

Többváltozós statisztika elméleti alapjai (BMNPS07500M)

Készítette: Soltész-Várhelyi Klára

Statisztikai alapfogalmak

1. Leíró statisztikák

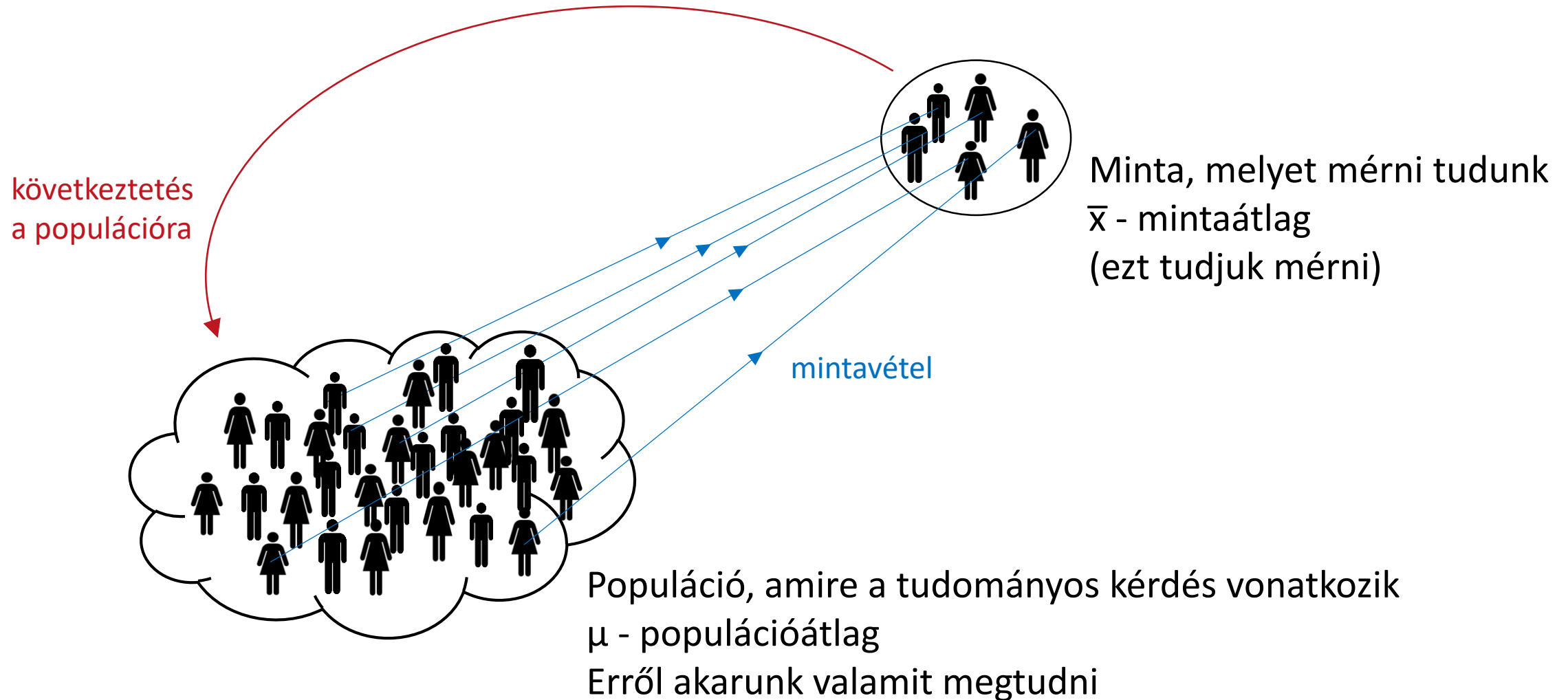
Populáció és minta

Miért csak valószínűségekről beszélhetünk?

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

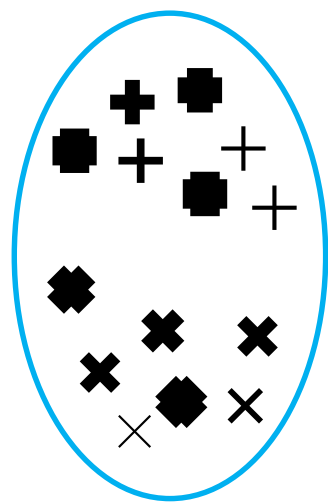
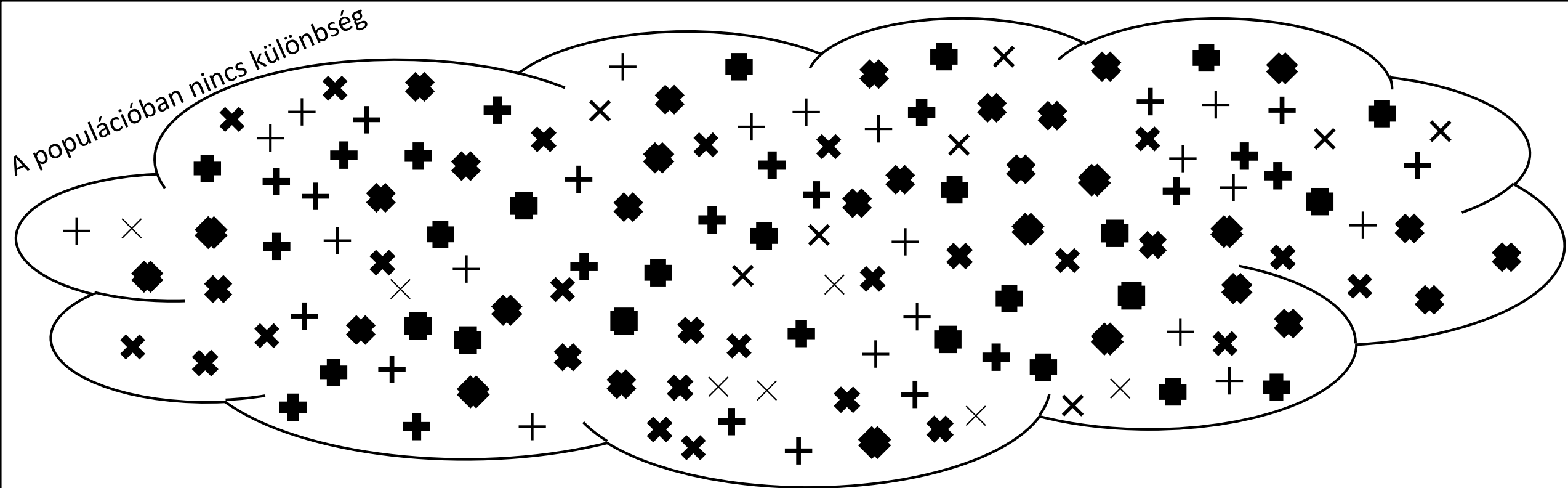
W. Edwards Deming

Populáció és minta

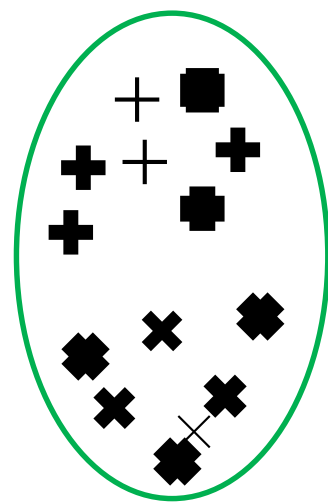


(az ábrát még kiegészítjük néhány diával később, ne ez alapján tanuld meg!)

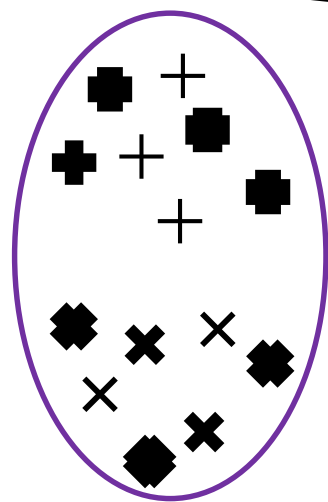
A populációban nincs különbség



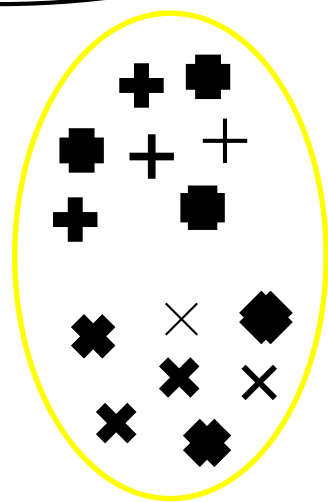
Nem találtunk
különbséget



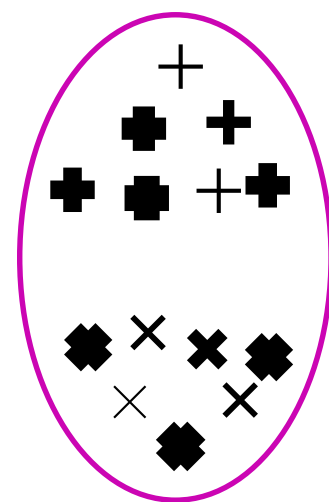
Nem találtunk
különbséget



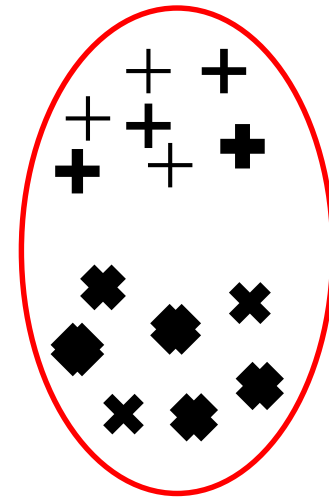
Nem találtunk
különbséget



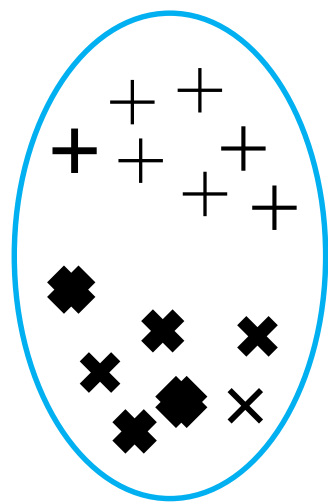
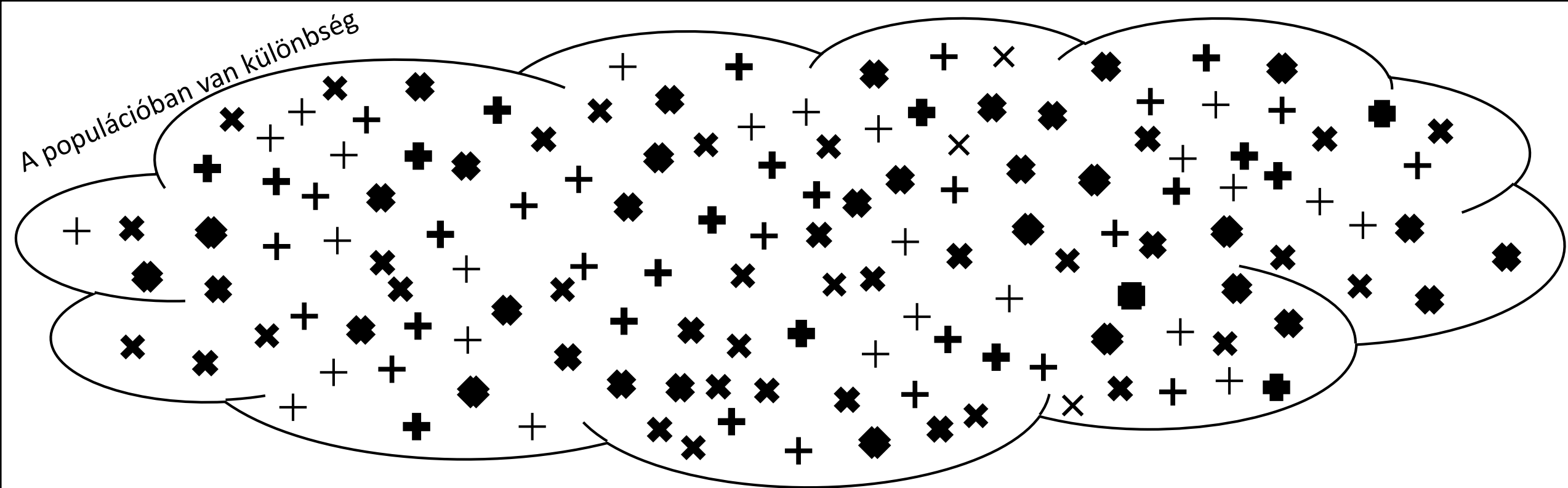
Nem találtunk
különbséget



