Leíró statisztika.
Ábrázolás az R-ben

Leíró statisztika

Definíciója: populáció egy ismert részhalmazára vonatkozó megfigyelések leírása és összegzése.

Jelentősége:

- nominális adatok esetén,
- exploratív tanulmányokban, ahol nincsenek konkrét hipotéziseink,
- adatok elsődleges felmérése,
- tesztek létjogosultságának ellenőrzése (pl. eloszlás).

Jellemzők

- Gyakoriság,
- eloszlás,
- középérték,
- szóródás (NEM szórás!).

Ábrázolás táblázatban vagy grafikonokon.

Gyakoriság

- abszolút érték (ha elemszámok megegyeznek),
- arány (elemszám/összes), százalékos arány (arány*100),
- kumulatív gyakoriság: előfordulás bizonyos érték ALATT.
- Értékeket gyakran csoportokba, azaz kategóriákba vonjuk össze.

A gyakoriságot gyakran csoportokra adják meg, pl. a 21, 23, 35, 43 évesek 21–30, 31–40, 45–50 stb. éves csoportokba rendezve.

R-funkciók:
`table(x)`, `table(x/length(x))`, `table(x/length(x)*100)`, `prop.table(x)`, `cumsum(table(x))`

Ábrázolás

- kördiagram,
- oszlopdiagram,
- hisztogram.

R-funkciók:
`pie(table(x))`, `barplot(table(x))`, `hist(x)`

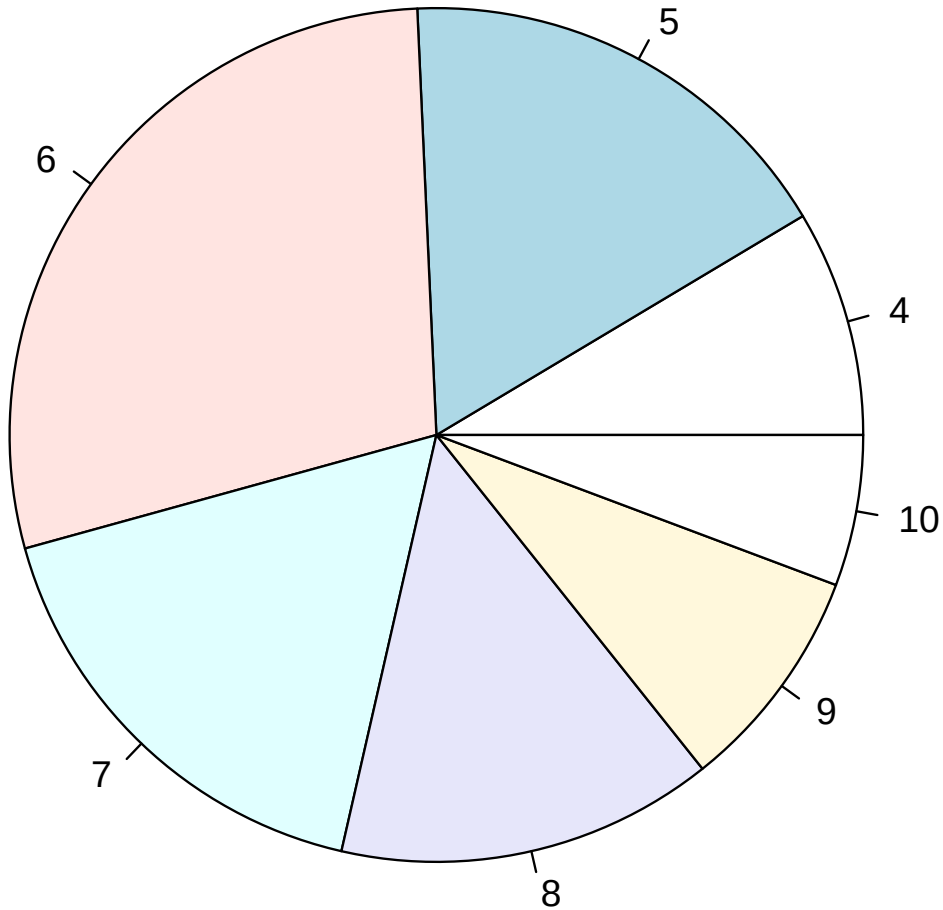
Példa

Angol növény- és állatnevek hosszúsága betűkben megadva.

