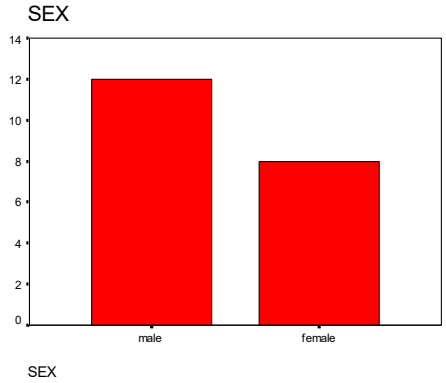
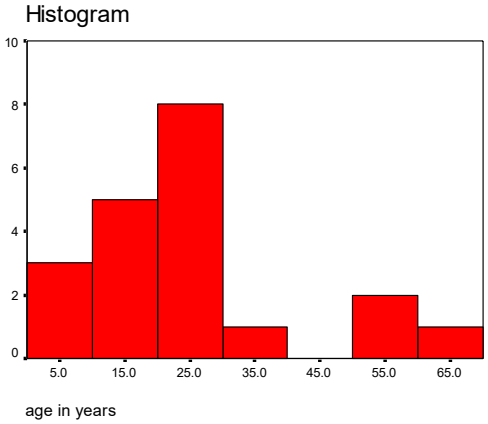
Biostatisztika
Összefoglalás

A változók eloszlása típus szerint, összefoglalás

- **Diszkrét**: Egy diszkrét változó eloszlása megadja, hogy melyek a lehetséges értékek, és ezek milyen gyakran fordulnak elő.
- **Folytonos**: egy folytonos változó eloszlása megadja, hogy melyek a lehetséges értékek, és ezek milyen gyakran esnek bizonyos intervallumokba .



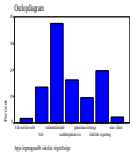


1.4. Folytonos eloszlások jellemzése számokkal, statisztikai jellemzők

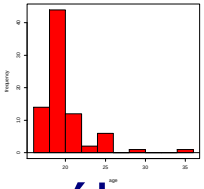
- **A közép jellemzése**:
  - Átlag: számtani átlag
  - Medián: az a szám, amelynél az adatok fele kisebb ,
  - (módusz: leggyakoribb érték)
- **A szóródás jellemzői :**
  - Terjedelem: maximum - minimum
  - interkvartilis terjedelem: harmadik és első kvartilis különbsége, ebben az intervallumban van az adatok középső 50%-a
  - variancia: szórásnégyzet
  - standard deviáció, SD (v. szórás): az adatok szóródása az átlag körül
- **Egyéb jellemzők:**
  - variációs együttható: átlag/SD
  - egy egyed (x) helyzetének jellemzése(rang, z-érték): z=(x-atlag)/SD

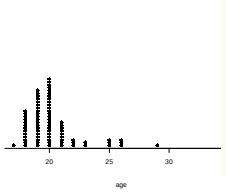
Az eloszlás ábrázolása a statisztikai jellemzők alapján

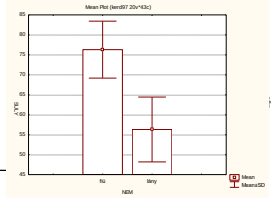
- **Kategórikus változók**
  - oszlopdiagram
  - kördiagram

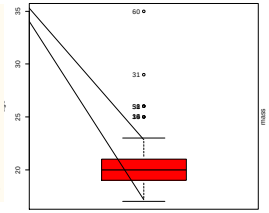


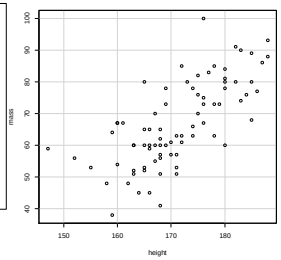
- **Folytonos változók**
  - hisztogram
  - (pont-ábra)
  - átlag-szóródás ábra
  - Doboz ábra (box diagram - box-whisker plot)
  - Szóródási diagram





