stathelp.hu

Készítette: Soltész-Várhelyi Klára

2. Leíró statisztika

Leíró statisztikák

Mi *van* a mintában?

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

A mért változók

- **Változó definíciója**:
  - Statisztikában a minta egyedeinek kutatás szempontjából érdekes tulajdonságát megmérjük, és a változóban tároljuk.
  - A változó értéke egyedről egyedre változhat
  - Lehet konkrét számadat (például magasság) vagy kategória-tagság jelölése számmal (például hajszín)

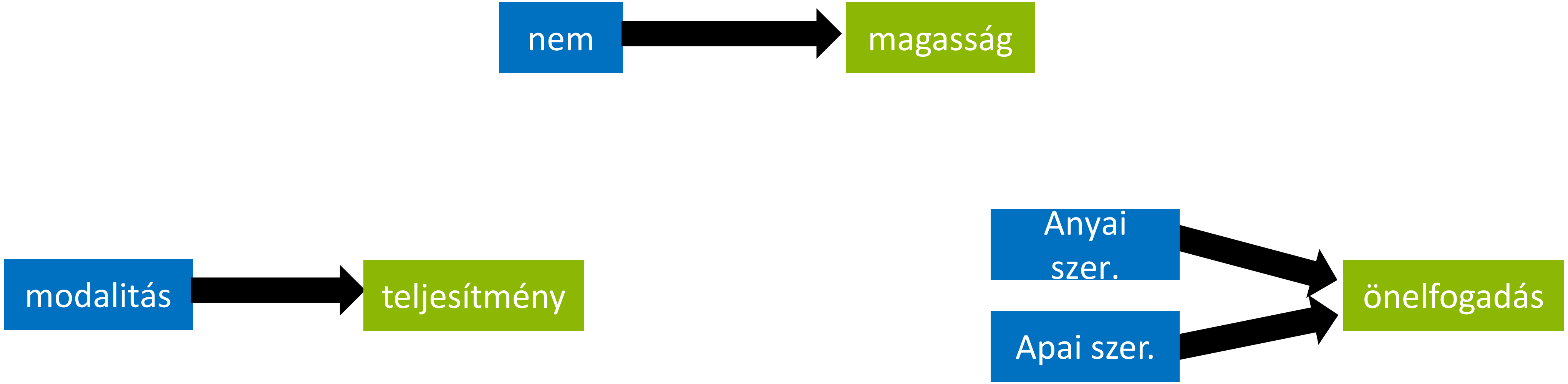
Azonosító	**Vérnyomás**	**Nem**	**Dohányzás**	**Alkohol**	**Egyetem**
VSZ1	90/60	férfi	nem	Soha	Bölcsész
VSZ2	95/78	nő	igen	Hetente	Mérnök
VSZ3	117/80	nő	nem	Havonta	Mérnök
VSZ4	125/73	nő	nem	Hetente	Mérnök
VSZ5	80/70	férfi	igen	Soha	Gazdasági
VSZ6	100/60	férfi	igen	Naponta	Gazdasági
VSZ7	101/50	férfi	nem	Hetente	Gazdasági
VSZ8	100/72	nő	igen	Havonta	Bölcsész
VSZ9	104/67	nő	igen	Naponta	Bölcsész
VSZ10	110/68	nő	nem	Naponta	Gazdasági
VSZ11	102/63	férfi	igen	Naponta	Bölcsész

Azonosito	**Szisztole**	**Diasztole**	**Nem**	**Dohanyzas**	**Alkohol**	**Egyetem**
VSZ1	90	60	1	0	1	1
VSZ2	95	78	2	1	3	2
VSZ3	117	80	2	0	2	2
VSZ4	125	73	2	0	3	2
VSZ5	80	70	1	1	1	3
VSZ6	100	60	1	1	4	3
VSZ7	101	50	1	0	3	3
VSZ8	100	72	2	1	2	1
VSZ9	104	67	2	1	4	1
VSZ10	110	68	2	0	4	3
VSZ11	102	63	1	1	4	1

- **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)

- **Mérési szint szerint**:
  - **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, versenyeredmény)
  - **Skála** típusú változó – az elemek sorba rendezhetőek, az elemek közötti különbség, arány is kifejezhető
    - **Intervallum skála** – nincs természetes nulla pont (Celsius-skála, fény hullámhossz, dátumok)
    - **Arányskála** – van természetes nulla pont (magasság, darabszám, Kelvin-skála, teljesítmény)
  - **Egyéb**:
    - **Dichotóm** / **bináris** – csak kétféle értéket vehet fel
    - **Dummy** / **indikátor** változók – egy kategoriális változó több dichotómra bontva
  - Egyéb fogalmak: skála típusú, legalább intervallum szintű, Likert-skála, kvázi intervallum típusú

- **Szerep szerint**:
  - **Függő** vagy **kimeneti változó**
    - amely változóra azt gondoljuk, hat a többi változó, amelynek az értékét be szeretnénk jósolni más változókkal.
    - *Például ha férfiak és nők magassága közötti különbséget vizsgáljuk, akkor a függő változó a magasság.*
    - A különböző statisztikai próbáknál különböző neve van, például t-próbánál függő változónak szokás nevezni, regresszióban kimeneti változónak

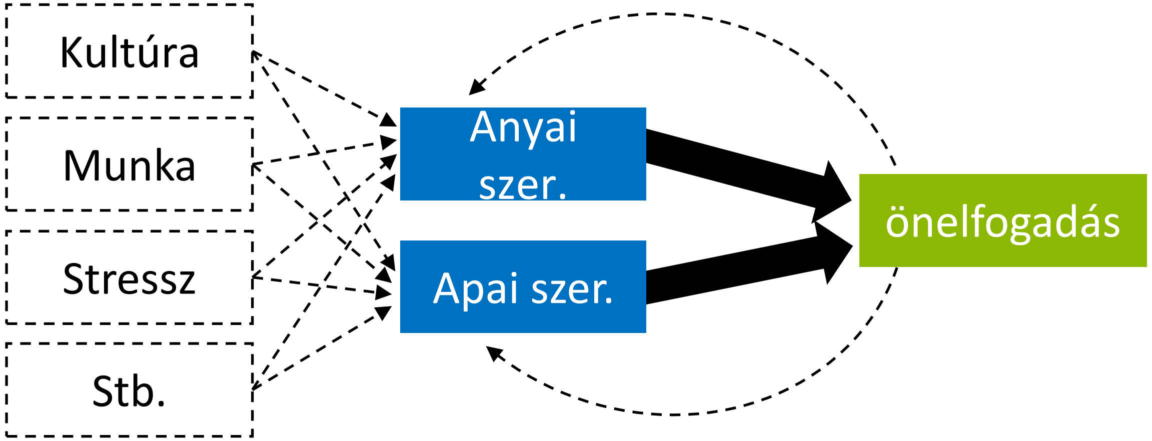

nem → magasság

modalitás → teljesítmény

Anyai szer. → önelfogadás
Apai szer. → önelfogadás

- **Szerep szerint** (folyt.):
  - **Független** (**prediktor**) **változó** – az a változó, melyről feltételezzük, hogy a hat a függő változóra, amelyből be szeretnénk jósolni a függő változót. *Például ha férfiak és nők magassága közötti különbséget vizsgáljuk, akkor a független változó a nem.*

    - Valódi **független változó** – melyre nem hat más változó. *Például a vizsgálatban egyik csoport hallotta, másik csoport olvasta a szólistát, és azt vizsgáljuk, melyik csoport emlékszik jobban.*
    - **Magyarázó** (**prediktor**) **változó** – melyre hat(hat) más változó (akár a függő is), de az adott elemzésben a prediktor értékéből szeretnénk a kimeneti értékét bejósolni. *Például regressziós modellben az apai és anyai szeretet mértékéből szeretnénk a gyerek önelfogadását bejósolni.*



- **Szerep szerint** (folyt.):
  - **Egyéb**
    - **Kontrol változó** – olyan változó, melynek a hatását stabilan szeretnénk tartani. *Például, ha a fény hatását vizsgáljuk a növény növekedésére, akkor a föld minőségét és a víz mennyiségét kontrol alatt kell tartani.*
    - **Csoportosító változó** – mely alapján csoportok hozhatók létre, általában független változónál használjuk. *Például, ha férfiak és nők magassága közötti különbséget vizsgáljuk, akkor a csoportosító változó a nem.*

Lásd még: valtozo_tipusok.pdf

Skálázás

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 előadás 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: átlag, medián, kvartilisek, percentilisek, módusz
    - Szóródási mutatók: terjedelem, variancia, szórás, relatív szórás, interkvartilis terjedelem
    - 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)
    - Grafikonok

- **Következtető statisztika** (inferential stats)
  - Célja a populációból választott mintából a populációra való (vissza)következtetés
  - Valószínűségekkel dolgozik

- A kettő között nincs éles határ

Elemszám, minimum, maximum

- Példa: statisztika tanárok barátai: 2, 3, 1, 3, 4

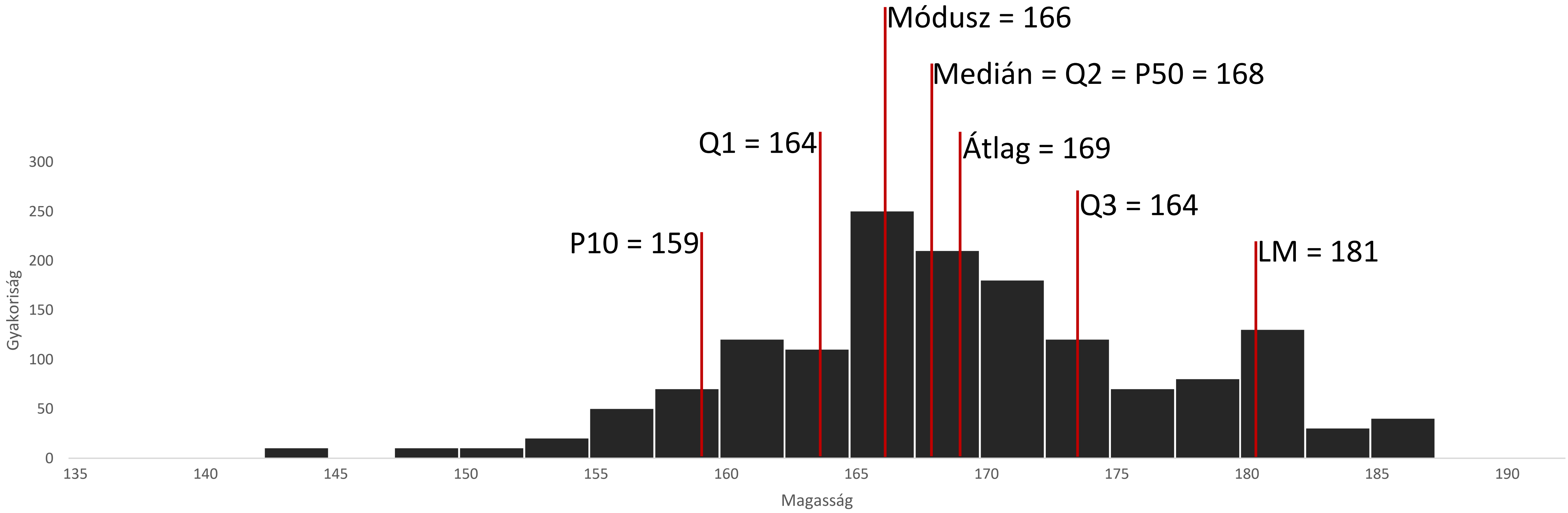
- **Elemszám**
  - Minta esetében jelölése ***n*** vagy ***N***
  - Példában: *N* = 5

- **Minimum**
  - Jelölése: ***Min***
  - Példában: *Min* = 1

- **Maximum**
  - Jelölése: ***Max***
  - Példában: *Max* = 4

Elhelyezkedési mutatók

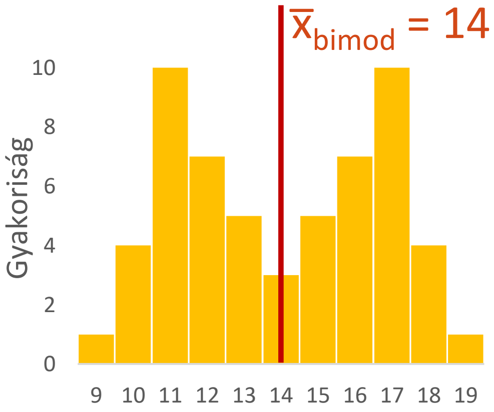
- A minta leírása egyetlen reprezentatív értékkel
  - Tanult mutatók:
    - **Számtani átlag**, **medián**, **kvartilisek**, **percentilisek**, **módusz**, **lokális maximum**



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 ismerjünk.
    - 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 (egy csúcsú) 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



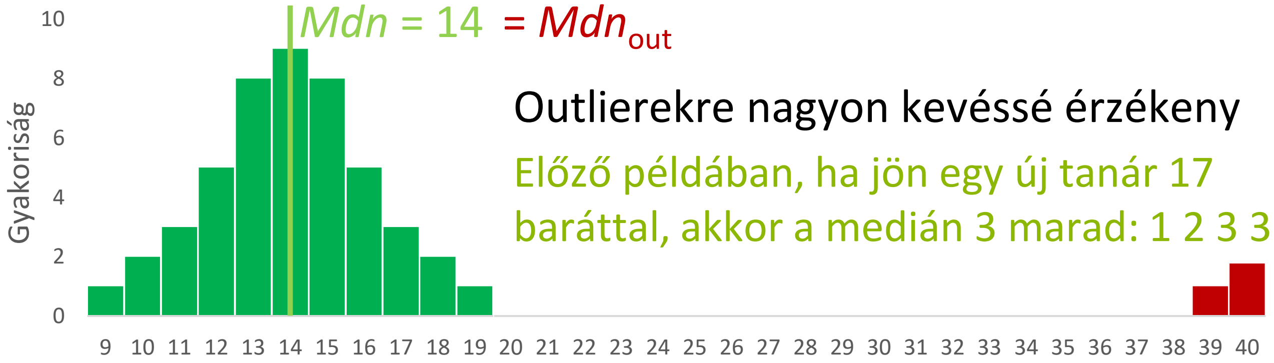
Elhelyezkedési mutatók: Medián

- **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
    - Pszichológia tanárok barátainak száma: 3, 5, 3, 4, 2, 6 -> Sorba rendezve: 2, 3, 3, 4, 5, 6

