Leíró statisztika.
Adatok ábrázolása 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.

Jellemzők

- gyakoriság
- eloszlás
- középérték
- szóródás

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)`

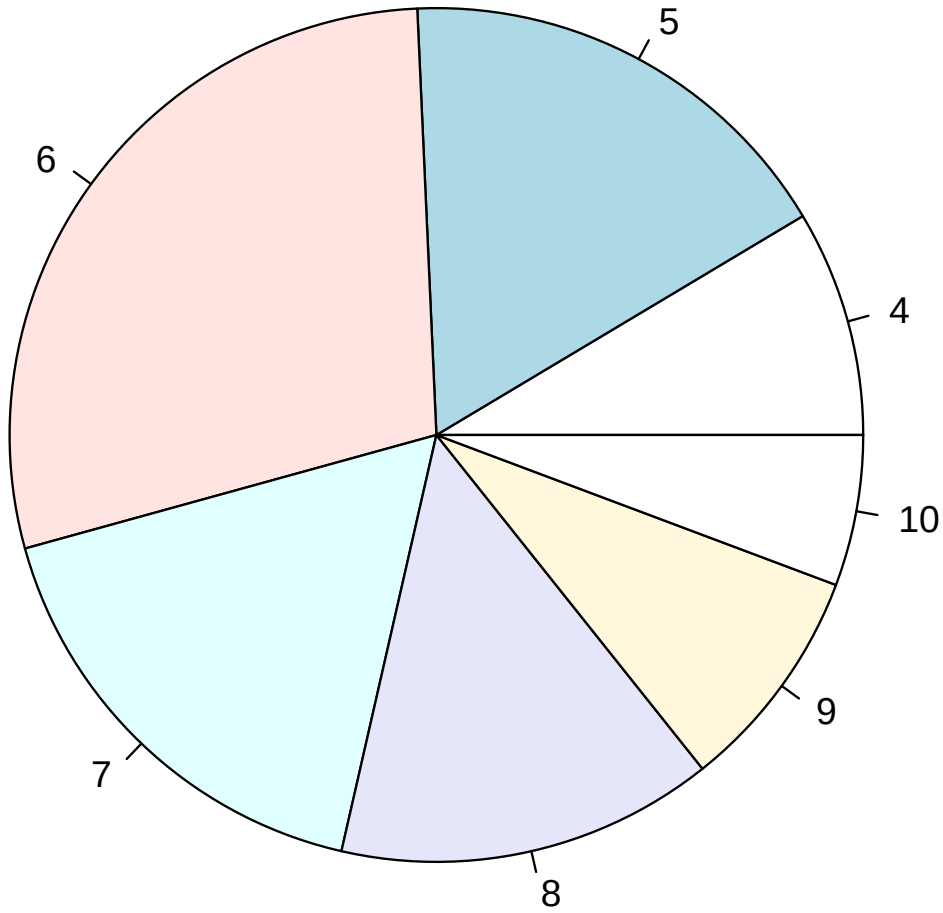
Példa

Angol növény- és állatnevek hosszúsága betűkben megadva.