típus	elemszám
növény	35
állat	46

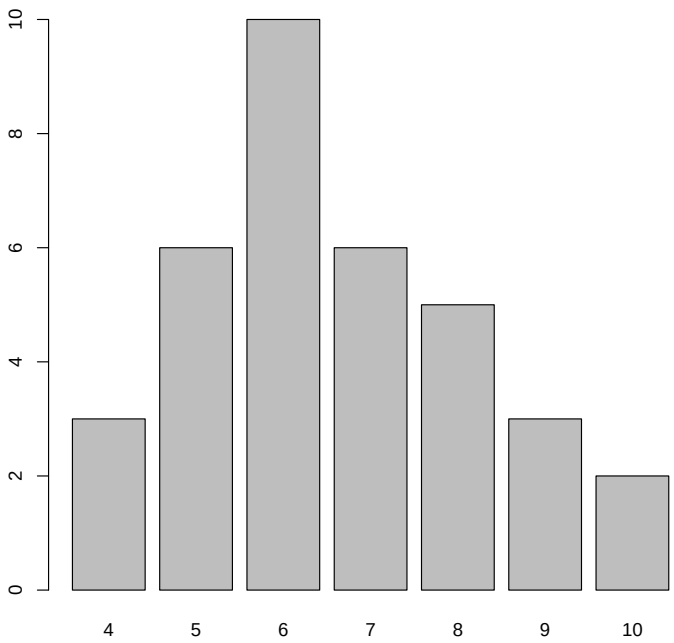
Kördiagram

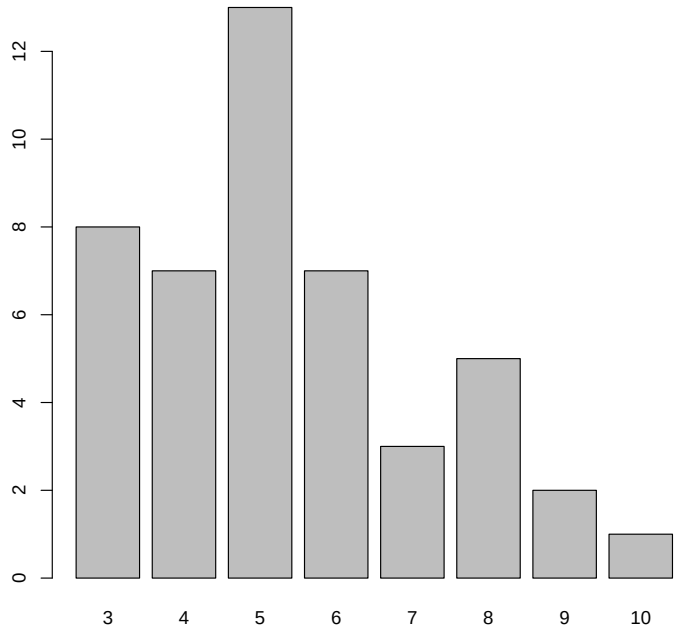
Növénynevek betűszámának gyakoriságai



Oszlopdiagram

Növény- (bal), és állatnevek (jobb) betűszámának gyakoriságai (hány betűből áll a szó):