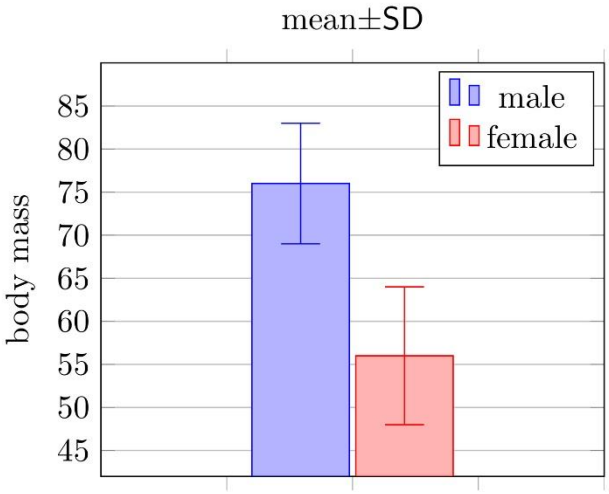


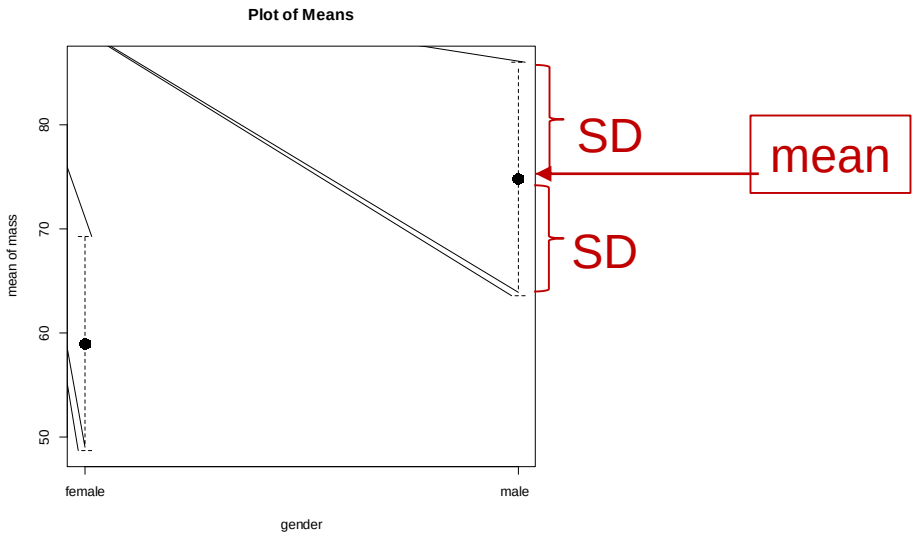




Átlag-szóródás ábrák

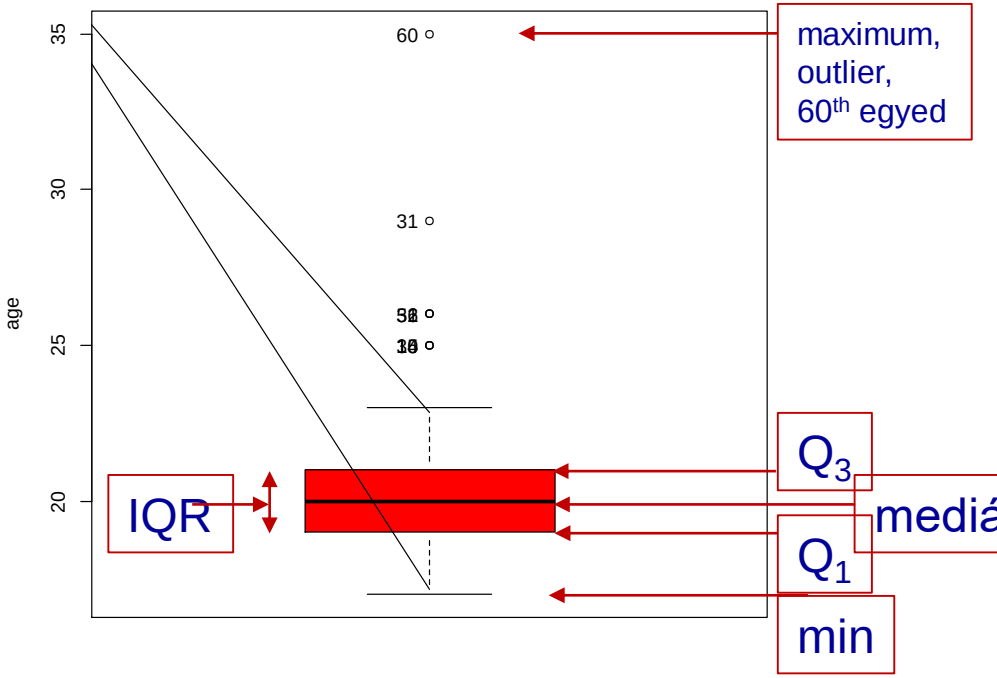
- Általában egy oszlop, melynek magassága éppen az átlag értéke, a rárajzolt kalapácsok hossza SD. Az átlagot gyakran csak egy ponttal jelzik.





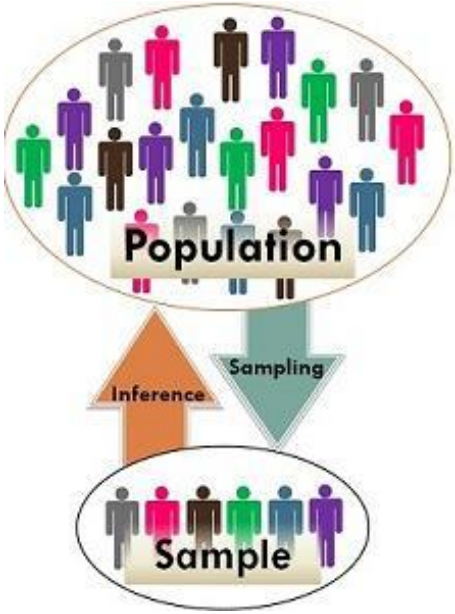
Doboz-ábra, box diagram vagy Box-whisker plot

Doboz-ábra (box diagram). A **kvartiliseken** alapul (minimum, első kvartilis, medián, második kvartilis, maximum).
Ha a doboz hosszának (=IQR) 1.5-szeresénél messzebb eső adat van, azok kiugró értékekként jelennek meg. Ekkor a kalapácsok vége a Q3 +1.5 IQR alatti legnagyobb adat, illetve a Q1-1.5 IQR feletti legkisebb adat. A kalapácsok feletti illetve alatti adatok kiugró értékeknek minősülnek. Ezeket külön jellel jelölik, kis körrel vagy csillaggal.



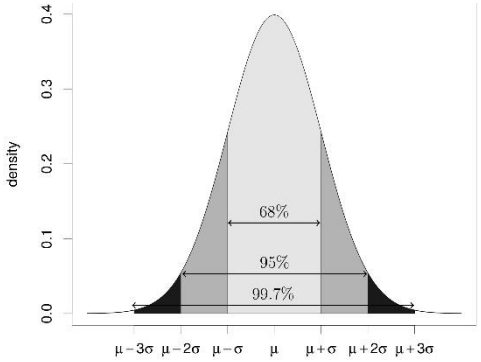
Populáció, minta

- **Populáció**: azoknak az egyedeknek, objektumoknak az összessége, amelyről egy vizsgálat során információt kívánunk nyerni.
- **Minta**: a sokaság azon részhalmaza, amelyet éppen vizsgálunk
- A minta kiválasztásakor arra törekszünk, hogy lehetőleg reprezentálja az egész populációt, vagy legalábbis következtetni lehessen a populációra. Követelmény a mintaelemek függetlensége is.