típus	elemszám
növény	35
állat	46

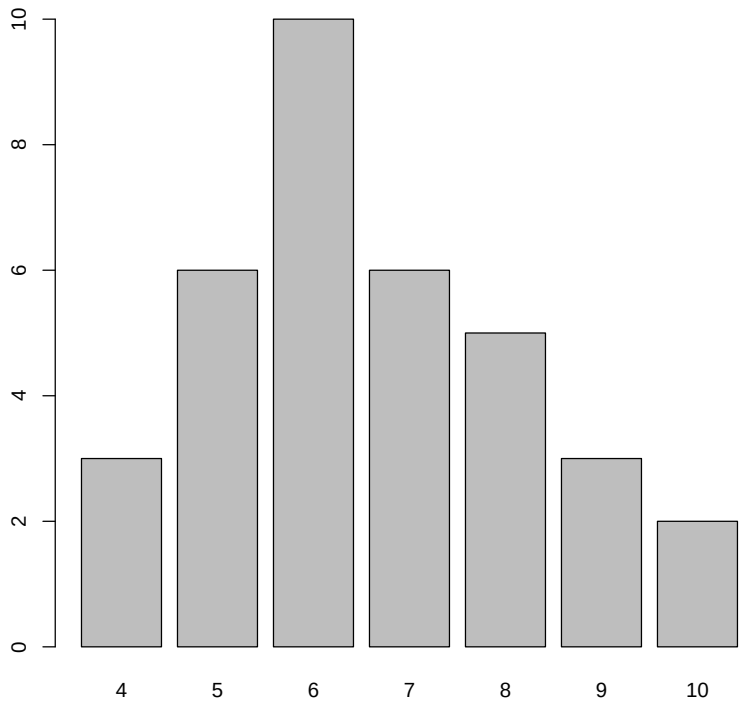
Kördiagram

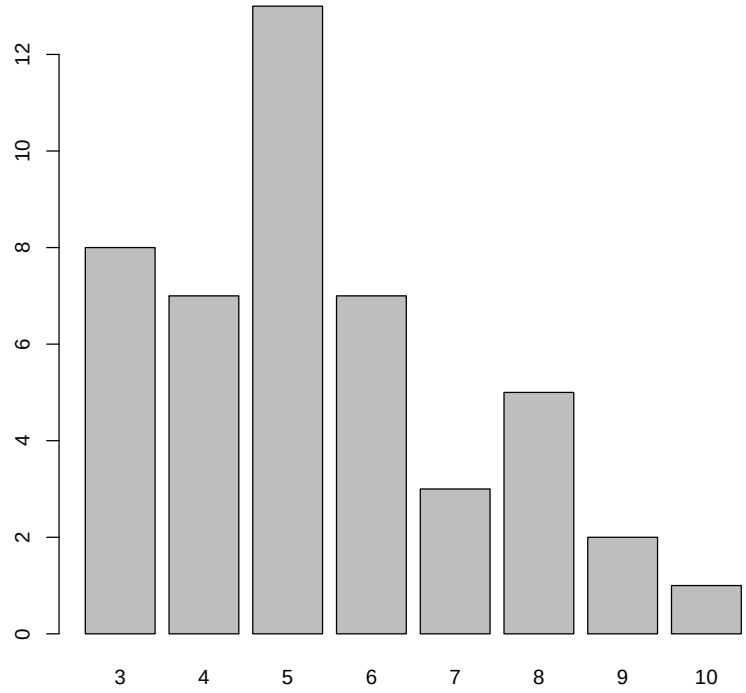
Növénynevek betűszámának gyakoriságai



Oszlopdiagram

