

Összefüggések statisztikai elemzése

BBNPS01000

Készítette: Soltész-Várhelyi Klára

Statisztikai alapfogalmak:
Leíró statisztikák

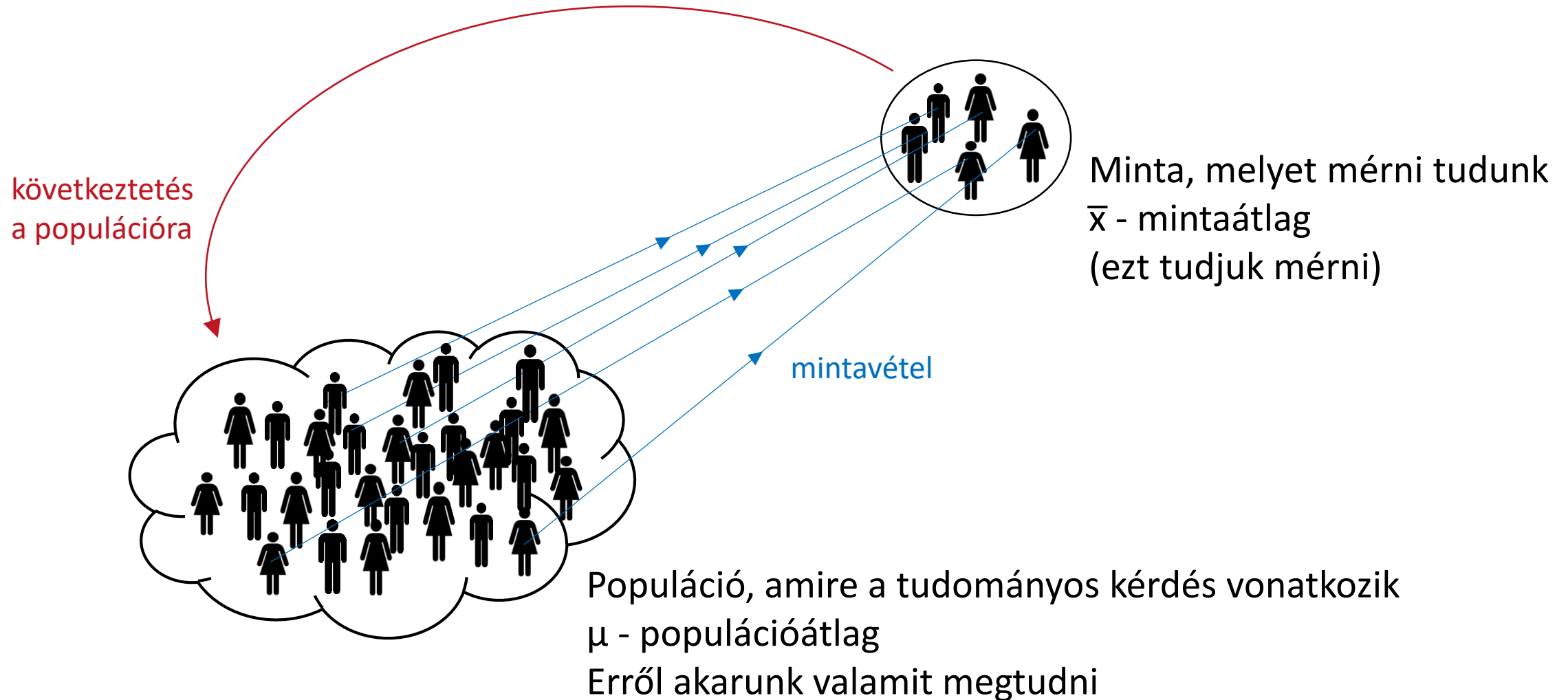
Populáció és minta

Miért csak valószínűségekről beszélhetünk?

“In God we Trust, all others bring data.”

W. Edwards Deming

A kutatások legnagyobb dilemmája



(az ábrát még kiegészítjük néhány diával később, ne ez alapján tanulj meg!)

Populáció és minta

- **Populáció**

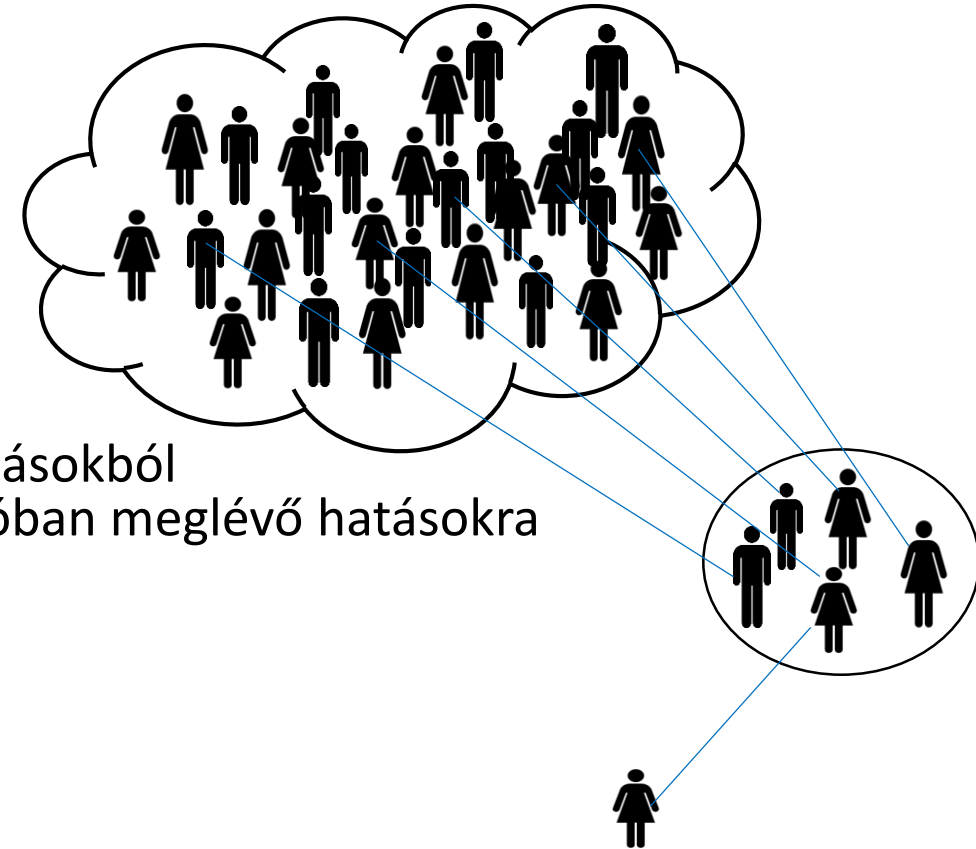
- A kutatási kérdésben meghatározott **tulajdonságoknak megfelelő összes elemek halmaza**
- Hipotézis: Különbség van férfiak és nők magassága között -> Populáció: A világ összes férfijának és nőjének magassága
- **A populációt általában nem lehet vizsgálni**, mert
 - **túl nagy** ahhoz, hogy minden elemét lemérjük
 - **nem minden eleme elérhető** számunkra

- **Minta**

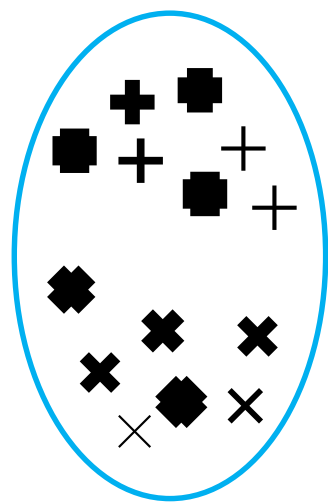
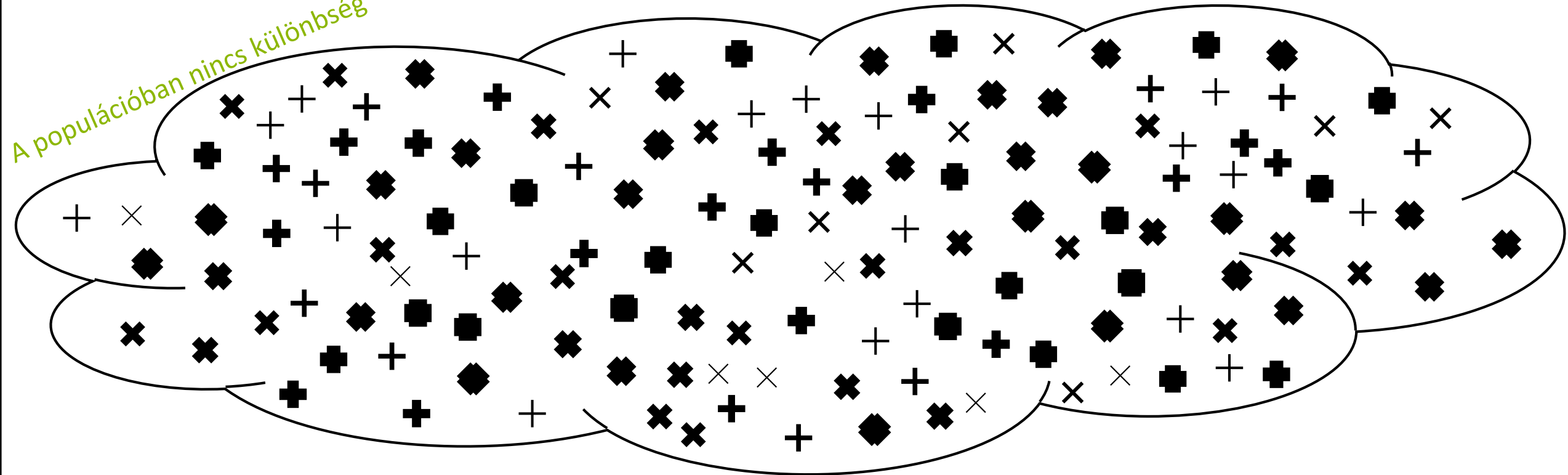
- **Populációból választott kezelhető méretű rész**
- **A minta reprezentálja a populációt:** a mintában mért hatásokból (különbségek, összefüggések) következtetünk a populációban meglévő hatásokra
- Fontos, hogy a minta jól reprezentálja a populációt