$Mdn = 3$

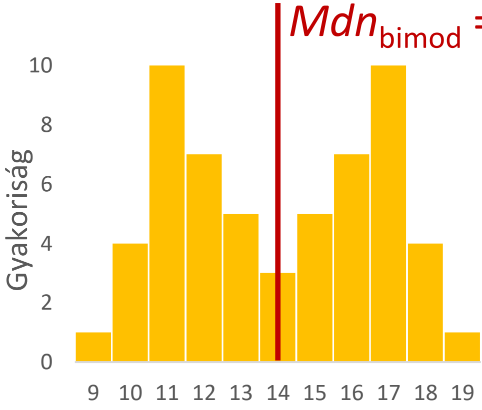
$Mdn = \frac{3+4}{2} = 3,5$

  - Mikor előnyösebb a használata az átlagnál?
    - Ordinális adatok (de működik folytonos adatokon is), ferde eloszlás, outlierek esetén



Outlierekre nagyon kevéssé érzékeny

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

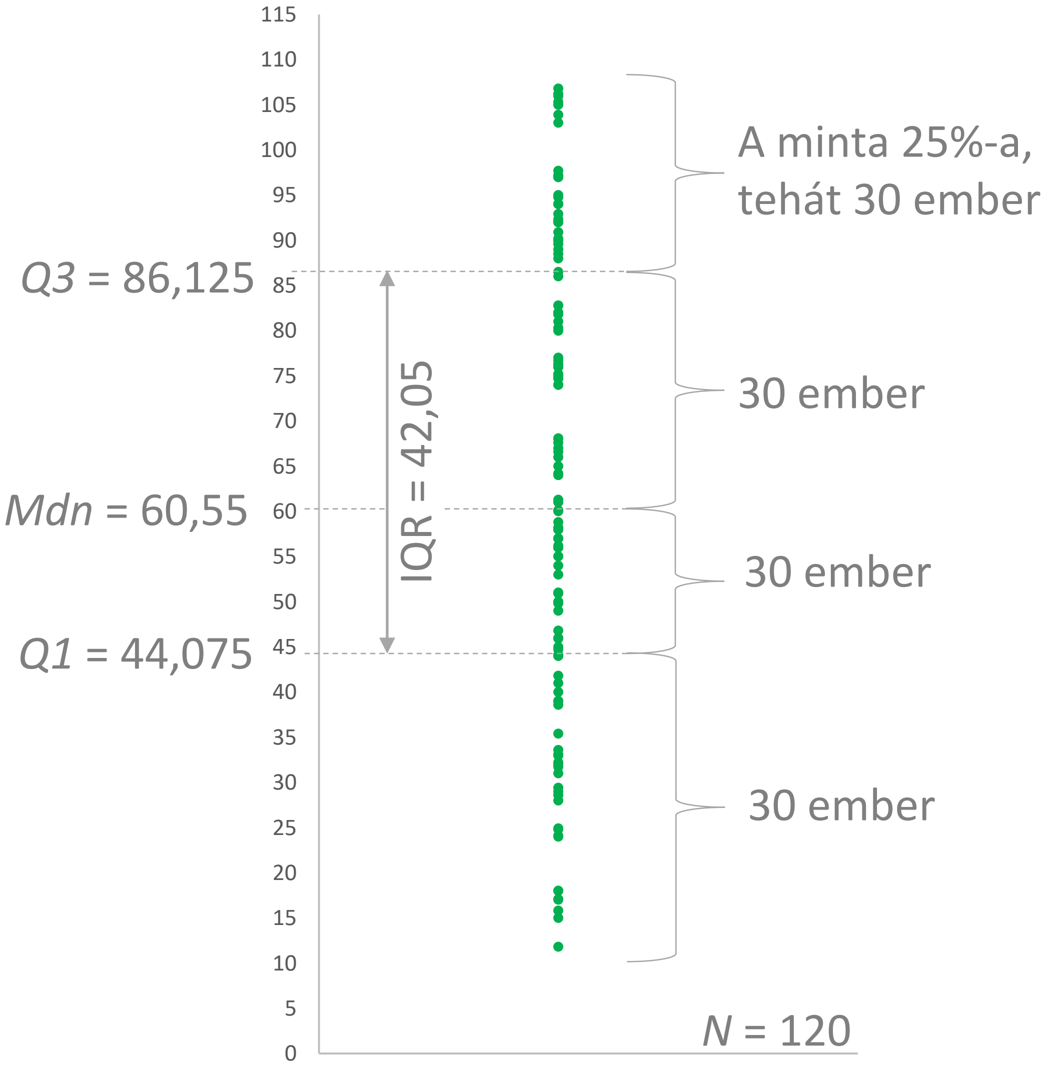


Multimodális eloszlásra ez sem jó választás

Elhelyezkedési mutatók: Kvartilisek, percentilisek

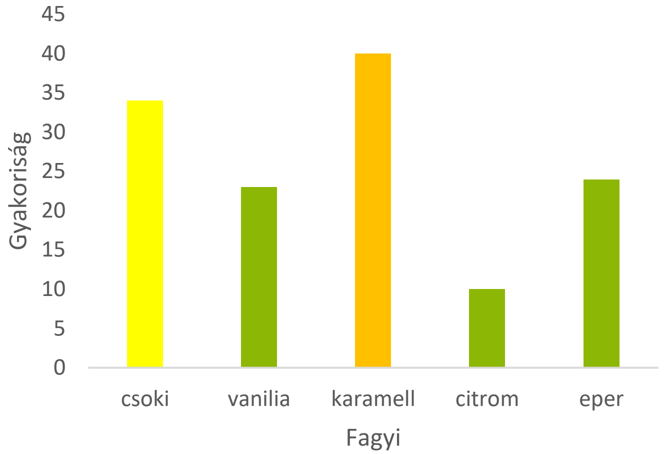
- **Kvartilisek**
  - Negyedelő pontok
  - **Q1** – első kvartilis (alsó negyedelő pont): melynél a minta maximum 25%-a vesz fel alacsonyabb, 75%-a magasabb értéket
  - **Q2** – második kvartilis (felező pont): medián
  - **Q3** – harmadik kvartilis (felső negyedelő pont): melynél a minta maximum 75%-a 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)

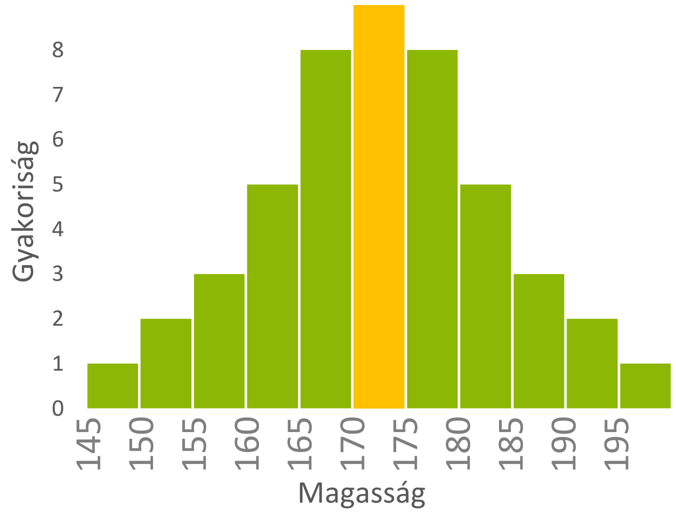
- **Percentilisek**
  - Századoló pontok.
  - Pl 95%-os percentilis: a minta 95%-a alacsonyabb, 5%-a magasabb értéket vesz fel.



Elhelyezkedési mutatók: Módusz

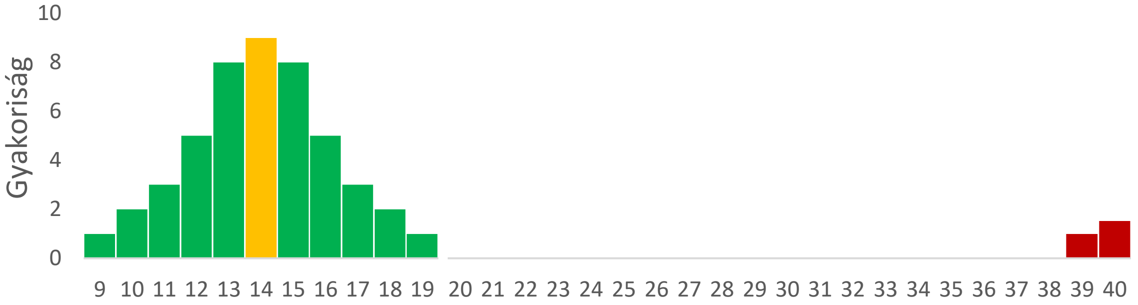
- **Módusz**
  - A legdivatosabb, **leggyakrabban előforduló érték**.
  - Jelölése: ***Mo***
  - Számolása:
    - Diszkrét adatok esetén: a leggyakrabban előforduló érték. Példa: 2, 3, 1, 3, 4 -> *Mo* = 3
    - Folytonos változók esetén: a folytonos változót értelmes méretű intervallumokra bontjuk, és 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.



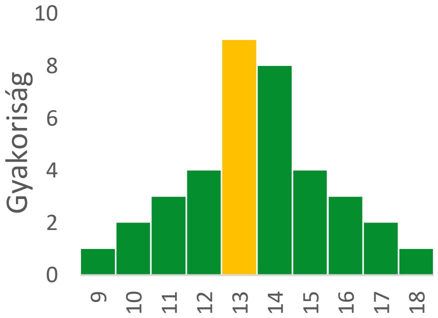


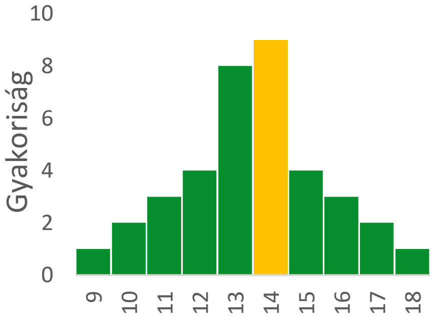
Elhelyezkedési mutatók: Módusz

- **Módusz jellegzetességei:**
  - Outlierekre nem érzékeny

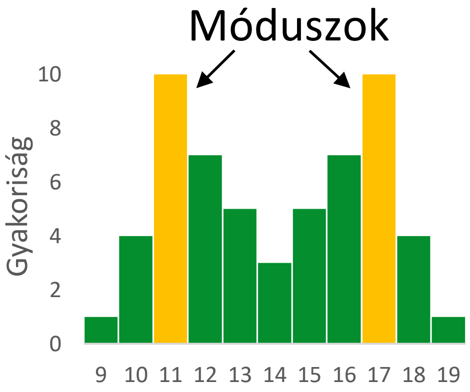


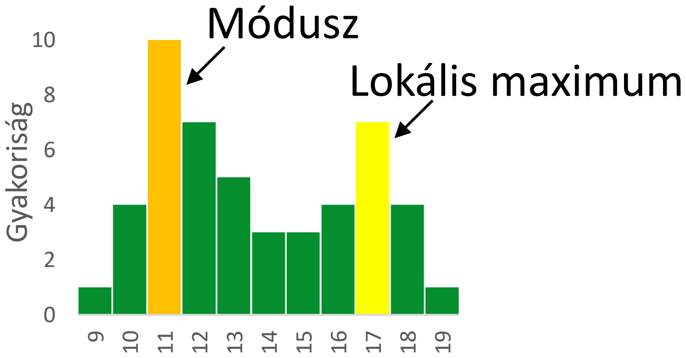
  - Annál inkább a kicsi különbségek esetlegességére,
    és folytonos változók esetén az intervallum szélességére.





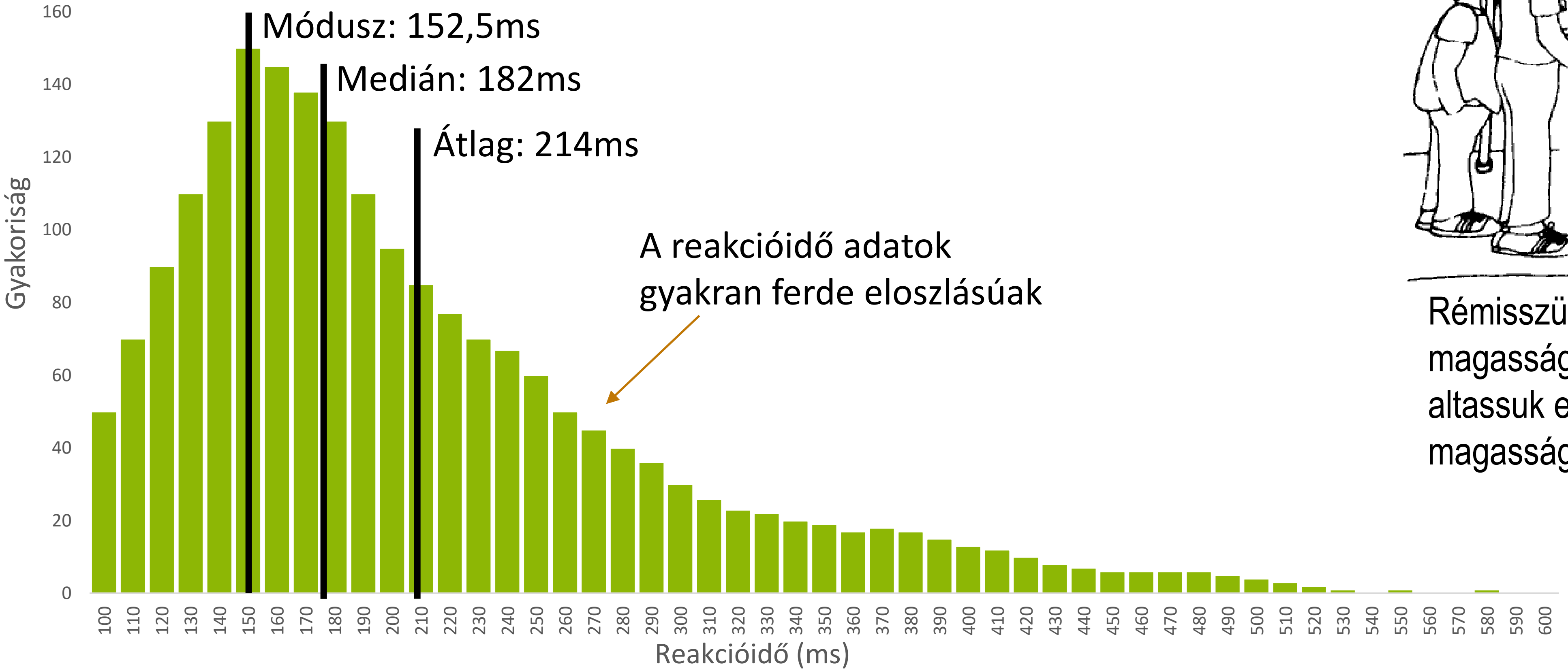
  - Értelmezhetőek az **unimodális** eloszlás mellett **multimodális** (pl. bimodális, trimodális) eloszlások is.
    Sőt, sokszor a **lokális maximumok**at is móduszként közlik.





