Leíróstatisztikák

A statisztikai modell

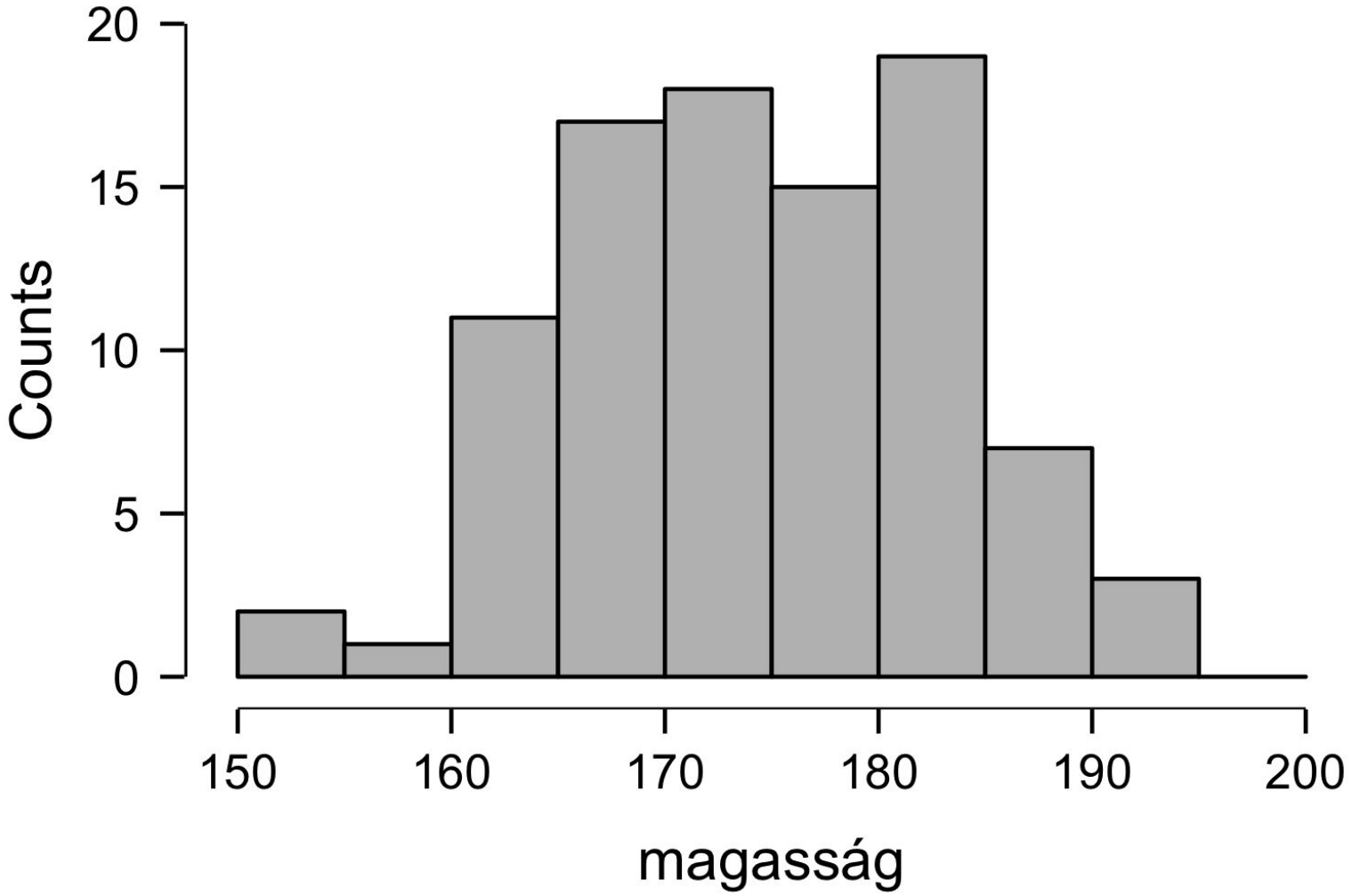
- A cél: a minta paramétereiből megbecsülni a populáció paramétereit
- A probléma: a mintavételi hiba

Eloszlás (disztribúció)

- A becsléshez használt adatok:
  - központi érték, mint pl az átlag
  - az adatok variabilitását jelző érték, pl. a terjedelem (max - min)
  - és az adatok eloszlása

- Eloszlás:
  - a változó egyes értékeinek gyakorisága
  - intervallum/arányskála és ordinális változókra értelmezhető - amiknek az értékei növekvő sorrendbe rendezhetők
  - hisztogrammal ábrázolható (distribution plot)
    - értékek a vízszintes tengelyen, gyakoriságok a függőleges tengelyen

JASP > Descriptives > Plots > Distribution plots. Variables: magasság



A normál eloszlás

- Középen van a leggyakoribb érték (**módusz, mode**)
- Középről a két oldal szimmetrikusan, haranggörbe alakban ível lefelé

A magasság normál eloszlást mutat a populációban. Miért van mégis két csúcsa a hisztogramnak? (**multimodális**)

Teszt: JASP Split

A normál eloszlás folyt.