A normális eloszlás tulajdonságai

- Harang alakú
- Szimmetrikus
- Paraméterei: μ - elméleti átlag, σ - elméleti szórás
- Középpontja: μ = átlag=medián=módusz
- Szóródás (szélesség):σ
- Inflexiós pontok: μ-σ, μ+σ
- „A 68-95-99.7 szabály”
  - A megfigyelések 68% -a esik a μ középtől egyszeres σ távolságra
  - A megfigyelések 95% -a esik a μ középtől kétszeres σ távolságra
  - A megfigyelések 99.7% -a esik a μ középtől 3-szoros σ távolságra



Standard deviáció vagy standard error??

- Standard deviáció, SD: a minta szórása, a mintaadatok szóródása az átlag körül.
  Normális eloszlás esetén az átlag ±2SD-n belül van az adatok kb. 95%-a
- Standard error (SE=SD/√n): azt fejezi ki, hogy a populáció-átlagot mennyire jól becsültük. A mintaátlag(ok) szóródását mutatja az (ismeretlen) populáció átlag körül (az elképzelt ismétlések során).
  Normális eloszlás esetén az átlag ± 2SE-n belül van az igazi átlag (populáció átlag) kb. 95%-os valószínűséggel.
- Mindegyiket alkalmazhatjuk, csak egyértelműen adjuk meg, hogy a „±” jel után mi következik.

Konfidencia-intervallum

- Az átlagra számolt 95%-os konfidencia-intervallum jelentése
- 95% bizonyosak vagyunk abban, hogy az igazi átlag (populációátlag) ebben az intervallumban van.

Hipotézisvizsgálatok

- Hipotézis: állítás a populációról
- A mintaadatok alapján az egész jelenségre (populációra) következtetünk
- Azt vizsgáljuk, hogy az általunk tapasztalt különbség nagyobb-e, mint amit a véletlen önmagában okoz.

A hipotézisvizsgálat lépései

- **Step 1-2**. A H_0 nullhipotézis és a H_a alternatív hipotézis felállítása
- **Step 3**. Az első fajta hiba valószínűségének rögzítése. Leggyakrabban *α =0.05* vagy α =0.01.
- **Step 4**. A minta elemszámának meghatározása (*n)*
- **Step 5.** Mintavétel.
- **Step 6**. Döntési szabály kiszámítása –függ a kísérleti elrendezéstől, az adatoktól, feltételektől, stb…
  - Átlagok összehasonlítása – t-próbák, ANOVA
  - Varianciák összehasonlítása: F-próba
  - Gyakoriságok összehasonlítása: khi-négyzet próbák
- **Step 7**. Döntés.
- a) Elvetjük a nullhipotézist, azaz
  - elfogadjuk H_a -t
  - A különbség szignifikáns α100% szinten.
- b) Nem vetjük el a nullhipotézist,
  - elfogadjuk H_0 -t
  - A különbség nem szignifikáns α100% szinten.

A próba megválasztása függ az adatok típusától, a kísérleti elrendezéstől és az összehasonlítás céljától Csoportosítás az adathalmazban szereplő változók szerint

A kurzus során tanult próbák (a nem tanultak áthúzva):

~~**Egy (folytonos) változó** adott értékhez való hasonlítása: **egymintás t-próba**~~

Két változó:

1) mindkettő folytonos (egyanazon változó ugyanazokon az egyedeken kétszer mérve):
   a) az átlagok összehasonlítása (a változás összehasonlítása): **páros t-próba** (egymintás t-próba a különbségekre)
   b) a változók közötti kapcsolat vizsgálata: k**orrelácó, regresszió**
2) egy folytonos függő változó egy másik, kategorikus változó szerinti csoportjaiban az átlagok összehasonlítása:
   a) a csoportok száma=2**: kétmintás t-próba** (Independent t-test)
   b) ~~a csoportok száma>2**: One-way ANOVA** (variancia analízis~~)
3) mindkettő kategorikus: kontingencia táblázatok értékelése **chi-square test** (khi-négyzet próba)

1) és 2) esetén feltétel, hogy a minták normális eloszlású populációból származnak. 2) esetén még az is, hogy a varianciák azonosak az egyes csoportokban. ~~Ha a normalitás feltétele nem teljesül, de az adataink folytonosak, akkor a fenti próbák helyett rangszámokon alapuló ún. nemparaméteres próbákat alkalmazhatunk – további feltételekkel.~~

A próba megválasztása
Csoportosítás a próba célja szerint

A kurzus során tanult próbák (a nem tanultak áthúzva):

Átlagok (várható értékek) hasonlítása:
~~*Egy változó*: **One-Sample t-test (egymintás t-próba)**~~
Két változó: (ugyanazokon az egyedeken kétszer mérték ugyanazt): **Páros t-próba**
Két független minta (mindenkit csak egyszer mértek egyik vagy másik csoportban): **kétmintás t-próba** (Independent t-test)
~~*Több független minta*: **One-way ANOVA** (variancia analízis)~~
Kapcsolat, függetlenség vizsgálata:
Két folytonos változó (lineáris kapcsolat): k**orrelácó, regresszió**
Két kategorikus változó: **chi-square test,** khi-négyzet próba