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



Módusz: 152,5ms

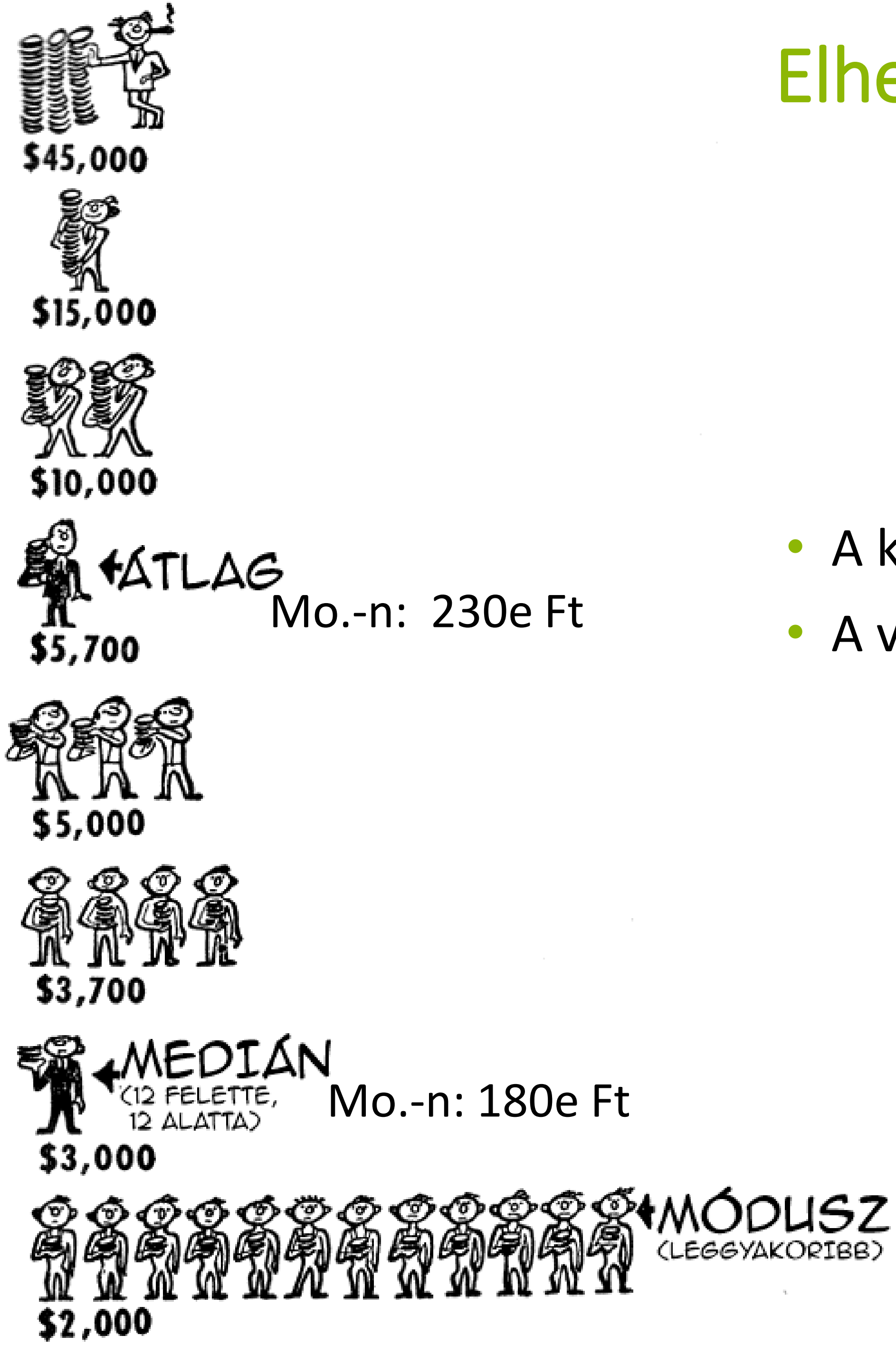
Medián: 182ms

Átlag: 214ms

A reakcióidő adatok gyakran ferde eloszlásúak

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

Elhelyezkedési mutatók: Átlag vs Medián vs Módusz



Mo.-n: 230e Ft

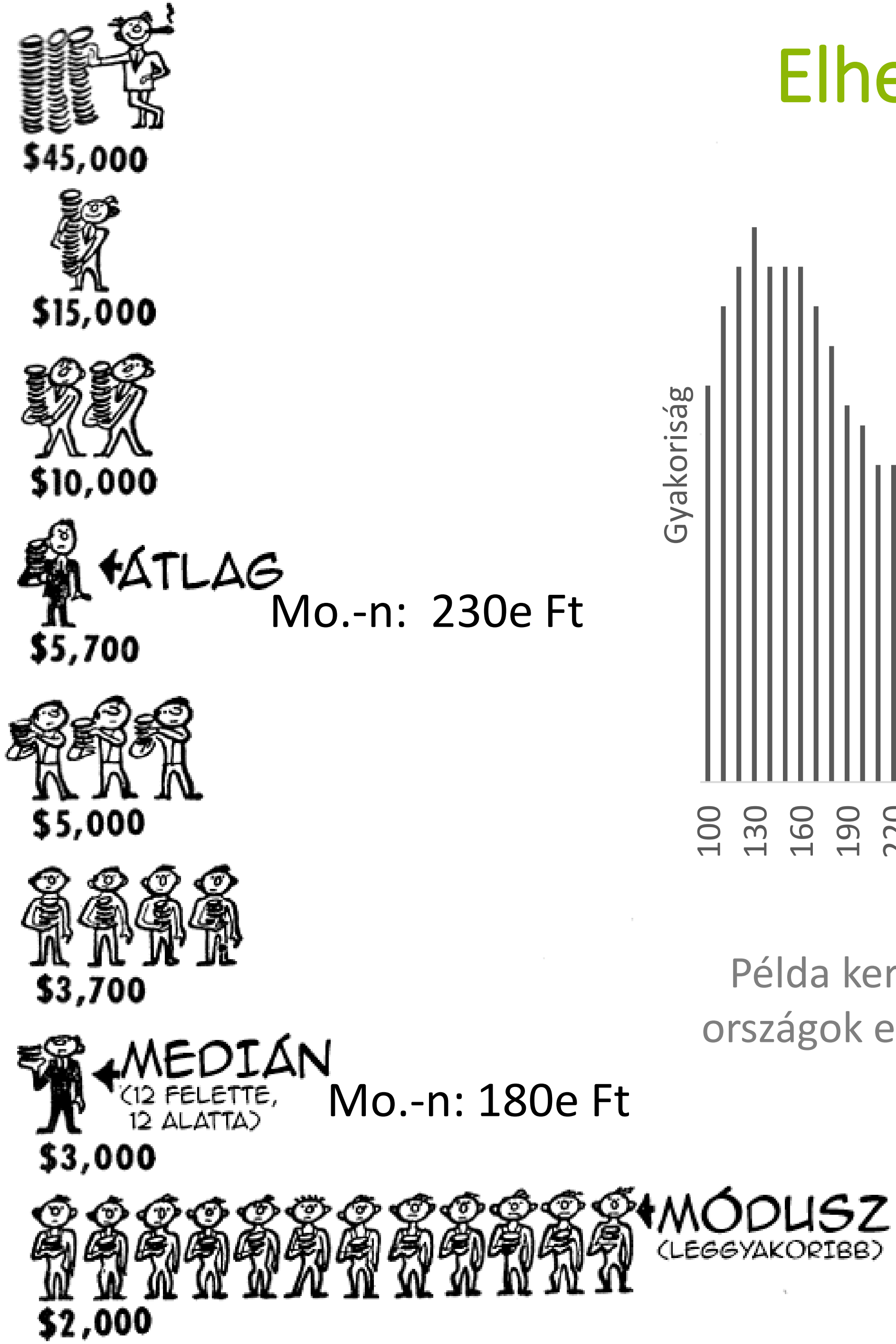
Mo.-n: 180e Ft

Mo.-n: 130e Ft

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

(bruttóban értve havonta, a fizetesek.hu felmérése alapján)

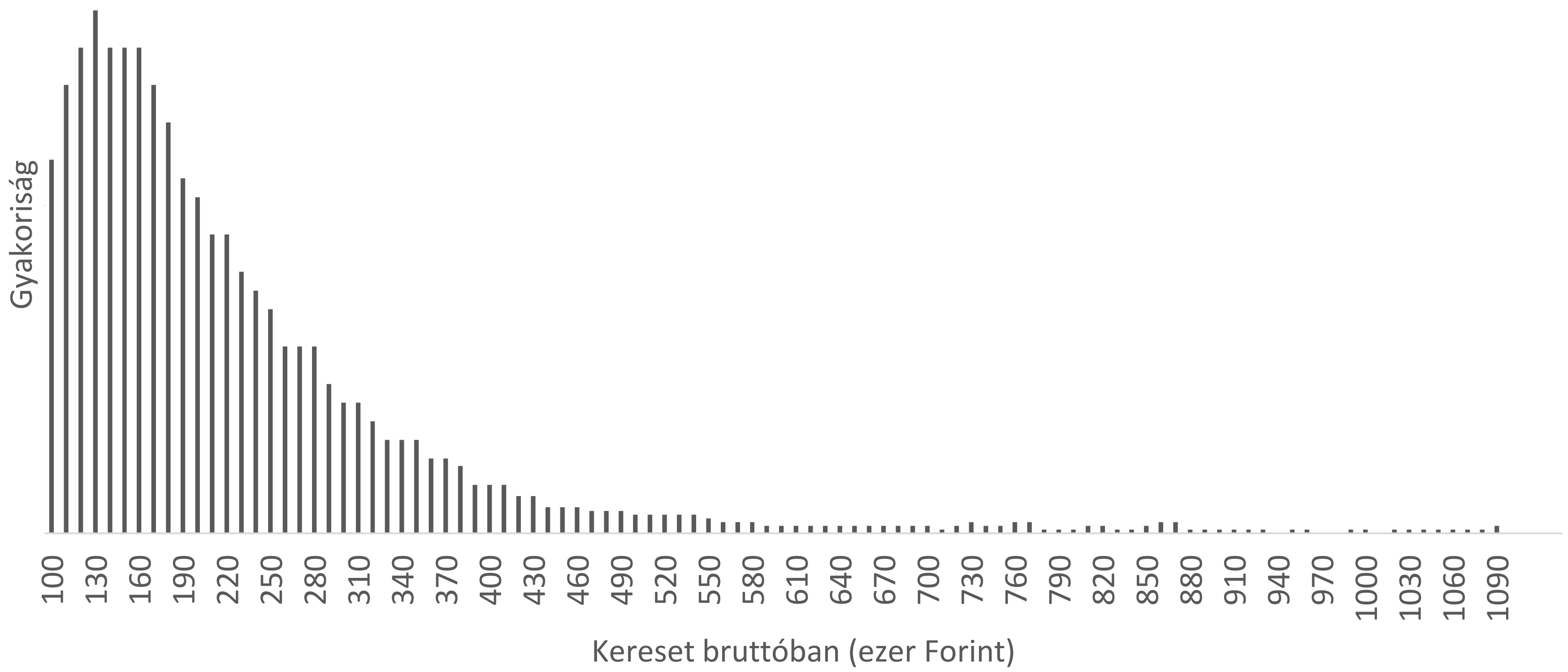
Elhelyezkedési mutatók: Átlag vs Medián vs Módusz



Mo.-n: 230e Ft

Mo.-n: 180e Ft

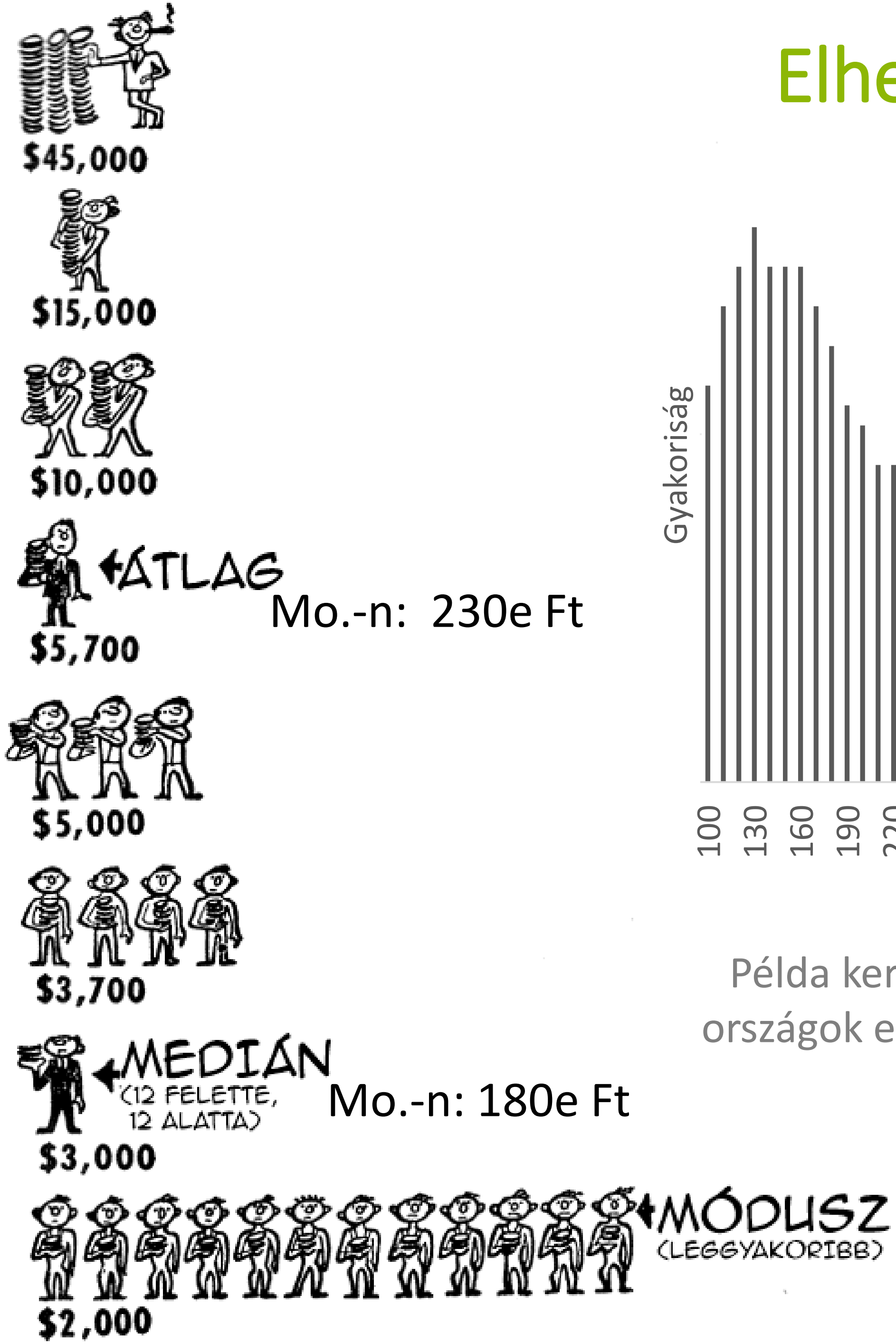
Mo.-n: 130e Ft



Példa keresetek eloszlására. Az adatok fiktívek, a hisztogram a középértékek és más országok elérhető kereset-eloszlása alapján készült. Célja csupán a statisztikai jelenség demonstrálása, a valóságot nem feltétlenül fedi.