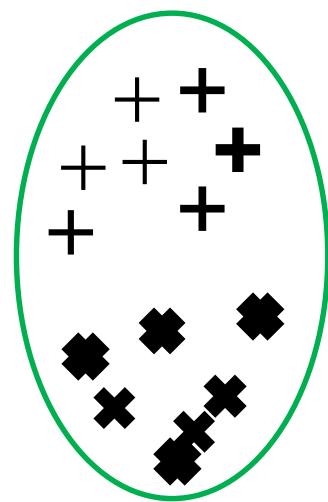
Nem találtunk
különbséget



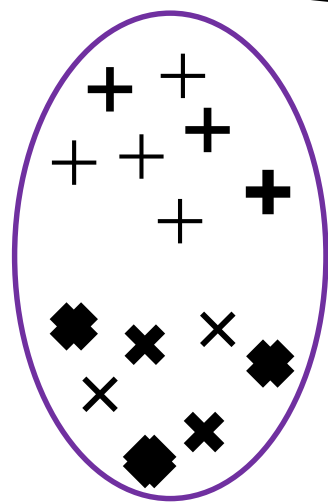
Különbséget találtunk,
és tévedtünk



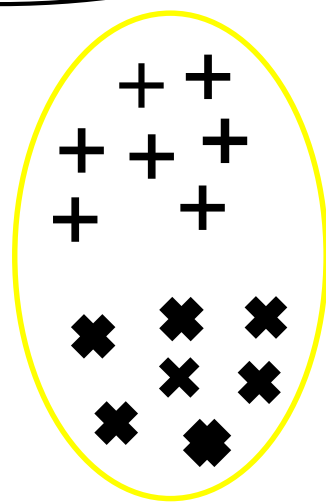
Kimutatjuk a különbséget



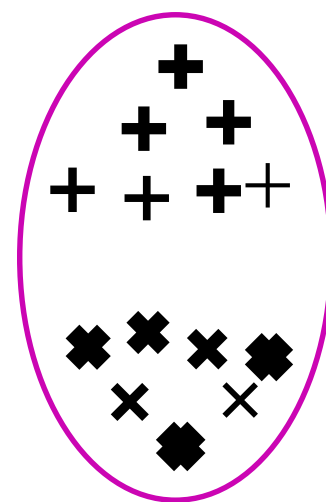
Kimutatjuk a különbséget



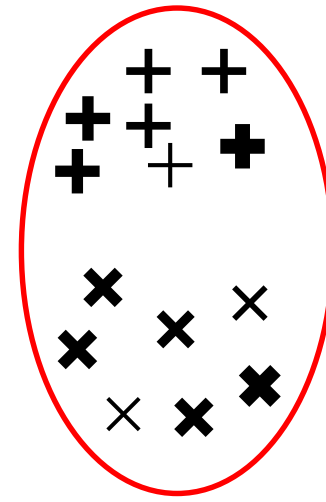
Kimutatjuk a különbséget



Kimutatjuk a különbséget

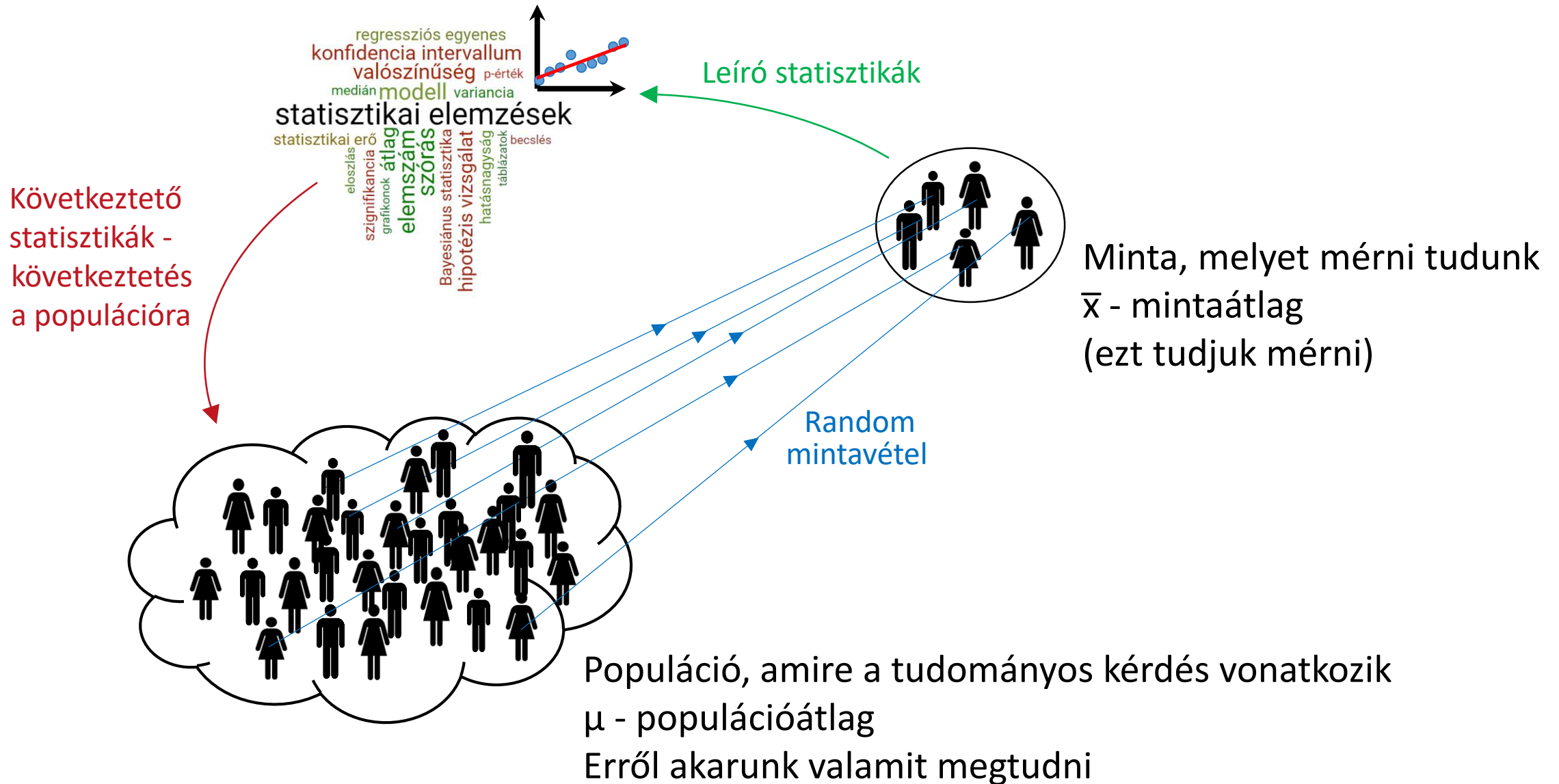


Kimutatjuk a különbséget?



Nem találunk különbséget, és tévedtünk

Populáció és minta

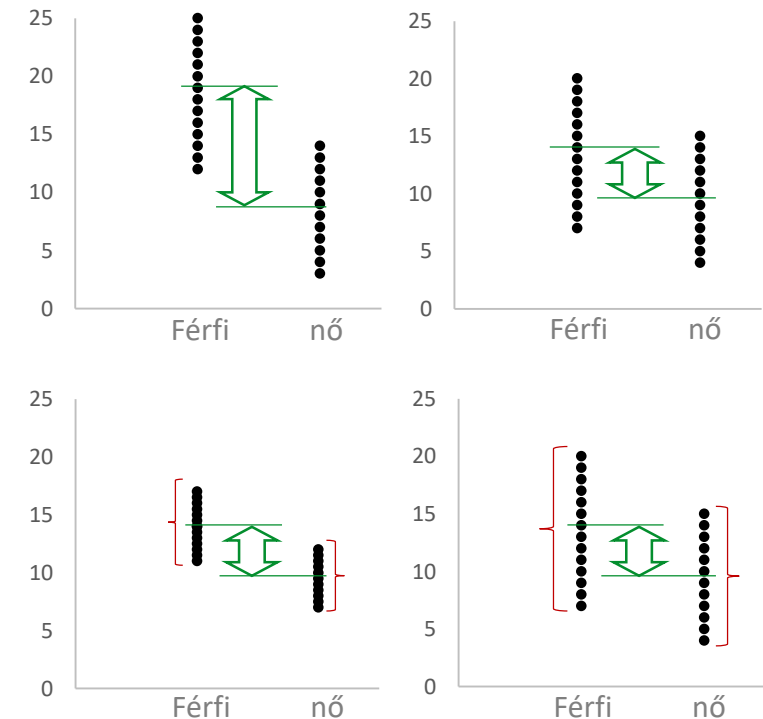
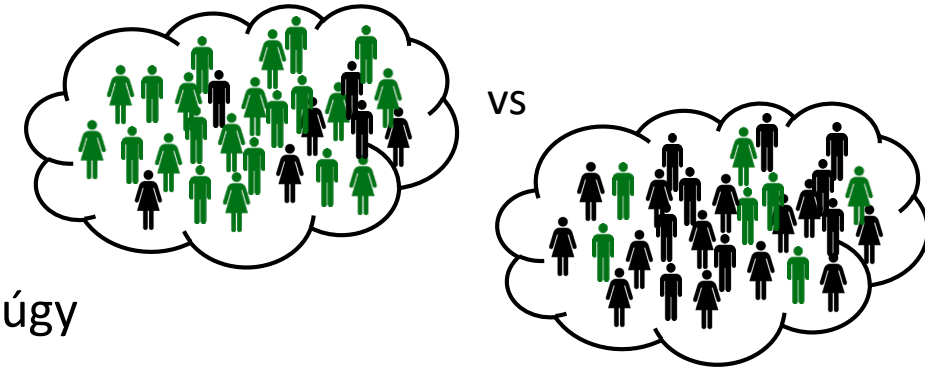


- Fontos, hogy a minta jól reprezentálja a populációt.
 - A) **Random mintavételezés**
 - A minta kiválasztásakor törekedni kell arra, hogy a **populáció minden tagjának egyforma esélye legyen a mintába való kerülésre.**
 - A random mintavételezés szinte **soha nem valósul meg tökéletesen**
 - *(brazil nő a mintámban)*
 - Törekedni kell a populáció arányos reprezentálására **a mérést meghatározó tényezők mentén!**
 - Törekedni kell, hogy a mintavételezéssel ne vigyünk a mérésbe **szisztematikus torzítást!**
 - *(férfiak és nők matematika tudása, nők pszichológia, férfiak a informatika szakról)*
 - Korlátozott mintavételezés esetén **át kell gondolni, valójában mi az a populáció, melyre következtethetünk!**
 - Magyarok, tinik, egyetemisták
 - *(férfiak magasabb empátia értéke – pszichológia szakos minta)*

- Fontos, hogy a minta jól reprezentálja a populációt.
- B) Statisztikai elemzések
 - *Igaz-e, hogy jelen van a hatás?*
 - Annak megállapítására, hogy a **mintában kapott hatásokat mennyire hihetjük el a populációban is létezőnek.**
 - Mennyire valószínű, hogy a mintában látott hatások (különbségek, összefüggések) a populációban is jelen vannak.
 - Pontosabban fogalmazva éppen megfordítva történik: **p érték annak valószínűsége** lesz, hogy a **mintában látott hatás** csak a **véletlen mintavételezésből** adódik, és a **populációban nincs jelen.**
 - *Igaz-e, hogy nincs jelen a hatás?*
 - Klasszikus hipotézisbecsléssel nem alkalmas annak eldöntésére, hogy a populációban nincs jelen a hatás vagy csak rosszul sikerült a mérés (később részletesen beszélünk róla).
 - Bayesiánus tesztelés
 - Hatásnagyság elemzés

Populáció és minta

- **Mi befolyásolja, mennyire hihetjük el a mintában kapott hatásokat a populációban is létezőnek?**
 - Sok minden. Ebből egyelőre hármat emelünk ki:
 - **Minta elemszáma**
 - evidens, minél jobban lefedi a minta a populációt, annál valószínűbb, hogy pontosan reprezentálja azt
 - Annál kisebb a valószínűsége, hogy csak véletlenül állt úgy össze a minta, mintha lenne hatás benne
 - **A hatás nagysága**
 - minél nagyobb a hatás, annál kevésbé valószínű, hogy a véletlen mintavételezés eredménye.
 - **A zaj/hiba nagysága**
 - minél kisebb a zaj (pl. a csoporton belüli különbségek), annál jobban el tud különülni a két minta,
 - annál valószínűbb, hogy a mintában talált hatás nem a véletlen mintavételezés eredménye.