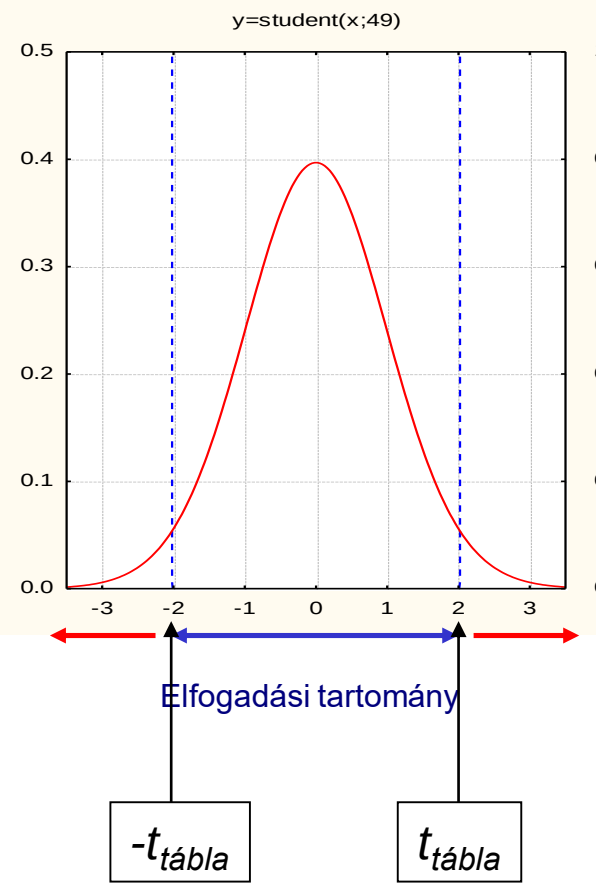
Az utolsó eset kivételével feltétel, hogy a minták normális eloszlású populációból származnak. Ha normalitás feltétele nem teljesül, de az adataink folytonosak, akkor a fenti próbák helyett rangszámokon alapuló ún. nemparaméteres próbákat alkalmazhatunk – további feltételek figyelembe vételével.

A szignifikancia megállapítása, páros t-próba

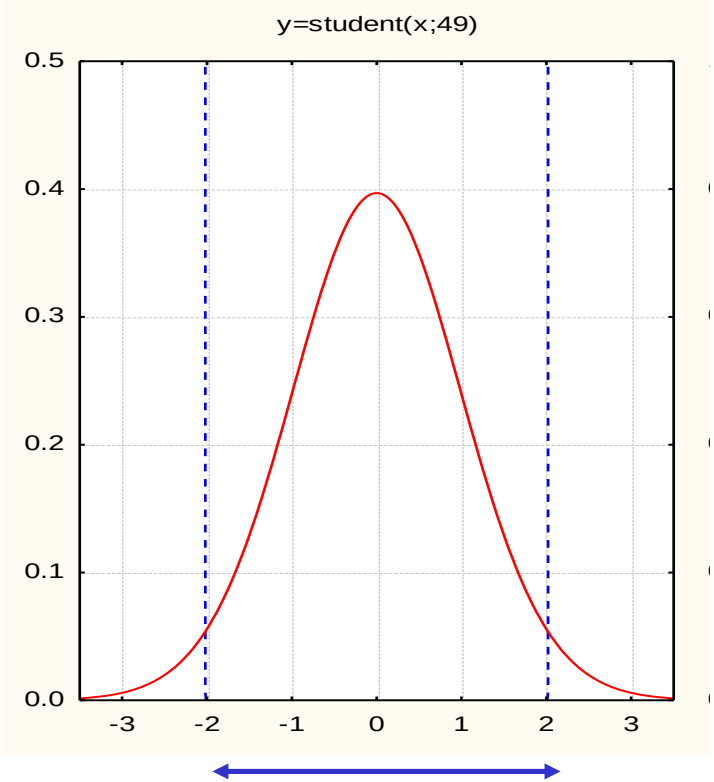
- Adott két összetartozó minta (tipikusan kezelés előtt és után mért adatok).
- H_0: $\mu_{diff}=0$. A nullhipotézis azt állítja, hogy a populációban nincs változás, vagyis az átlagos különbség 0 populációban.
- Ki tudjuk számolni a *t*-értéket a következő formulával:

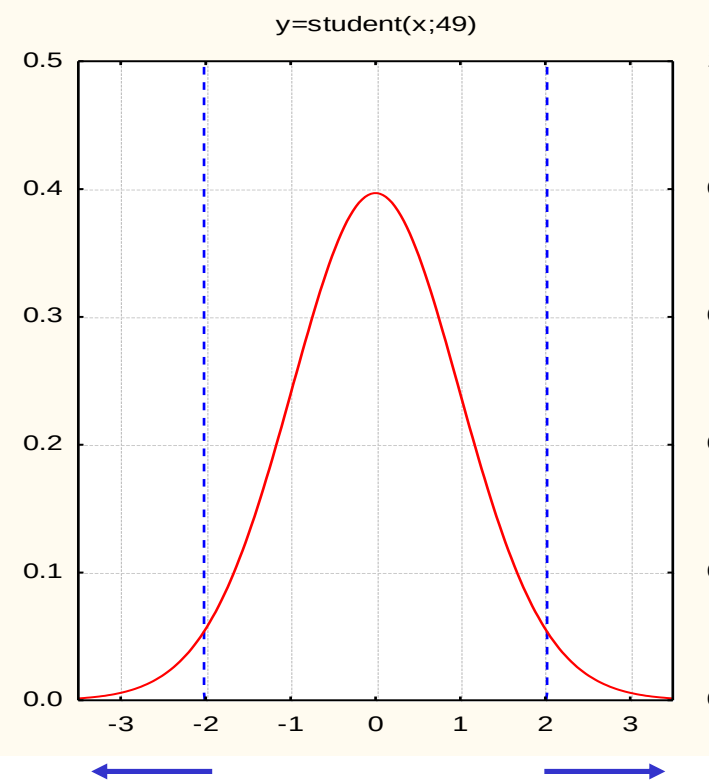
$$t = \frac{\bar{x}}{SE}$$

- Ha igaz a nullhipotézis, akkor ismerjük a számított *t*-érték (próbastatisztika) eloszlását (*t*-eloszlás *n*-1 szabadságfokkal). Tehát meg tudjuk mondani, hogy a számított *t*-érték mely intervallumba esik (1-α) valószínűséggel: ez az elfogadási tartomány, melynek határait a táblázatbeli kritikus értékek adják.