(bruttóban értve havonta, a fizetesek.hu felmérése alapján)

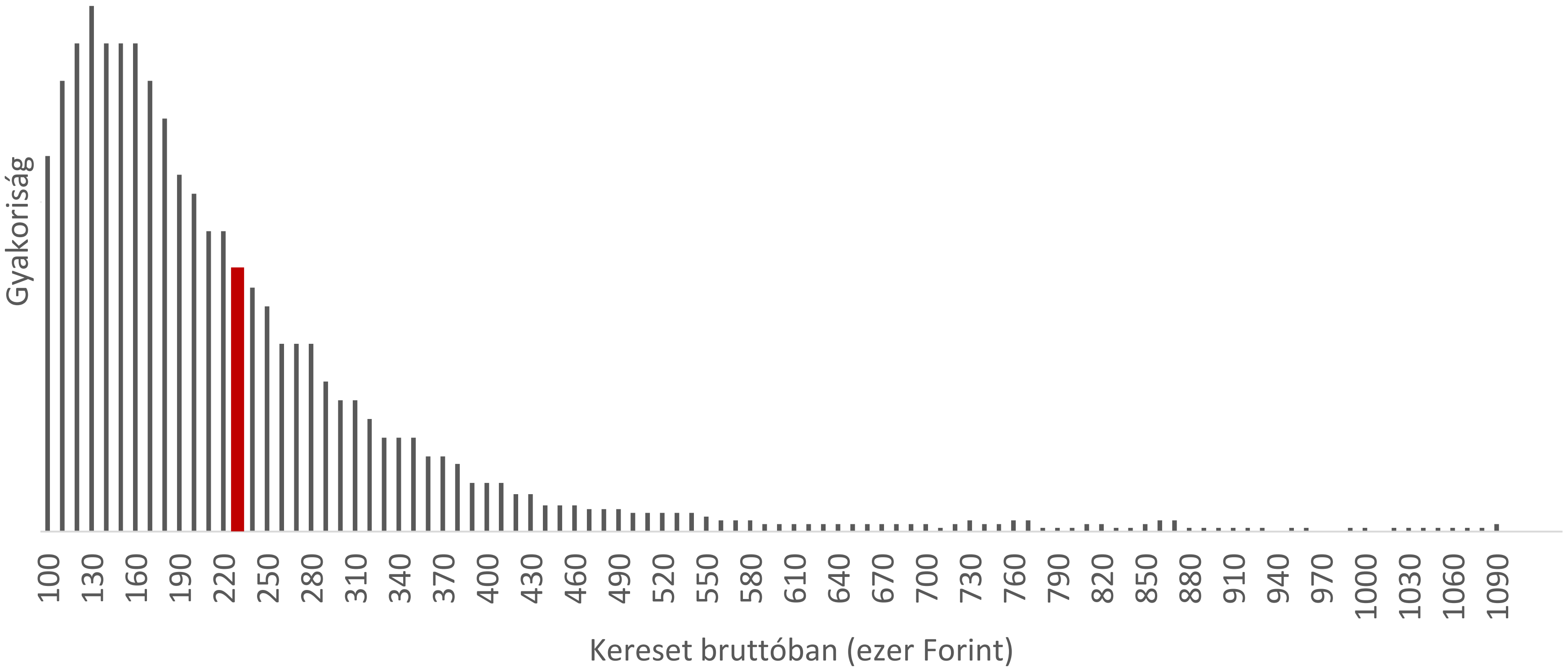
Elhelyezkedési mutatók: Átlag vs Medián vs Módusz



Mo.-n: 230e Ft

Mo.-n: 180e Ft

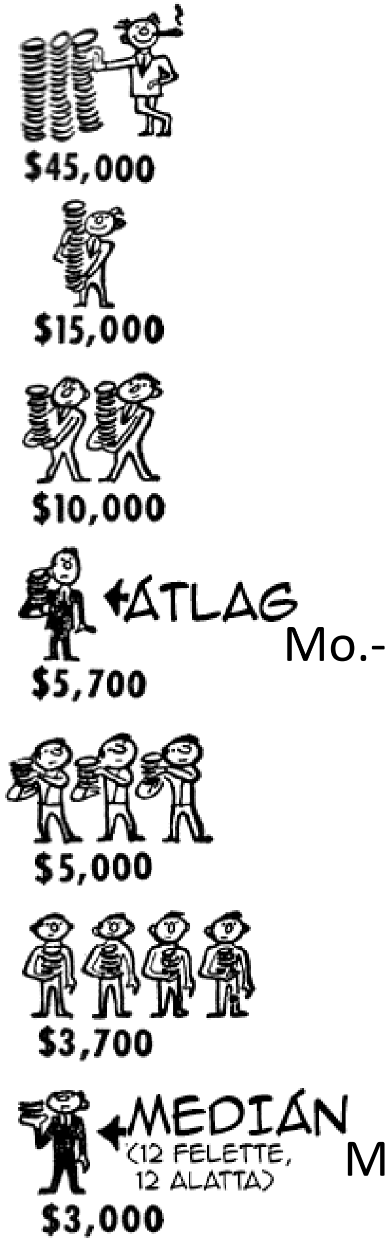
Mo.-n: 130e Ft



Példa keresetek eloszlására. Az adatok fiktívek, a hisztogram a középértékek és más országok elérhető kereset-eloszlása alapján készült. Célja csupán a statisztikai jelenség demonstrálása, a valóságot nem feltétlenül fedi.

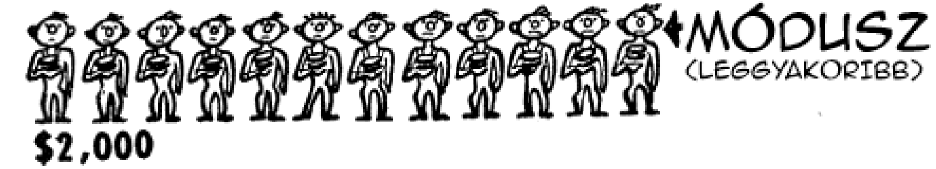
(bruttóban értve havonta, a fizetesek.hu felmérése alapján)

Elhelyezkedési mutatók: Átlag vs Medián vs Módusz

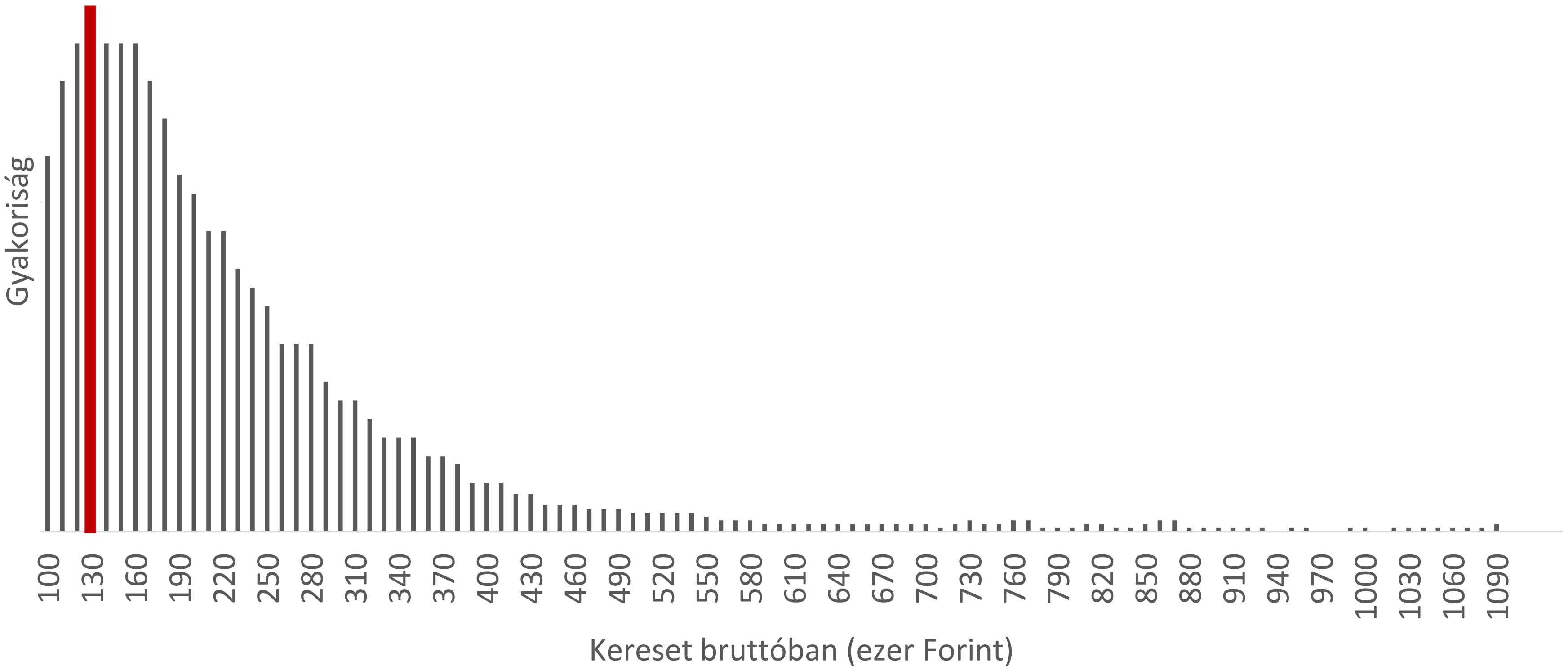


Mo.-n: 230e Ft

Mo.-n: 180e Ft



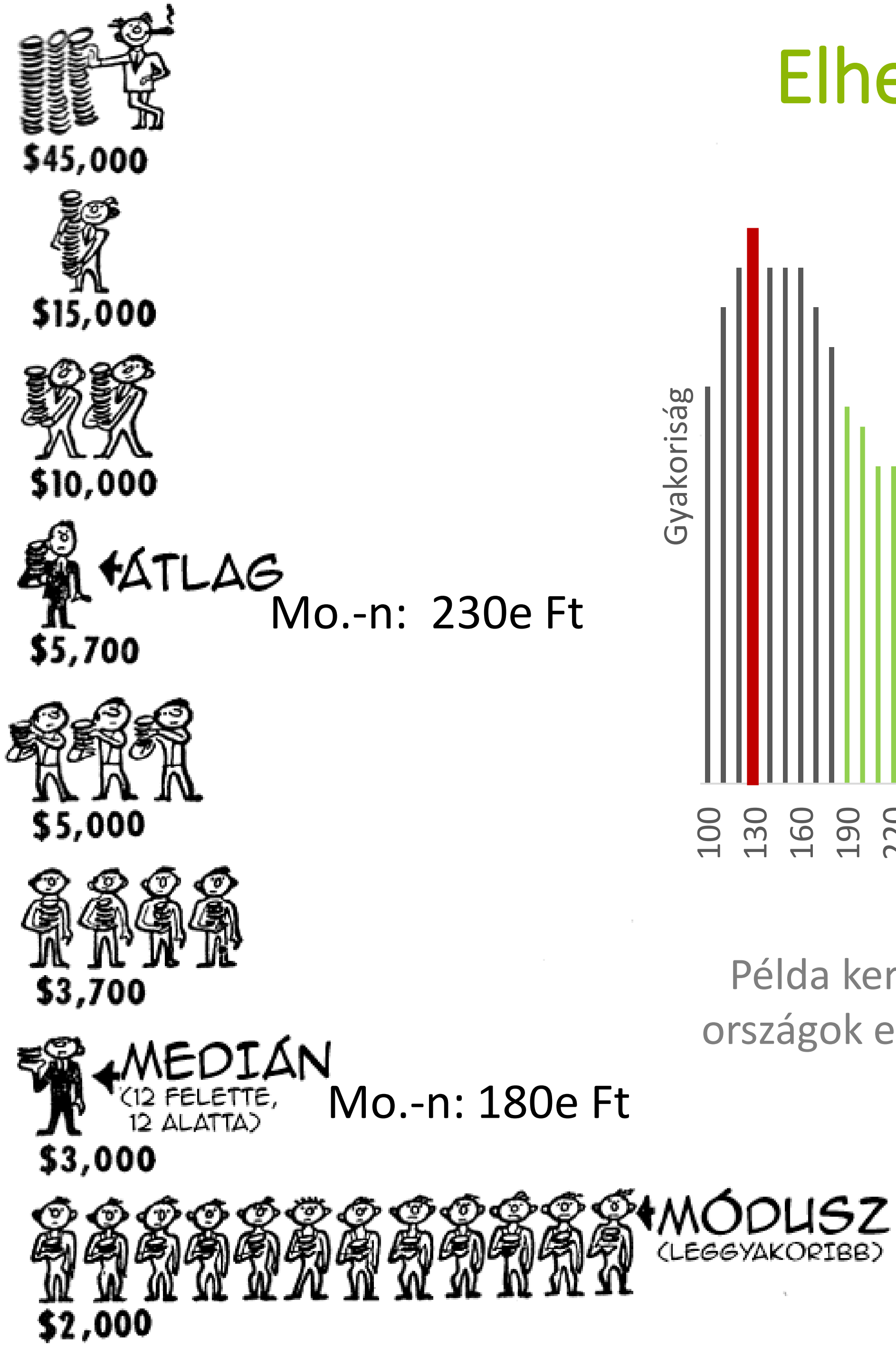
Mo.-n: 130e Ft



Példa keresetek eloszlására. Az adatok fiktívek, a hisztogram a középértékek és más országok elérhető kereset-eloszlása alapján készült. Célja csupán a statisztikai jelenség demonstrálása, a valóságot nem feltétlenül fedi.