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)
- **Folytonosság szerint:**
 - **Diszkrét** – az ilyen változó csak bizonyos pontokban vehet fel értéket, a mérőeszköznek van legnagyobb felbontási egysége (gyerekek száma)
 - **Folytonos** – minden ponton értelmezhető (legalább egy intervallumon belül), a skála tetszőleges finomsággal felbontható (magasság)
- **Jokerek:**
 - **Dichotóm / bináris** – csak kétféle értéket vehet fel
 - **Dummy** változók – egy kategoriális változó több dichotómra bontva

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 |

Leíró és következtető statisztika

- **Leíró statisztika** (descriptive stats)
 - Célja a mért adathalmaz jellemzése, az rendelkezésre álló információ tömörítése
 - A mért mintát írja le, következtetéseket nem tartalmaz
 - Fogalmak:
 - Elemszám, minimum, maximum
 - Elhelyezkedési mutatók: súlyozott és súlyozatlan átlag, medián, kvartilisek, percentilisek, módusz
 - Szóródási mutatók: terjedelem, variancia, szórás, relatív szórás
 - Minta eloszlása: ferdeség és csúcsosság (de ezt később beszéljük meg)
 - Bár már a következtető statisztikához tartozna, de itt beszélünk róla: standard error, relatív standard error, konfidencia intervallum
 - Bár ide tartozna, de majd csak a következtető statisztikánál beszélünk róla: hatásnagyság (effect-size)
- **Következtető statisztika** (inferential stats)
 - Célja a populációból választott mintából a populációra való (vissza)következtetés
 - Valószínűségekkel dolgozik
- A kettő között nincs éles határ

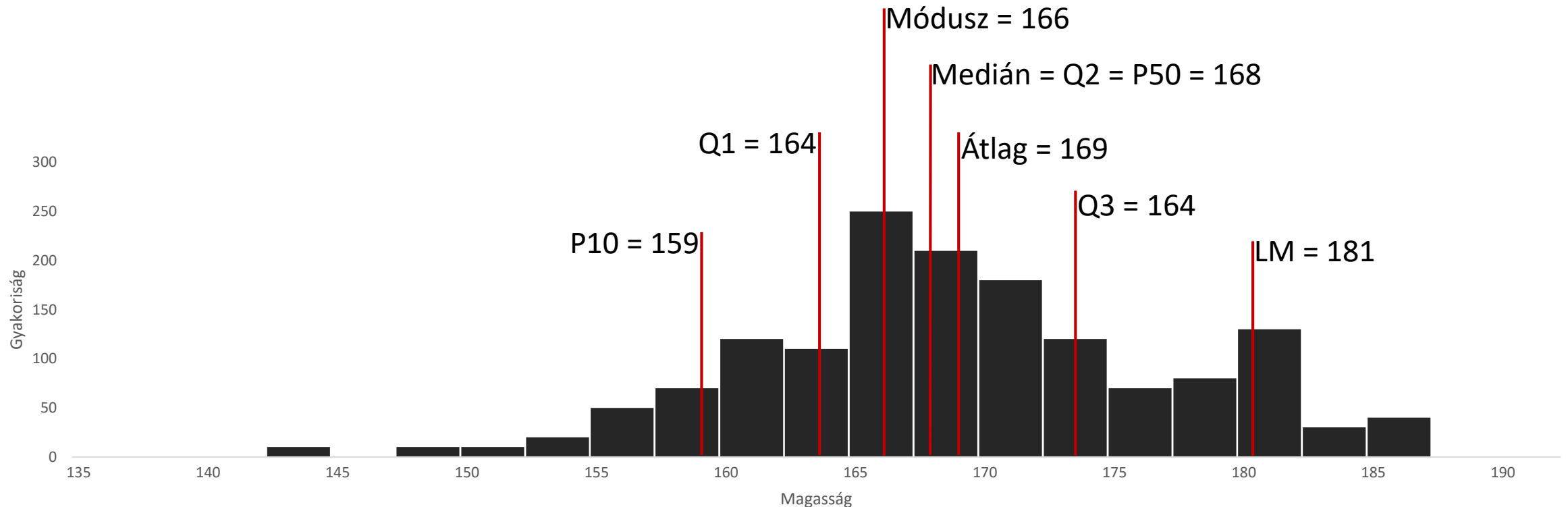
Elemzés, minimum, maximum

- 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

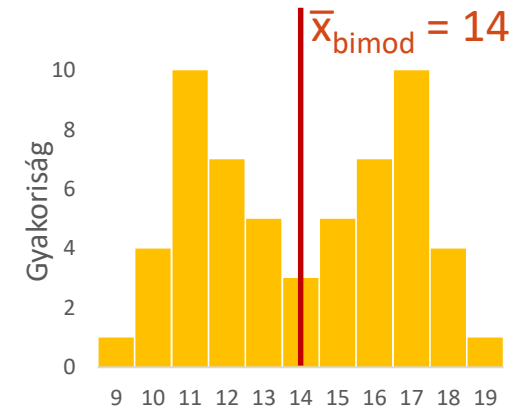
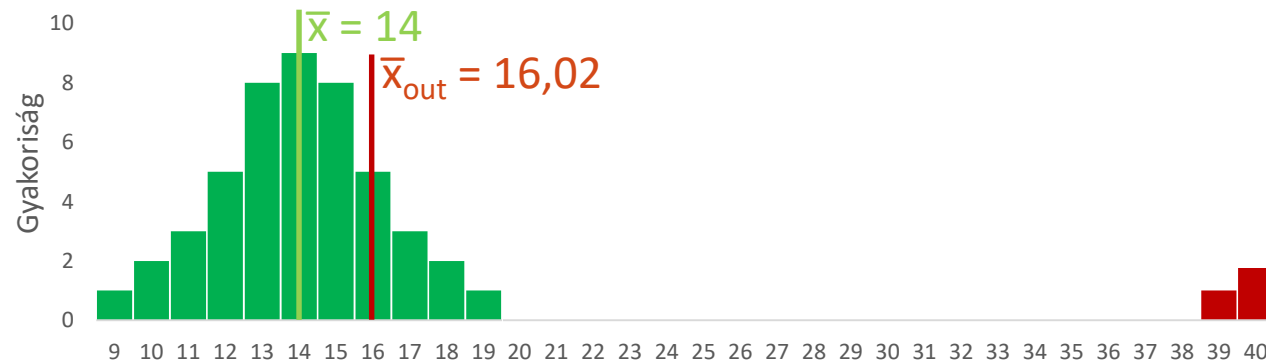
- A minta leírása egyetlen reprezentatív értékkel
 - Tanult mutatók:
 - Számtani átlag, súlyozott számtani átlaga, súlyozatlan számtani átlag (Estimated Marginal Means), medián, kvartilisek, percentilisek, módusz, lokális maximum



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, unimodális (egycsúcsú) eloszlások esetén.



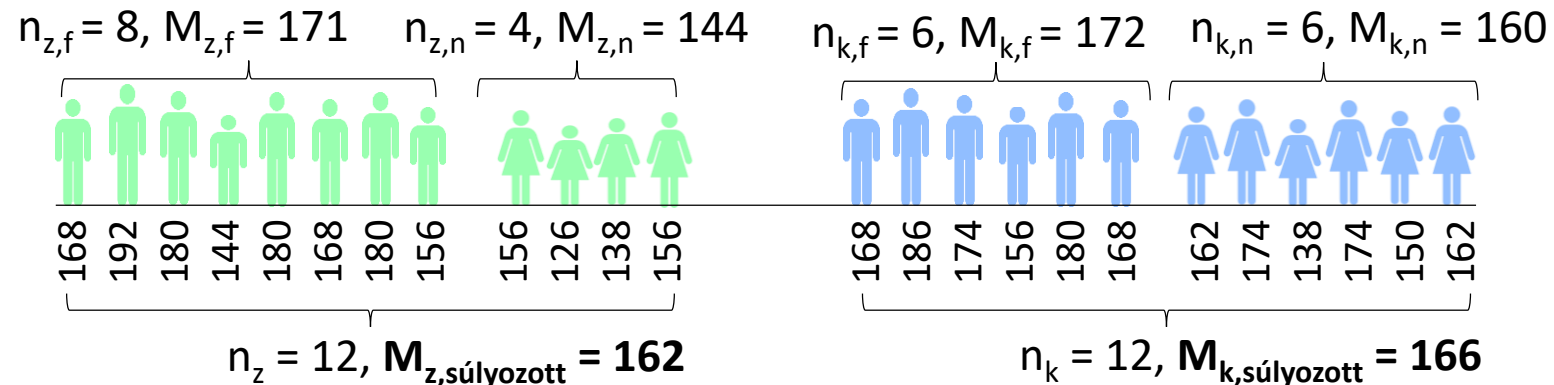
Elhelyezkedési mutatók: Súlyozott és súlyozatlan átlag

- Több szempontos vizsgálatban kétféleképp lehet egy szempont egy szintjének átlagát kiszámolni:
- Az elemszámmal súlyozottan:**
 - Figyelembe veszi, hogy egy szempontban a másik szempont szerinti mintáknak eltérő az elemszáma (és feltételezi, hogy ez a populációban is fennáll): **a nagyobb minta nagyobb súllyal kerül az átlagba**
 - Például a zöld átlag: a zöld férfi és női csoportok átlagának átlaga a két csoport elemszámával súlyozva

$$M_{\text{zöld}} = \frac{n_{z,f} * M_{z,f} + n_{z,n} * M_{z,n}}{n_{z,f} + n_{z,n}} = \frac{8 * 171 + 4 * 144}{8 + 4} = 162$$

- Az előzővel megegyező eredményt ad az összes zöld ember átlaga függetlenül attól, hogy nő vagy férfi

$$M_{\text{zöld}} = \frac{\sum x_{z,i}}{n_z} = \frac{168 + 192 + 180 + 168 + 180 + 156 + 126 + 138 + 156}{12} = 162$$

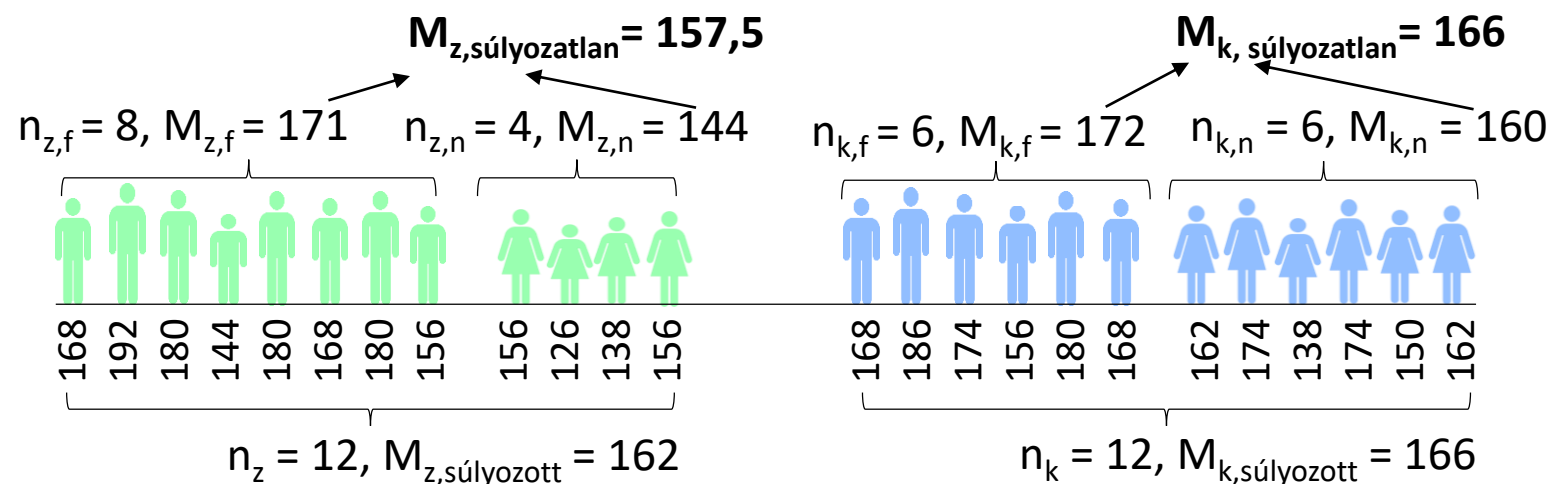


Elhelyezkedési mutatók: Súlyozott és súlyozatlan átlag

- Több szempontos vizsgálatban kétféleképp lehet egy szempont egy szintjének átlagát kiszámolni:
- Súlyozatlanul:**
 - Egy szempont átlagának kiszámolásakor **a másik szempont csoportjainak elemszámbeli különbségeit nem veszi figyelembe**. Nem veszi figyelembe a másik szempont szerinti különbségeket.