- **Egyed**

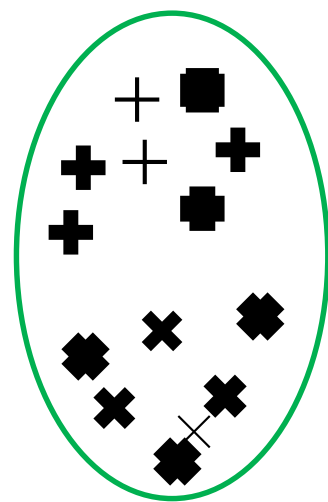
- A populáció egyedeiből épül fel
- A populációból kiválasztott egyedek alkotják a mintát, őket mérjük le



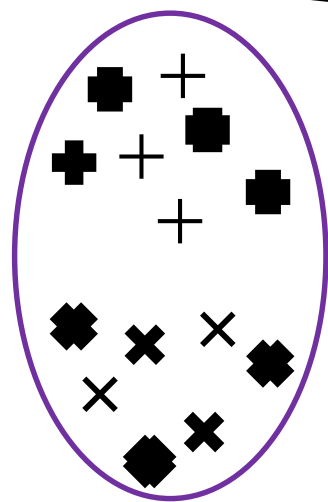
A populációban nincs különbség



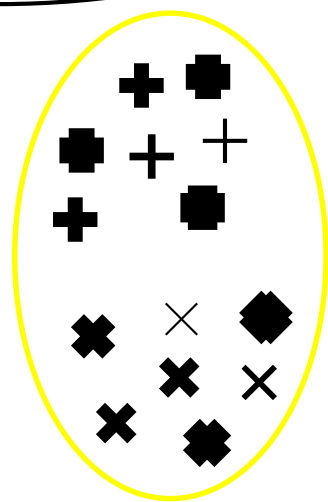
Nem találtunk
különbséget



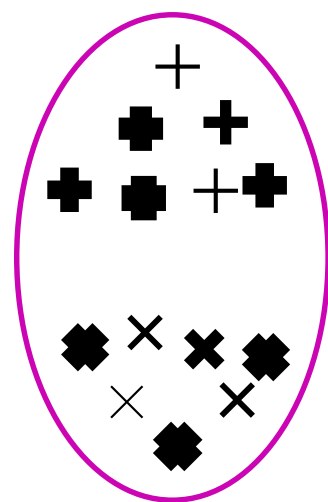
Nem találtunk
különbséget



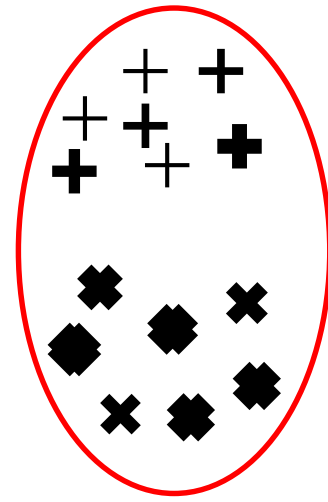
Nem találtunk
különbséget



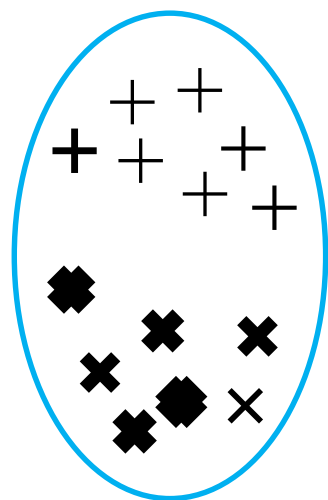
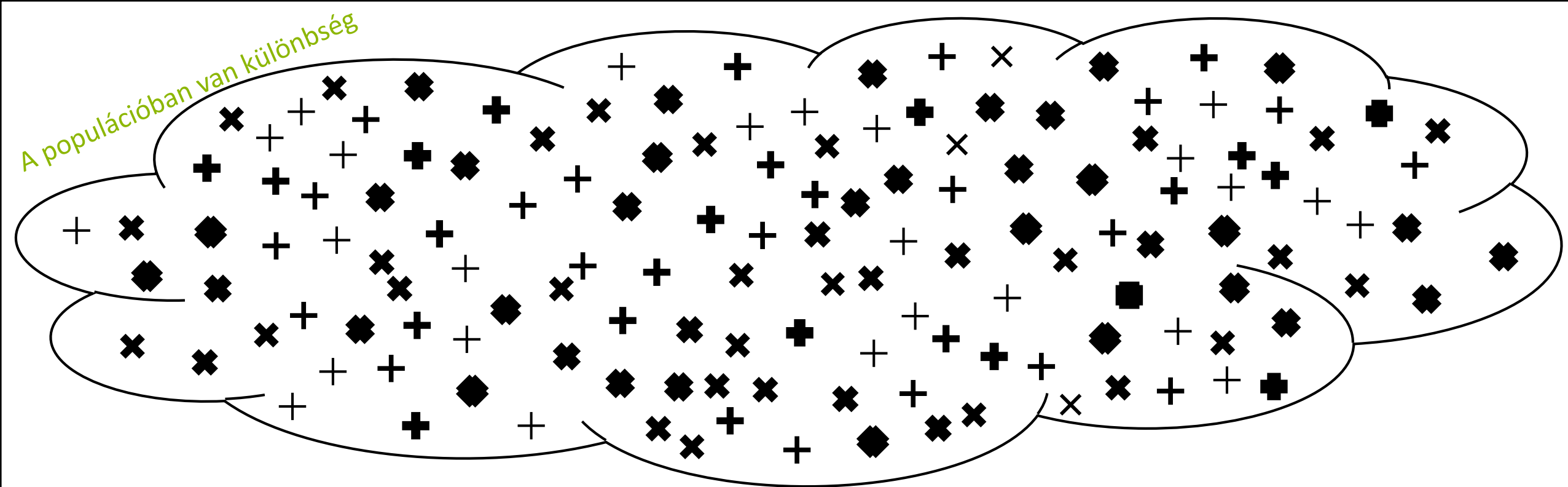
Nem találtunk
különbséget