Döntés *t*-érték alapján



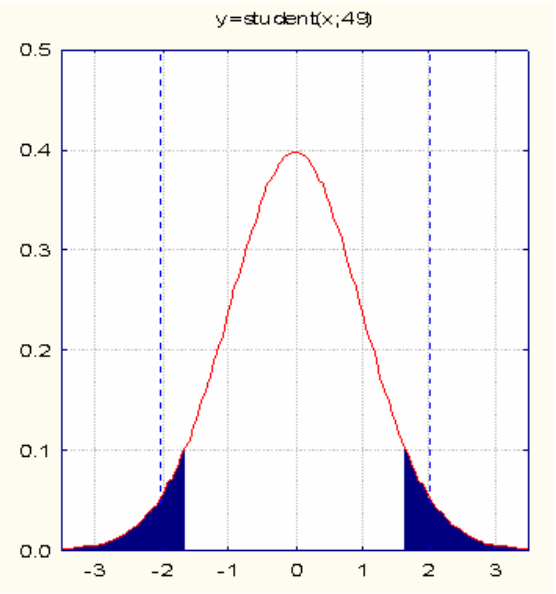


Ha $|t|<t_{\text{tábls}}$, vagyis
a számított *t* az elfogadási tartomány**ba** esik,
elfogadjuk H0-t és azt miondjuk, hogy
a különbség **nem szignifikáns** α szinten
(ebben az esetben $|t|$ **kicsi**, kisebb, mint a kritikus érték)

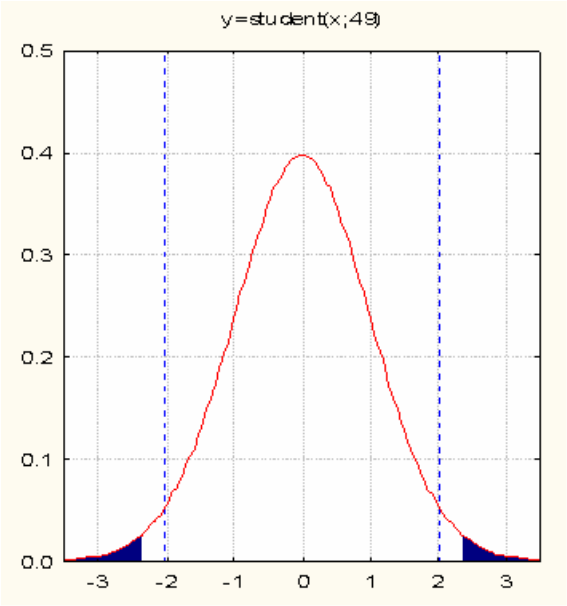
Ha $|t|>t_{\text{tábla}}$, vagyis
a számított *t* az elfogadási tartományon **kívül** esik,
elvetjük H0-t és azt mondjuk, hogy
a különbség **szignifikáns** α szinten
(ebben az esetben *t* **nagy** (abszolút értékben), nagyobb, mint a kritikus érték)

Döntés *p*-érték alapján

- *p*-érték: az általunk számított próbastatisztika (*t*-érték) által a H0-nak megfelelő eloszlás két széléből levágott terület nagysága. Annak valószínűsége, hogy ha igaz a nullhipotézis, akkor legalább ekkora eltérést kapjunk.



$p > \alpha$, a különbség nem szignifikáns α szinten



$p < \alpha$, a különbség szignifikáns α szinten

A szignifikancia megállapítása

- **Próbastatisztika alapján** (számolt t-érték, F-érték, $\chi 2$ érték) – szükség van a statisztikai táblázatra, hogy a kritikus értéket megtaláljuk a szabadságfok és α függvényében.
- Ha $|t|<t_{tábla}$ vagy $F<F_{tábla}$ vagy $\chi^2<\chi^2_{tábla}$,
  - A különbség nem szignifikáns α szinten
  - Nem vetjük el H_0–t (elfogadjuk H_0-t)
- Ha $|t|>t_{tábla}$ vagy $F>F_{tábla}$ vagy $\chi^2>\chi^2_{tábla}$,
  - A különbség szignifikáns α szinten
  - Elvetjük H_0–t (elfogadjuk H_A-t)

- ***p*-érték alapján** nincs szükség táblázatra, a p-értéket elegendő α-val összehasonlítani .
- Ha $p>\alpha$ [$|t|< t_{tábla}$],
  - A különbség nem szignifikáns α szinten
  - Nem vetjük el H_0–t (elfogadjuk H_0-t)
- Ha $p<\alpha$ [$|t|>t_{tábla}$],
  - A különbség szignifikáns α szinten
  - Elvetjük H_0–t (elfogadjuk H_A-t)

A tanulmányozott statisztikai próbák, nullhipotéziseik és próbastatisztikák

Egymintás *t*-próba

- Cél: az átlagot egy adott *c* konstanshoz hasonlítjuk
- Feltétel: normalitás
- H0: μ=*c*, a populáció-átlag = *c*
- Ha: μ≠*c*, a populáció-átlag ≠ *c*
- Próbastatisztika: $t = \frac{\bar{x} - c}{SE}$
- Szabadságfok=*n*-1

Páros *t*-próba

- Cél: két összetartozó minta átlagának összehasonlítása (az átlagos különbség 0-hoz hasonlítása)
- Feltétel: a különbség-minta normális eloszlásból származik
- H0: $\mu_1=\mu_2$ vagy $\mu_{diff}=0$, a populáció-átlagok egyenlők
- Ha: $\mu_1 \neq \mu_2$ vagy $\mu_{diff} \neq 0$, a populáció-átlagok különbözők
- Próbastatisztika: $t = \frac{\bar{x}}{SE}$, különbség átlag/különbség SE