$$M_{zöld} = \frac{M_{z,f} + M_{z,n}}{2} = \frac{171 + 144}{2} = 157,5$$

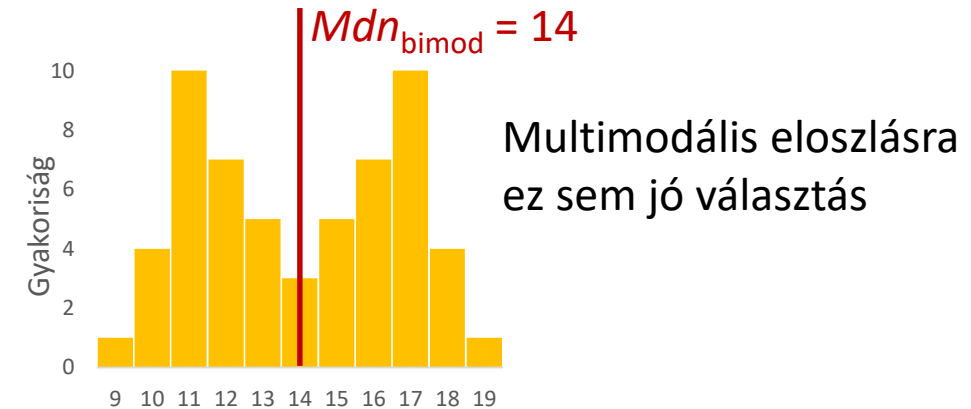
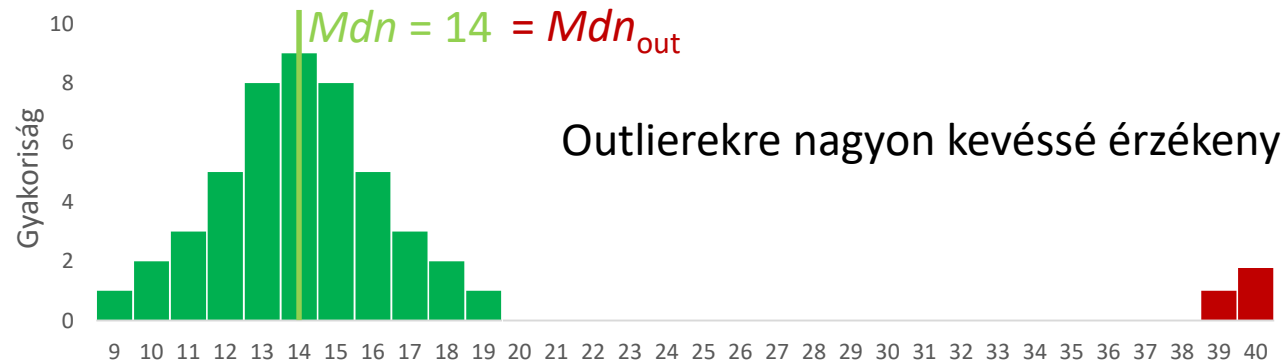
- Az elemzések során **Estimated Marginal Means**-ként találkozunk a súlyozatlan átlaggal.
- Ez történik pl. az ANOVÁban (bár a modell paraméterei között beállítható más módszer is)
- Azért előnyös, ha egyenlő elemszámmal dolgozol, mert akkor a súlyozott és súlyozatlan átlag megegyezik (lásd kék emberek).



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



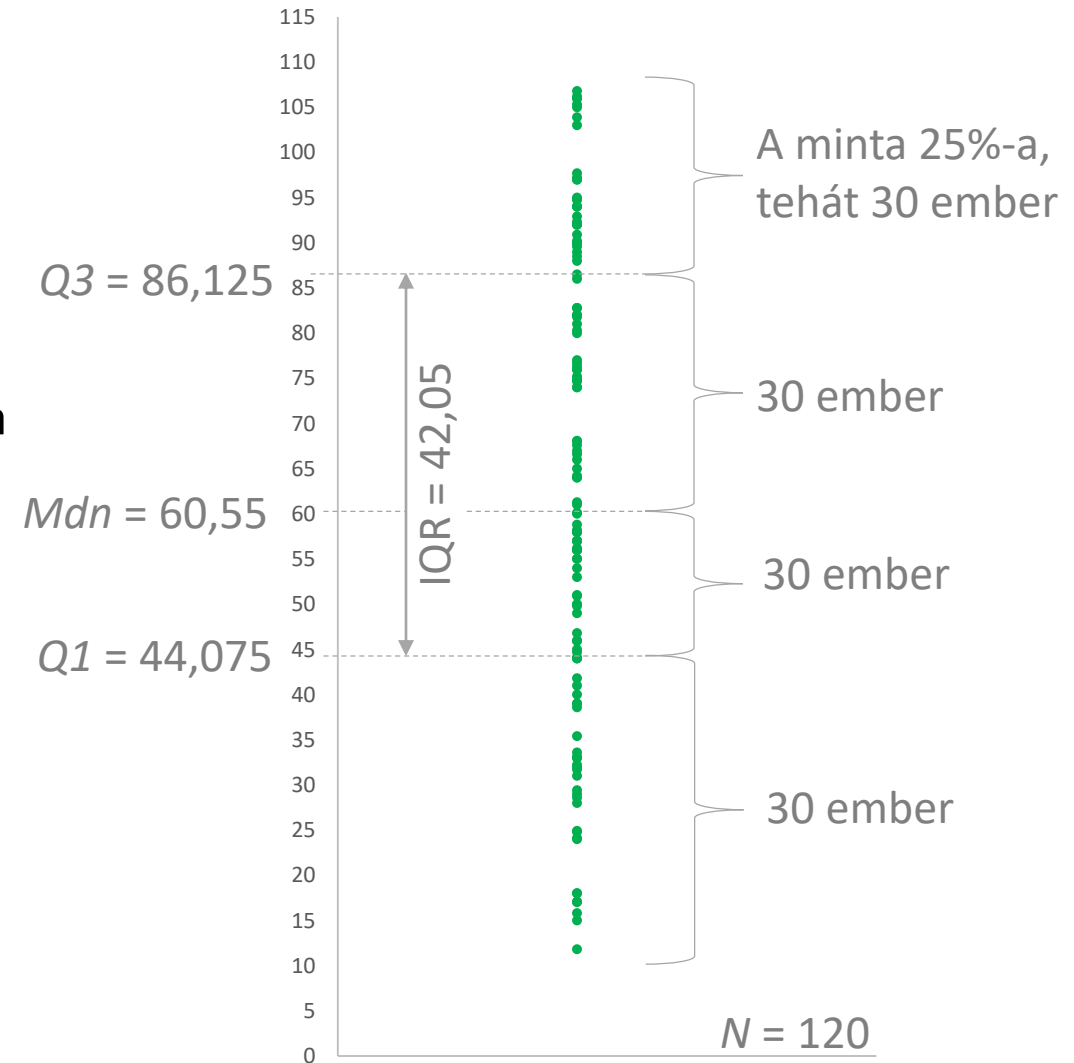
Elhelyezkedési mutatók: Kvartilisek, percentilisek

- **Kvartilisek**

- Negyedelő pontok
- **Q1** – első kvartilis (alsó negyedelő pont): melynél a minta maximum 25%-a vesz fel alacsonyabb, 75%-a magasabb értéket
- **Q2** – második kvartilis (felező pont): medián
- **Q3** – harmadik kvartilis (felső negyedelő pont): melynél a minta maximum 75%-a vesz fel alacsonyabb, és 25%-a magasabb értéket.
- **Interkvartilis terjedelem** – *IQR* (interquartile range)
Q3 és Q1 távolsága, a középső 50% szélessége
(valójában inkább szóródási mutató)

- **Percentilisek**

- Századoló pontok.
- **Pl 95%-os percentilis:** a minta 95%-a alacsonyabb, 5%-a magasabb értéket vesz fel.



Elhelyezkedési mutatók: Módusz

- **Módusz**

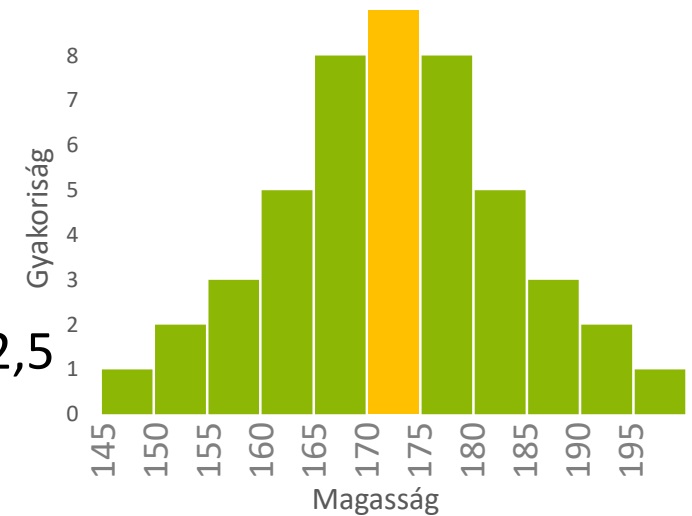
- A legdivatosabb, **leggyakrabban előforduló érték**.
- Jelölése: **Mo**

- Számolása:

- Diszkrét adatok esetén: a leggyakrabban előforduló érték. Példa: 2, **3**, 1, **3**, 4 -> $Mo = 3$
- Folytonos változók esetén: a folytonos változót értelmes méretű intervallumokra bontjuk, és megkeressük a leggyakrabban előforduló intervallumot. A módusz az intervallum alsó és felső határának átlaga lesz.

- Példa: Magasság esetén az intervallum legyen 5cm!

Leggyakoribb a 170-175-ös intervallum, $Mo = \frac{170+175}{2} = 172,5$



- 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.