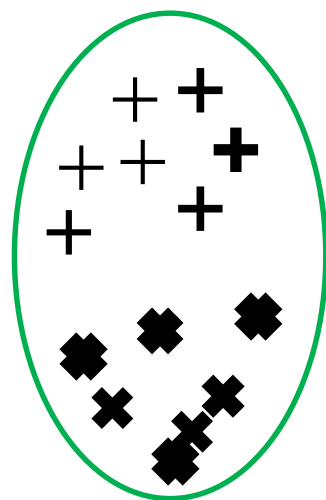
Nem találtunk
különbséget



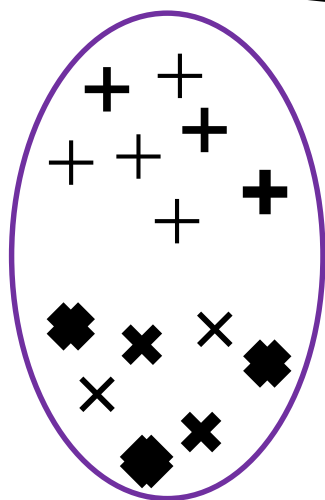
Különbséget találtunk,
és tévedtünk



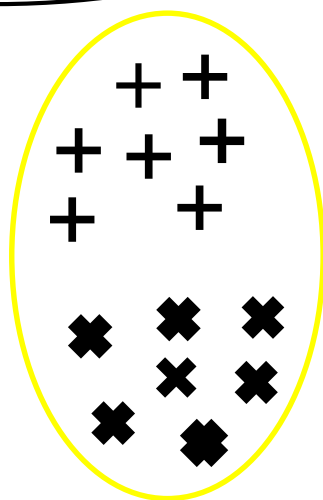
Kimutatjuk a különbséget



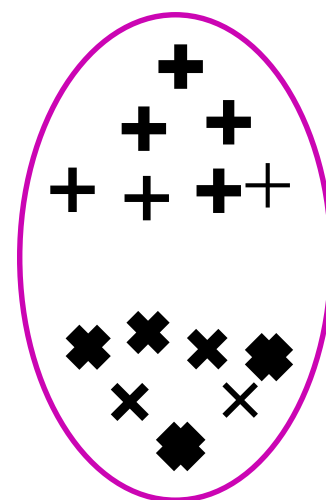
Kimutatjuk a különbséget



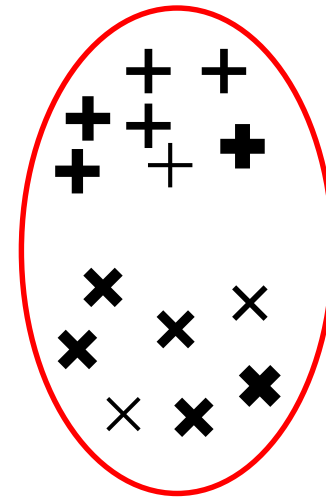
Kimutatjuk a különbséget



Kimutatjuk a különbséget

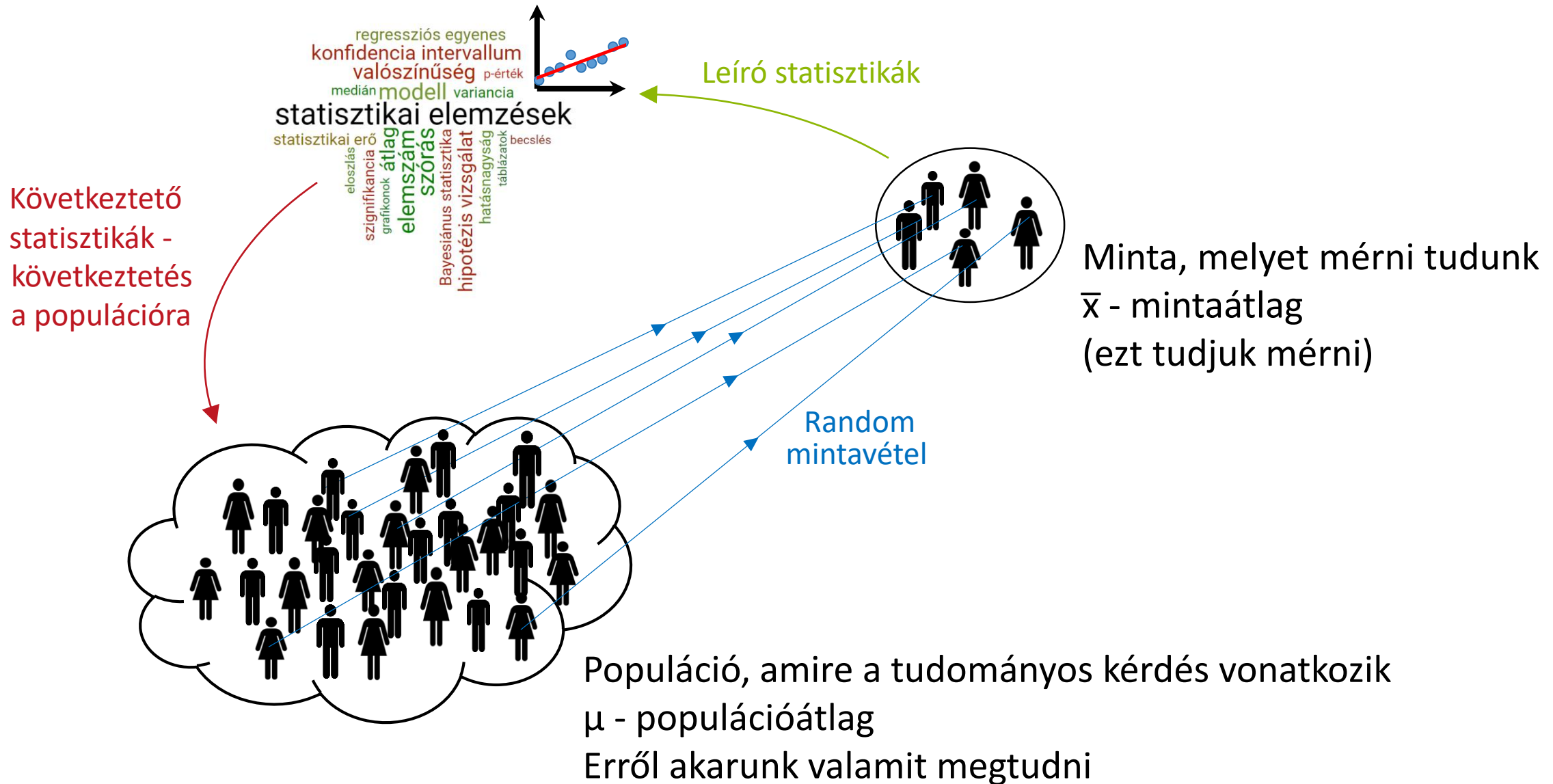


Kimutatjuk a különbséget?



Nem találunk különbséget, és tévedtünk

Populáció és minta



Leíró statisztikák

Mi *van* a mintában?

Statisticians are pleasant folks – even the mean ones are quite nice.

A mért változók

- **Mérési szint:**
 - **Nominális / kategoriális** – diszkrét kategóriák, melyek nem rendezhetőek semmilyen sorrendbe (nem)
 - **Ordinális** – nagyság szerint sorrendbe rakható, de nincs információ a rangsor tagjai közötti különbségekről (elvégzett iskolák)
 - **Skála** típusú változó – az elemek sorba rendezhetőek, az elemek közötti különbség, arány is kifejezhető
 - **Intervallum skála** – nincs természetes nulla pont (dátumok)
 - **Arányskála** – van természetes nulla pont (magasság)

Lásd még: [valtozo_tipusok.pdf](#)



Milyen skála?

(nominális, ordinális, intervallum, arány)

- 1. A személyek neme
- 2. A személyek kora
- 3. Hőmérséklet Celsiusban
- 4. A személyek neve
- 5. Iskolai osztályzat
- 6. IQ
- 7. Katonai rang
- 8. Hőmérséklet Kelvinben
- 9. Betegség stádiuma
- 10. Tömeg
- 11. Foglalkozás
- 12. Részecskeszám
- 13. Elvégzett iskolák
- 14. Telefonszám
- 15. Hány A betűvel kezdődő szót tudsz mondani 1 perc alatt
- 16. Az óra kezdete óta eltelt idő
- 17. Fizetés
- 18. Naptári napok



Milyen skála?

(nominális, ordinális, intervallum, arány)

- | | |
|--|---------|
| • 1. A személyek neme | • 1. N |
| • 2. A személyek kora | • 2. A |
| • 3. Hőmérséklet Celsiusban | • 3. I |
| • 4. A személyek neve | • 4. N |
| • 5. Iskolai osztályzat | • 5. O |
| • 6. IQ | • 6. O |
| • 7. Katonai rang | • 7. O |
| • 8. Hőmérséklet Kelvinben | • 8. A |
| • 9. Betegség stádiuma | • 9. O |
| • 10. Tömeg | • 10. A |
| • 11. Foglalkozás | • 11. N |
| • 12. Részecskeszám | • 12. A |
| • 13. Elvégzett iskolák | • 13. O |
| • 14. Telefonszám | • 14. N |
| • 15. Hány A betűvel kezdődő szót tudsz mondani 1 perc alatt | • 15. A |
| • 16. Az óra kezdete óta eltelt idő | • 16. A |
| • 17. Fizetés | • 17. A |
| • 18. Naptári napok | • 18. I |

Elemzés, minimum, maximum

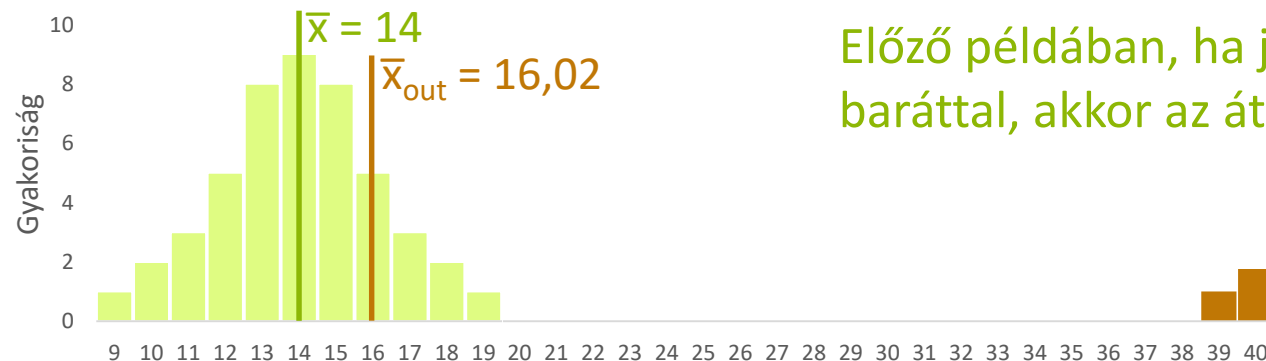
- Példa: statisztika tanárok barátai: 2, 3, 1, 3, 4
- **Elemzés**
 - Minta esetében jelölése **n** vagy **N**
 - Példában: $N = 5$
- **Minimum**
 - Jelölése: **Min**
 - Példában: $Min = 1$
- **Maximum**
 - Jelölése: **Max**
 - Példában: $Max = 4$



Elhelyezkedési mutatók: Átlag

- **Átlag**

- Statisztikában általában a számtani átlagot használjuk.
- A **mintaátlag** jelölése: a képletekben $\bar{x}$, a statisztikai közlésben **M** (mint Mean)
- Számolása: $\bar{x} = \frac{\sum x_i}{n}$ Példában: $\bar{x} = \frac{2+3+1+3+4}{5} = 2.6$
- A mintaátlagból következtetünk a populációátlagra, melynek értékét nem ismerjük.
 - Populációátlag: a populáció központi tendenciáját leíró paraméter, jelölése μ (mű)
- Mikor működik jól az átlag?
 - Csak skála típusú változókon számolható (ordinális változókon nem).
 - Szimmetrikus eloszlások esetén működik.
 - Szélsőséges értékekre érzékeny, előfordulhat, hogy azok túlzottan befolyásolják az értékét.

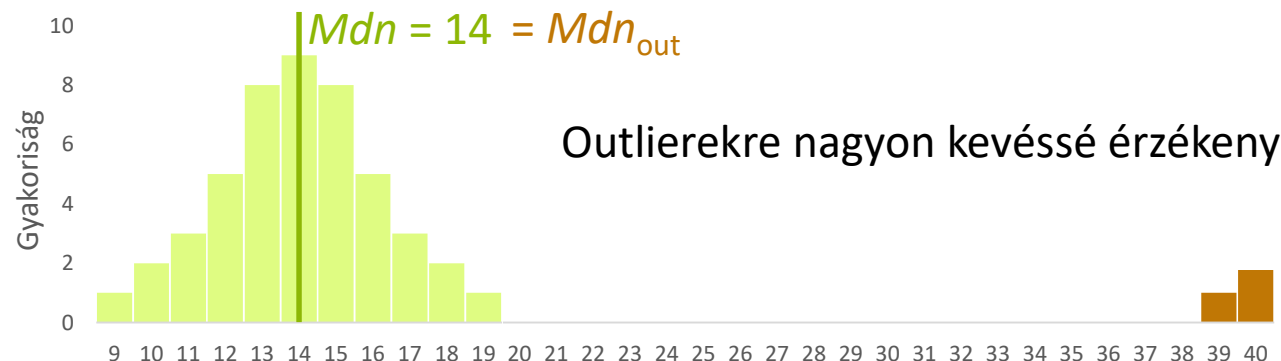


Előző példában, ha jön egy új tanár 17 baráttal, akkor az átlagot felviszi 5-re

Elhelyezkedési mutatók: Medián

- **Medián**

- Az az érték, melynél az **elemek legfeljebb 50%-a nagyobb és legfeljebb 50%-a kisebb**
- Jelölése: **Mdn**
- Számolása: Sorba rendezzük a minta elemeit, majd ha páratlan számú a minta, akkor a medián a középső elem, ha páros, akkor a két középső elem átlaga
 - Statisztika tanárok barátainak száma: 2, 3, 1, 3, 4 → Sorba rendezve: 1, 2, 3, 3, 4
Mdn = 3
 - Pszichológia tanárok barátainak száma: 3, 5, 3, 4, 2, 6 → Sorba rendezve: 2, 3, 3, 4, 5, 6
 $Mdn = \frac{3+4}{2} = 3,5$
- Mikor előnyösebb a használata az átlagnál?
 - Ordinális adatok (de működik folytonos adatokon is), ferde eloszlás, outlierok esetén