- Szabadságfok=*n*-1

Kétmintás *t*-próba vagy független mintás *t*-próba

- Cél: két független minta átlagának **összehasonlítása**
- Feltételek: mindkét minta normális eloszlásból származik, a varianciák egyenlők
- H0: $\mu_1=\mu_2$ vagy $\mu_{diff}=0$, a populáció-átlagok egyenlők
- Ha: $\mu_1 \neq \mu_2$ vagy $\mu_{diff} \neq 0$, a populáció-átlagok különbözők
- Student–*t*, próbastatisztika egyenlő varianciák esetén

$$t = \frac{\bar{x}-\bar{y}}{SD_p\sqrt{\frac{1}{n}+\frac{1}{m}}} = \frac{\bar{x}-\bar{y}}{SD_p}\cdot\sqrt{\frac{nm}{n+m}}$$

$$SD_p^2 = \frac{(n-1)\cdot SD_x^2 + (m-1)\cdot SD_y^2}{n+m-2}$$

szabadságfok=*n*+*m*-2

- Welch-próba, próbastatisztika különböző varianciák esetén

$$d = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{SD_x^2}{n}+\frac{SD_y^2}{m}}}$$

Szabadságfok:

$$g = \frac{\frac{SD_x^2}{n}}{\frac{SD_x^2}{n}+\frac{SD_y^2}{m}} \qquad \frac{(n-1)\cdot(m-1)}{g^2\cdot(m-1)+(1-g^2)\cdot(n-1)}$$

- Választás a kétféle képlet közül:
  - Ha nincs előzetes feltevésünk a varianciákról, használjuk inkább a Welch próbát
  - Ha tudjuk korábbi tanulmányokból, hogy különbözők a varianciák, használjunk Welch próbát
  - Ha tudjuk korábbi tanulmányokból, hogy egyenlők a varianciák, hazsnáljunk Studnt-t próbát

F-próba

- Cél: két független minta varianciájának összehasonlítása
- Feltételek: mindkét minta normális eloszlásból származik
- H0: $\sigma_1^2 = \sigma_2^2$, a populáció-varianciák egyenlők
- Ha: $\sigma_1^2 < \sigma_2^2$ vagy $\sigma_1^2 > \sigma_2^2$, a populáció-varianciák különbözők (egyoldalas alternatíva)
- Próbastatisztika

$$F = \frac{\max(SD_x^2, SD_y^2)}{\min(SD_x^2, SD_y^2)} = \frac{\text{nagyobb mintavarianci a}}{\text{kisebb mintavarianci a}}$$

- Szabadságfok=(két szabadságfok van)
  1. Nagyobb varianciájú minta elemszáma -1
  2. Kisebb minta varianciája -1

Khi-négyzet próba, függetlenségvizsgálat

- Cél: két kategórikus változók eloszlásainak összehasonlítása kereszt-osztályozás alapján
- Feltétel: nagy elemszám, ami a várható gyakoriságokkal van kifejezve: 5-nél kisebb várt gyakoriság maximum a cellák 20%-ában lehet
- H0: függetlenség, a két változó független (az egyik változó kategóriái szerint a másik változó eloszlása ugyanaz)
- Ha: a két változó nem független
- Próbastatisztika: $X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ χ^2 eloszlású
- szabadságfok=(sorok száma-1)(oszlopok száma-1)

Korreláció-regresszió

- Cél: két folytonos változó lineáris kapcsolatának jellemzése
- Feltétel: normális eloszlás
- Korrelációs együttható (r): a lineáris kapcsolat szorosságát jellemző statisztika
  - -1 és 1 között van
  - -1 és 1 jellemzi a tökéletes lineáris kapcsolatot
  - 0 körüli értékek a kapcsolat hiányát vagy nemlineáris kapcsolatot
- A korreláció szignifikanciája
  - H0: függetlenség, a két változó között nincs lineáris kapcsolat a populációban
  - Tesztelése a korreláció t-próbájával, melynek szabadságfoka n-2
- Lineáris regresszió: x->y irányú lineáris kapcsolat esetén a legjobban illeszkedő egyenes meghatározása
  - H0: a regressziós együttható (=az egyenes meredeksége) 0 a populációban, azaz y nem függ lineárisan x-től
  - Tesztelése ekvivalens a korreláció t-próbájával
- Determinációs együttható (r^2): az y változó szóródásának hány %-a magyarázható az x-től való lineáris függésével

Példák az orvosi irodalomból

Pain Physician 2013; 16:E491-E504 • ISSN 2150-1149

Randomized Trial

A Randomized, Double-Blind, Active-Controlled Trial of Fluoroscopic Lumbar Interlaminar Epidural Injections in Chronic Axial or Discogenic Low Back Pain: Results of 2-Year Follow-Up

Laxmaiah Manchikanti, MD[1,2], Kimberly A. Cash, RT[1], Carla D. McManus, RN, BSN[1], Vidyasagar Pampati, MSc[1], and Ramsin M. Benyamin, MD[3]

Demográfiai és klinikai jellemzők bemutatása két csoportban.
A p-értékek vajon milyen próbák eredményei és mit jelentenek?

Fluoroscopic Lumbar Interlaminar Epidural Injections in Chronic Axial or Discogenic Low Back Pain

Table 1. *Baseline demographic and clinical characteristics.*

		Group I (60)	Group II (60)	*P* value
Gender	Male	23% (14)	40% (24)	0.077
	Female	77% (46)	60% (36)	
Age	Mean ± SD	41.2 ± 11.9	42.7 ± 11.4	0.477
Weight	Mean ± SD	211.2 ± 60.9	168.6 ± 40.6	**0.000**
Height	Mean ± SD	65.8 ± 3.7	66.4 ± 4.1	0.430
Duration of Pain (months)	Mean ± SD	104.2 ± 106.5	129.0 ± 90.9	0.173
Onset of Pain	Gradual	67% (40)	70% (42)	0.845
	Injury	33% (20)	30% (18)	
Pain Distribution	Unilateral	20% (12)	25% (15)	0.662
	Bilateral	80% (48)	75% (45)	
Back Pain Distribution	Back pain only	15% (9)	20% (12)	0.849
	Back pain worse than leg pain	65% (39)	60% (36)	
	Leg pain worse than back pain	5% (3)	3% (2)	
	Both equal	15% (9)	17% (10)	
Numeric Rating Score	Mean ± SD	8.0 ± 1.0	7.7 ± .9	0.082
Oswestry Disability Index	Mean ± SD	30.7 ± 4.5	29.2 ± 5.2	0.096

*Multiple patients presented with disc herniation at more than one level.