- A normál eloszlású adatokat tudjuk a legmegbízhatóbban statisztika becslésre használni
- mert kiszámítható az egyes értékek előfordulási valószínűsége a populációban, pl. annak a valószínűsége, hogy egy nő 185 cm magas.
- Ami ehhez kell
  - normál eloszlás
  - az átlag
  - a **szórás**

A normál eloszlás folyt.

- Az átlag ≈ a módusz
- A szórás (Standard Deviation, Std.Dev, SD)

$$SD = \sqrt{\frac{\Sigma(x - M)^2}{N - 1}}$$

x = az adathalmaz egy eleme
M: átlag (mean)
N = elemek száma

A szórás

- Az átlagtól való átlagos eltérés
- Az eredeti mértékegységben mér
- Nagyobb szórás → nagyobb különbség az egyének között

JASP > Descriptives > Statistics

Használhatjuk standardizálásra:

$$z = \frac{x - M}{SD}$$

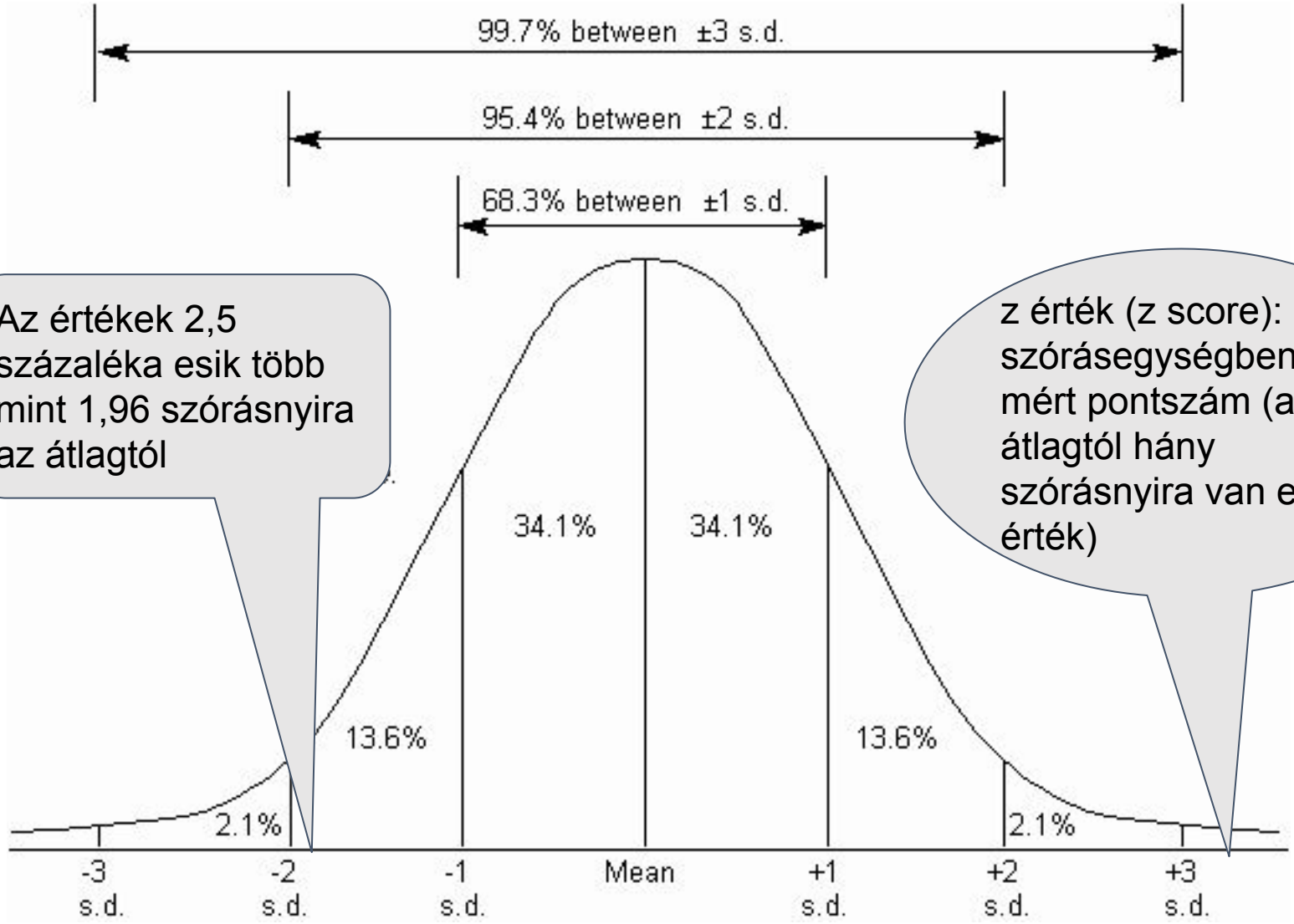
(Milyen magas lenne, ha ellenkező nemű lenne?)