(bruttóban értve havonta, a fizetesek.hu felmérése alapján)

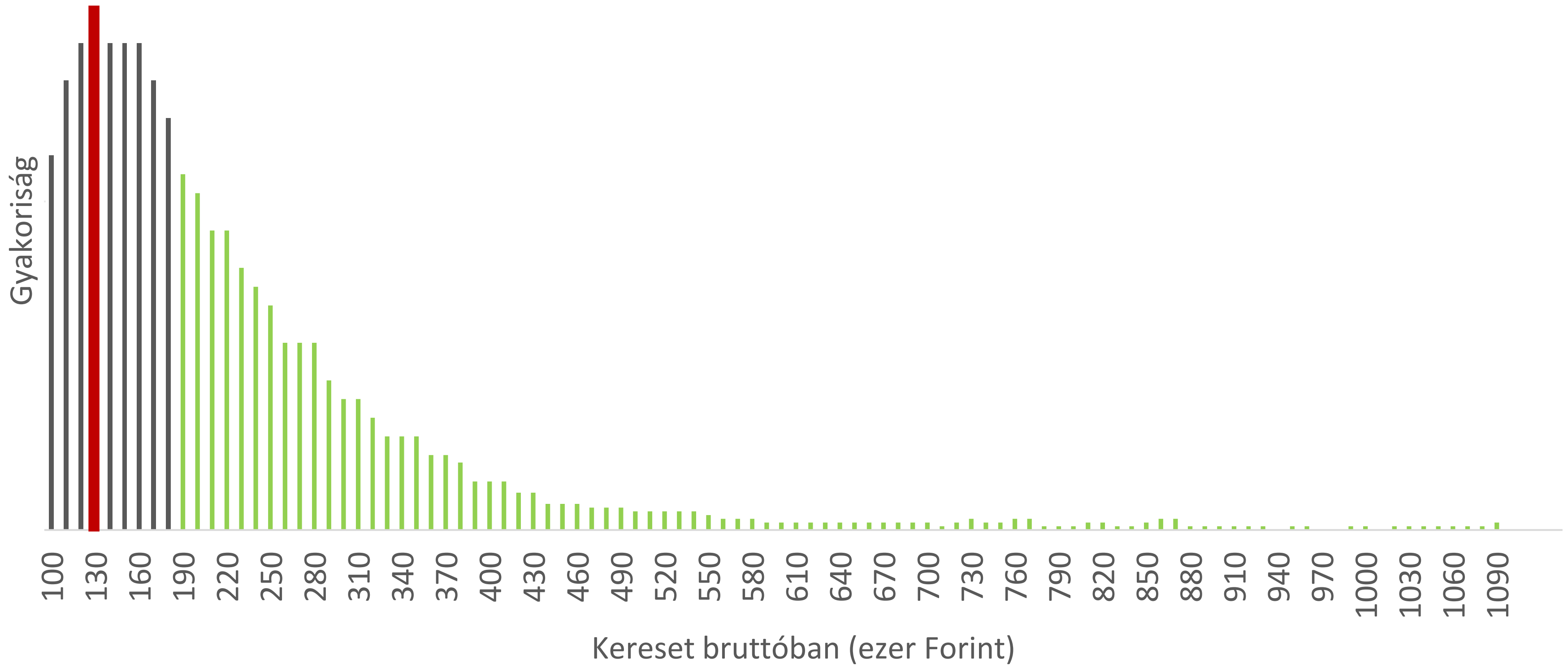
Elhelyezkedési mutatók: Átlag vs Medián vs Módusz



Mo.-n: 230e Ft

Mo.-n: 180e Ft

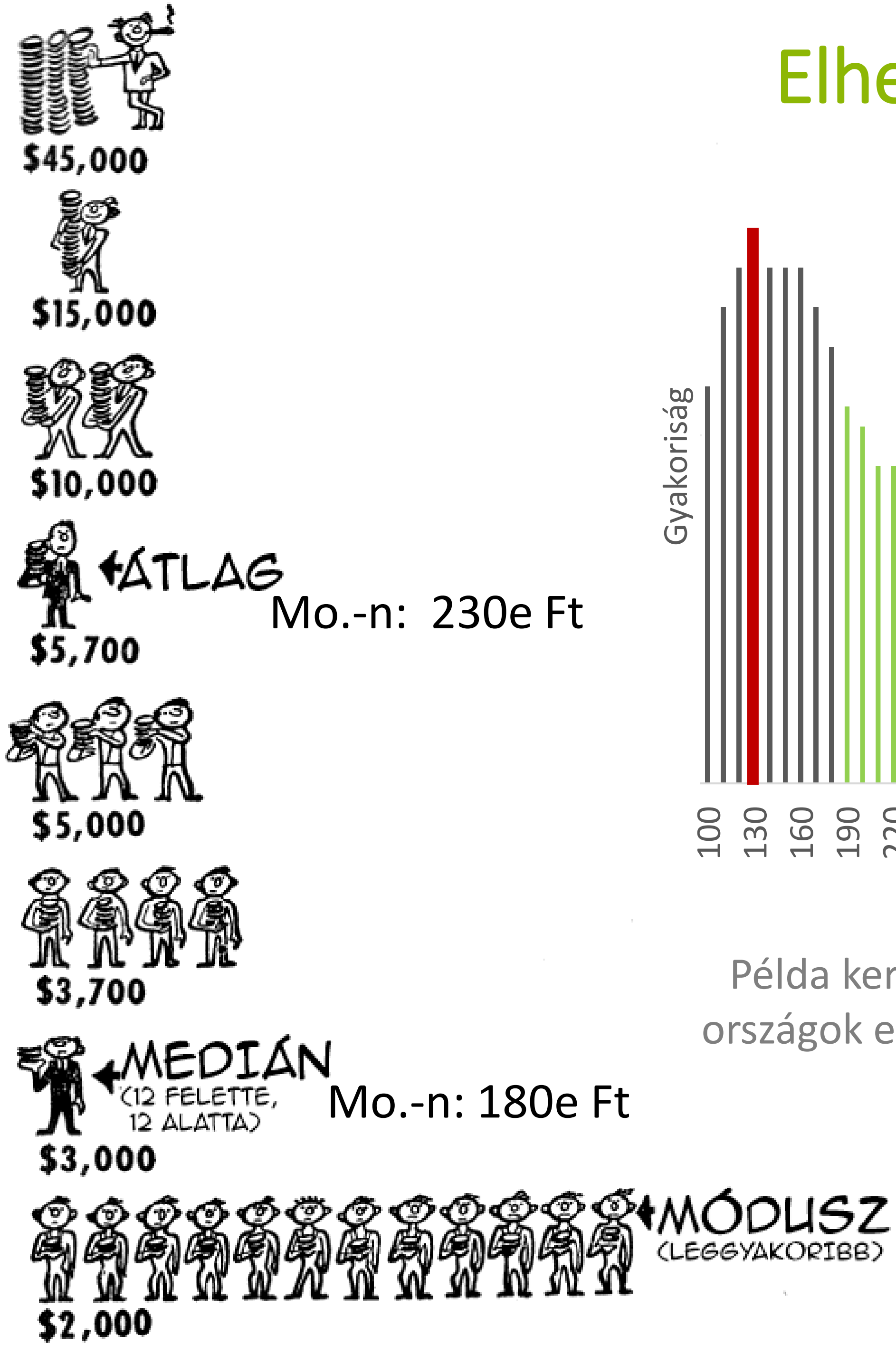
Mo.-n: 130e Ft



Példa keresetek eloszlására. Az adatok fiktívek, a hisztogram a középértékek és más országok elérhető kereset-eloszlása alapján készült. Célja csupán a statisztikai jelenség demonstrálása, a valóságot nem feltétlenül fedi.

(bruttóban értve havonta, a fizetesek.hu felmérése alapján)

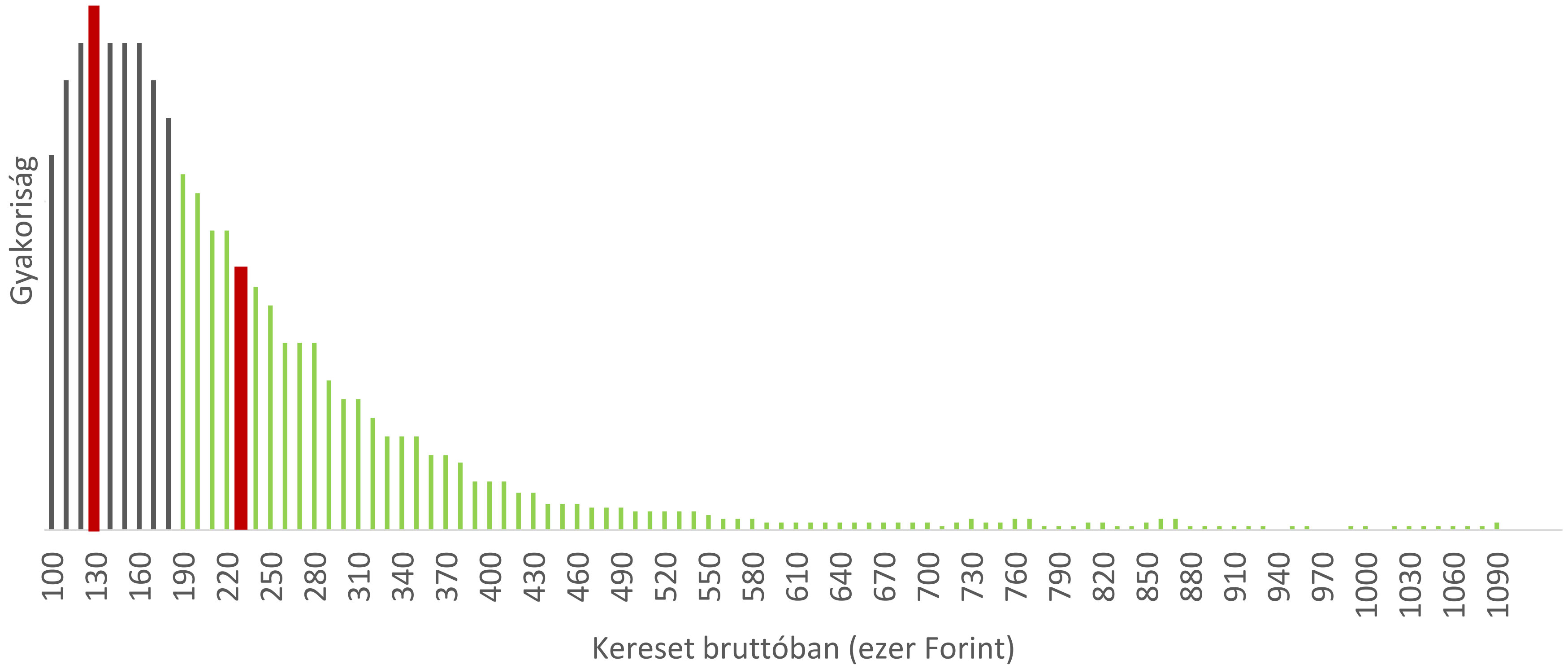
Elhelyezkedési mutatók: Átlag vs Medián vs Módusz



Mo.-n: 230e Ft

Mo.-n: 180e Ft

Mo.-n: 130e Ft

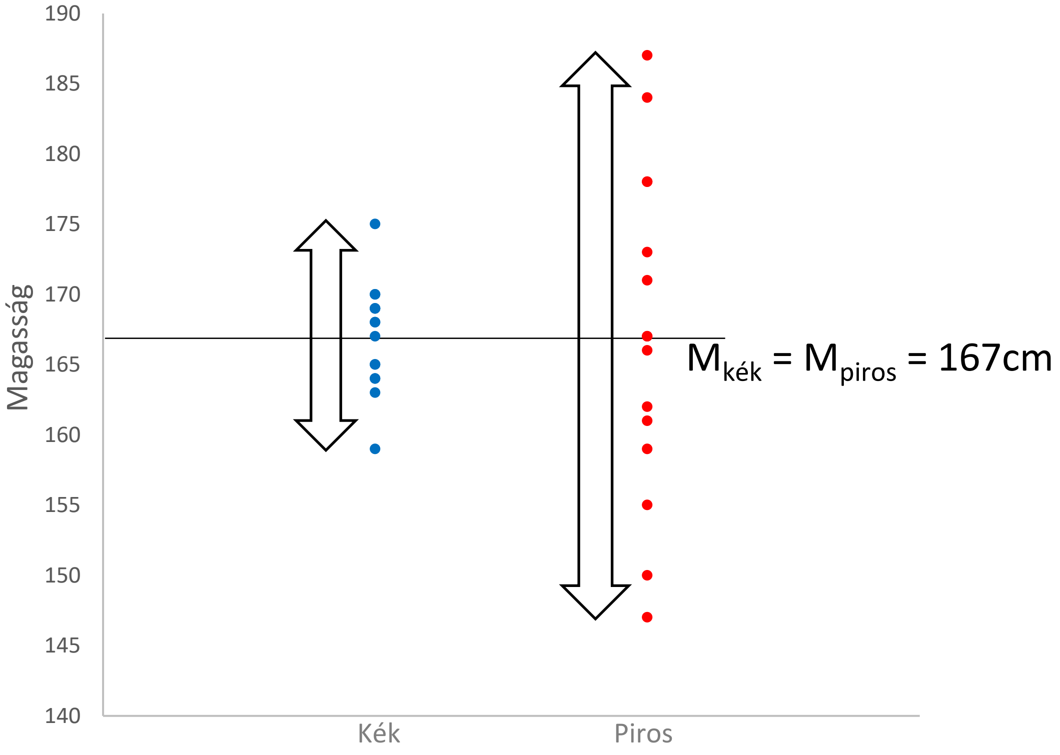


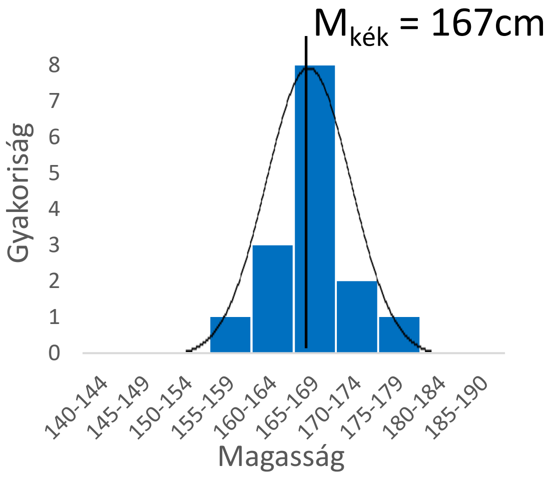
Példa keresetek eloszlására. Az adatok fiktívek, a hisztogram a középértékek és más országok elérhető kereset-eloszlása alapján készült. Célja csupán a statisztikai jelenség demonstrálása, a valóságot nem feltétlenül fedi.