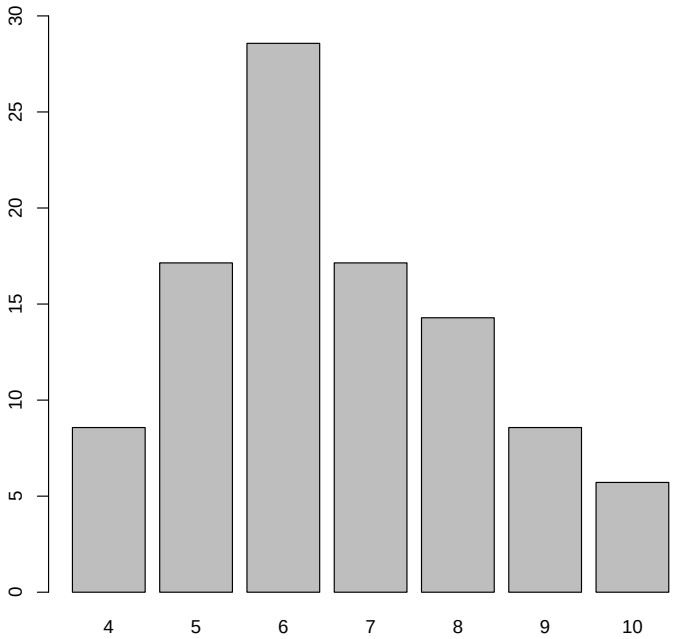


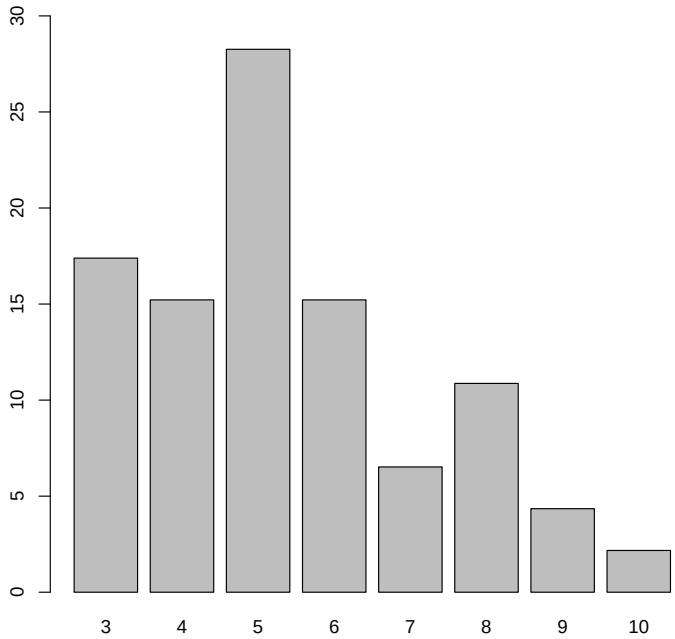
Jellemző felhasználás: nominális adatok, ordinális diszkrét adatok, kategorizált adatok.

Probléma: két csoportban eltérő elemszám! ($n_n = 35$, $n_a = 46$)