Descriptive Statistics

	magasság	
	férfi	nő
Valid	46	47
Missing	0	0
Mean	181.2	168.9
Mode [a]	181.0	162.0
Std. Deviation	6.027	6.782

[a] More than one mode exists, only the first is reported

A normál eloszlás és a valószínűségek



Az értékek 2,5 százaléka esik több mint 1,96 szórásnyira az átlagtól

z érték (z score): szórásegységben mért pontszám (az átlagtól hány szórásnyira van egy érték)

Példa

- Tegyük fel, hogy a Margitsziget körbefutásának ideje normál eloszlású.
- Sára 45 perc alatt futja körbe a Margitszigetet
- A nők átlaga 38 perc
- A nők szórása 6,35 perc

$$z = \frac{(X - M)}{SD} = \frac{45 - 38}{6.35} = 1.1$$

- A standard normál eloszlás szerint, ha z = 1.1, akkor a nők 86,43%-a ennyi idő alatt vagy gyorsabban futja körbe a szigetet.
- Más szóval: 86,43% annak az esélye, hogy egy nő ennyi idő alatt vagy gyorsabban fussa körbe a szigetet.

- Egy idegen 17 perc alatt futotta körbe a szigetet.
- Mennyi annak az esélye, hogy egy nő ennyi idő alatt (vagy gyorsabban) fussa körbe?
- Innen már csak egy lépés: mennyi annak az esélye, hogy az idegen nő?

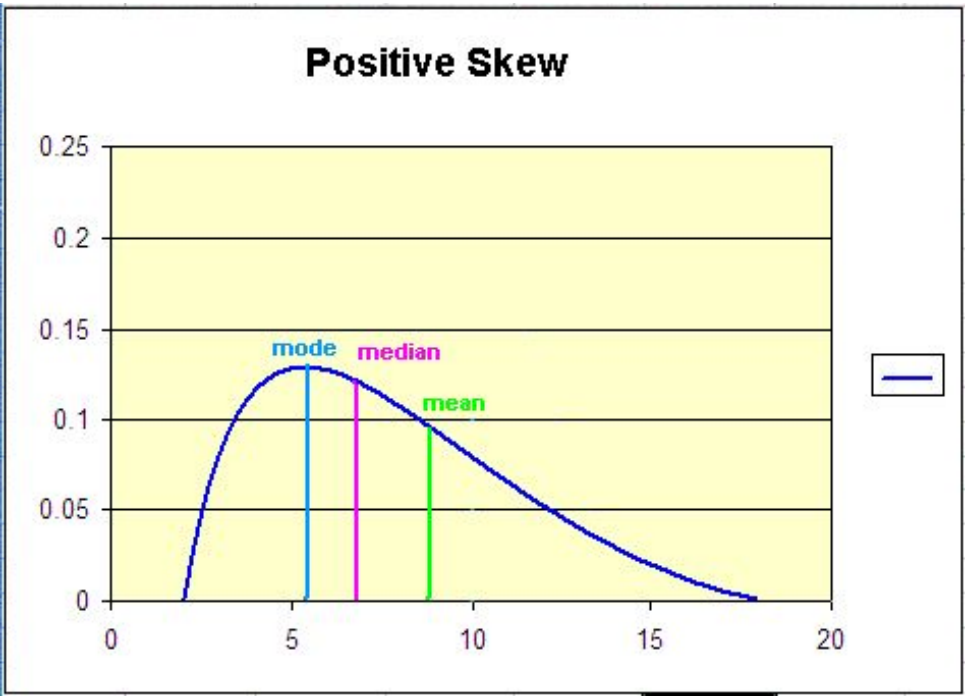
(A nők átlaga 38 perc, a szórás 6,35 perc)