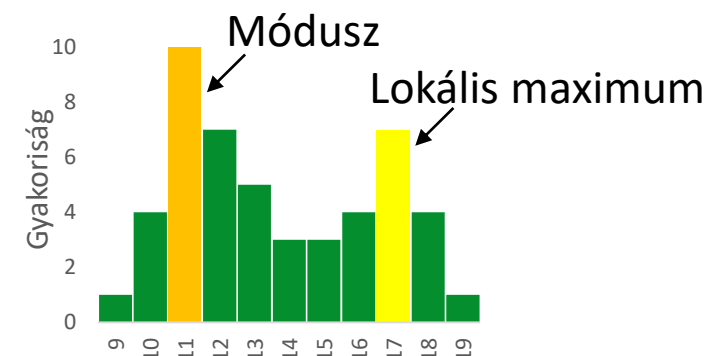
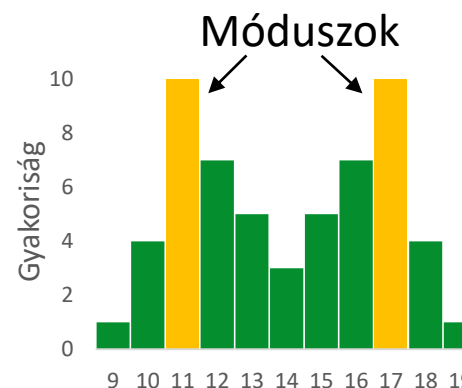
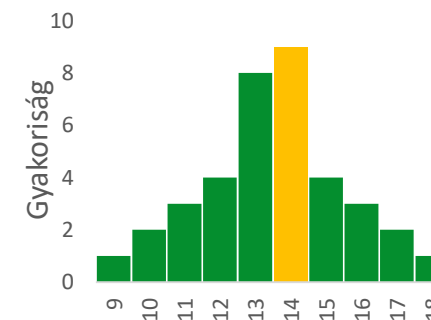
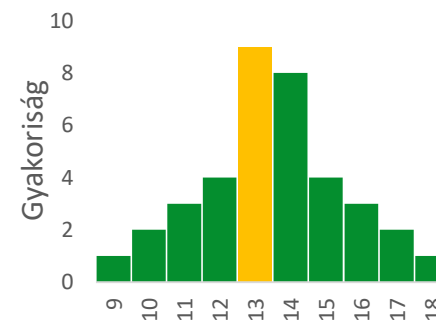
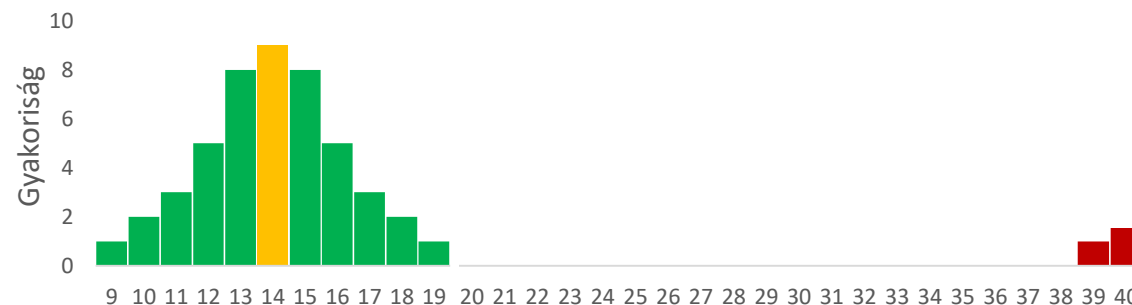
Elhelyezkedési mutatók: Módusz

- **Módusz jellegzetességei:**

- Outlierekre nem érzékeny

- Annál inkább a kicsi különbségek esetlegességére, és folytonos változók esetén az intervallum szélességére.

- Értelmezhetőek az **unimodális** eloszlás mellett **multimodális** (pl. bimodális, trimodális) eloszlások is. Sőt, sokszor a **lokális maximumok** is móduszként közlik.

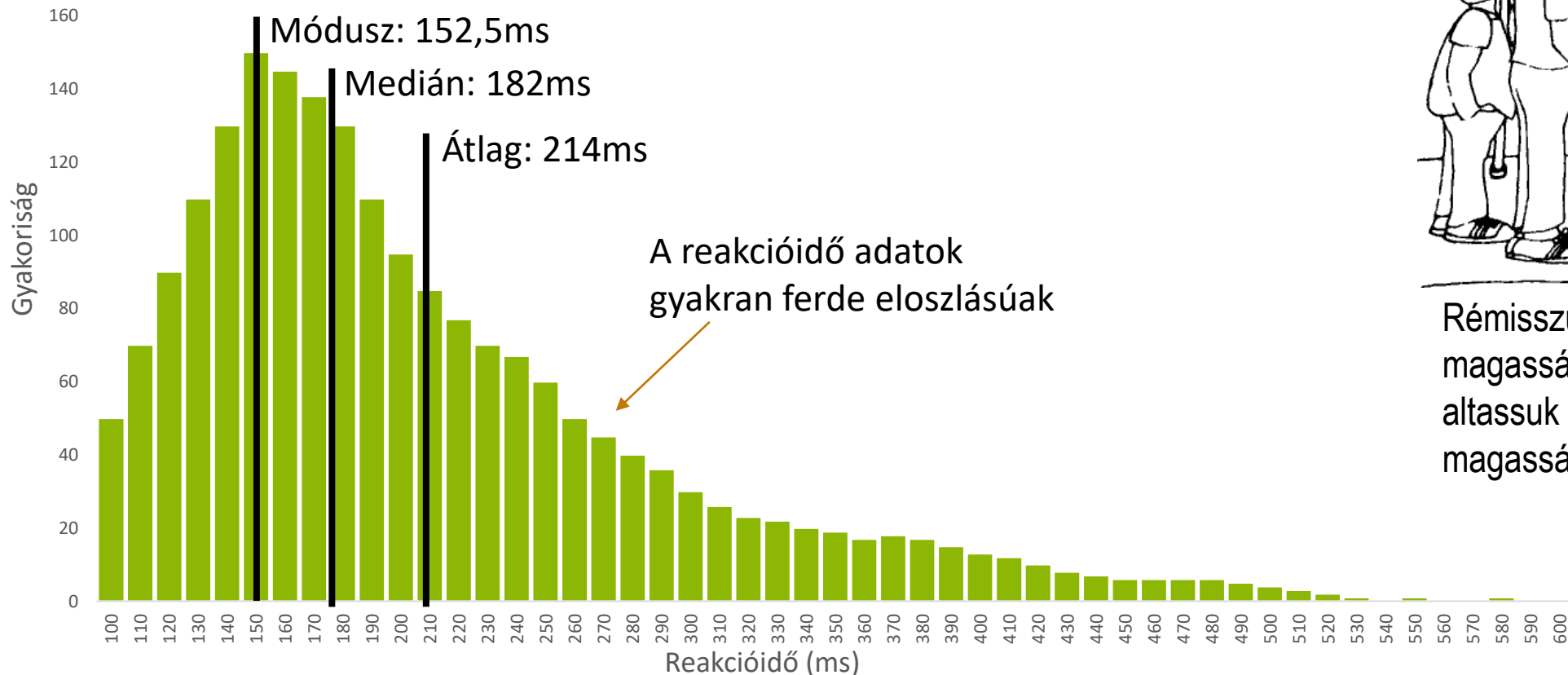


Elhelyezkedési mutatók: Átlag vs Medián vs Módusz

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

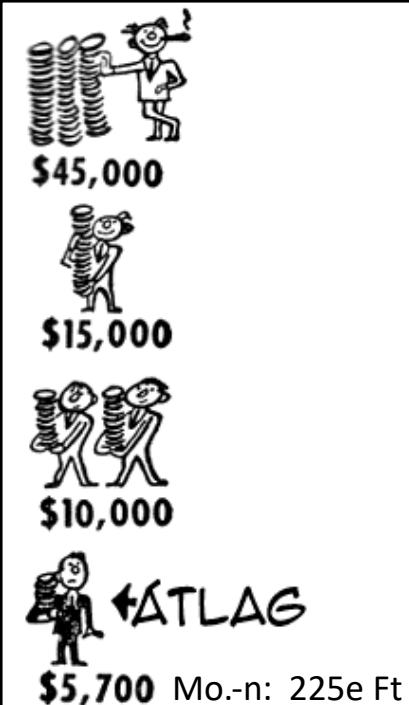
Medián: ferde eloszlású vagy szélsőséges értékekkel rendelkező adatokon a legjobb, például reakcióidőadatokról.

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?

Elhelyezkedési mutatók: Átlag vs Medián vs Módusz

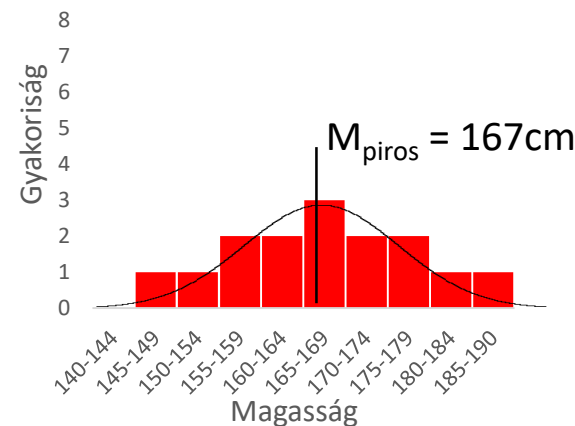
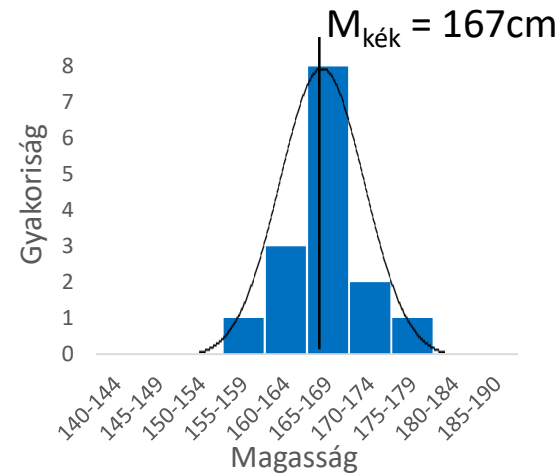
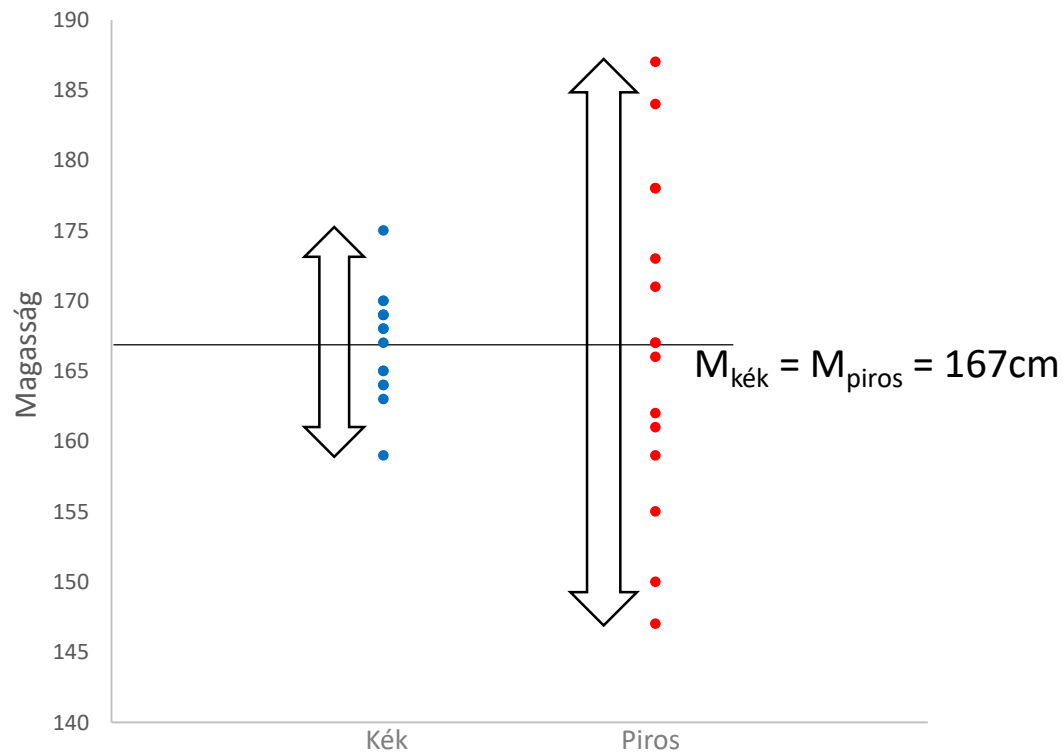


- A kérdéstől függ, mi a megfelelő középérték mutató.
- A világ nagyon más aspektusát ragadják meg.