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



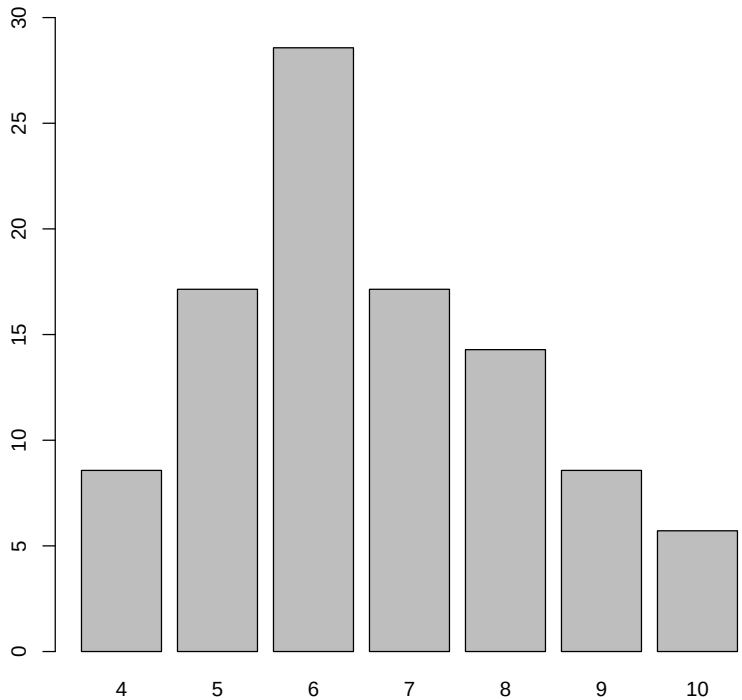


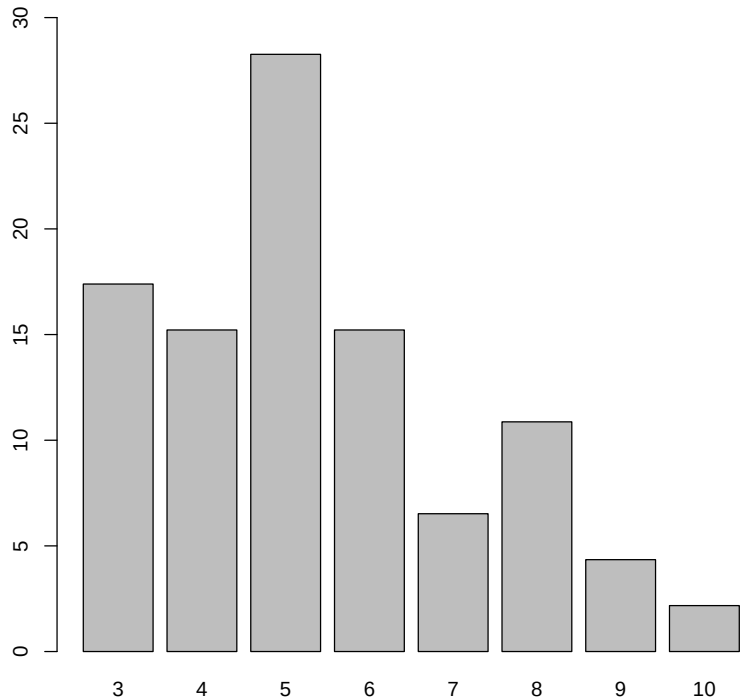
Jellemző felhasználás: nominális adatok, ordinális diszkrét adatok, kategorizált adatok.