Demográfiai és klinikai jellemzők bemutatása két csoportban.
A p-értékek vajon milyen próbák eredményei és mit jelentenek?

Fluoroscopic Lumbar Interlaminar Epidural Injections in Chronic Axial or Discogenic Low Back Pain

Table 1. *Baseline demographic and clinical characteristics.*

		Group I (60)	Group II (60)	*P* value
Gender	Male	23% (14)	40% (24)	0.077
	Female	77% (46)	60% (36)	
Age	Mean ± SD	41.2 ± 11.9	42.7 ± 11.4	0.477
Weight	Mean ± SD	211.2 ± 60.9	168.6 ± 40.6	**0.000**
Height	Mean ± SD	65.8 ± 3.7	66.4 ± 4.1	0.430
Duration of Pain (months)	Mean ± SD	104.2 ± 106.5	129.0 ± 90.9	0.173
Onset of Pain	Gradual	67% (40)	70% (42)	0.845
	Injury	33% (20)	30% (18)	
Pain Distribution	Unilateral	20% (12)	25% (15)	0.662
	Bilateral	80% (48)	75% (45)	
Back Pain Distribution	Back pain only	15% (9)	20% (12)	0.849
	Back pain worse than leg pain	65% (39)	60% (36)	
	Leg pain worse than back pain	5% (3)	3% (2)	
	Both equal	15% (9)	17% (10)	
Numeric Rating Score	Mean ± SD	8.0 ± 1.0	7.7 ± .9	0.082
Oswestry Disability Index	Mean ± SD	30.7 ± 4.5	29.2 ± 5.2	0.096

*Multiple patients presented with disc herniation at more than one level.

kétmintás t-próba

khi-négyzet próba

Demográfiai és klinikai jellemzők bemutatása két csoportban.
A p-értékek vajon milyen próbák eredményei és mit jelentenek?

Fluoroscopic Lumbar Interlaminar Epidural Injections in Chronic Axial or Discogenic Low Back Pain

Table 1. *Baseline demographic and clinical characteristics.*

		Group I (60)	Group II (60)	*P* value
Gender	Male	23% (14)	40% (24)	0.077
	Female	77% (46)	60% (36)	
Age	Mean ± SD	41.2 ± 11.9	42.7 ± 11.4	0.477
Weight	Mean ± SD	211.2 ± 60.9	168.6 ± 40.6	**0.000**
Height	Mean ± SD	65.8 ± 3.7	66.4 ± 4.1	0.430
Duration of Pain (months)	Mean ± SD	104.2 ± 106.5	129.0 ± 90.9	0.173
Onset of Pain	Gradual	67% (40)	70% (42)	0.845
	Injury	33% (20)	30% (18)	
Pain Distribution	Unilateral	20% (12)	25% (15)	0.662
	Bilateral	80% (48)	75% (45)	
Back Pain Distribution	Back pain only	15% (9)	20% (12)	0.849
	Back pain worse than leg pain	65% (39)	60% (36)	
	Leg pain worse than back pain	5% (3)	3% (2)	
	Both equal	15% (9)	17% (10)	
Numeric Rating Score	Mean ± SD	8.0 ± 1.0	7.7 ± .9	0.082
Oswestry Disability Index	Mean ± SD	30.7 ± 4.5	29.2 ± 5.2	0.096

*Multiple patients presented with disc herniation at more than one level.

Csak a súly esetén volt szignifikáns különbség 5%-os szinten az átlagok között

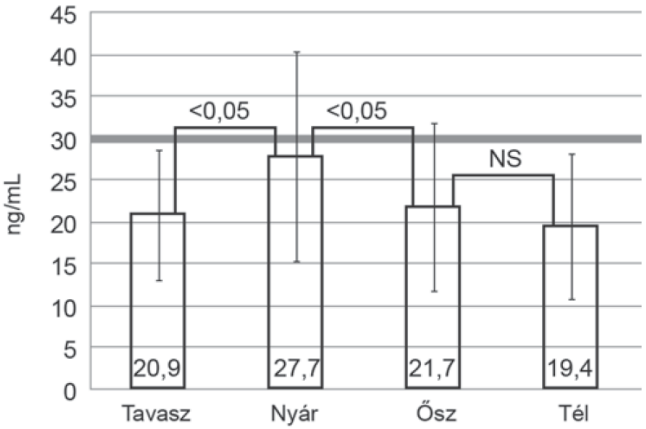
Orvosi Hetilap

D-vitamin-szint mérése hazai gyulladásos bélbetegekben

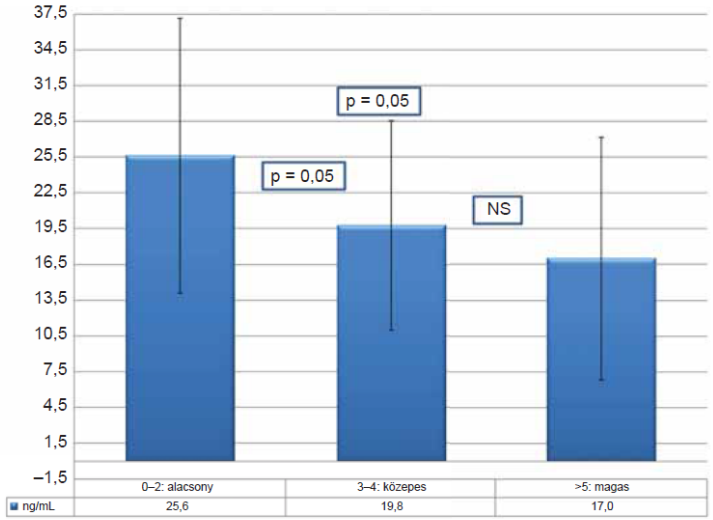
Lőrinczy Katalin dr.[1] ▪ Lakatos Péter László dr.[2] ▪ Tóth Miklós dr.[1]
Salamon Ágnes dr.[3] ▪ Nemes Adrienn dr.[3] ▪ Csontos Ágnes Anna dr.[1]
Fekete Bálint oh.[1] ▪ Terjék Orsolya oh.[1] ▪ Herszényi László dr.[1]
Juhász Márk dr.[1] ▪ Tulassay Zsolt dr.[1] ▪ Miheller Pál dr.[1]