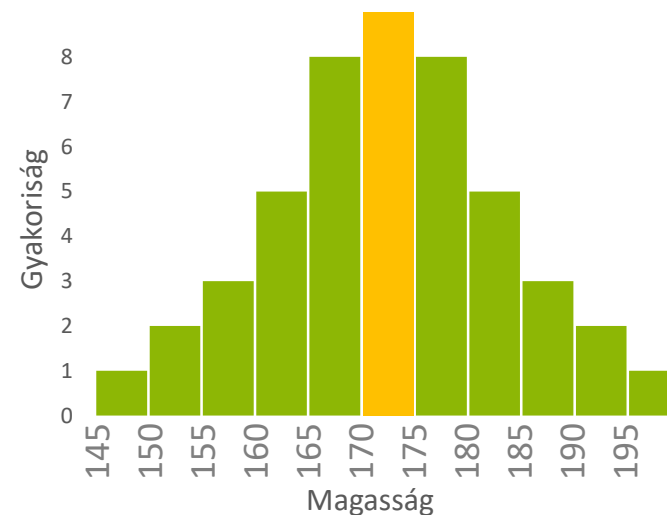
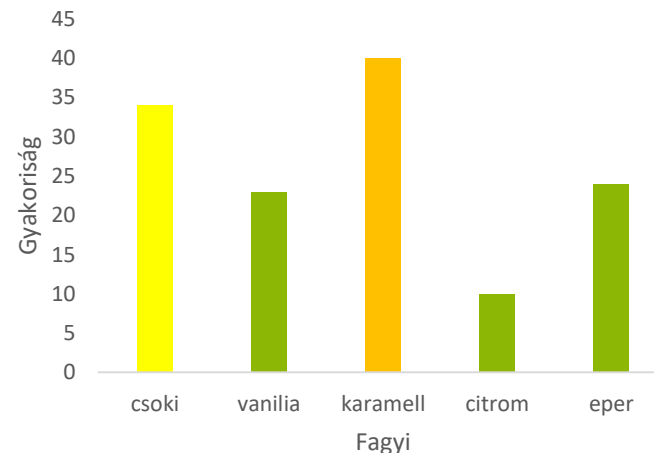


Előző példában, ha jön egy új tanár 17 barátával, akkor a medián 3 marad
1 2 3 3 4 17

Elhelyezkedési mutatók: Módusz

- **Módusz**

- A legdivatosabb, **leggyakrabban előforduló érték**.
- Jelölése: **Mo**
- Számolása:
 - Lehet számolni folytonos változókra is, de legnagyobb haszna diszkrét adatok esetén van:
a leggyakrabban előforduló érték. Példa: 2, **3**, 1, **3**, 4 -> **Mo** = 3
- Mikor előnyös a használata?
 - Nominális változók esetén is értelmezhető: emberek kedvenc színe.
 - Ha olyan a kérdés: egy cipőgyárnak a leggyakrabban vásárolt méretből kell sokat gyártania, nem az átlagos vagy medián méretből.

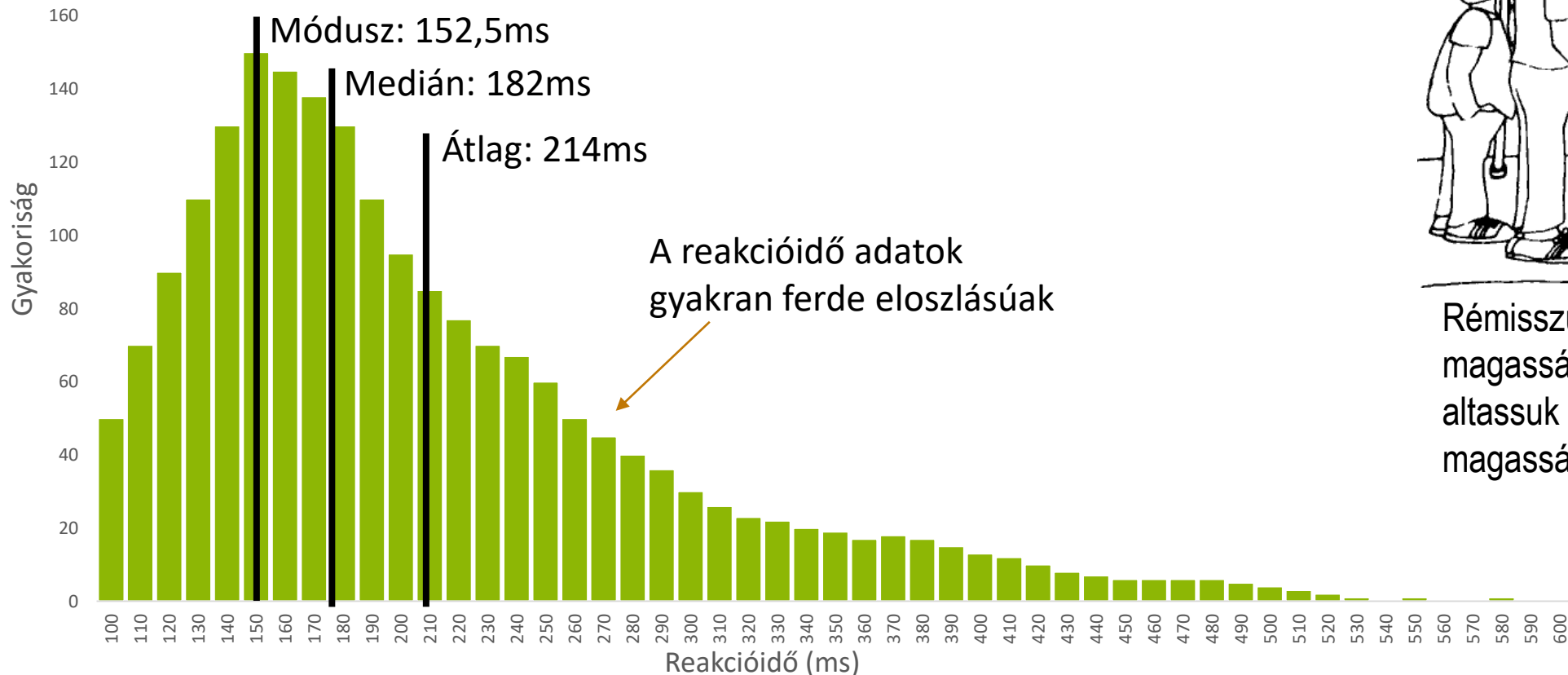


Elhelyezkedési mutatók: Átlag vs Medián vs Módusz

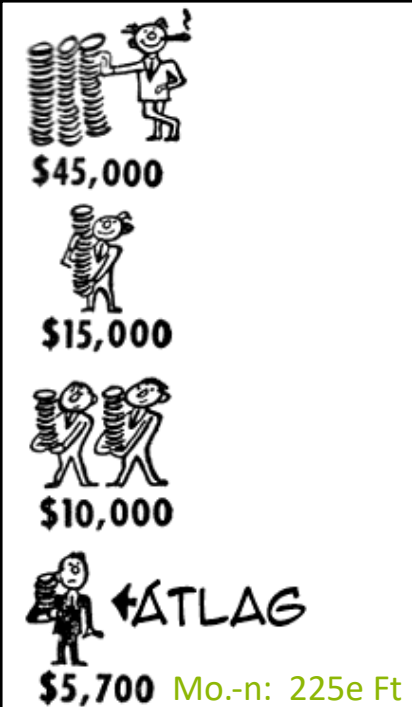
Átlag: szimmetrikus, szélsőséges értékektől mentes adatokon a legjobb, az alábbi példában túlzottan elhúzzák a ritkán előforduló, lassú reakcióidők

Medián: ferde eloszlású vagy szélsőséges értékekkel rendelkező adatokon a legjobb, például reakcióidőadatokról.

Módusz: túlzottan (teljesen) figyelmen kívül hagyja a lassabb reakcióidőket.



Rémísszük meg őket a csapatunk magasságának átlagával, vagy altassuk el a figyelmüket a csapat magasságának mediánjával?

