus leírását adja.
A következő részekből áll a sodrófadiagram. (A dián az ábrán használt jelölt paramétereket dőlt betűkkel jelöltem.)
A sodrófa szeme, amely általában a medián, ritkábban az átlag értéke. Szimmetrikus 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*.

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



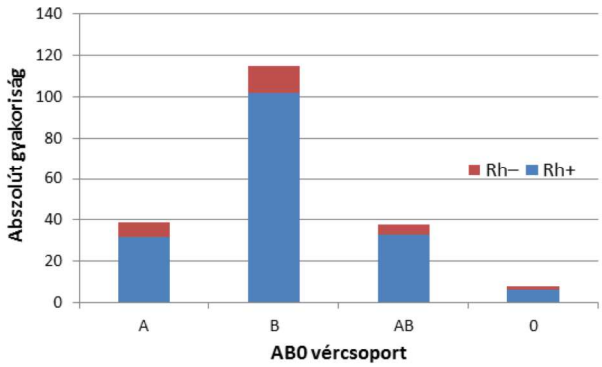
További paramétereket is meghatározhatunk, ezek a *momentumok* és *centrális momentumok*.
A számszerű adatok további jellemzőinek egy csoportja az eloszlásgörbe alakját írja le, azaz megmutatják, hogy hol van az adatok „nagyobb tömege”.
Ehhez a csoporthoz alapvetően két jellemző tartozik (amelyek kiszámítása a 3., illetve 4. centrális momentumokon alapul): a *ferdeség* és a *csúcsosság*. A ferdeség egy szimmetrikus eloszláshoz képesti (vízszintes) eltolódást írja le, amíg a csúcsosság a görbe laposságát-csúcsosságát, illetve a „farkak” (kvázi kiszóró értékek) súlyosságát mutatja. A pozitív kurtózisú (súlyosabb „farkú”) eloszlást leptokurtotikusnak, a negatív kurtózisú eloszlást („vállas”) platikurtotikusnak nevezik.

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***

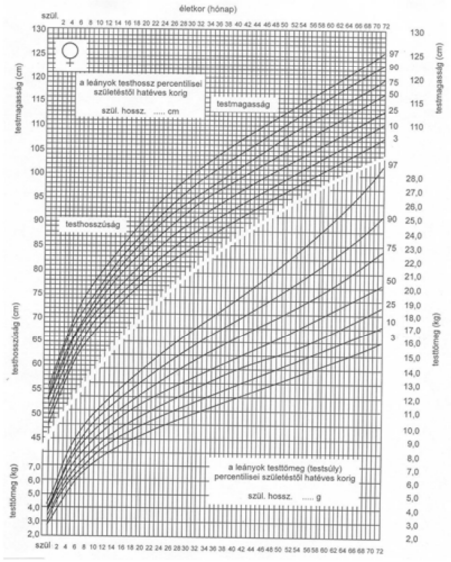


Több változó együttes leírása igen bonyolult. A következőkben csak néhányat emelek ki ezek közül. Több kvalitatív változó analitikus jellemzésére a *kontingencia táblázatokat* szoktuk használni. Grafikus megjelenítésre pedig kiválóan alkalmasak a *mozaik ábrák*. Ismét rámutatnék arra, hogy mennyivel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

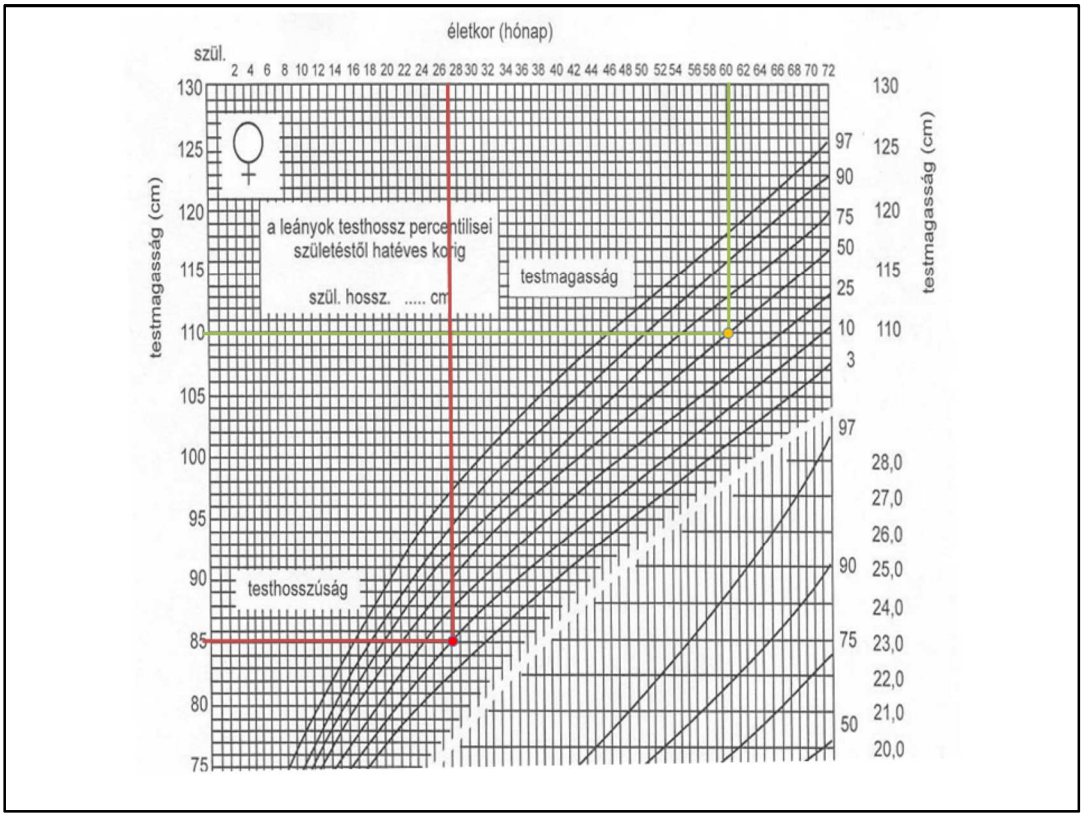
Több kvantitatív változó jellemzése

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.)

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**

Elemszám számítások, randomizáció…
- kérdezze statisztikusát/választható kurzusok

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

Ezen az utolsó dián néhány lényeges, de sokszor elhanyagolt adatgyűjtési és adatrögzítési tényre hívnám fel a figyelmet.
Adatgyűjtésnek célja van és nem változói!
Változó legyen a lehető legmagasabb skálájú.
Adatrögzítés
- olyan formában, amely könnyen rendszerezhető, alakítható - excel
- változókat külön-külön
- kódolás egyértelmű legyen (változó minősége, eltérő kategóriák) – például a változó mérési skálája nem változik attól, hogy számokkal kódoljuk

Az adatgyűjtés során további lényeges szempontokat is figyelembe kell venni, mint például az elemszám kérdése, a randomizáció és egyebek. Ezekről a későbbi évek szabadon választható kurzusain hallhattok – vagy kérdezzétek statisztikusotokat.

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?

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

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?