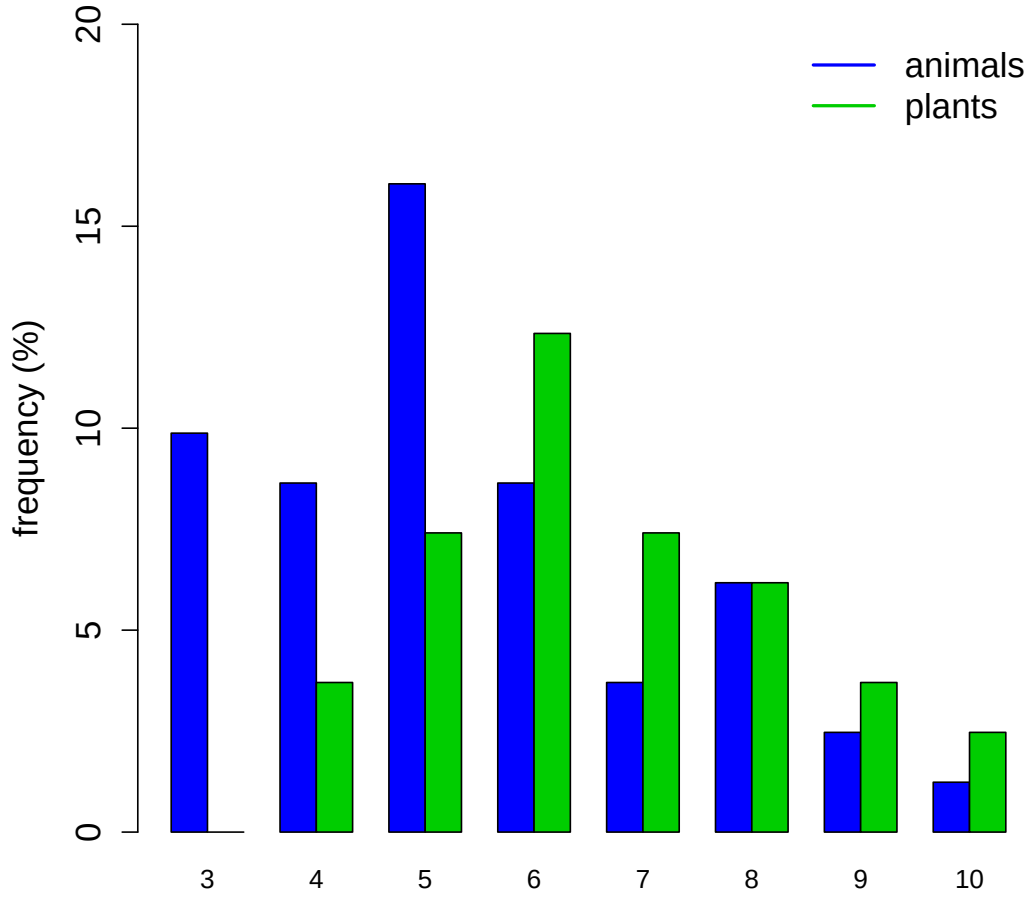
Oszlopdiagram százalékos arányokkal

Növény- (bal), és állatnevek (jobb) betűszámának gyakoriságainak százalékos aránya:





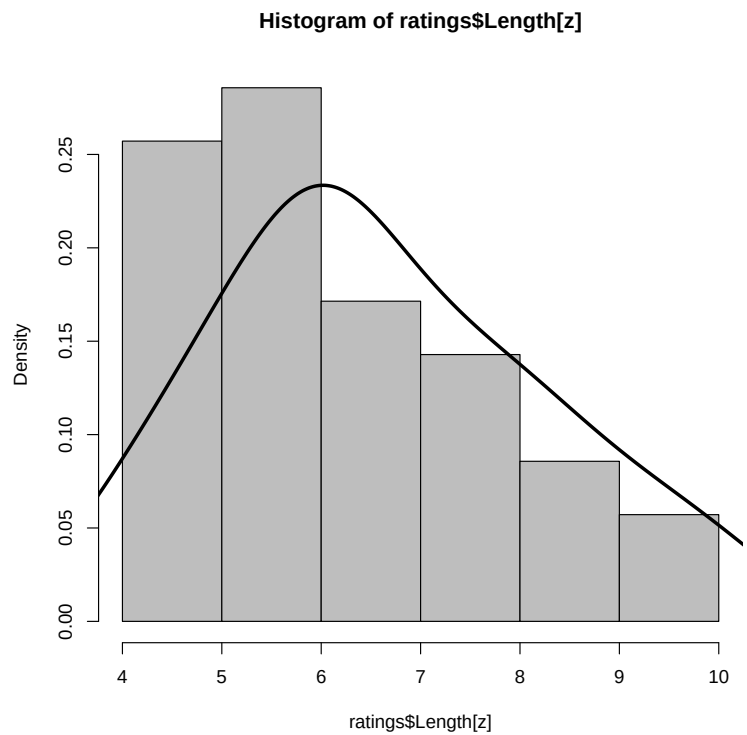
Két minta egy oszlopdiagramban

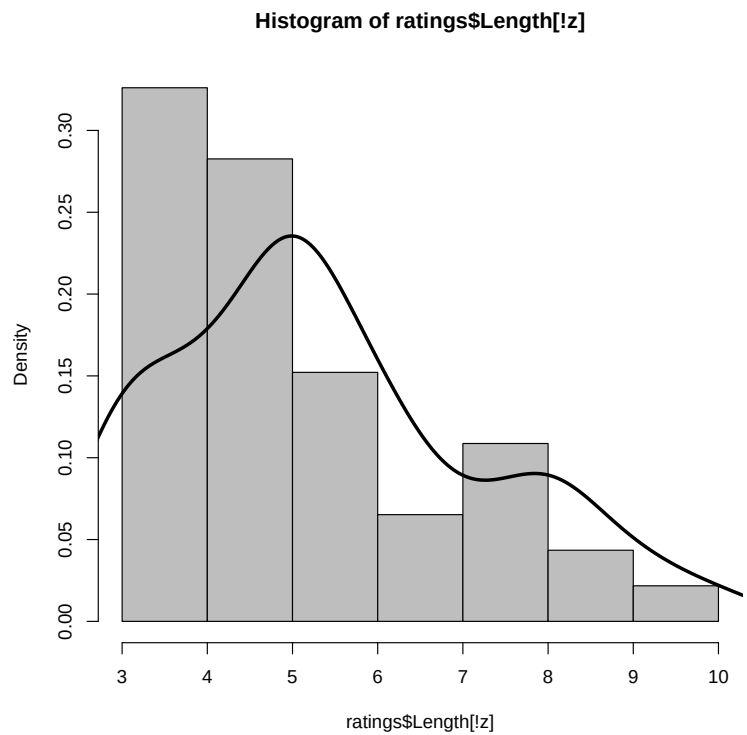


Előny: adatok jobb összehasonlíthatósága.

Hisztogram

Növény- (bal), és állatnevek (jobb) betűszámának sűrűsége:





Felhasználás: legalább ordinális skála.

Eloszlás

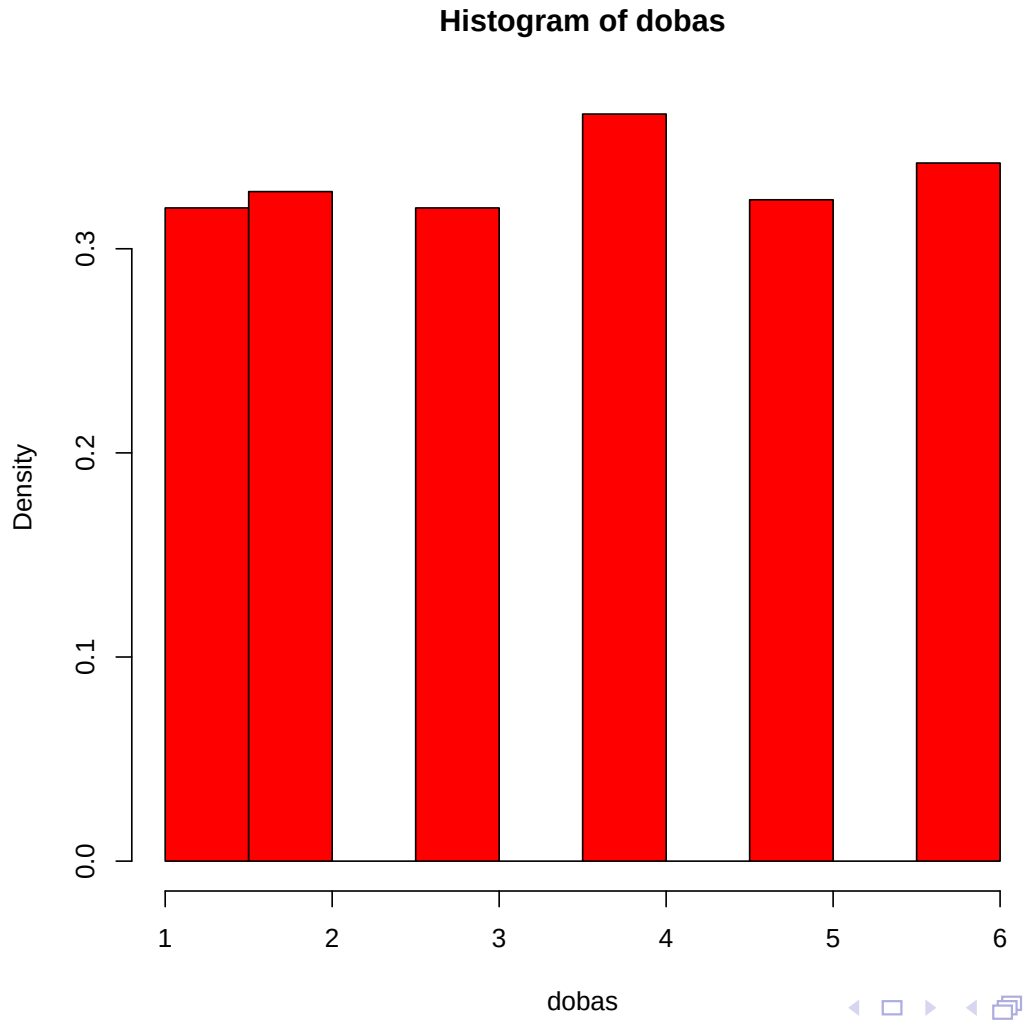
- **Definíció:** sorrendbe állított elemek milyen gyakorisággal fordulnak elő.
- **Felhasználás:** ordinális skálától felfelé.
- **Előállítás:** folytonos vagy diszkrét értékek közötti interpoláció.
- **Jelentőség:** valószínűségi statisztikai elemzés alapja.