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

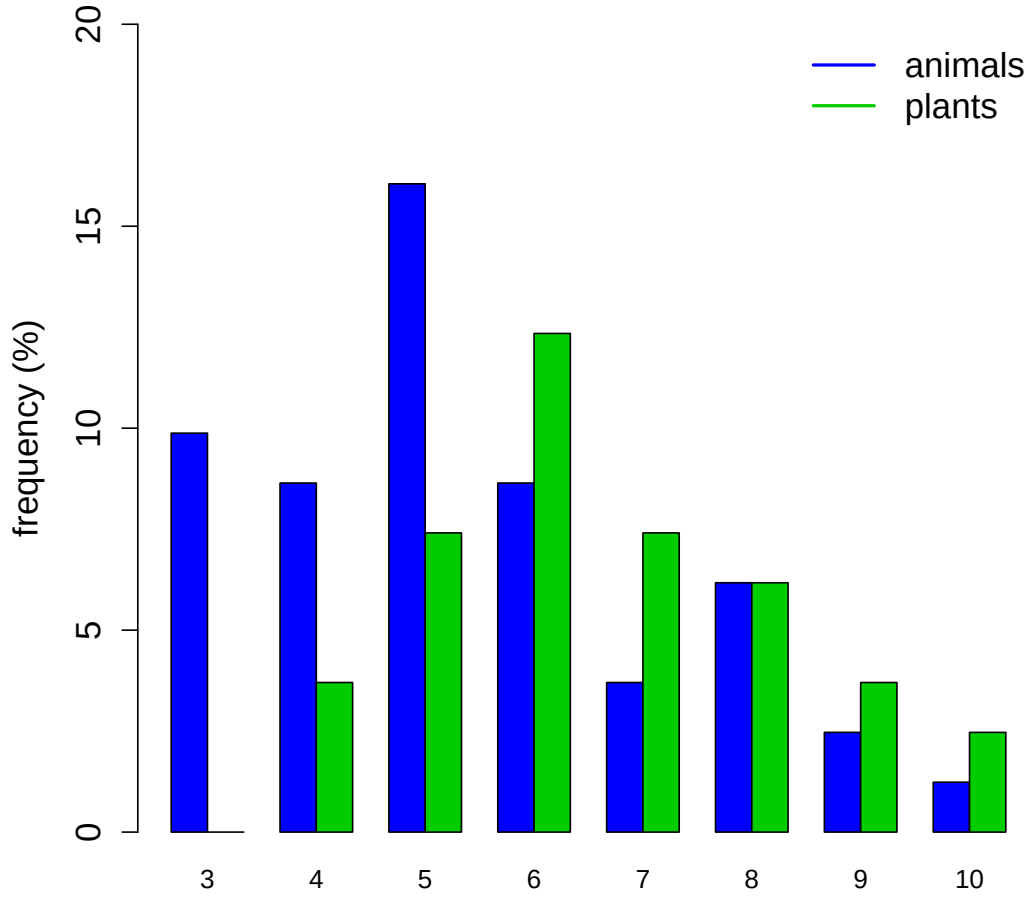
Oszlopdiagram százalékos arányokkal

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





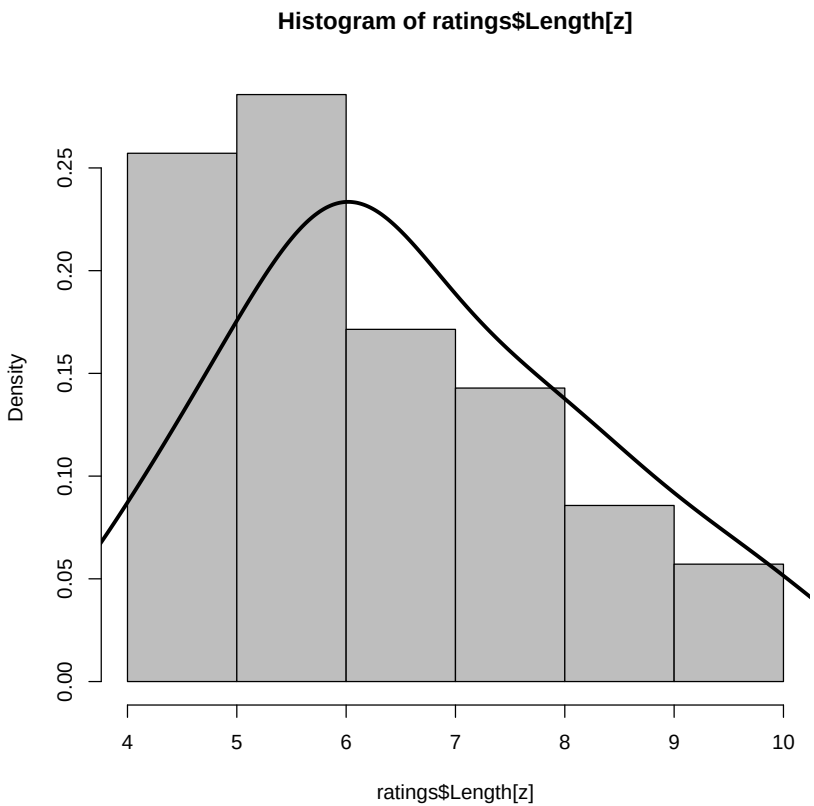