Szóródási mutatók

- Miért nem elég az átlag?
 - A minta szóródását nem jellemzi
 - Tanult szóródási mutatók: **terjedelem**, **átlagtól való (négyzetes) eltérés**, **variancia**, **szórás**, **relatív szórás**, **standard error**, **relatív standard error**, **konfidencia intervallum**

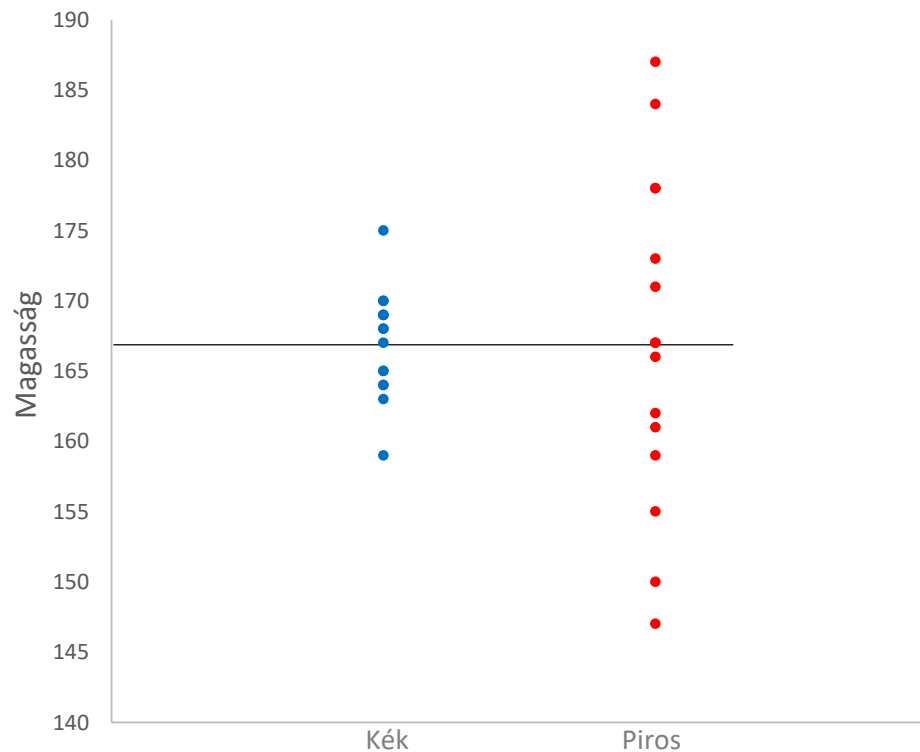


Kékek	Pirosak
163	155
169	178
168	187
169	166
170	178
168	159
165	171
164	150
164	167
170	147
159	184
175	162
165	167
169	161
167	173

Szóródási mutatók: Terjedelem

- **Terjedelem**

- A legkisebb és legnagyobb érték távolsága
- Jelölése: **range**
- Számítása: Max-Min Példában: $\text{Range}_{\text{kék}} = 175 - 159 = 16$ és $\text{Range}_{\text{piros}} = 187 - 147 = 40$
- APA szabvány szerint értéke helyett a minimumot és maximumot adjuk meg: $\text{range}_{\text{kék}} = 159 - 175$



Kékek

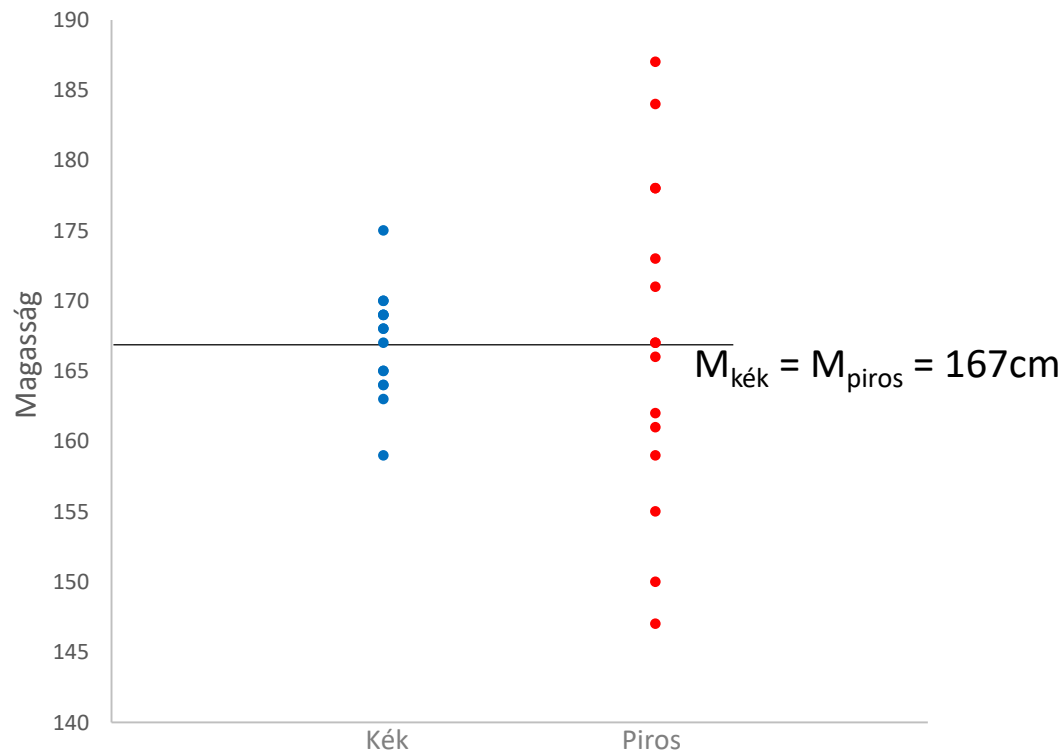
163
169
168
169
170
168
165
164
164
170
159
175
165
169
167

Pirosak

155
178
187
166
178
159
171
150
167
147
184
162
167
161
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: x_{di} vagy D_i Számolása: $D_i = x_i - \bar{x}$
- **Átlagtól való négyzetes eltérés (D^2 , vagy angol **squared deviances**)**
 - 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**

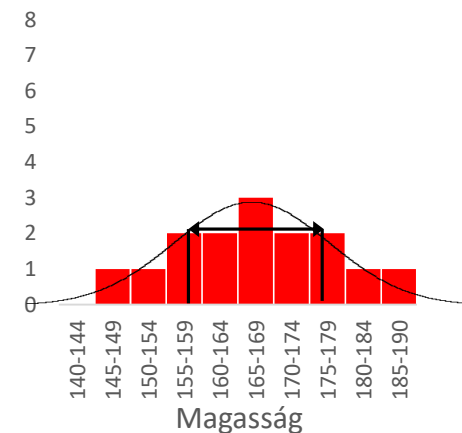
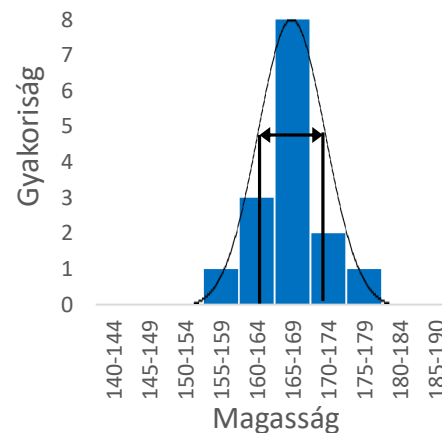
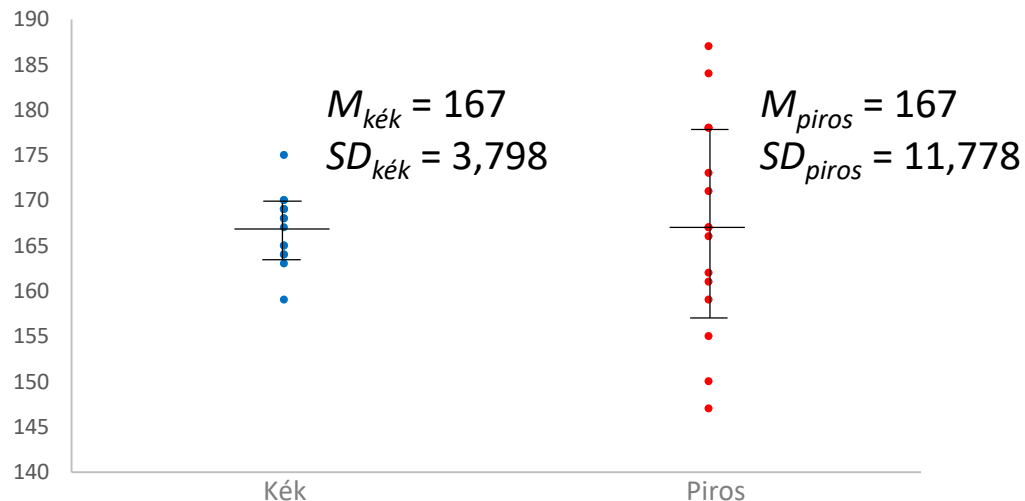
- Számolása: $Var = \frac{\sum D_i^2}{df} = \frac{\sum (x_i - \bar{x})^2}{n-1}$ Pl.: $Var_{kék} = \frac{16+4+1+4+9+1+4+9+9+9+64+64+4+4+0}{15-1} = 14,429$
Szabadságfok (következő dián)

- **Szórás**

- Átlagtól való átlagos eltérés

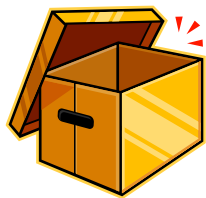
- Jelölése: **SD** (standard deviation), **s** vagy **σ**

- 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

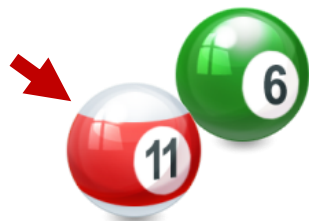
1.



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: $df = N - 1$
- Matematikai értelme (Bessel korrekció):
 - a szórás számolásakor felhasználom a tagok mellett az átlagot is, mely nem független a tagoktól, így tagok száma = N , összefüggések száma = 1, szabadságfok $N - 1$
- Gyakorlati haszna:
 - A mért értékek a mintaátlaghoz közelebb lesznek, mint a populációátlaghoz, pedig mi a populációátlagot akarjuk becsülni.
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.



Variancia és szórás: akkor miért van wikin más képlet?

Statisztika:

$$\text{Variancia: } s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Valószínűségszámítás:

$$\text{Variancia: } s^2 = E(X^2) - E^2(X)$$

Máshogy néz ki, mégis ugyanaz. Lássuk, hogyan számolódnak!

A megértéshez kell az átlag és várható érték fogalma, számolása. (Példa a dobókockával való dobálás)

$$\begin{array}{c} \text{Átlag: } \frac{\sum x_i}{n} \\ \bar{x} = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5 \end{array}$$

$$\begin{array}{c} \text{Várható érték: } E(X) \\ E(X) = \frac{1}{6} * 1 + \frac{1}{6} * 2 + \frac{1}{6} * 3 + \frac{1}{6} * 4 + \frac{1}{6} * 5 + \frac{1}{6} * 6 = 3.5 \end{array}$$

1. Írjuk fel a mi képletünket a valószínűség számítás eszközeivel:

1.1. Vegyük az átlagot! $E(X)$

1.2. Vonjuk ki minden X értékből az átlagot: $X - E(X)$

1.3. Emeljük négyzetre a különbségeket: $(X - E(X))^2$

1.4. Átlagoljuk a különbségek négyzetét: $E((X - E(X))^2)$



Variancia és szórás: akkor miért van wikin más képlet?

Statisztika:

$$\text{Variancia: } s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Valószínűesszámítás:

$$\text{Variancia: } s^2 = E(X^2) + E^2(X)$$

Máshogy néz ki, mégis ugyanaz. Lássuk, hogyan számolódnak!