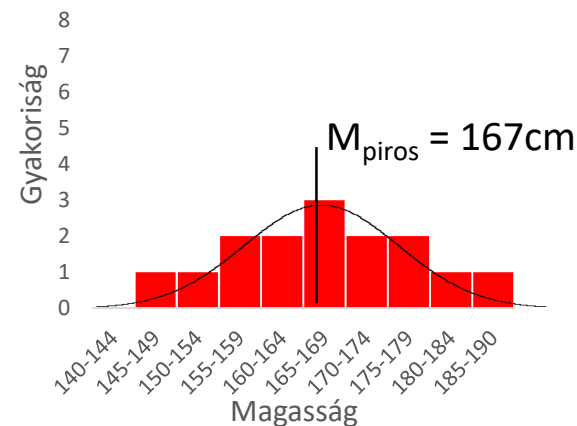
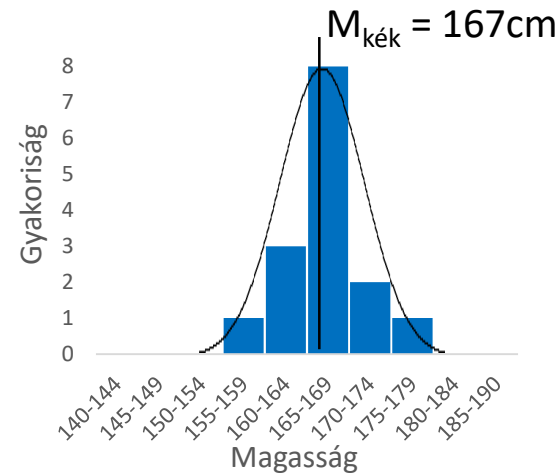
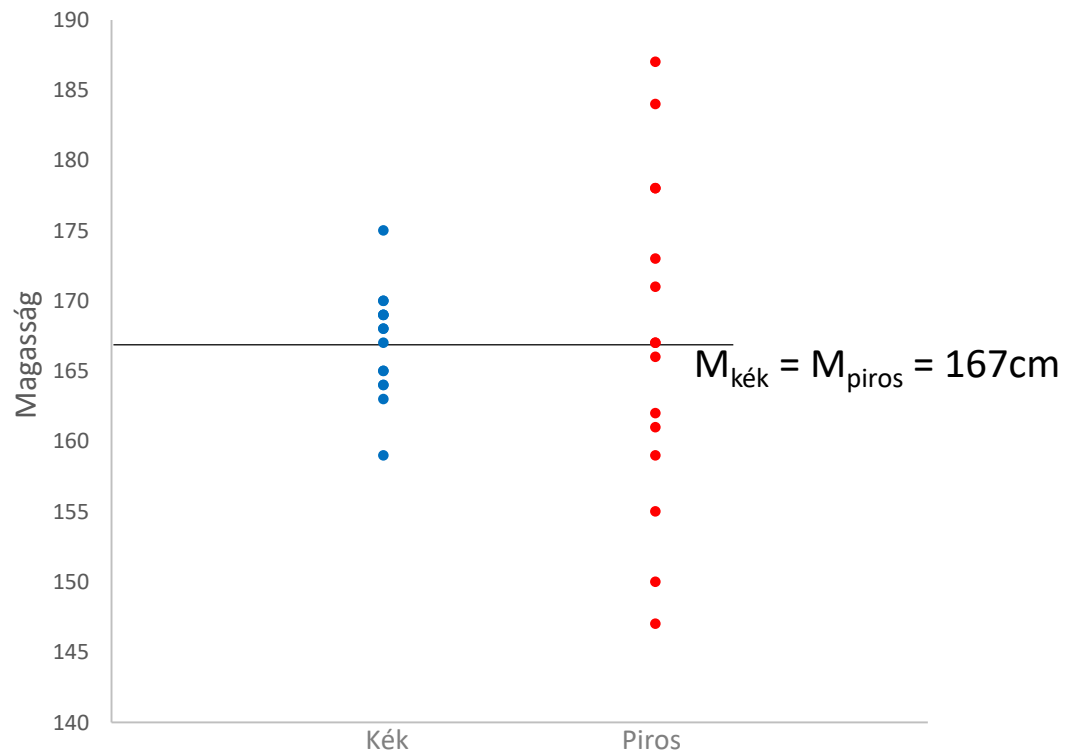


- A kérdéstől függ, mi a megfelelő középérték mutató.
- A világ nagyon más aspektusát ragadják meg.



Szóródási mutatók

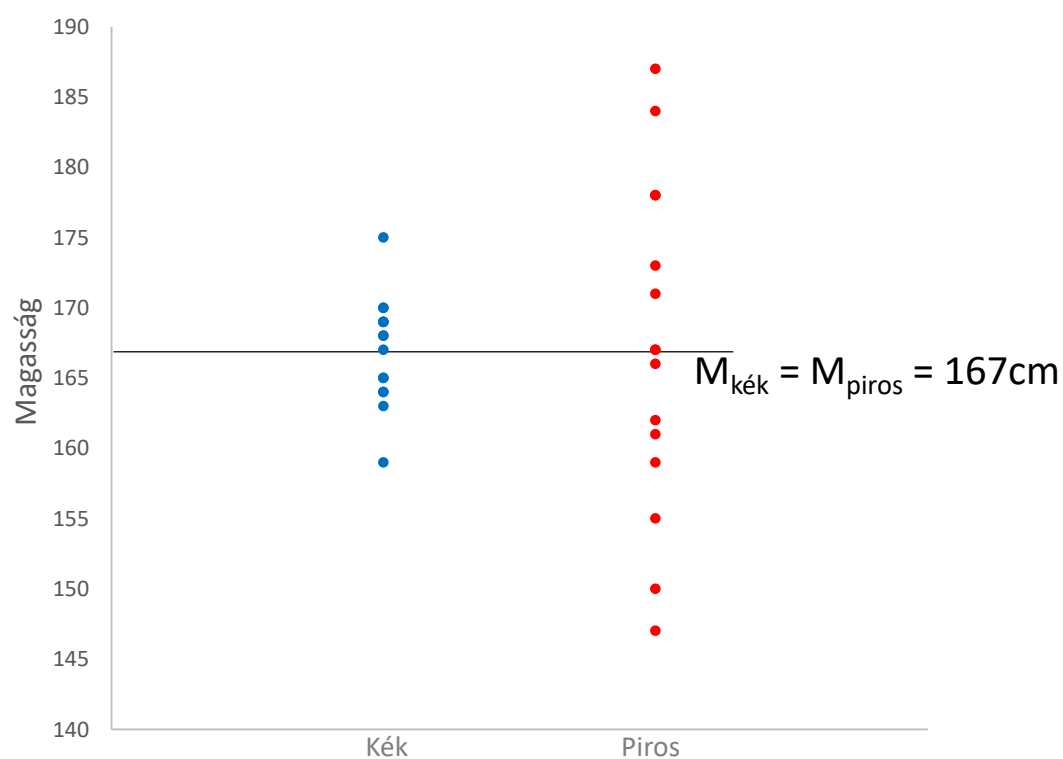
- Miért nem elég az átlag?
 - A minta szóródását nem jellemzi



Kékek	Pirosak
163	155
169	178
168	187
169	166
170	178
168	159
165	171
164	150
164	167
170	147
159	184
175	162
165	167
169	161
167	173

Szóródási mutatók: Átlagtól való eltérés

- **Átlagtól való eltérés (D, az angol deviance szóból)**
 - A **mért értékek és az átlag távolsága**, az átlaggal való predikció **hibája**, pontatlansága.
 - Jelölése: D_i Számolása: $D_i = x_i - \bar{x}$
- **Átlagtól való négyzetes eltérés (D^2)**
 - Csak a különbség nagysága kell, iránya (előjele) nem.
 - Abszolútérték vagy **négyzetre emelés**? Számolása: D_i^2



Kékek	$D_{\text{kék}}$	$D^2_{\text{kék}}$	Pirosak	D_{piros}	D^2_{piros}
163	$163-167 = -4$	$(-4)^2 = 16$	155	-12	144
169	$169-167 = 2$	$2^2 = 4$	178	11	121
168	$168-167 = 1$	$1^2 = 1$	187	20	400
169	2	4	166	-1	1
170	3	9	178	11	121
168	1	1	159	-8	64
165	-2	4	171	4	16
164	-3	9	150	-17	289
164	-3	9	167	0	0
170	3	9	147	-20	400
159	-8	64	184	17	289
175	8	64	162	-5	25
165	-2	4	167	0	0
169	2	4	161	-6	36
167	0	0	173	6	36

Szóródási mutatók: variancia és szórás

- **Variancia**

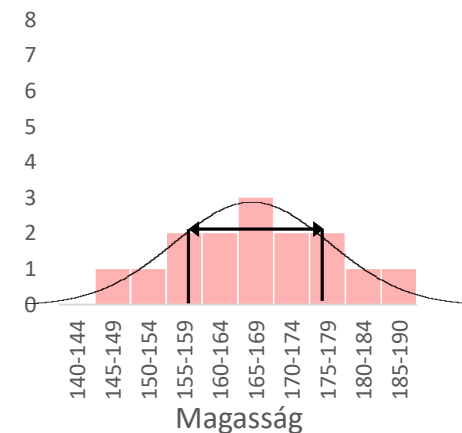
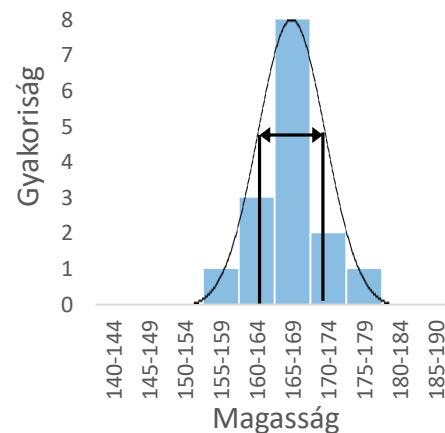
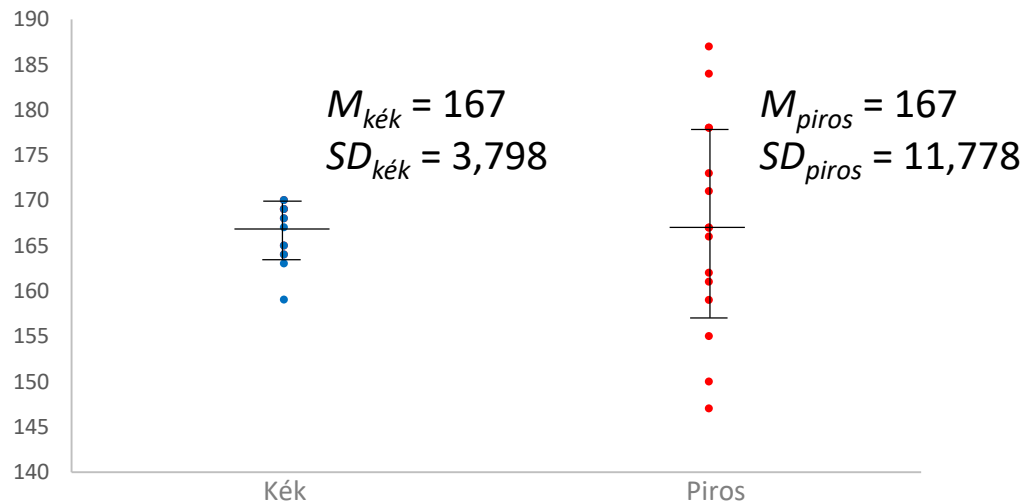
- Az átlagtól való négyzetes eltérés átlaga, jelölése: **Var**, **s^2** vagy **σ^2**

- Számolása: $Var = \frac{\sum D_i^2}{df} = \frac{\sum (x_i - \bar{x})^2}{n-1}$ Pl.: $Var_{kék} = \frac{16+4+1+4+9+1+4+9+9+9+64+64+4+4+0}{15-1} = 14,429$
Szabadságfok (következő dián)

- **Szórás**

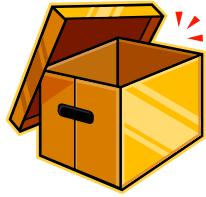
- Átlagtól való átlagos eltérés, jelölése: **SD** (standard deviation), **s** vagy **σ**

- Számolása: $SD = \sqrt{Var} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$ Pl.: $SD_{kék} = \sqrt{14,429} = 3,798$



Egy kis kitérő: szabadságfok

1.