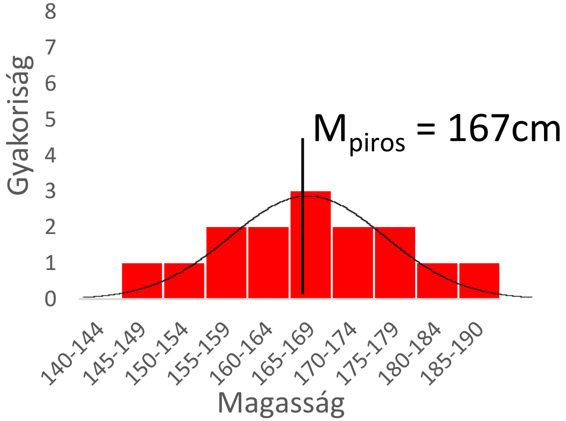
(bruttóban értve havonta, a fizetesek.hu felmérése alapján)

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**



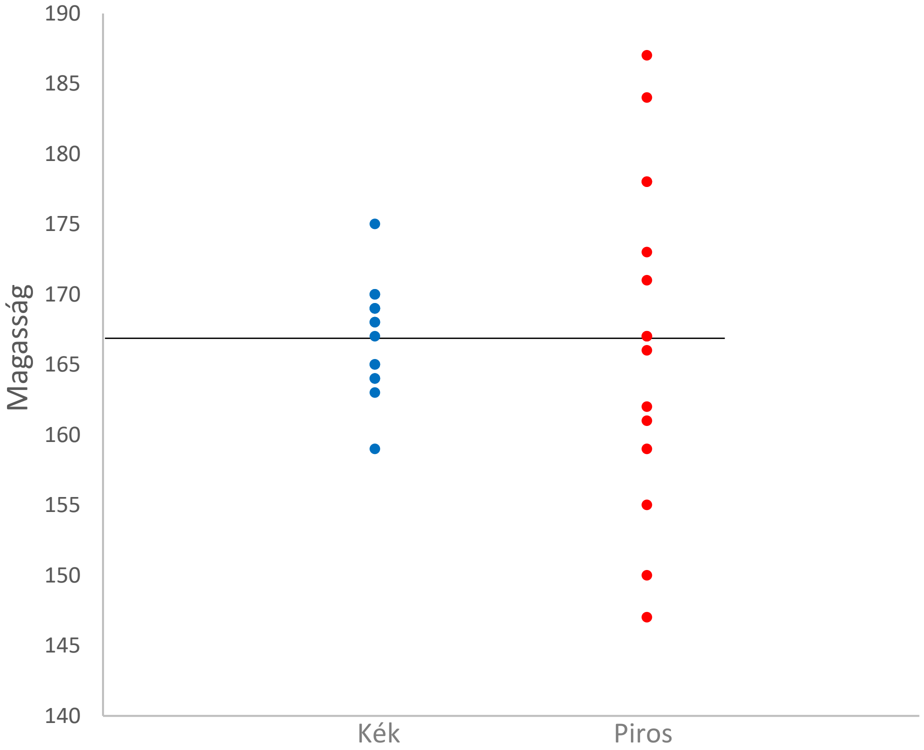




Kékek	**Pirosak**
163	155
169	178
168	187
169	166
170	178
168	159
165	171
164	150
164	167
170	147
159	184
175	162
165	167
169	161
167	173

Szóródási mutatók: Terjedelem

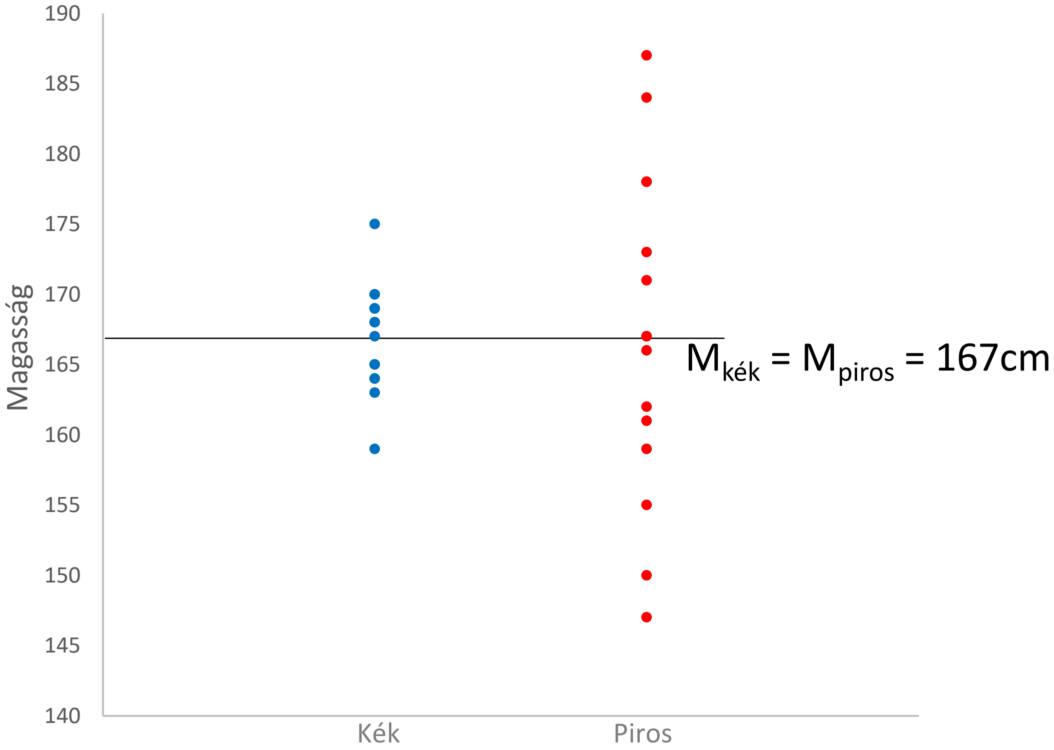
- **Terjedelem**
  - A legkisebb és legnagyobb érték távolsága
  - Jelölése: **range**
  - Számítása: Max-Min Példában: $\text{Range}_{\text{kék}} = 175-159 = 16$ és $\text{Range}_{\text{piros}} = 187-147 = 40$
  - APA szabvány szerint értéke helyett a minimumot és maximumot adjuk meg: $\text{range}_{\text{kék}}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



Kékek	**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 **hibá**ja, pontatlansága.
  - Jelölése: D_i Számolása: $D_i = x_i - \bar{x}$
- **Átlagtól való négyzetes eltérés** (**D^2**, angolul squared deviances)
  - Csak a különbség nagysága kell, iránya (előjele) nem.
  - Abszolútérték vagy **négyzetre emelés**? Számolása: ${D_i}^2$

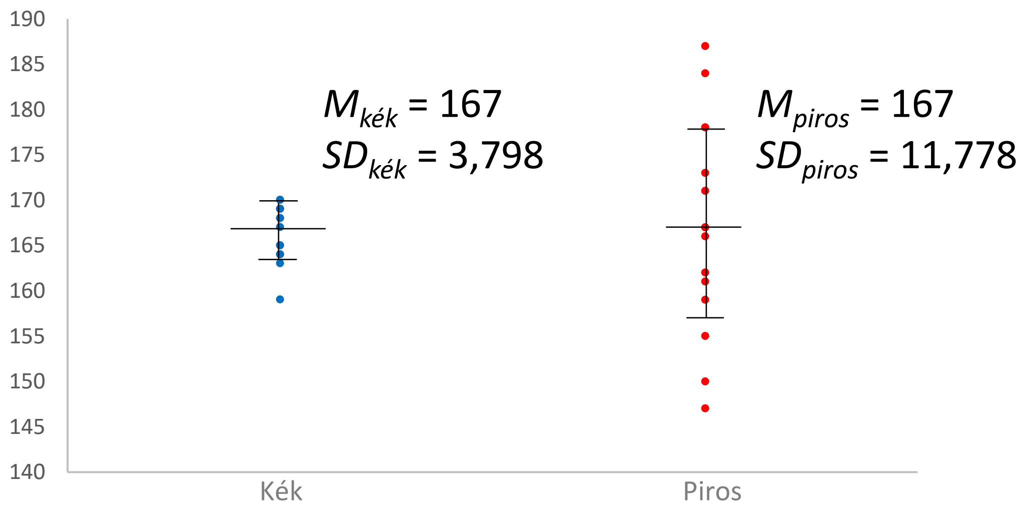


Kékek	**$D_{kék}$**	**$D^2_{kék}$**	**Pirosak**	**D_{piros}**	**D^2_{piros}**
163	163-167 = -4	$(-4)^2 = 16$	155	-12	144
169	169-167 = 2	$2^2 = 4$	178	11	121
168	168-167 = 1	$1^2 = 1$	187	20	400
169	2	4	166	-1	1
170	3	9	178	11	121
168	1	1	159	-8	64
165	-2	4	171	4	16
164	-3	9	150	-17	289
164	-3	9	167	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

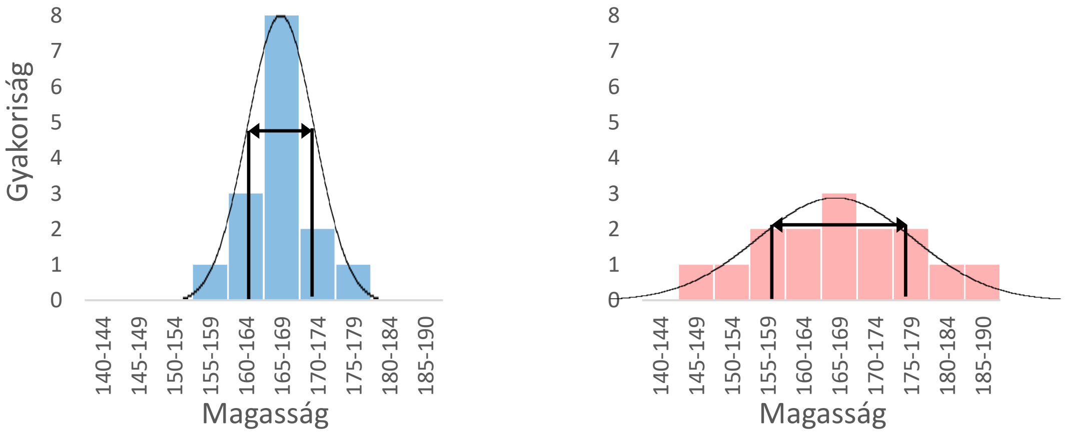
- **Variancia**
  - **Az átlagtól való négyzetes eltérés átlaga**, jelölése: ***Var***, ***s^2*** vagy ***σ^2***
  - Számolása: $Var = \frac{\sum {D_i}^2}{df} = \frac{\sum(x_i - \bar{x})^2}{n-1}$ Pl. : $Var_{kék} = \frac{16+4+1+4+9+1+4+9+9+9+64+64+4+4+0}{15-1} = 14{,}429$

    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.: $SD_{kék} = \sqrt{14{,}429} = 3{,}798$



$M_{kék}$ = 167
$SD_{kék}$ = 3,798

M_{piros} = 167
SD_{piros} = 11,778



Egy kis kitérő: szabadságfok

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

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{N}$$

Példa minta: 1, 2, 6 -> $\bar{x}$ = 3

$$3 = \frac{x_1 + x_2 + x_3}{3}$$

1, 1, 7 vagy 2, 2, 5
vagy 3, 3, 3 stb.

$$3 = \frac{1 + x_2 + x_3}{3}$$

1, 1, 7 vagy 1, 2, 6
vagy 1, 4, 4 stb.

$$3 = \frac{1 + 2 + x_3}{3}$$

1, 2, 6

Szabadságfok (**degrees of freedom**, **df**)

- **Egymástól független, szabadon változtatható tagok száma**
- Számolása: a tagok száma mínusz az összefüggések száma
- Szórás / variancia számolásakor értéke: *df = N-1*
- Matematikai értelme (Bessel korrekció):
  - A minta varianciájának és szórásának számolásakor felhasználom a tagok mellett az mintaátlagot is, mely *nem* független a tagoktól.
  - A tagok száma = *N*, az összefüggések száma = 1
    ebből adódik, hogy a 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.