A megértéshez kell az átlag és várható érték fogalma, számolása. (Példa a dobókockával való dobálás)

$$\begin{array}{c} \text{Átlag: } \frac{\sum x_i}{n} \\ \bar{x} = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5 \end{array}$$

$$\begin{array}{c} \text{Várható érték: } E(X) \\ E(X) = \frac{1}{6} * 1 + \frac{1}{6} * 2 + \frac{1}{6} * 3 + \frac{1}{6} * 4 + \frac{1}{6} * 5 + \frac{1}{6} * 6 = 3.5 \end{array}$$

2. Rendezzük az egyenletet!

2.1. Bontsuk fel a négyzetre emelést: $E(X^2 - 2 * X * E(X) + (E(X))^2)$

2.2. Bontsuk fel a külső átlagolást és rendezzük:

KÉK: a) hozzuk ki a kettes konstans, hiszen egy konstans átlaga önmaga. b) Az átlag átlaga az átlag, ezért az $E(E(X)) = E(X)$

LILA: Az átlag négyzetének átlaga megegyezik az átlag átlagának négyzetével, ami az átlag négyzete, ezért a lila: $E^2(X)$

2.3. Vonjuk össze a tagokat, és megkaptuk a valószínűség számítás képletét 😊

$$\begin{array}{l} E(X^2) - \underbrace{E(2 * X * E(X))}_{2 * E(X) * E(E(X)) = 2 * E^2(X)} + \underbrace{E((E(X))^2)}_{E((E(X))^2) = (E(X))^2 = E^2(X)} = E(X^2) - 2 * E^2(X) + E^2(X) = E(X^2) - E^2(X) \end{array}$$

Mekkora a nagy szórás? Relatív szórás fogalma

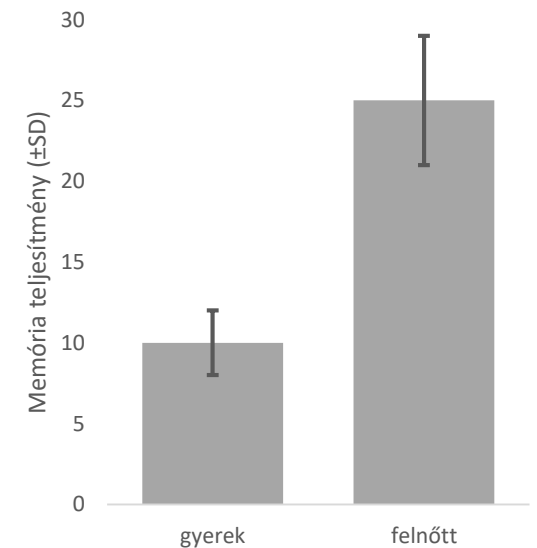
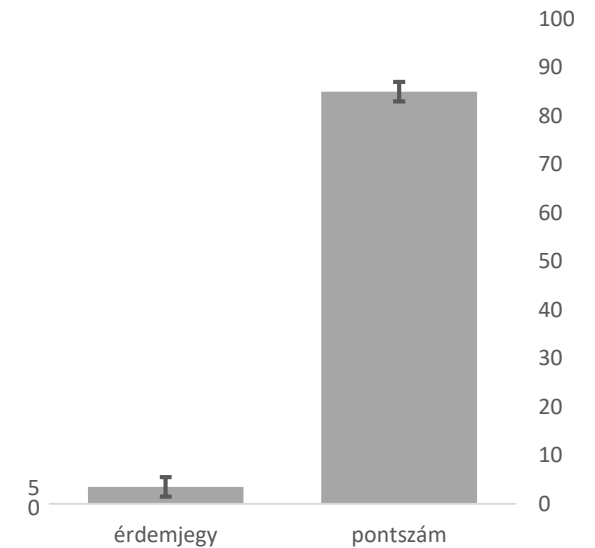
- Mekkora a „nagy” szórás?
 - Nincs kritérium érték, mert függ a skálázástól!
 - Két egységnyi szórás érdemjegy esetén (1-5-ig terjedő skála) vagy zh pontszám szerint (0-100-ig terjedő skála) mást jelent

- **Relatív szórás** vagy **variációs együttható**

- Jelölése: **RSD** (relative standard deviation) vagy **CV**(coefficient of variation)
- Számolása: $RSD = s / \bar{x}$ százalékban kifejezve
- Mire használjuk?
Két eltérő átlagú mintán mért adatok precizitásának összehasonlítására.

CSAK ARÁNYSKÁLA ESETÉN HASZNÁLHATÓ!!

- Gyerek minta: átlag: 10 ; szórás: 2
- Felnőtt minta: átlag: 25 ; szórás: 4
- $RSD_{\text{gyerek}} = 20\%$ míg $RSD_{\text{felnőtt}} = 16\%$ csak

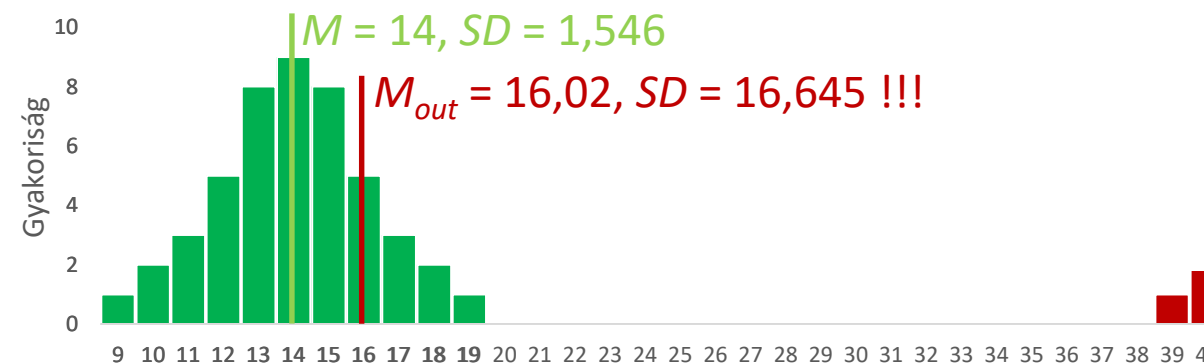
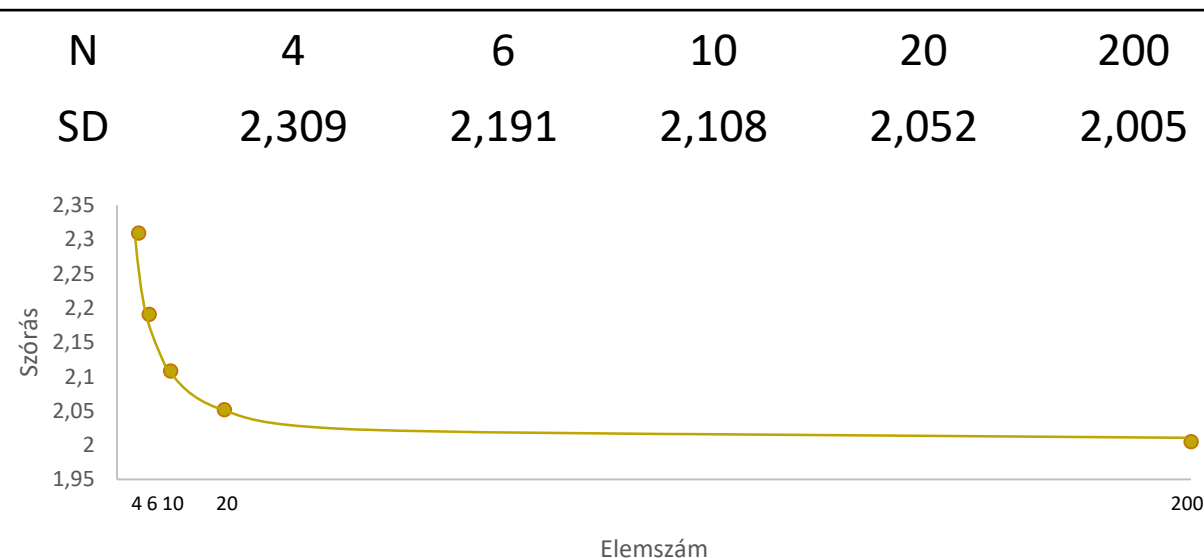


Miért nagy a szórás?

- Miért nagy a szórás?
 - **Kicsi az elemszám**
 - A szabadságfokkal való korrekció okozza.
 - $SD = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$
 - A korrekciónak csak kis elemszám esetén van jelentős hatása, nagy elemszám esetén a 1 eltörpül az N-hez képest.
 - Előnyös, hogy függ a szabadságfoktól, mivel minél nagyobb a minta, annál valószínűbb, hogy pontosan reprezentálja a populációt.
 - **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**

Példa elemszám hatására:

Minta, ami csak 1-eseket és 5-ösöket tartalmaz. Átlag: 3



Szóródási mutatók: Standard error

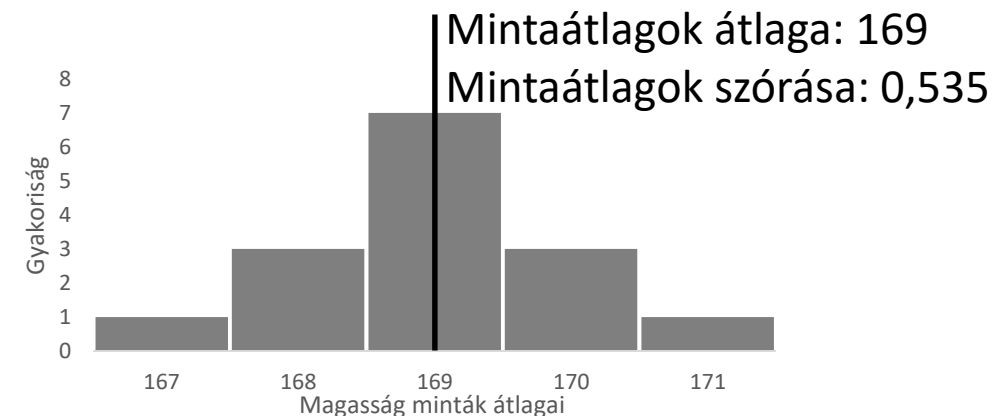
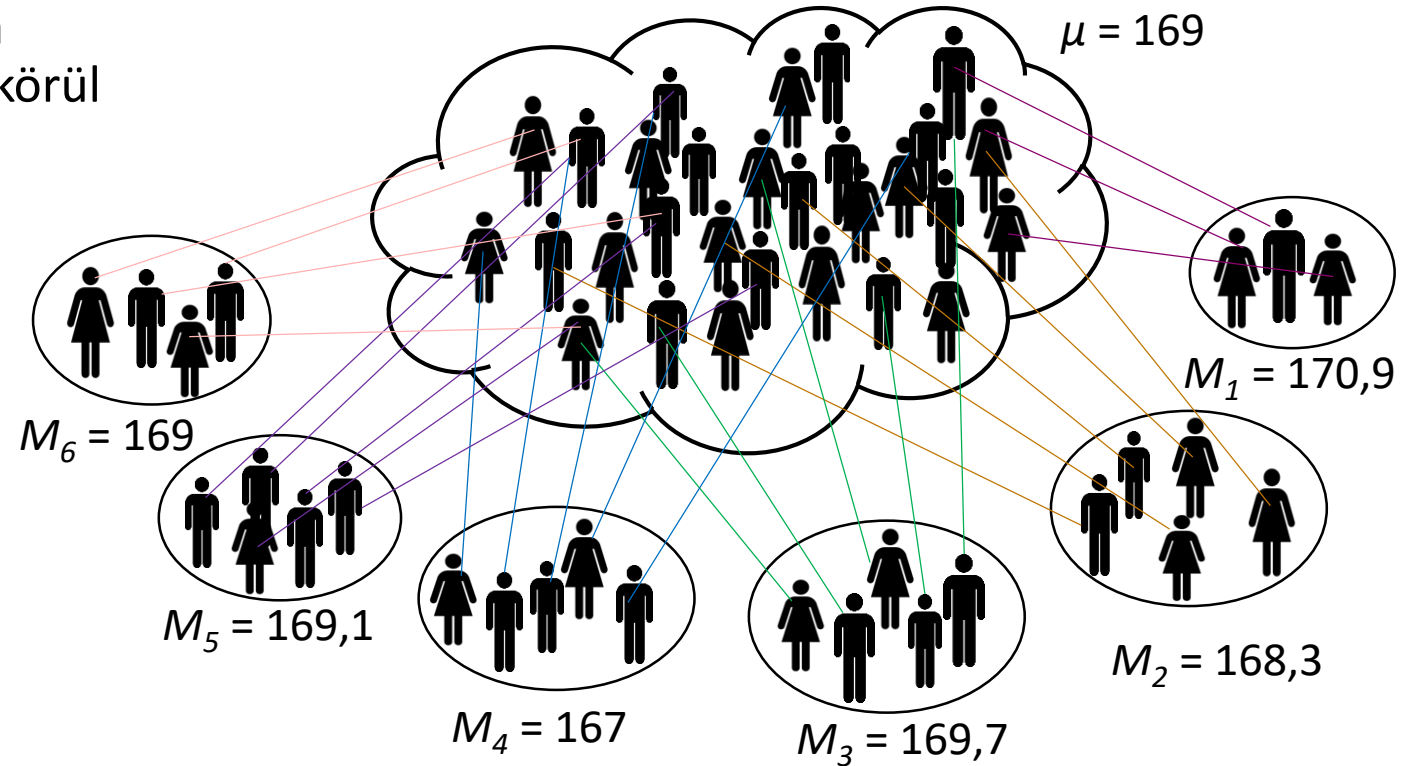
- Ha sok mintát vennénk a populációból, akkor a mintaátlagok ($\bar{x}_i$) valahol a populációátlag (μ) körül lennének, de nem lennének azonosak, minden mintának kicsit máshol lenne az átlaga