R-funkciók:
`hist(x,frequency=FALSE)`: arányos gyakoriságok,
`plot(density())`: sűrűségfüggvény.

Eloszlás típusai

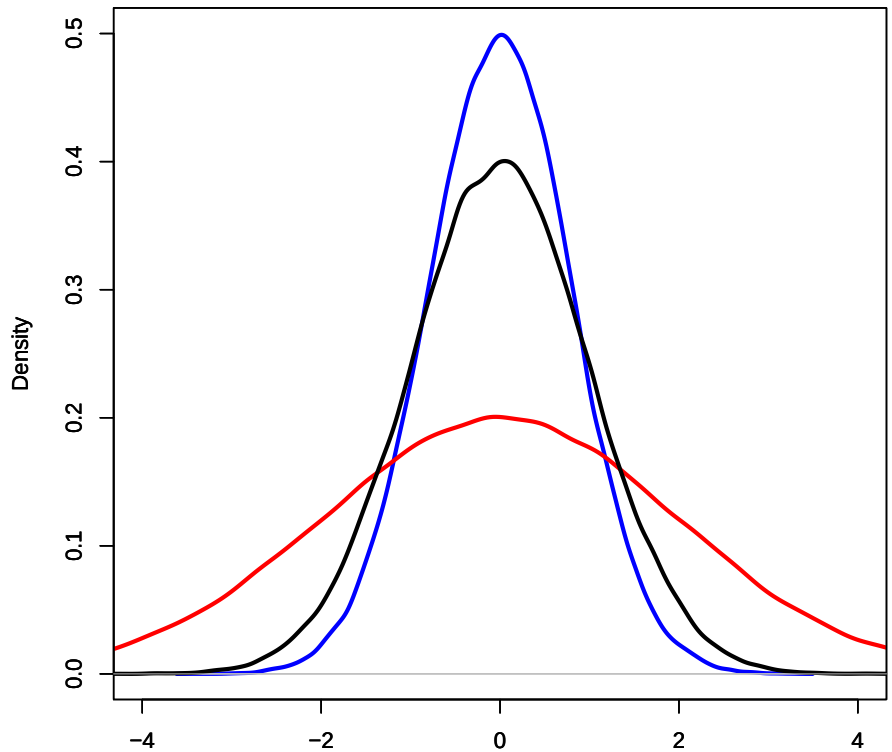
Egyenletes eloszlás

pl. dobott számok gyakorisága



Eloszlás típusai

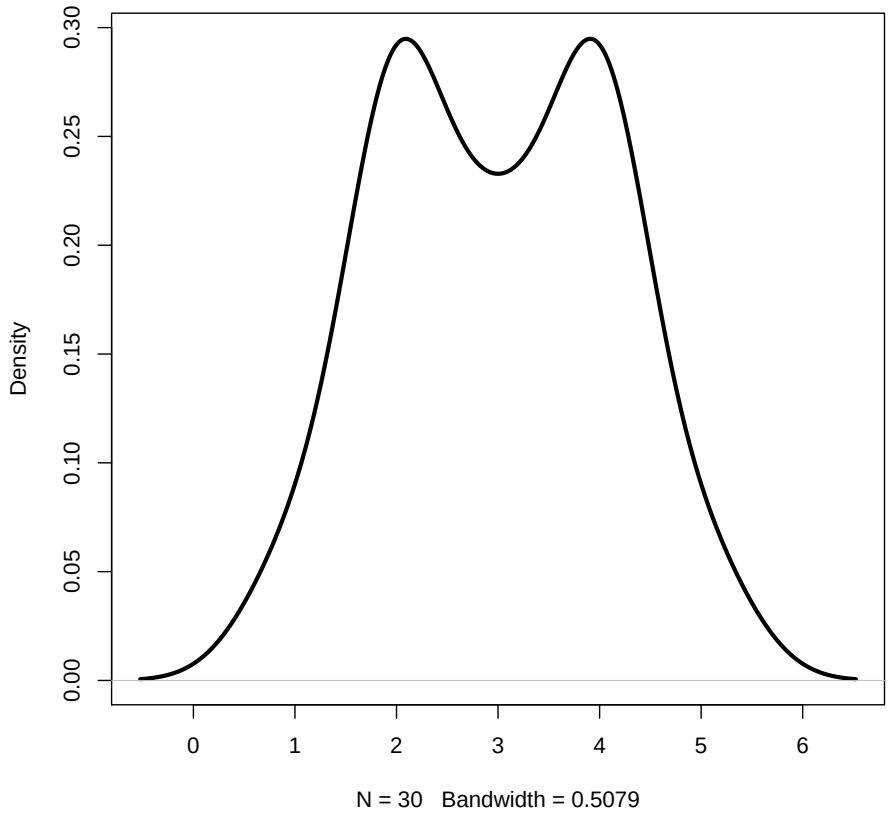
Unimodális: egy módusza van.