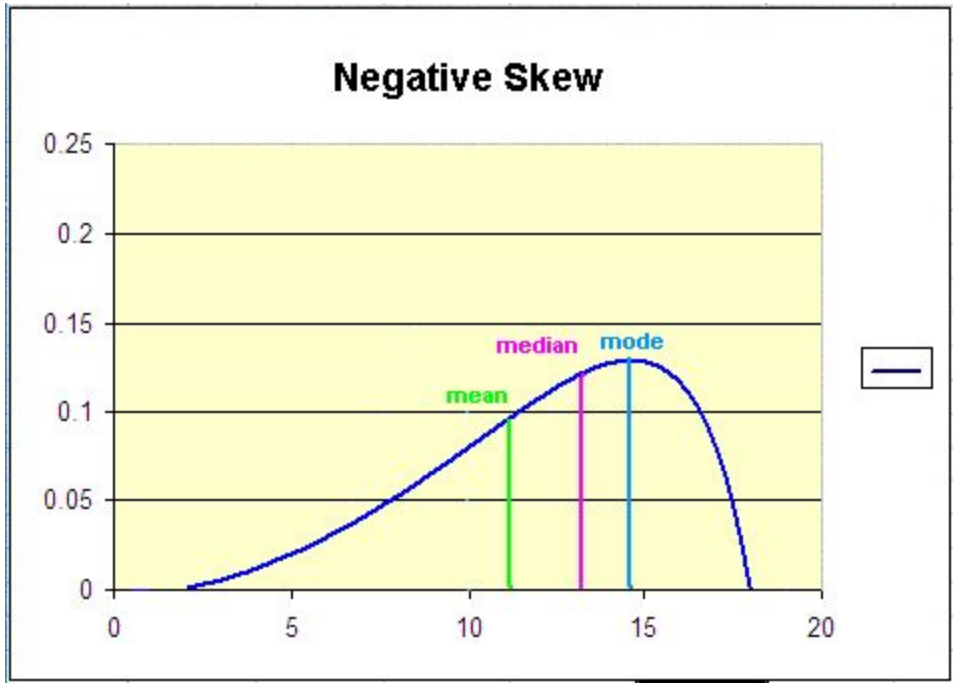
Példák normál és nem-normál eloszlásra

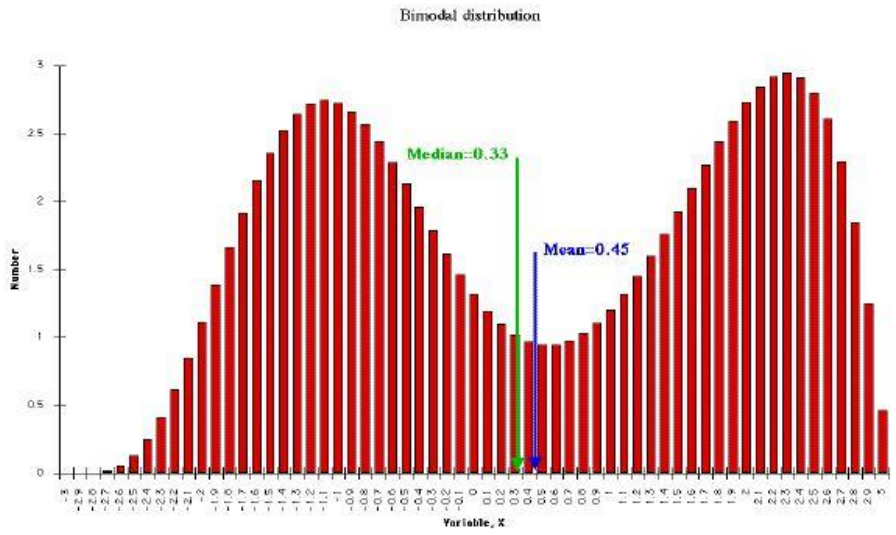
- Tipikusan normál eloszlás a populációban:
  - IQ
  - Személyiségtesztek
  - Reakcióidő

- Tipikusan nem normál eloszlás
  - Alkoholfogyasztás
    - valószínű **padlóhatás (**ellentéte: **plafonhatás**)
  - Facebook ismerősök száma

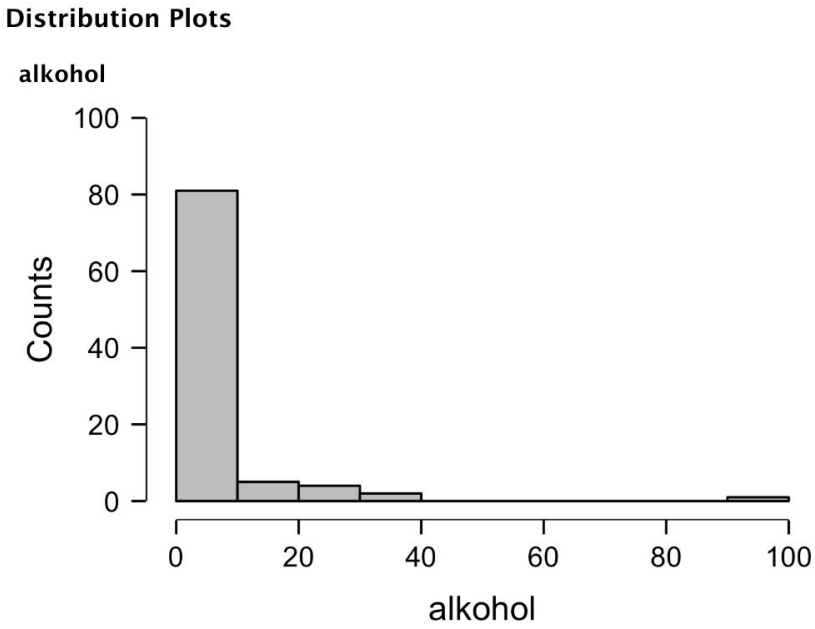
    Ha nem normál az eloszlás, nem tudjuk az előfordulási valószínűségeket