- Standard error:**

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

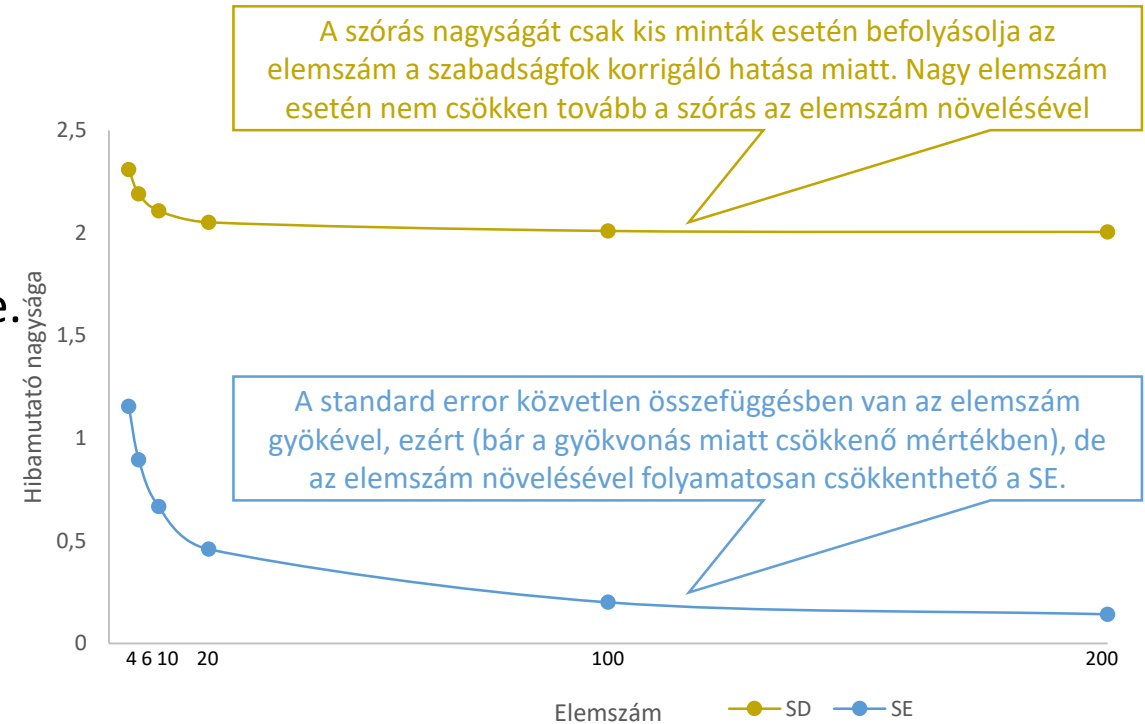
$$SE = \frac{s}{\sqrt{N}}$$

- 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)

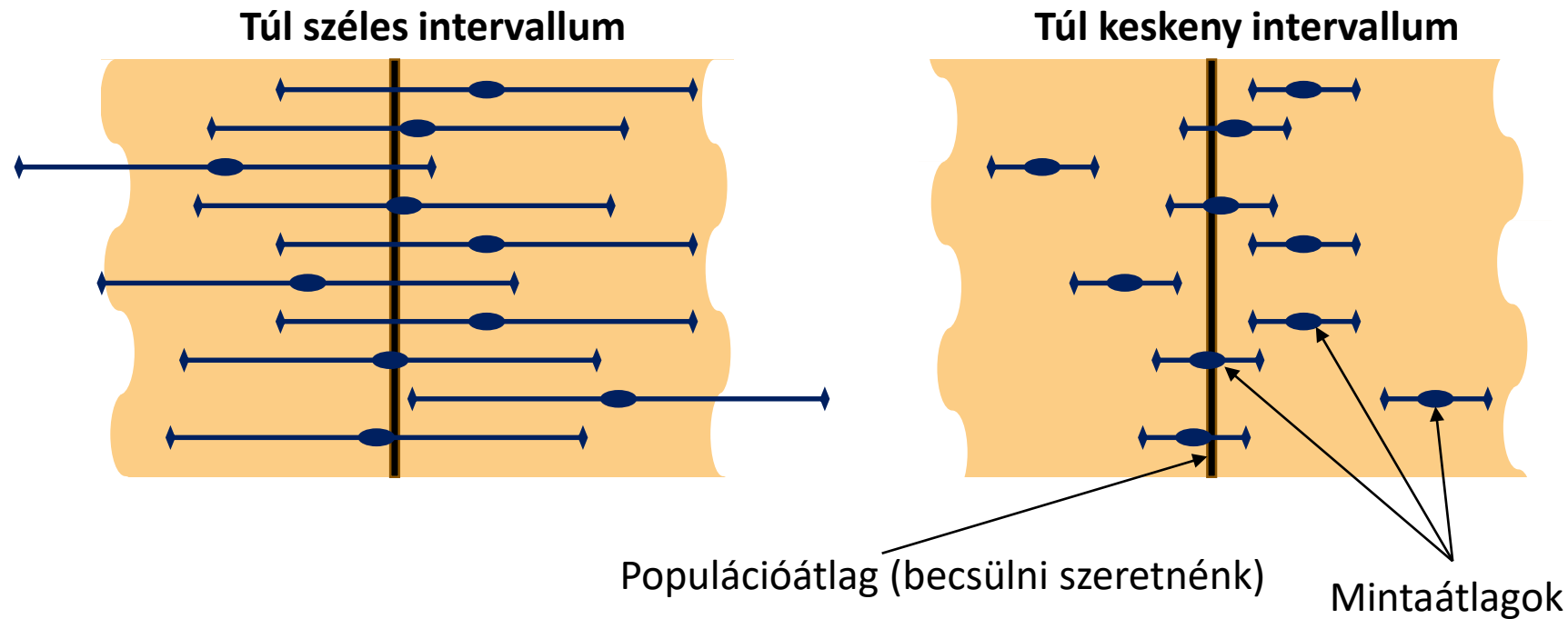


Szóródási mutatók: Standard error

- Mi számít nagy standard errornak?
 - A szóráshoz hasonlóan ez is skálafüggő.
- **Relatív standard error (RSE):**
 - Számítása: $RSE = SE / \bar{x}$ százalékban kifejezve.
 - A relatív szóráshoz hasonlóan csak arányskála esetén használható!!
 - 25% alatt elfogadható a mérés
- Mitől függ a standard error értéke?
 - $SE = \frac{s}{\sqrt{N}}$
 - Szórás (így mindentől, ami a szórást befolyásolja) – minél nagyobb a szórás, annál nagyobb lesz a SE értéke
 - Elemszám - Minél nagyobb az elemszám, annál kisebb lesz az SE értéke
 - A szórással ellentétben az SE értéke folyamatosan csökken az elemszám növelésével, hiszen minél nagyobb mintákat választunk a populációból, annál kisebb a véletlen mintavételezés hatása, tehát várhatóan annál hasonlóbbak lesznek a mintaátlagok.



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.

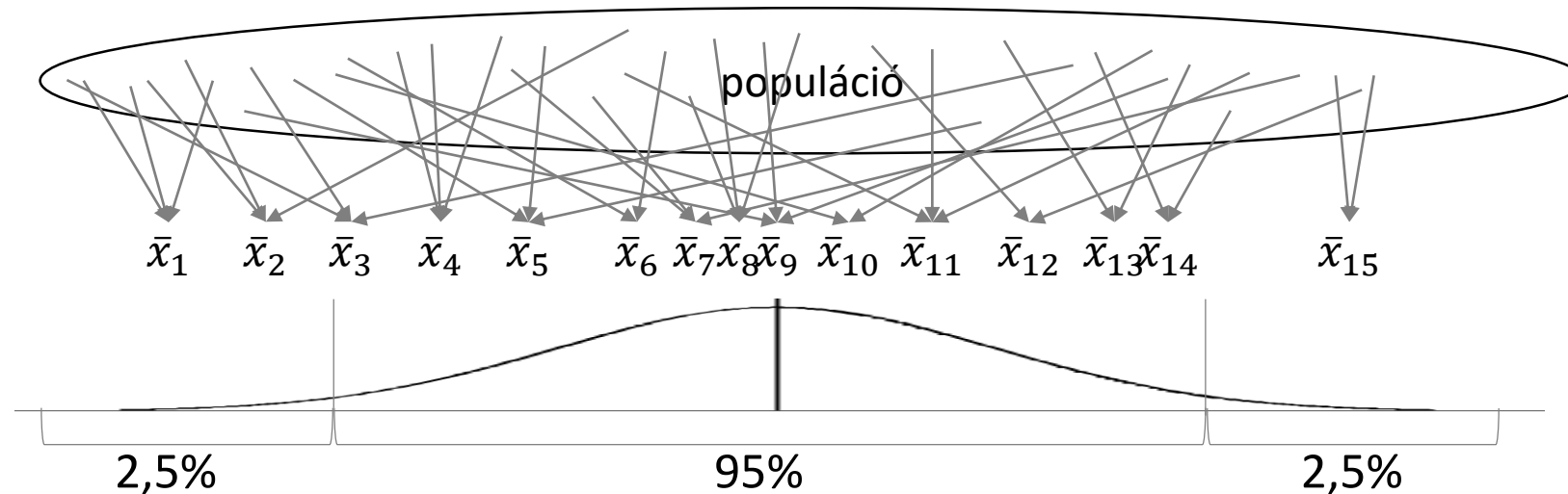
Kellene egy olyan határ, amibe nagy (pl. 90% vagy 95%-os) valószínűséggel beleesik a populációátlag.

A CI már valószínűségekkel számoló mutató, ezért valójában nem a leíró statisztika része.

Szóródási mutatók: Konfidencia intervallum / megbízhatóság

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

1. Vegyünk a populációból sok mintát, és számoljuk ki az átlagaikat! Azt látjuk, hogy a mintaátlagok normál eloszlást követnek.
2. 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.**

Szóródási mutatók: Konfidencia intervallum / megbízhatóság

Hogyan számoljuk ki egyetlen mért minta esetén?

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

Ebből egyenletrendezéssel:

Az átlag átlaga az átlag marad, tehát ez nem változik

± 1.96 (95%-os bizonyosság esetén)

A határérték, amit keresünk

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

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

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

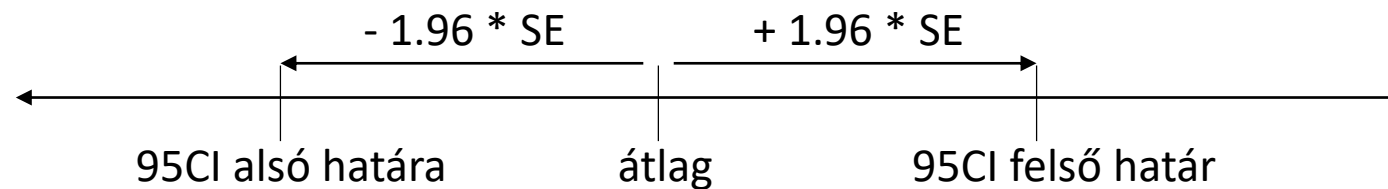
A 95%-os konfidencia intervallum felső határa: átlag + 1.96 * SE

A 95%-os konfidencia intervallum alsó határa: átlag (- 1.96) * SE

Szóródási mutatók: Konfidencia intervallum / megbízhatóság

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 errorrt: $X = \bar{X} + z * \frac{s}{\sqrt{N}}$



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: nevezőben 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