2.



3.



4.



- **Szabadságfok** (degrees of freedom, **df**)
 - Szabadon, **egymástól függetlenül választható tagok száma**
 - Számolása: a tagok száma mínusz az összefüggések száma
 - Szórás számolásakor értéke: $df = N - 1$
- Gyakorlati haszna (Bessel korrekció):
 - Mivel $df < N$, és df a nevezőben van, így a szórás egy kicsit nagyobb lesz szabadságfokkal számolva, mint a minta elemszámával, ezáltal korrigál a mérés pontatlanságára.
 - A korrekció kis elemszám esetén jelentősebb, ami logikus, hiszen nagy elemszám esetén valószínűbb, hogy pontosan becsüljük a populációátlagot.

$$VAR = \frac{\sum (x_i - \bar{x})^2}{N - 1}$$

Miért nagy a szórás?

- **Mekkora a „nagy” szórás?**

- Nincs kritérium érték, mert függ a skálázástól!
 - Két egységnyi szórás érdemjegy esetén (1-5-ig terjedő skála) vagy zh pontszám szerint (0-100-ig terjedő skála) mást jelent

- **Miért nagy a szórás?**

- **Kicsi az elemszám**

- A szabadságfokkal való korrekció okozza. A korrekciónak csak kis elemszám esetén jelentős a hatása, nagy elemszám esetén a 1 eltörpül az N-hez képest

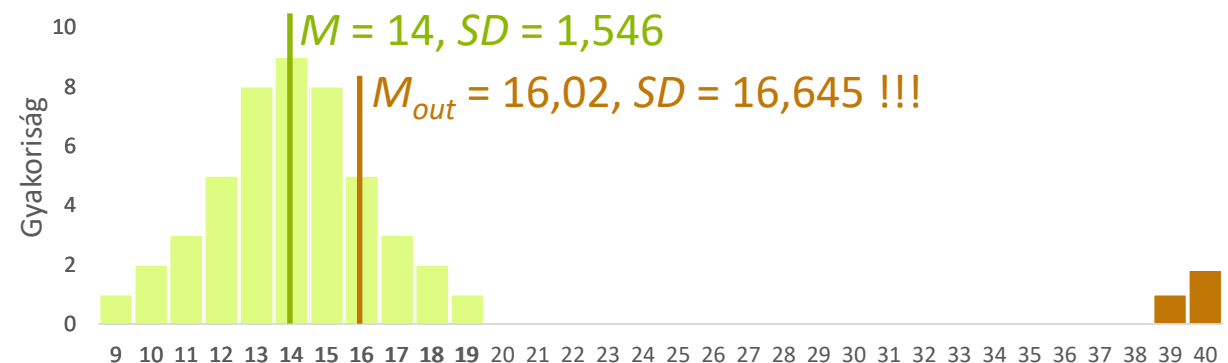
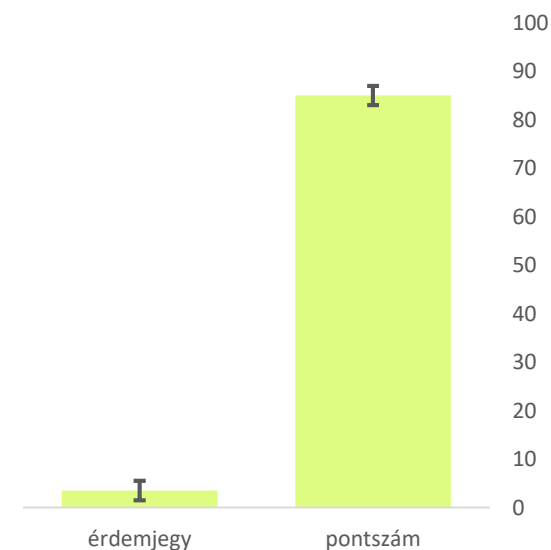
- $$SD = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

- **Outlierek**

- A szélsőséges értékek megnövelik a szórást, mert a többi értékhez képest jóval nagyobb az átlagtól való távolságuk

- **Csúcsosság** nem normális (később)

- **A tulajdonság nem stabil**



Szóródási mutatók: Standard error

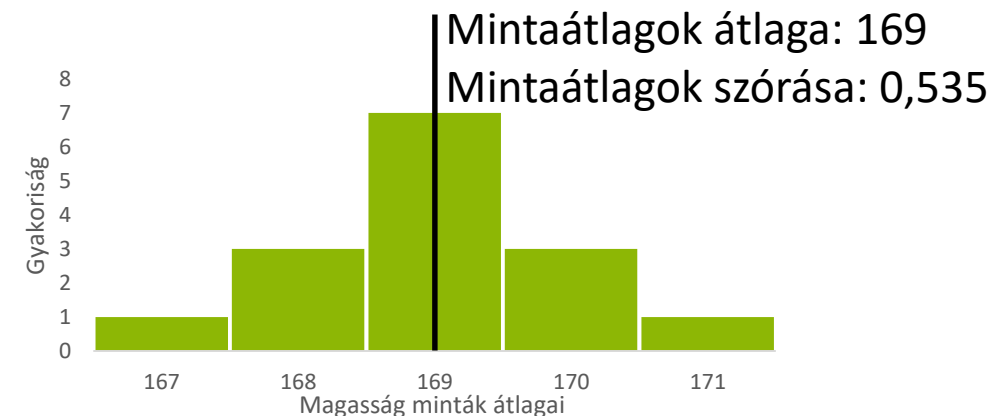
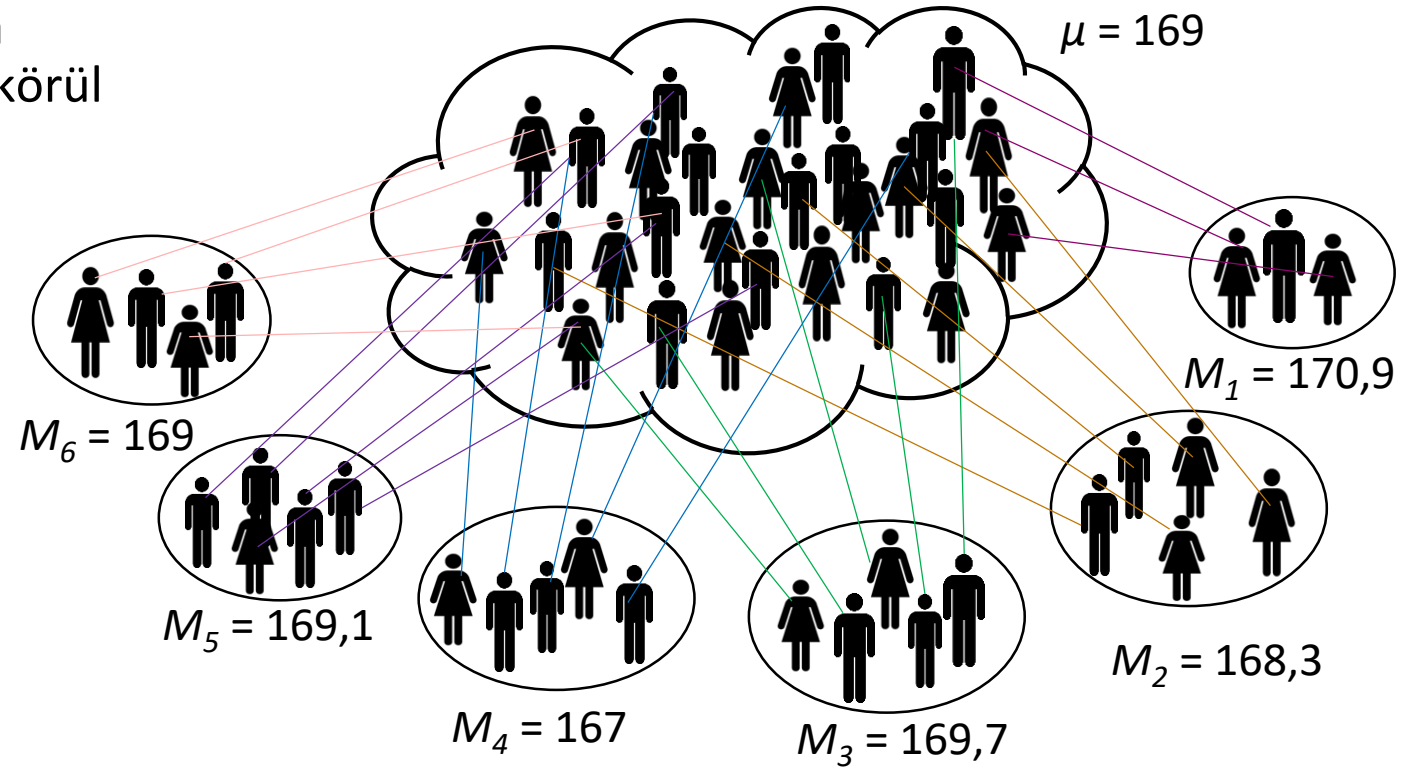
- Ha sok mintát vennénk a populációból, akkor a mintaátlagok ($\bar{x}_i$) valahol a populációátlag (μ) körül lennének, de nem lennének azonosak, minden mintának kicsit máshol lenne az átlaga