Positive Skew
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
mode
median
mean


Negative Skew
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
median
mode
mean


Bimodal distribution
Median=0.33
Mean=0.45
Number
Variable, X

Javítható eloszlás



- **Kiugró értékek** (outlier): egy-egy olyan érték, ami valamilyen irányban erőser eltér a többi értéktől (pl. alkohol)
  - elkülönül a hisztogramon
  - mérhető szórásegységben: több mint 2 vagy 3 szórásra az átlagtól
  - leggyakrabban kvartilisekhez viszonyítva mérik

Kvartilisek (quartiles)

- A középérték és variabilitás másik mérési rendszere
- Használható ordinális adatokra is
- Megbízhatóbb nem-normál eloszlás esetén

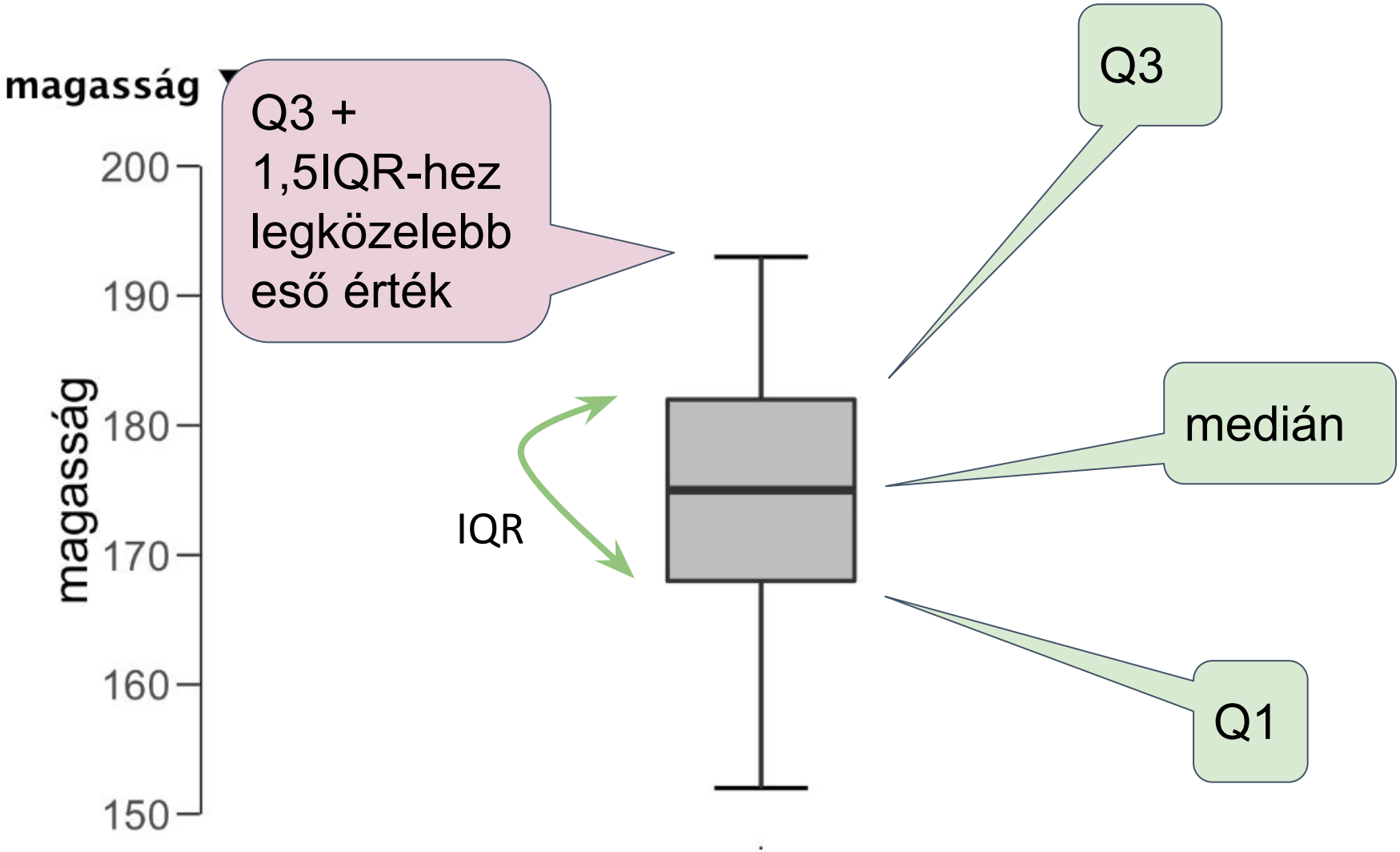
- A **medián (**median**):** középső érték
- Percentilisek: az 1. percentilis az az érték, ami alá az adatok egy százaléka esik
- Első **kvartilis (Q1)** = 25. percentilis: az az érték, ami alá az adatok egynegyede esik
- Második kvartilis (Q2) = 50. percentilis = medián
- I**nterkvartilis tartomány** (Interquartile Range, IQR): az első és harmadik kvartilis különbsége (az adatok középső 50%-a esik bele)