Az eloszlás lehet szimmetrikus vagy aszimmetrikus, laposabb vagy csúcsosabb.

Eloszlás típusai

Bimodális: két módusza van.



Bi- és multimodális eloszlásra a legtöbb statisztikai teszt nem végezhető el!

Szóródás: terjedelem

Szóródás/diszperzió: az adatok egymástól való távolsága. Jelzi az eloszlás szélességét. Pl. az unimodális eloszlást szemléltető görbék közül a piros a legnagyobb szóródású, a kék a legkisebb.

Terjedelem: a legkisebb és legnagyobb érték különbsége.
Ordinális és metrikus skálára egyaránt alkalmazható, de érzékeny a szélső értékekre.

Átlagos Facebook-felhasználó ismerőseinek száma:
$\texttt{terjedelem} = 724 - 113 = 611$

Extravagáns Facebook-felhasználó ismerőseinek száma:
$\texttt{terjedelem} = 5439 - 11 = 5428$

Probléma: az első érték valószínűleg jobb becslése a populációra jellemző terjedelemnek, mert az 5000 fölötti ismerőssel rendelkező ismerősök ritkák.

Interkvartilis tartomány

- Jelentőség: ha ordinális skála vagy nem szimmetrikus eloszlású parametrikus adatok.
- Interkvartilis tartomány: az X változó értékskálájának az a középen elterülő övezete, ahol a populáció 50%-a található.
- Folytonos változó esetén: negyedelő vagy 1. kvartilis és felső vagy 3. kvartilis közé esik.
- Interkvartilis félterjedelem: (K3-K1)/2, vagyis az 1. és 3. kvartilis ÁTLAGA – csak szimmetrikus eloszlás esetén egyezik a mediánnal.

Kétféle Facebook-felhasználó:

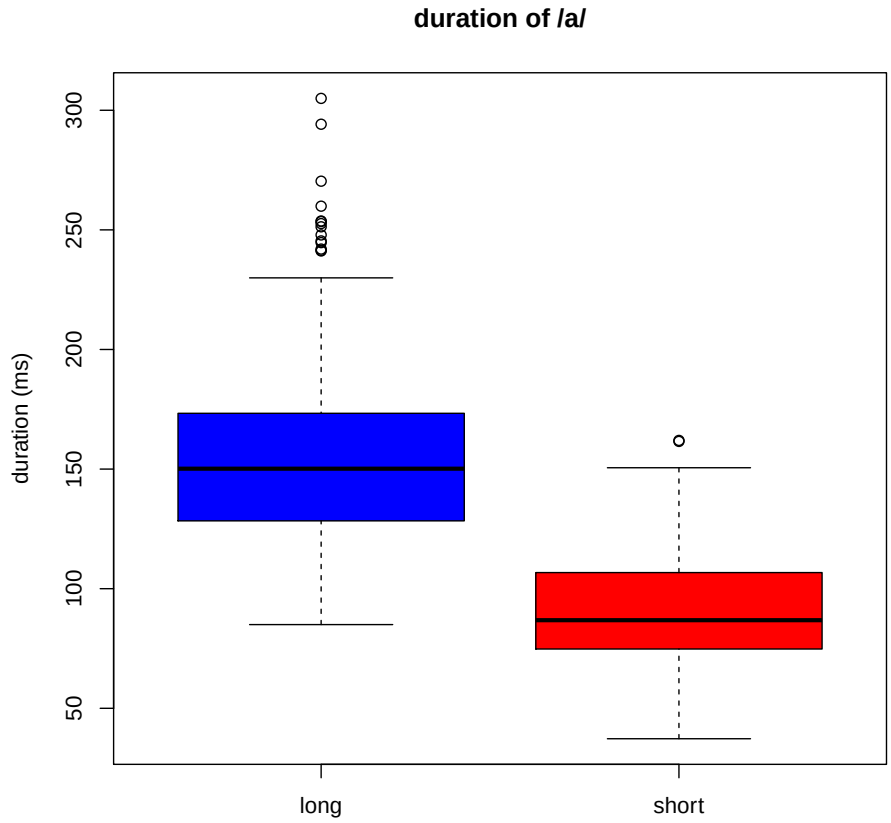
1.: 113 149 178 196 269 382 388 467 546 682 724