Két minta egy oszlopdiagramban

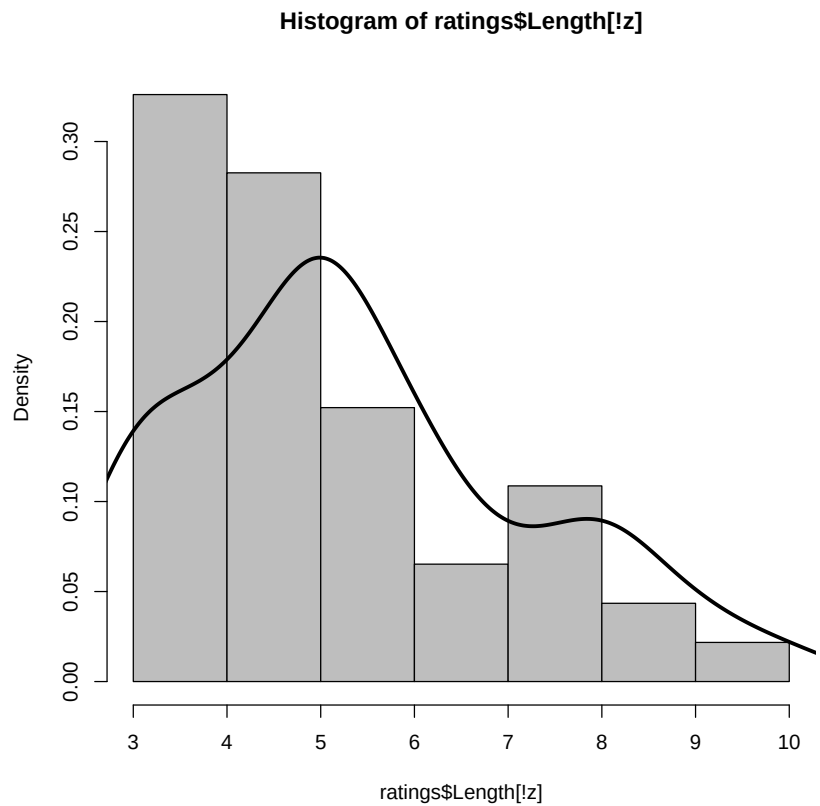


Előny: adatok jobb összehasonlíthatósága.

Hisztogram

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





Felhasználás: legalább ordinális skála.