JASP > Descriptives > Statistics

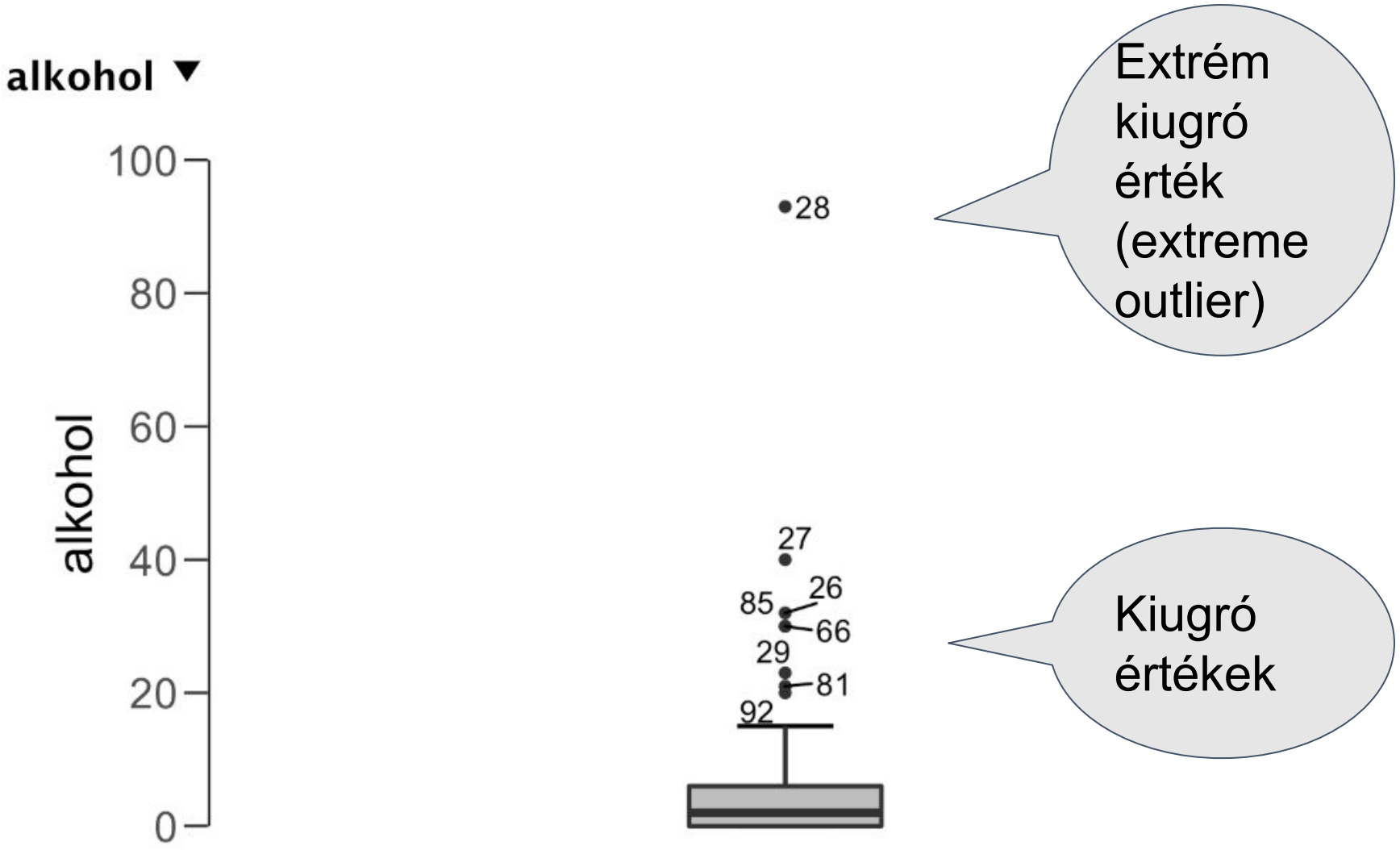
Descriptive Statistics ▼

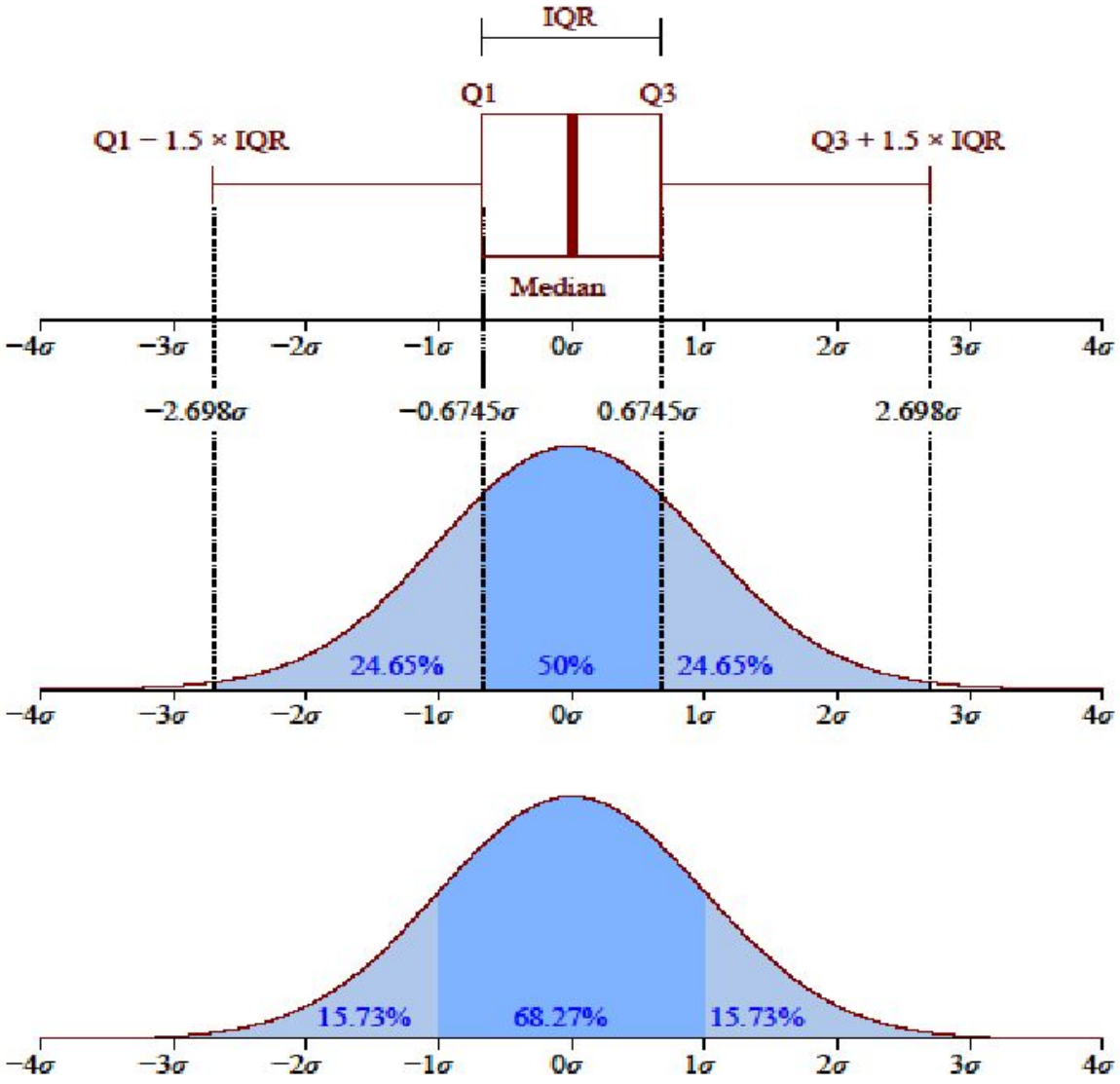
	alkohol
Valid	93
Missing	0
Mean	5.957
Median	2.000
Mode	0.000
Std. Deviation	11.91
IQR	6.000
25th percentile	0.000
50th percentile	2.000
75th percentile	6.000

A dobozdiagram (boxplot) (JASP > Descriptives > Plots)