- Standard error:**

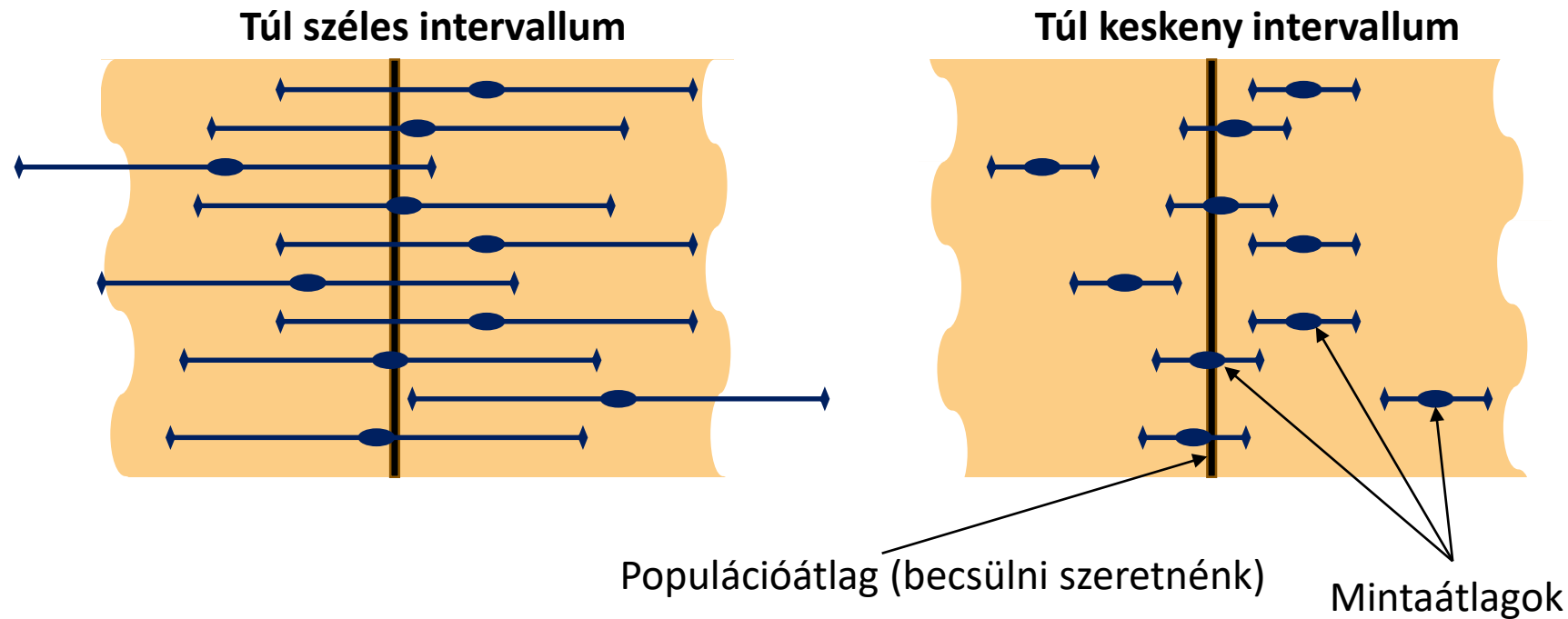
- A populációból vett minták átlagainak szórása**
- Jelölése: $s_{\bar{x}}$ vagy **SE**
- Számítása (ha csak egy mintánk van):

$$SE = \frac{s}{\sqrt{N}} = \frac{\frac{\sum (x_i - \bar{x})^2}{N - 1}}{\sqrt{N}}$$

- Az SE megadja, mennyire bizonytalan az általunk mért minta átlaga
 - Ha nagy a SE, akkor nem igazán reprezentatív a mintánk a populációra (lehet, holnap veszek egy másik mintát, és egészen más lesz az átlaga)



Szóródási mutatók: Konfidencia intervallum / megbízhatóság

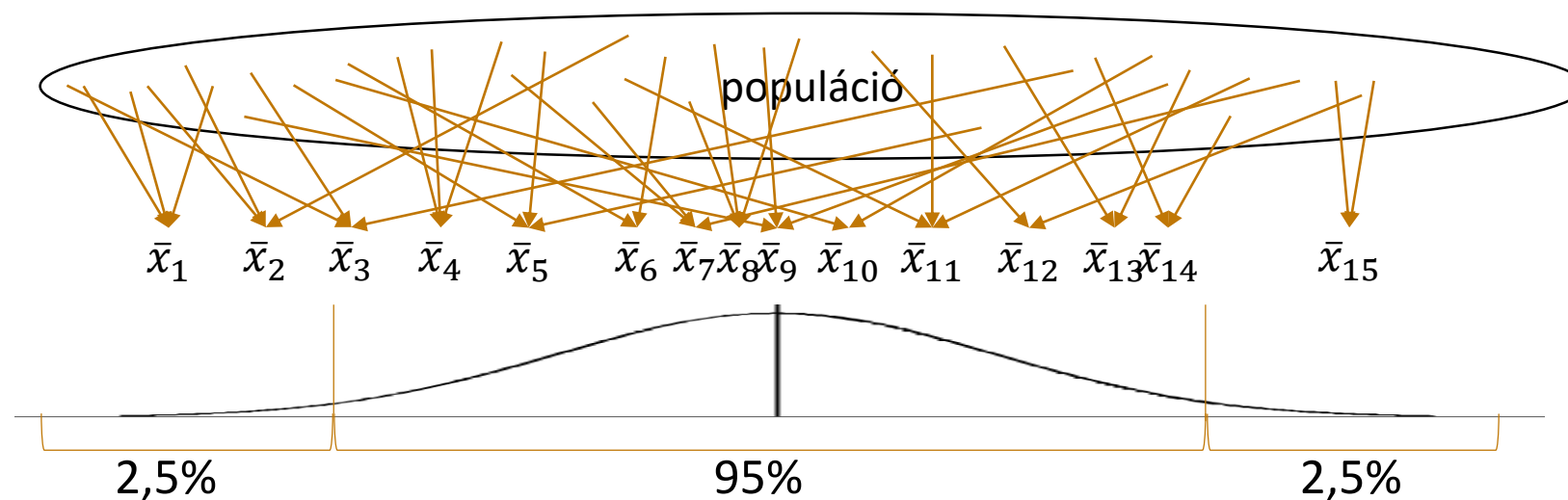


Adjunk meg a mintánk átlaga körül egy intervallumot, amin belül fog feltehetőleg a populációátlag is esni!
De mekkora legyen az intervallum? Ha túl nagy, nincs információértéke, ha túl kicsi, valószínű, hogy a populációátlag nem fog beleesni.

Kellene **egy olyan határ, amibe nagy (pl. 90% vagy 95%-os) valószínűséggel beleesik a populációátlag.**

Ezt és a következő két diát most kihagyjuk, és az eloszlásgörbék megértése után térünk vissza ide!!

- Vegyünk a populációból sok mintát, és számoljuk ki az átlagaikat! Azt látjuk, hogy a mintaátlagok normál eloszlást követnek.
- Határozzuk meg egy határértéket, amibe a mintaátlagok 95%-a beleesik, így feltételezhetjük, hogy a populáció átlag is 95%-os valószínűséggel benne lesz!



- Hol lesz ez a határérték?
 - Emlékezzünk! Z-értéknél kiemeltem a **± 1.96** -os z-értéket, melyről tudjuk, hogy a z-értékek 95%-a e kettő érték között van.
 - Tehát, ha kiszámolnánk a **mintaátlagok z-értékeit**, akkor az a két mintaátlag között, melyhez a **$+$ és -1.96 -os z-érték tartozik, lesz a mintaátlagok 95%-a.**

- Hogyan számoljuk ki a konfidencia intervallumot egyetlen mért minta esetén?

- Lássuk a z-értékek képletét még egyszer: $Z = \frac{x_i - \bar{x}}{s}$.

- Most a Z-értéket ismerjük: a ± 1.96 -os z-értékhez tartozó határértéket keressük

- Ebből egyenletrendezéssel:

Az átlag átlaga az átlag marad,
tehát ez nem változik

± 1.96 (95%-os bizonyosság esetén)

A határérték, amit keresünk

$$X = \bar{X} + z * s$$

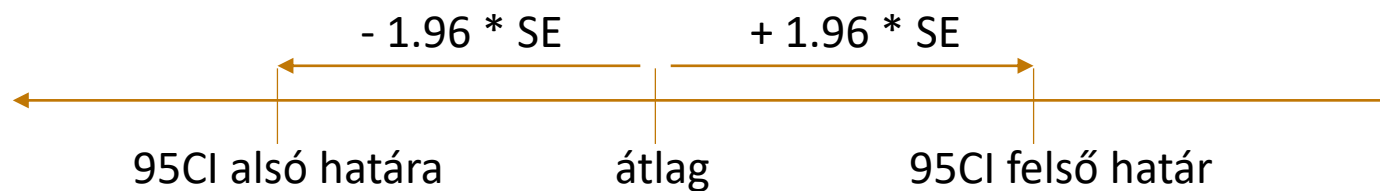
helyett SE

Minket a mintaátlagok szórása érdekel, ezért szórás helyett a
mintaátlagok szórásával, tehát standard errorral kell számolnunk

- Az így kialakult képlet:
$$X = \bar{X} + (\pm 1.96) * SE$$

- A 95%-os konfidencia intervallum felső határa: $\text{átlag} + 1.96 * SE$
- A 95%-os konfidencia intervallum alsó határa: $\text{átlag} - 1.96 * SE$