Eloszlás

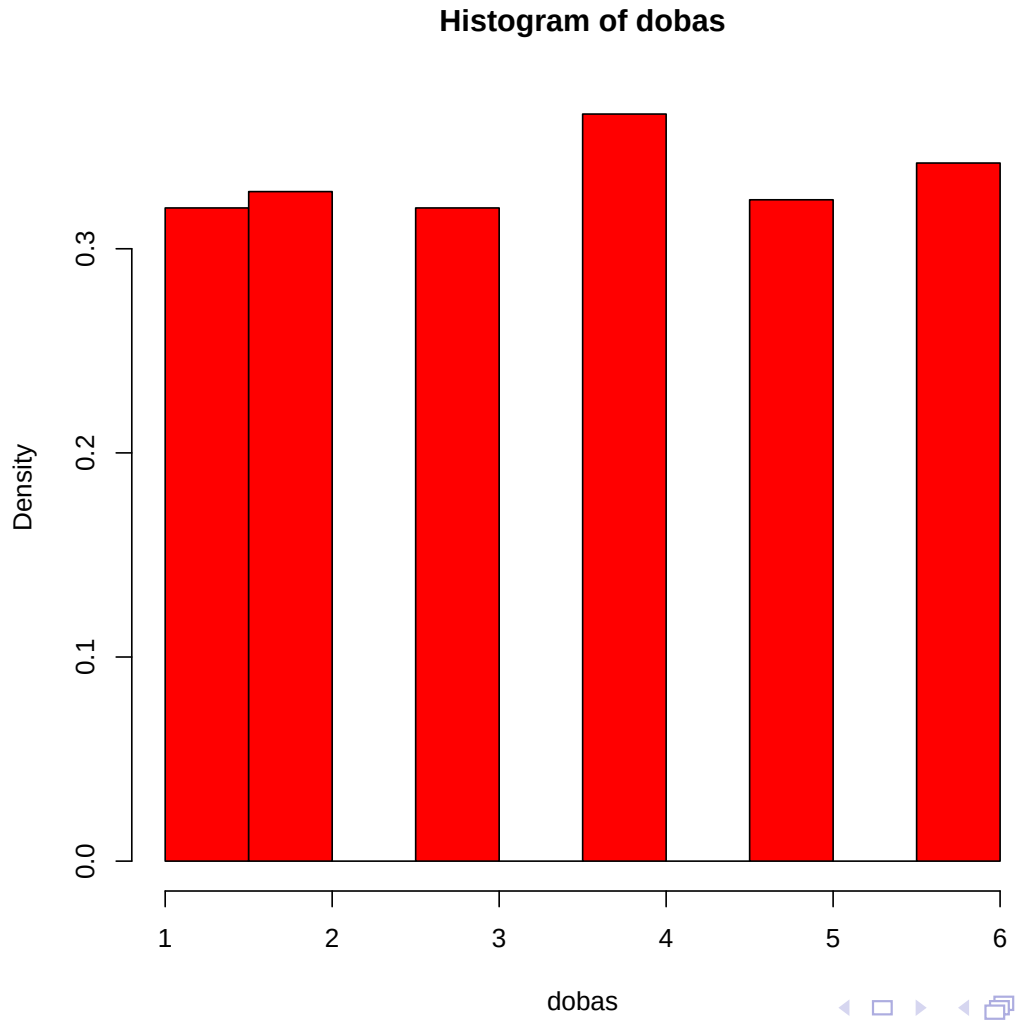
- **Definíció:** sorrendbe állított elemek milyen gyakorisággal fordulnak elő.
- **Felhasználás:** ordinális skálától felfelé.
- **Előállítás:** folytonos vagy diszkrét értékek közötti interpoláció.
- **Jelentőség:** valószínűségi statisztikai elemzés alapja.