Lőrinczy, K., Lakatos, P. L., Tóth, M., Salamon, Á., Nemes, A., Csontos, Á. A., Fekete, B., Terjék, O., Herszényi, L., Juhász, M., Tulassay, Zs., Miheller, P. (2013). [Vitamin D level in Hungarian patients with inflammatory bowel diseases]. Orv. Hetil., *154*(46), 1821–1828.

tettük. Az adatok statisztikai elemzése az SPSS 17,00 programmal készült. Egy- és kétmintás t-próba és Pearson-féle korrelációszámításokat végeztünk. Az eredményeket átlag±SD formában adtuk meg.

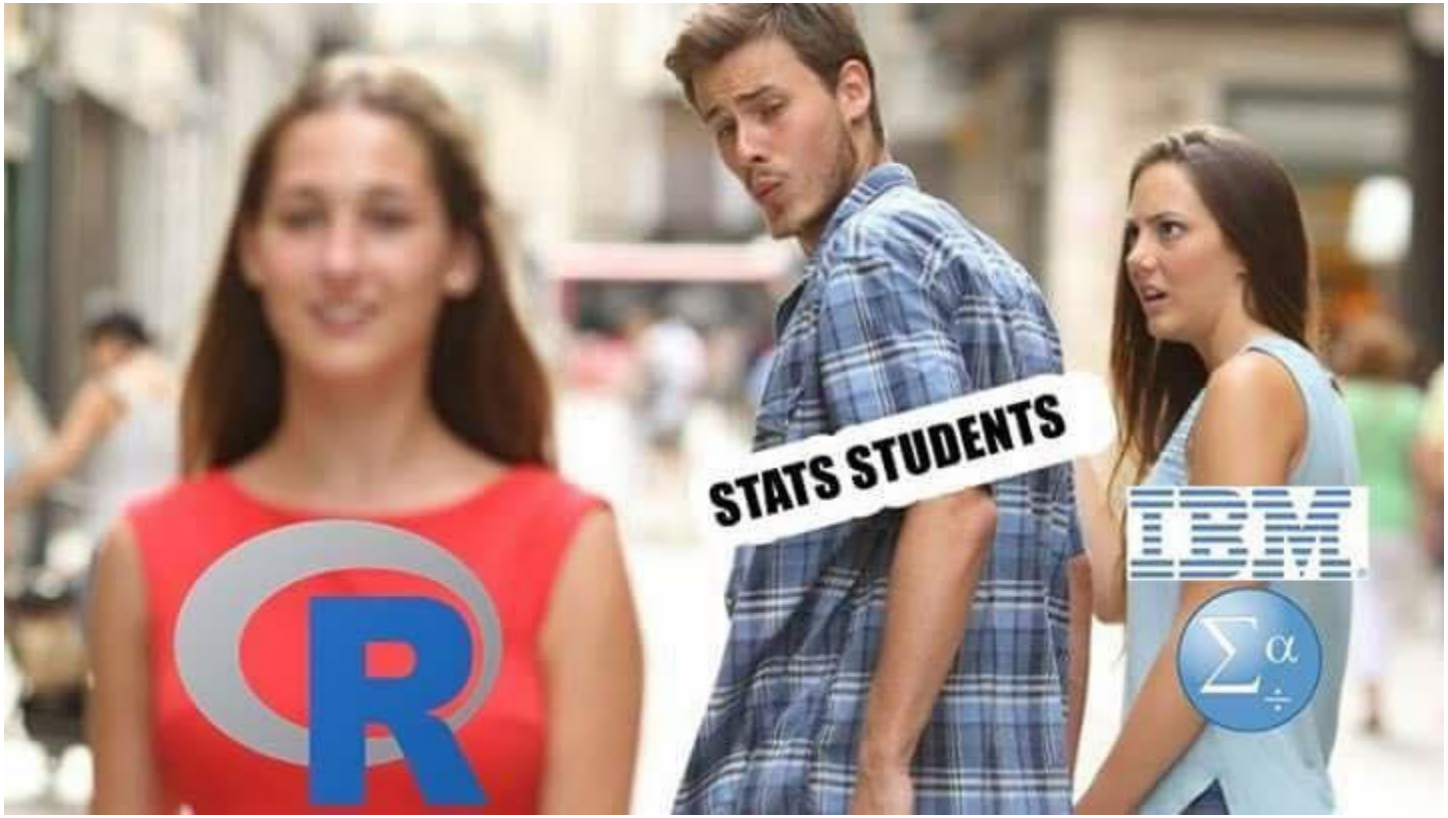


1. ábra | D-vitamin-szint IBD-betegekben évszakonként



2. ábra | D-vitamin-szint-különbség a kalkulált felszívódási zavarnak megfelelően

Statisztikai szoftverek ☺



https://www.facebook.com/thestatisticians/photos/a.326469871020760/955109458156795/?type=3&theater

SITTING IN STATISTICS CLASS
KEEP HEARING "INSTAGRAM" INSTEAD OF HISTOGRAM
quickmeme.com