- Mit mond el nekünk a konfidencia intervallum a minta és populáció kapcsolatáról? Mi befolyásolja, milyen magabiztosan következtethetünk a mintából a populációra?
- A konfidencia intervallum határainak képlete felbontva a standard error: $X = \bar{X} + z * \frac{s}{\sqrt{N}}$

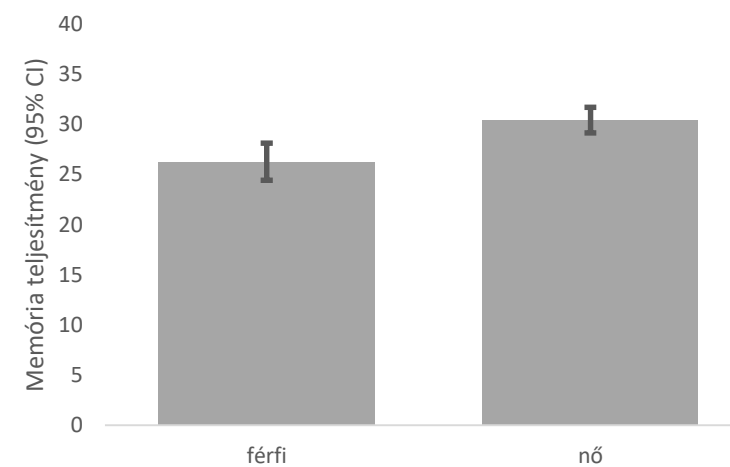
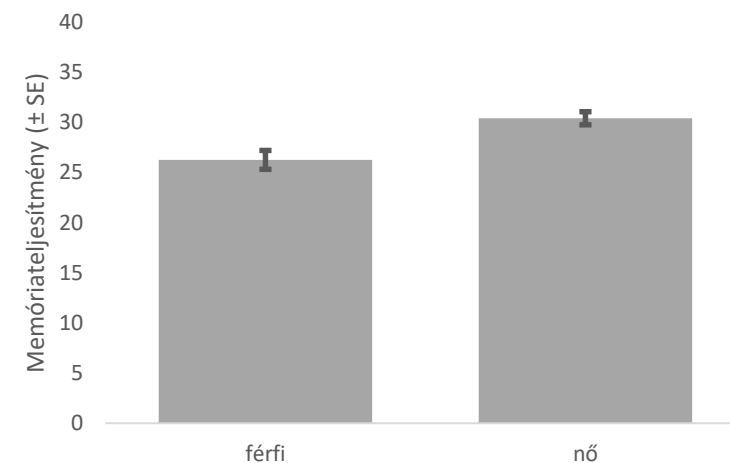
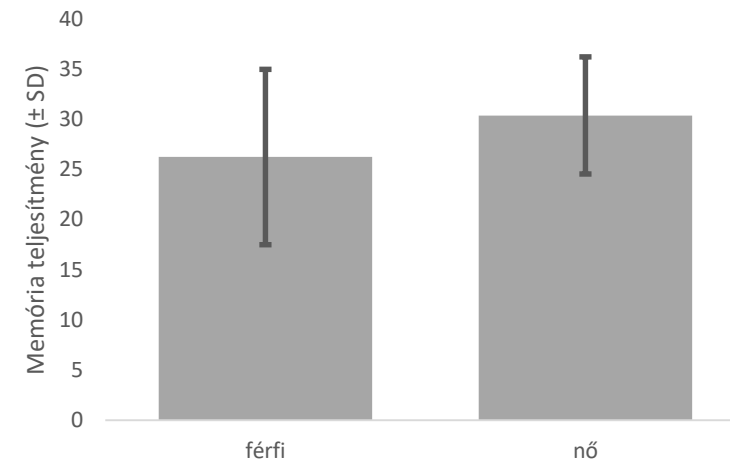
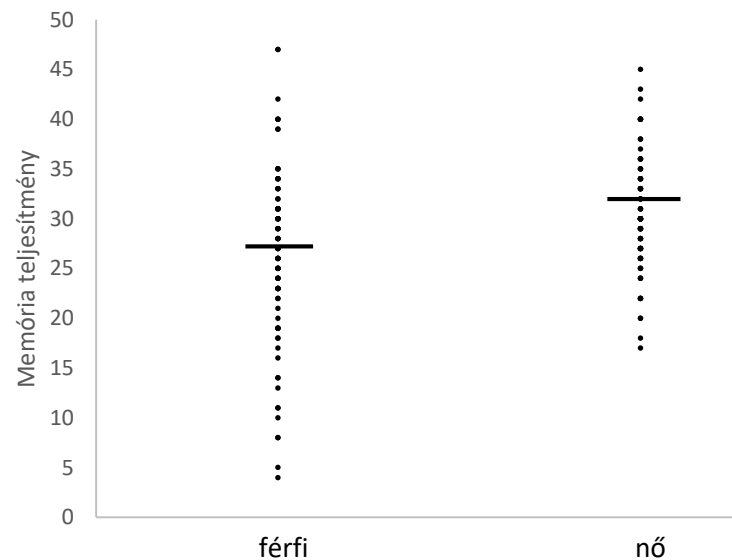


- **Mi van a képletben?**
 - **Átlag:** csak az offsetet adja meg, a CI szélességét nem befolyásolja
 - **Z érték:** ezzel szabályozhatjuk, mekkora valószínűséggel akarjuk a populációátlagot megtalálni. 95%-os valószínűséghez ± 1.96 -os z-érték tartozik. 99%-hoz ± 2.576 . 90%-hoz csak ± 1.645 . Láthatjuk, hogy minél „biztosabbra akarunk menni”, annál szélesebb CI kell.
 - **Szórás:** a számlálóban szerepel, tehát minél nagyobb a szórás, annál kevésbé lehetünk biztosak abban, hogy a mintaátlag jó reprezentálja a populációt, annál szélesebb lesz a CI.
 - **Elemsszám:** számlálóban szerepel. Minél nagyobb a minta, annál nagyobb részét fedtük le a populációnak, annál biztosabbak lehetünk abban, hogy a minta jól reprezentálja a populációt, annál keskenyebb lesz a CI.

Hibamutatók összefoglaló

- A grafikonokon a hibamutatók hibasávokkal (error bars vagy whiskers) jelezzük
- grafikonokon az átlag mellett mindig meg kell jeleníteni valamilyen hibasávot, és fel kell tüntetni, a három hibasáv közül melyiket alkalmaztuk
- Grafikonok olvasásánál is fontos figyelembe venni, milyen hibasávot látunk: a szórások mindig a legszélesebbek, míg a standard errorok a legszűkebbek.

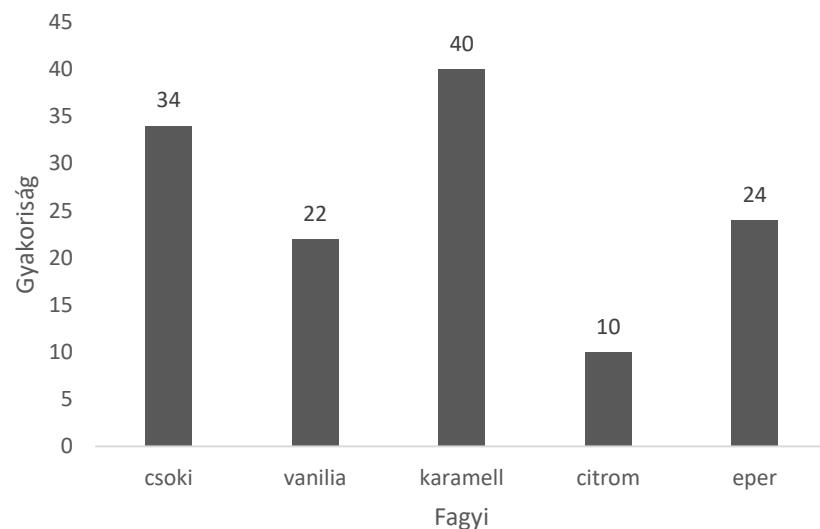
	Férfi	Nő
Átlag	26,244	30,388
Elemzés	86	80
SD	8,737	5,823
SE	0,942	0,651
CI	1,846	1,276



- Gyakoriság eloszlás, hisztogram

- Diszkrét változó esetében

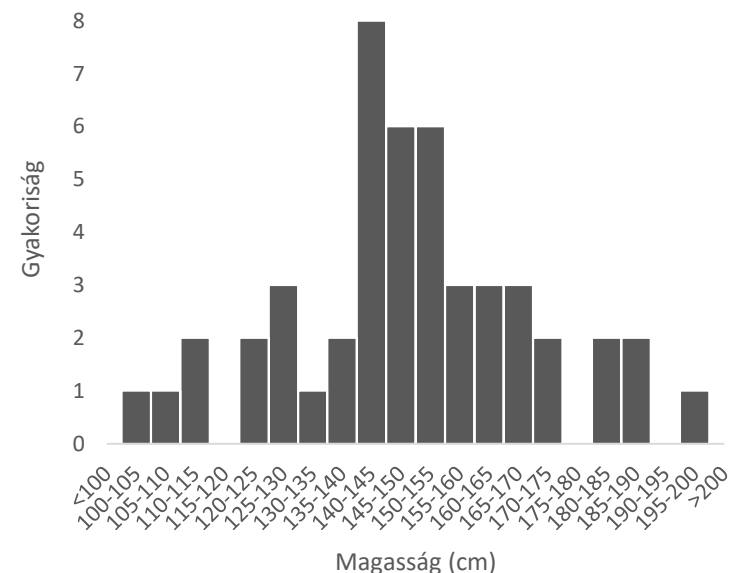
- A skála diszkrét értékeinek gyakorisága
- Az egyes értékek előfordulásának száma
- Például fagyik népszerűsége



Fagyi íz	CS	V	K	C	E
Gyakoriság	34	22	40	10	24

- Folytonos változó esetében

- A skálát egyforma széles intervallumokra bontva egy adott intervallumon belül eső értékek száma



Hisztogram lépésközének jelentősége

A hisztogram ábrázolásánál mindig fontos az optimális lépésköz (**bin**, részletesség) megtalálása.