2.: 11 149 178 196 269 382 388 467 546 682 5439

Interkvartilisek kevésbé érzékenyek a szélső értékekre.

Interkvartilisek ábrázolása

Dobozdiagram (boxplot).
Szerkezete: (1) megfigyelések sűrűsége a középső 50%-os tartományban, (2) eloszlás szimmetriája.



Pontok: szélső értékek.

R

Adatok beolvasása az R-be

Adatokat többnyire más szoftverrel állítottuk elő. Kezdő szinten például az Excelben.

Töltsünk le egy xlsx-fájlt innen:
http://clara.nytud.hu/~mady/courses/statistics/materials/soc.xlsx

Az R-be csak szöveges fájlokat tudunk beolvasni, MS-Office és más, saját kódolású fájlokat nem (ez minden más szoftverre is igaz az Office-on kívül). Ezért az Excel-ből csv fájlként (comma-separated values) mentve alakítjuk szöveges fájllá a táblázatot.

Fontos szempontok:

- Vannak-e oszlopnevek?
- A decimális pont vagy vessző?
- Az oszlopokat hogyan válasszuk el egymástól az átalakításnál?

Problémák

Közép-európai kódolásban a decimális vessző, a programnyelvekben viszont angol mintára pont.

A csv fájl, ahogy a neve is mondja, alapértelmezetten vesszővel választja el az oszlopokat - ez a magyarban összetéveszthető lenne a decimális vesszővel.

Tabulátor mint oszlopelválasztás miért nem jó? Ha egy cella üres, az R nem látja, hogy ott két tabulátor van. Mint a legtöbb programban, a whitespace (tabulátor, szóköz) egyszer számolódik, akárhány is van belőle.

Javaslat a közép-európai kódoláshoz: oszlopok elválasztása pontosvesszővel. Mentés: soc.csv
`read.csv2("soc.csv")`
Ez megfelel a következő beállításoknak:
`read.table(file, header = TRUE, sep = ";", dec = ",")`

Néhány hasznos függvény

`ls()`: R munkamemóriában tárolt objektumok (változók).

`names(soc)`: oszlopban tárolt változók neve.

`head(soc)`: első hat adatsor.

data.frame változóira (oszlopaira) hivatkozás: soc$valtozo, ahol valtozo az oszlop nevével azonos.

Feladat I

Adjuk meg a soc.csv fájl alapján:

- Hány személy adatait tartalmazza a táblázat?
- Hány nő és hány férfi szerepel benne?
- Mi a résztvevők életkorának átlaga és mediánja?

Helynevek gyakorisága a soc.csv adatai alapján. Ábrázolás kördiagrammal, oszlopdiagrammal (abszolút és százalékos értékek).

Kor eloszlása hisztogrammal. Szóródás ábrázolása dobozdiagrammal. Mekkora az 1. és 3. interkvartilis, a medián és az interkvartilis félterjedelem?

Feladat II

Két dobókockával való 10, 100, 1000 dobálás összege: módusz, medián, átlag, ezek eloszlásának ábrázolása oszlopdiagrammal, hisztogrammal és dobozdiagrammal.