Egy kis kitérő: szabadságfok

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

Példa minta$_1$: 5db 1-es, és 5db 2-es

Példa minta$_2$: 500db 1-es, és 500db 2-es

$$\bar{x}_1 = \bar{x}_2 = \bar{x} = 1{,}5$$

$$VAR'_1 = \frac{\sum(x_i - \bar{x})^2}{N_1} = \frac{2{,}5}{10} = 0{,}25$$

$$VAR_1 = \frac{\sum(x_i - \bar{x})^2}{N_1 - 1} = \frac{2{,}5}{9} = 0{,}277777$$

$$VAR'_2 = \frac{\sum(x_i - \bar{x})^2}{N_2} = \frac{250}{1000} = 0{,}25$$

$$VAR_2 = \frac{\sum(x_i - \bar{x})^2}{N_2 - 1} = \frac{250}{999} = 0{,}25025$$

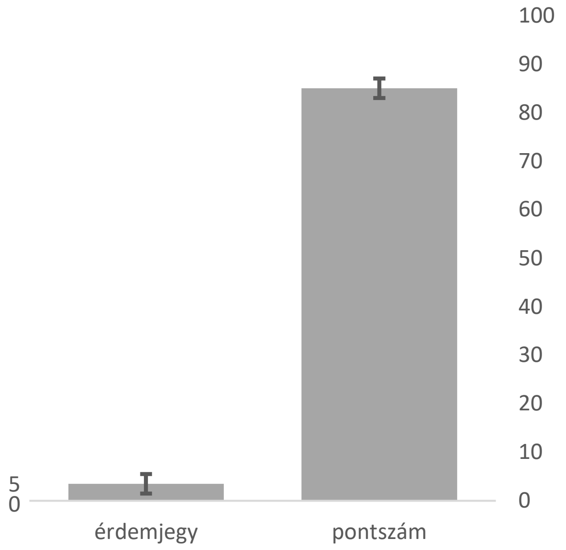
Szabadságfok (degrees of freedom, df)

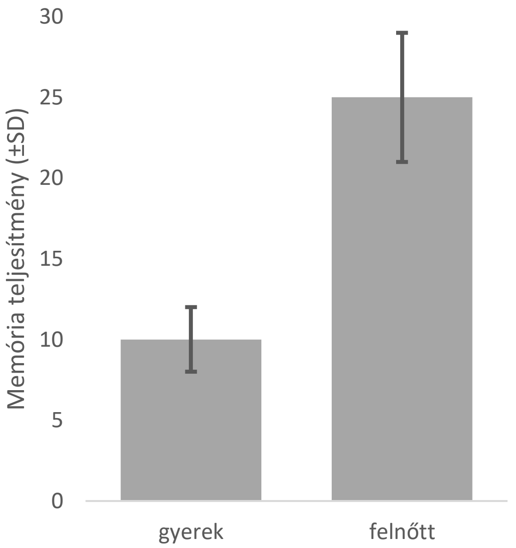
- **Egymástól független, szabadon változtatható tagok száma**
- Számolása: a tagok száma mínusz az összefüggések száma
- Szórás / variancia számolásakor értéke: *df = N-1*
- Matematikai értelme (Bessel korrekció):
  - A minta varianciájának és szórásának számolásakor felhasználom a tagok mellett az mintaátlagot is, mely *nem* független a tagoktól.
  - A tagok száma = *N*, az összefüggések száma = 1 ebből adódik, hogy a 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.

Mekkora a nagy szórás? Relatív szórás fogalma

- Mekkora a „nagy” szórás?
  - Nincs kritérium érték, mert függ a skálázástól!
    - Kétpontnyi szórás kisZH pontszám esetén (0-5-ig terjedő skála) és nagyZH pontszám esetén (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
    - $RSD_{gyerek} = 20\%$ míg $RSD_{felnőtt} = 16\%$ csak



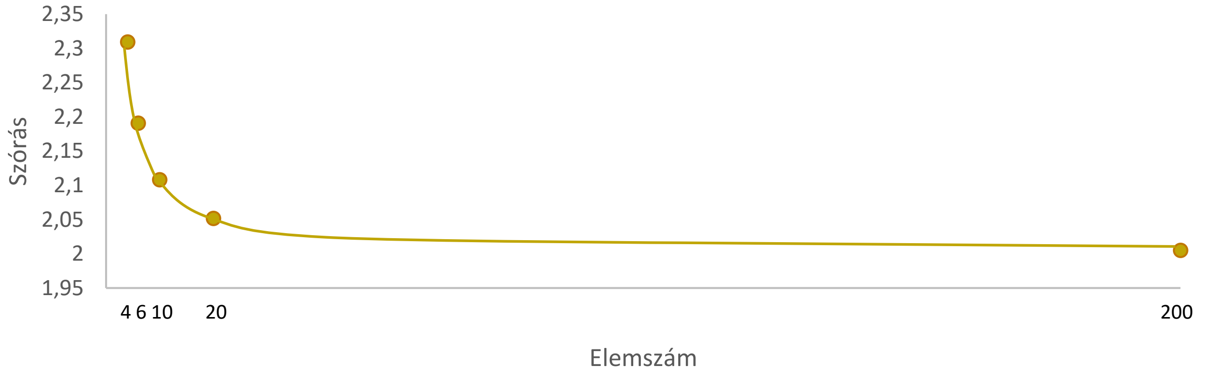


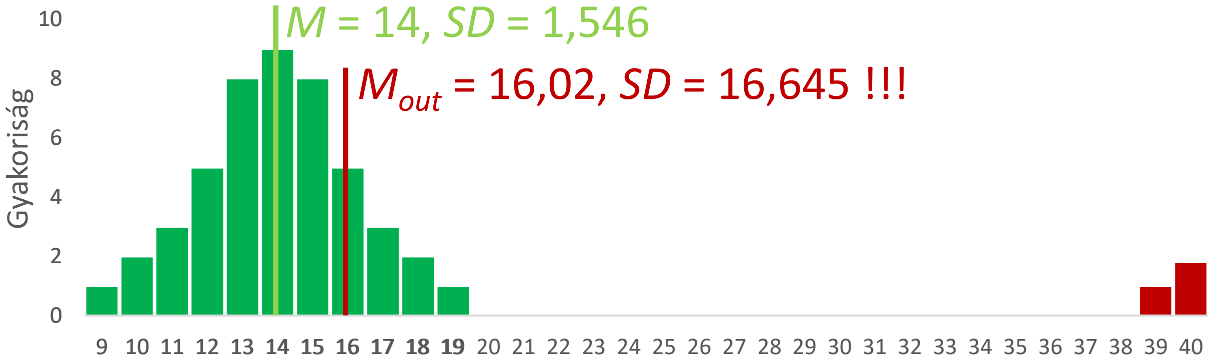
Miért nagy a szórás?

- 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)

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

N	4	6	10	20	200
SD	2,309	2,191	2,108	2,052	2,005





M = 14, *SD* = 1,546

M_{out} = 16,02, *SD* = 16,645 !!!

Variancia és szórás: akkor miért van wikin más képlet?

Statisztika:	**Valószínűségszámítás:**
Variancia: $\boldsymbol{s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \rightarrow \sigma^2 = \frac{\sum(x_i - \mu)^2}{N}}$	Variancia: $\boldsymbol{s^2 = \mathrm{E}\left(X^2\right) - E^2(X)}$
Máshogy néz ki, mégis ugyanaz. Lássuk, hogyan számolódnak!	
A megértéshez kell az átlag és várható érték fogalma, számolása. Példa: Fiamnak van 10db egy elem, 15db két elem, 20db három elem, 5db négy elem hosszú építőkockája.	
Átlag: $\mu = \frac{\sum x_i}{N} = \frac{\sum f_j * X_j}{\sum f_j}$ $\mu = \frac{10 \ * 1 + 15 * 2 + 20 * 3 + 5 * 4}{10 + 15 + 20 + 5} = 2{,}4$	Várható érték: $E(X) = \sum p_i * X_j$ $E(X) = \frac{10}{50} * 1 + \frac{15}{50} * 2 + \frac{20}{50} * 3 + \frac{5}{50} * 4 = 6{,}4$

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

1.1. Vegyük az átlagnak megfelelő várható értéket! E(X)

1.2. Vonjuk ki minden X értékből a várható értéket! X - E(X)

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

1.4. Vegyük a várható értékét (átlagoljuk) a négyzetes különbségeket! $E(\ (X- E(X))^2\ )$

Variancia és szórás: akkor miért van wikin más képlet?

Statisztika:	**Valószínűségszámítás:**
Variancia: $\boldsymbol{s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} \rightarrow \sigma^2 = \frac{\sum(x_i - \mu)^2}{N}}$	Variancia: $\boldsymbol{s^2 = \mathrm{E}(X^2) - E^2(X)}$

2. Itt tartunk: E($(X- E(X))^2$) Rendezzük az egyenletet!

2.1. Bontsuk fel a négyzetre emelést a $(a-b)^2= a^2 – 2ab + b^2$ matematikai azonosság szerint!

$E(\ X^2\ -\ 2*X*E(X)\ +(E(X))^2\ )$

2.2. Bontsuk fel a külső várható értéket, és rendezzük!

SÁRGA: $E(X^2)$

KÉK:

Hozzuk ki a mínusz kettes konstans, hiszen egy konstans várható értéke önmaga! $E(-2*X*E(X)) = -2 * E (X*E(X))$

Szorzat is felbontható. $= -2 * E (X) * E(E(X))$

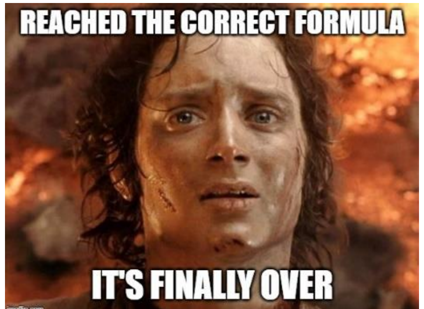
A várhatóérték várhatóértéke önmaga, ezért az E(E(X))=E(X). $= -2* E(X) * E(X) = -2 * (E(X))^2$

A várhatóérték négyzetének van egy szebb megjelenítési formája: $E^2(X)$ $= -2 * E^2(X)$

LILA:

A várhatóérték négyzetének várhatóértéke megegyezik a várhatóérték várhatóértékének négyzetével. A belül megjelent várhatóérték várhatóértékéről előbb láttuk, hogy az egyszerűen várhatóérték.

$E((E(X))^2) = (E(E(X)))^2 = (E(X))^2 = E^2(X)$

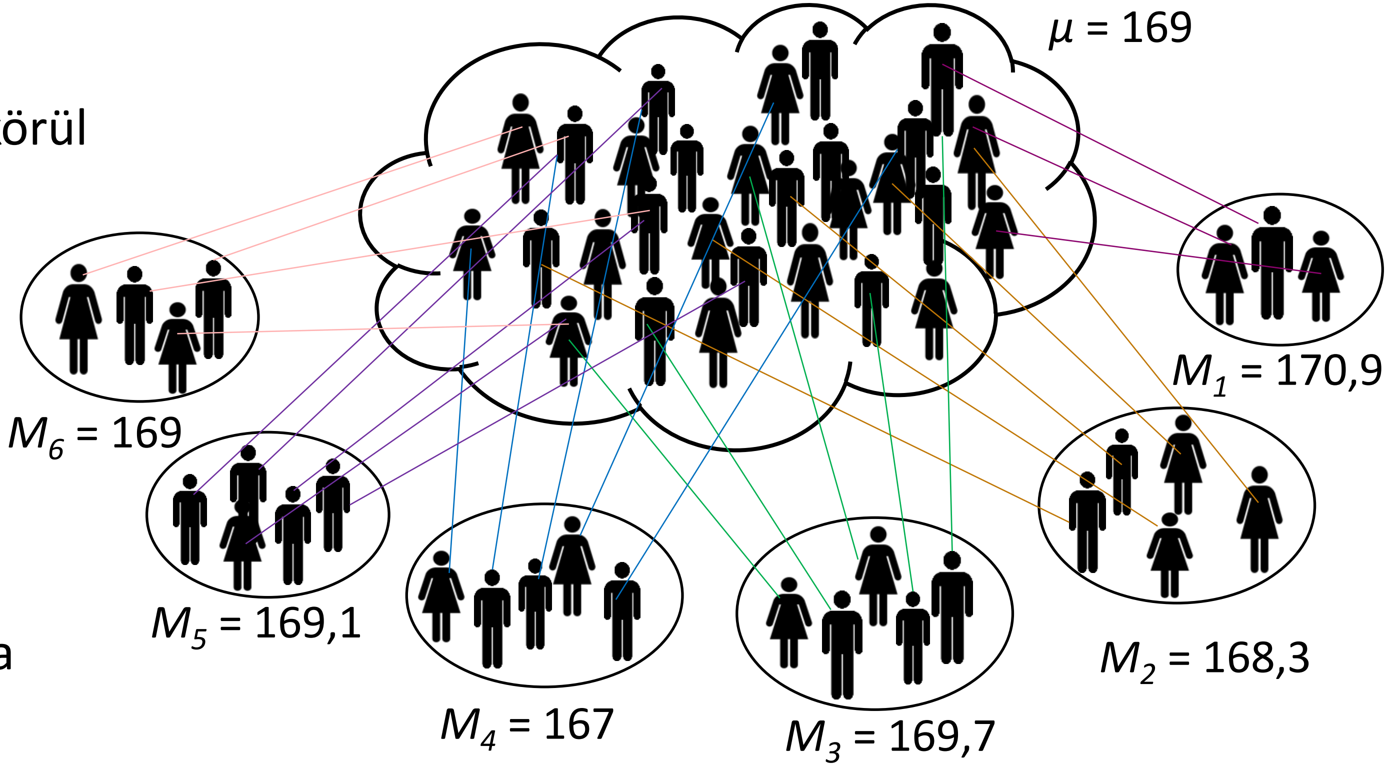


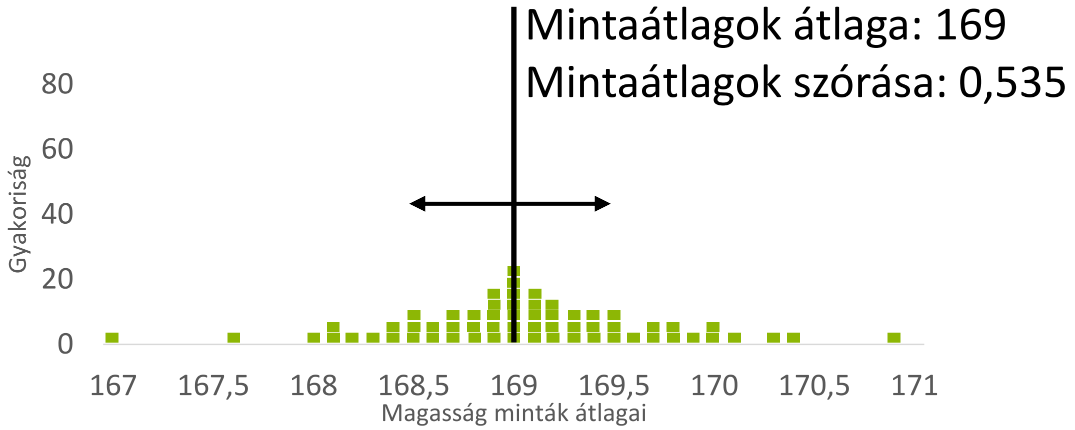
2.3. Vonjuk össze a tagokat, és megkaptuk a valószínűség számítás képletét:

$E(X^2)\ -\ 2*E^2(X)\ +\ E^2(X)\ =\ \mathbf{E(X^2)\ -\ E^2(X)}$

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.
- Ha elég sok mintát veszünk, akkor hisztogramon ábrázolhatjuk a mintaátlagok eloszlását is.
- Várható, hogy a mintaátlagok átlaga jó közelítését adja a populációátlagnak,
  a mintaátlagok szórása pedig kisebb lesz, mint a minta szórása.
- **A mintaátlaghoz tartozó standard error**:
  - **A populációból vett minták átlagainak szórása**
  - Jelölése: $s_{\bar{x}}$ vagy **SE**
- Standard errort nem csak a mintaátlaghoz lehet számítani, hanem más becsült mutatókhoz is, például a regressziós együtthatók meredekségéhez.