JASP > Descriptives > Plots - Label Outliers





A doboz-diagram és a normál eloszlás

Összefoglalva: középértékek

- Átlag:
  - Intervallum/arányskála adatok
  - Nagyon érzékeny a kiugró értékekre
- Medián:
  - A középső érték, 50. percentilis, 2. kvartilis
  - Intervallum/arányskála adatok és ordinális adatok
  - Kevésbé érzékeny a kiugró értékekre
- Módusz:
  - A leggyakoribb érték
  - Bármilyen adattípusra használható
  - Nem érzékeny a kiugró értékekre

Összefoglalva: variabilitás

- **Terjedelem (range)**:
  - A legalacsonyabb és legmagasabb érték közötti különbség
  - Intervallum/arányskála adatok
  - Nagyon érzékeny a kiugró értékekre
- **Szórás (Standard Deviation)**:
  - Átlagtól való átlagos eltérés
  - Intervallum/arányskála adatok
  - Elég érzékeny a kiugró értékekre
  - Normál eloszlás esetén az adatok kb 68%-a esik az átlagtól 1 szóráson belülre
- **Interkvartilis tartomány (Interquartile Range)**:
  - Az 1. és kvartilis és 3. kvartilis közötti különbség
  - Nem túl érzékeny a kiugró értékekre
  - Az adatok 50%-a esik a tartományba

Mi legyen a kiugró értékekkel

Ha a kiugró érték

- nagyon eltolja az átlagot (a mediánhoz és móduszhoz képest)
- eltorzítja a hisztogramot
- nélkül az eloszlás normál eloszlás lesz

Akkor ki lehet zárni a további elemzésekből

JASP szűrő (törölni nem szabad!)

A kiugró érték definiálható SD-vel:

alkohol < (mean(alkohol) + (3 * σ alkohol))

vagy IQR-on keresztül a dobozdiagramból:

ksz ≠ 28

Vissza a normál eloszláshoz és z értékekhez: a z teszt

Normál eloszlás ellenőrzése

- Ha lehet tudni, hogy a populáció normál eloszlású
- Ha a hisztogram szimmetrikus harangalakú, nem dől egyik irányba sem
- Ha a dobozdiagram szimmetrikus, és nem mutat extrém kiugró értékeket
- Ha a **Shapiro-Wilk** teszt szerint normál az eloszlás:

Normál eloszlás ellenőrzése: Shapiro-Wilk

- A **Shapiro-Wilk** teszt:
  JASP > Descriptives > Statistics > Dispersion > Shapiro-Wilk test
  - Matematikusokról elnevezve
  - Intervallum/arányskála adatok eloszlását teszteli
  - Null hipotézis: az adatok eloszlása nem tér el a normáltól

Descriptive Statistics ▼

	alkohol	magasság
Valid	93	93
Missing	0	0
Mean	5.957	175.0
Median	2.000	175.0
Std. Deviation	11.91	8.896
Shapiro-Wilk	0.4901	0.9841
P-value of Shapiro-Wilk	2.139e −16	0.3183

a p érték: annak valószínűsége, hogy az eloszlás NEM tér el a normáltól

Ismét z értékek és valószínűségek

- Ha normál eloszlásunk van, meg tudjuk becsülni annak a valószínűségét, hogy egy adott pontszám a minta által reprezentált populációból jön (pl a 17 perces futó a nők populációjából)
- De minket inkább az szokott érdekelni, hogy egy minta (a kísérleti csoport) egy adott populációból (a kontroll populációból) jön-e
- Tehát nem egyetlen pontszámot vetünk össze egy eloszlással, hanem egy mintának az átlagát a populációból származó minták átlagainak az eloszlásával
- A populációból származó minták átlagainak az eloszlását a populáció átlagából és szórásából és a minta elemszámából becsüljük meg.

A z teszt

A z teszt annak a valószínűségét becsüli meg, hogy egy minta egy adott populációból jön

A sima z érték képletét igazítjuk: $$z = \frac{x - M}{SD}$$

A populáció szórása helyett az ún. **Standard Hibát** (**Standard Error, SE**) vesszük, ahol SD a populáció szórása, N pedig a minta elemszáma:

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

És z, ahol M a minta átlaga és μ a populáció átlaga:

$$z = \frac{M - \mu}{SE}$$

Példa

Teszteljük le, hogy a mintánk különbözik-e a nagyvilágtól a Facebook ismerősök számát tekintve.

- A Facebook használók között az átlagos Facebook ismerősök száma 2018-ban: 338, SD = 224.
- Szűrjük ki a Facebook-ot nem használókat a mintánkból
- Ellenőrizzük az eloszlást
- Számoljuk ki a Standard Hibát (populáció SD osztva a megszűrt minta elemszámának gyökével)
- z = ((minta átlaga – 338)/SE)
- Mi a valószínűsége ennek a z értéknek?