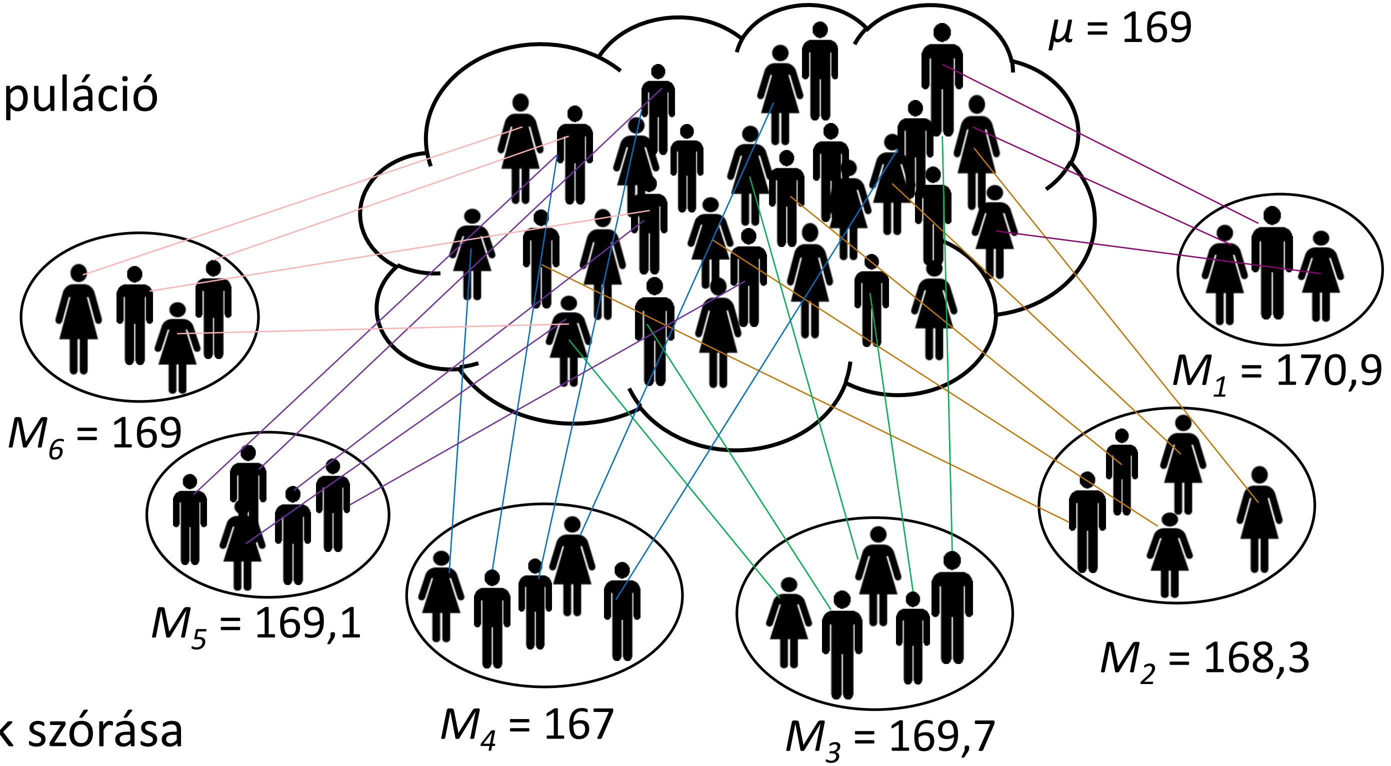

Mintaátlagok átlaga: 169
Mintaátlagok szórása: 0,535

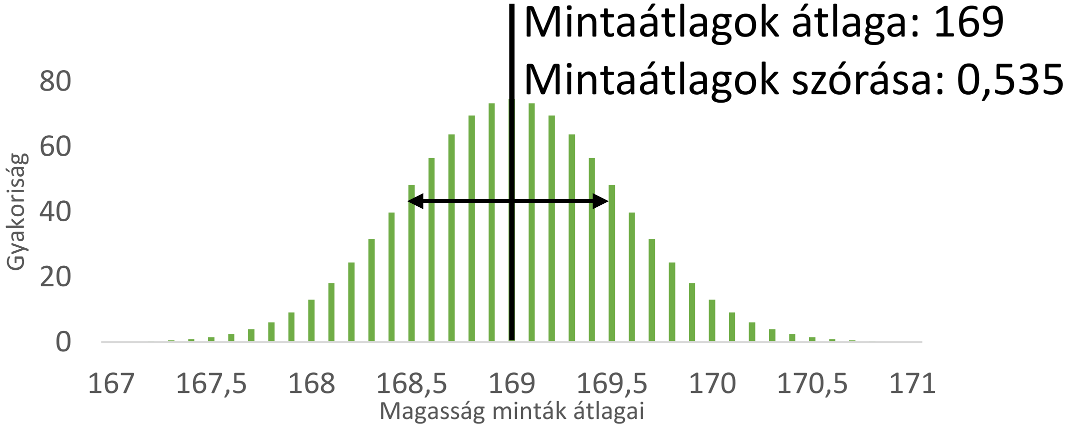
Szóródási mutatók: Standard error

- Számítása (ha csak egy mintánk van):
  - A mintaátlagok eloszlásának varianciája a populáció varianciájának és az elemszámnak a hányadosa.
  - Az elemszámmal való osztás oka az, hogy a minták elemszámának növelésével a mintaátlagok jobban a populációátlag köré csoportosulnak, így a mintaáltagok szórása is kisebb lesz.
  - A populáció varianciája azonban általában nem ismert, ezért a minta varianciájával becsüljük.
  - A mintaátlagok varianciájából a mintaátlagok szórása gyökvonással kapható.

$$\sigma_{\bar{x}}{}^2 = \frac{\sigma^2}{N} \rightarrow s_{\bar{x}}{}^2 = \frac{s^2}{N} \rightarrow \boldsymbol{s_{\bar{x}} = \sqrt{\frac{s^2}{N}} = \frac{s}{\sqrt{N}}}$$

- 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)



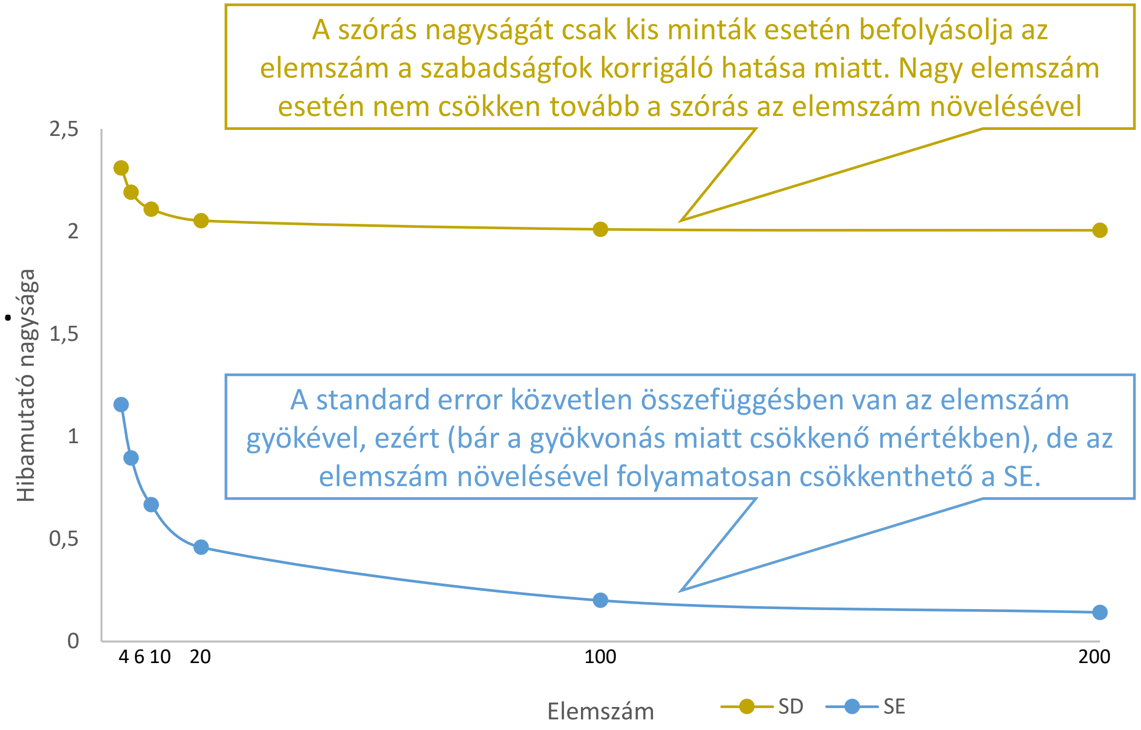

Mintaátlagok átlaga: 169
Mintaátlagok szórása: 0,535

Szóródási mutatók: Standard error

- **Mi számít nagy standard errornak?**
  - A szóráshoz hasonlóan ez is skálafüggő.

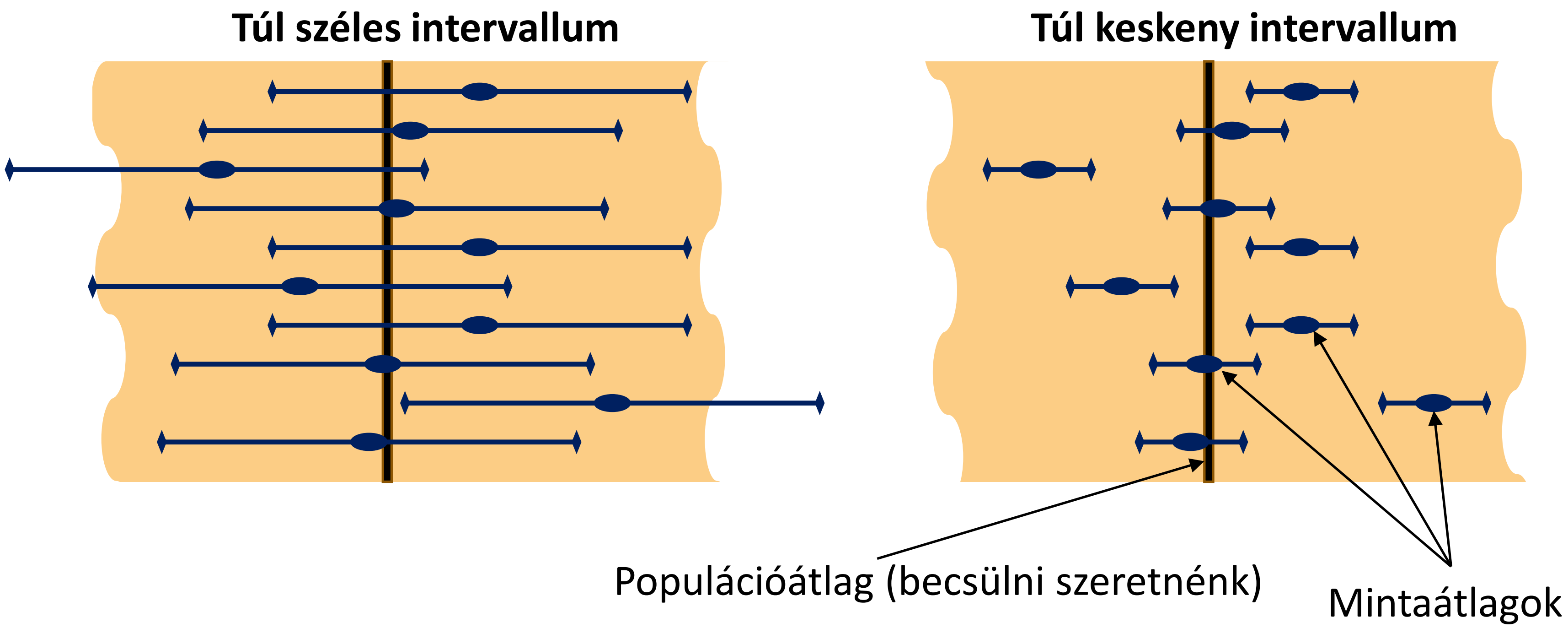
- **Relatív standard error** (**RSE**):
    - Számítása: $RSE = SE/\bar{x}$ százalékban kifejezve.
    - A relatív szóráshoz hasonlóan csak arányskála esetén használható!!
    - 25% alatt elfogadható a mérés

- **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 mintavételezésből adódó esetleges torzítások hatása, tehát várhatóan annál hasonlóbbak lesznek a mintaátlagok.


A szórás nagyságát csak kis minták esetén befolyásolja az elemszám a szabadságfok korrigáló hatása miatt. Nagy elemszám esetén nem csökken tovább a szórás az elemszám növelésével

A standard error közvetlen összefüggésben van az elemszám gyökével, ezért (bár a gyökvonás miatt csökkenő mértékben), de az elemszám növelésével folyamatosan csökkenthető a SE.

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. Megértéséhez ismerni kell a normál és t-eloszlást, így számolására az eloszlások megismerése után térünk vissza.

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
Elemszám	86	80
SD	8,737	5,823
SE	0,942	0,651
CI	1,846	1,276