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

Eloszlás típusai

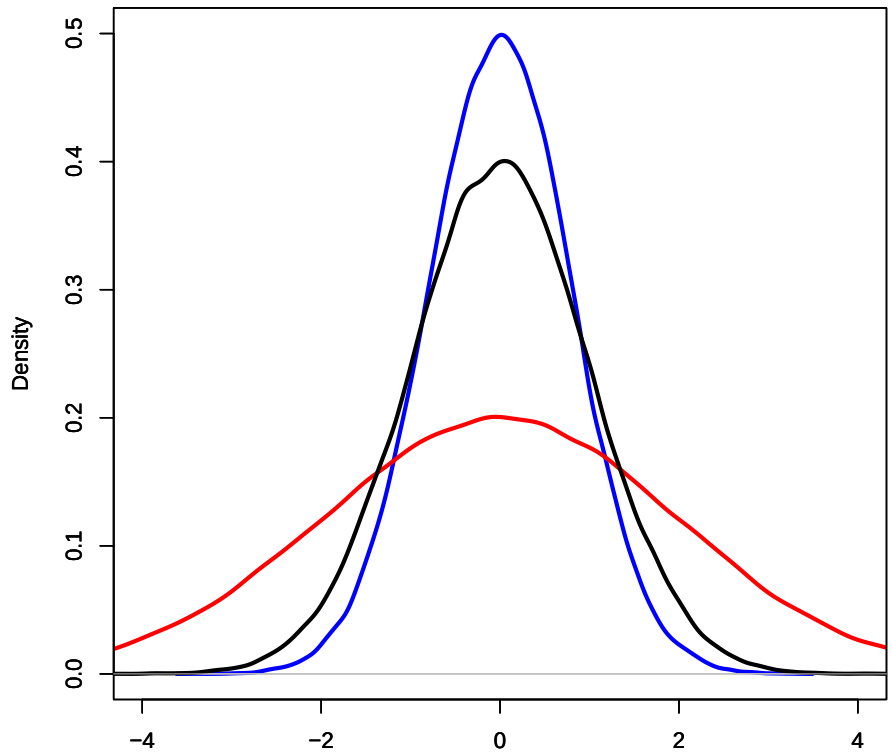
Egyenletes eloszlás

pl. dobott számok gyakorisága 



Eloszlás típusai

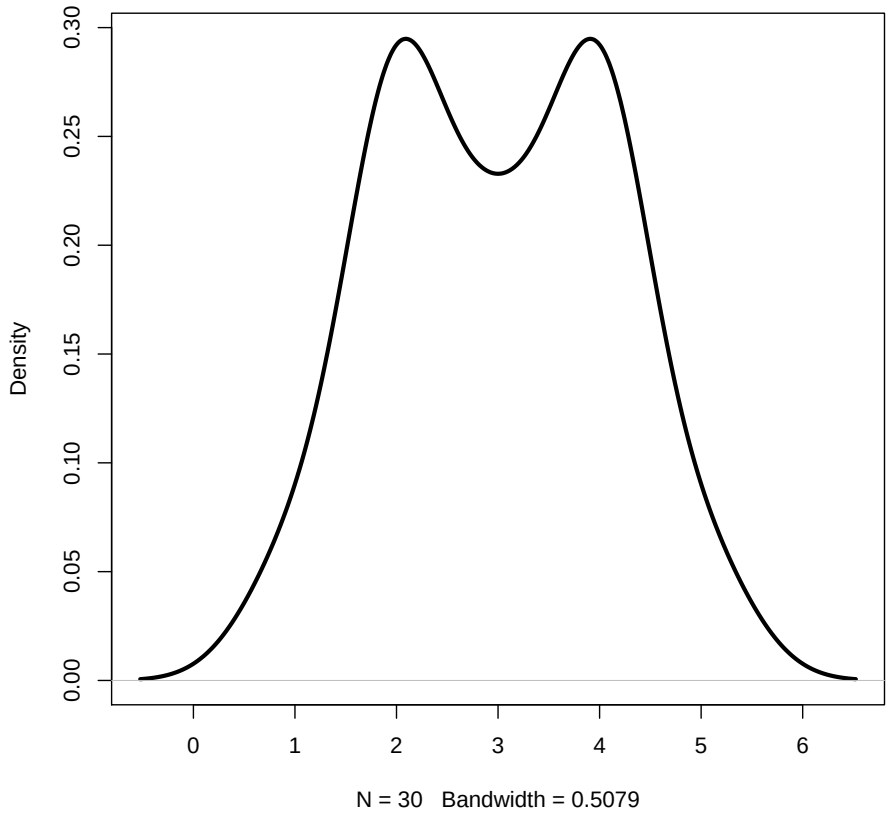
Unimodális: egy módusza van.



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 véletlenszerűen kiválasztott 11 ismerősének száma:
$\texttt{terjedelem} = 724 - 113 = 611$

Híres embereket ismerő Facebook-felhasználó 11 ismerősének 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.

Kétféle Facebook-felhasználó:

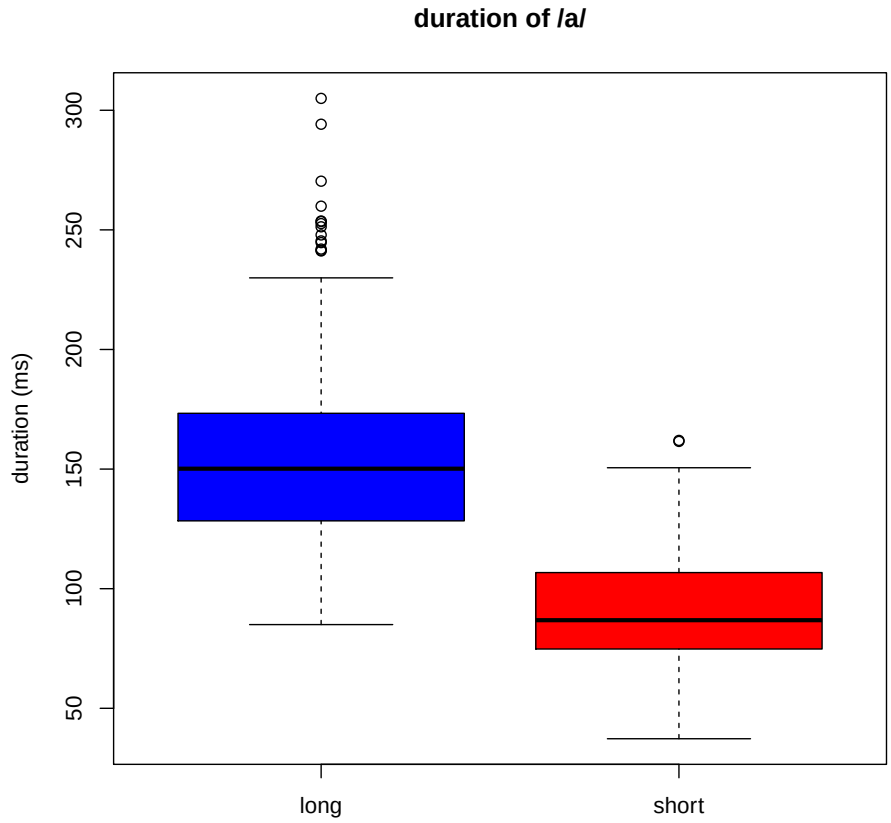
1.: 113 149 178 196 269 382 388 467 546 682 724

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 (E-Prime, Praat, manuális lejegyzés stb). Ezek beolvasása:
`read.table()`
`read.table(file, header = FALSE, sep = "", dec = ".")`
`header`: ha első sor eggyel kevesebb elemet tartalmaz, automatikus
`sep`: szóköz vagy tab, problémás lehet, ha vannak üres cellák. Pontosvessző megbízhatóbb.
`dec`: ha közép-európai kódolású szoftvert használunk, a decimális vessző! tehát `dec = ","`

Feladat: töltsünk le egy adatfájlt innen:
http://clara.nytud.hu/~mady/courses/statistics/materials/soc.dat

Érdemes a felhasználói név alatt létrehozni egy R könyvtárat erre a célra!

Fájl beolvasása Linuxban

Adatfájl helye: `/home/user/R/kurzus/soc.dat` (tetszés szerinti könyvtár). Beolvasás:

```
soc=read.table("/home/user/R/kurzus/soc.dat",
header=T,sep=";")
```

Ezzel a `soc` változóba (objektumba) írtuk a soc.dat fájl tartalmát.

Idézőjel szerepe: ha nincs, R a munkamemóriában tárolt változót (objektumot) keres!

Linux előnye: R bármelyik könyvtárból megnyitható az R parancs beírásával. Ha soc.dat-ot ide mentettük, elég a `read.table("soc.dat",...)` függvényt beírni.

Gyakorlati haszon: R-fájlokat projekteknek megfelelő könyvtárban tudjuk tárolni.

Grafikus felület (Mac, Windows)

Betöltés nem lehetséges közvetlen elérési útvonallal. Ehelyett:

(1) R-konzolban (ablak) File $>$ Change directory... megkeressük a könyvtárat, ahova soc.dat-ot mentettük.

```
soc=read.table("C:/Users/Users/Downloads/soc.dat",
header=T,sep=";")
```

VAGY

(2) aktuális munkamemória: getwd(). Betöltendő fájl helyének megadása: setwd("konyvtar").

Fontos: Windows-ban is / jelet használunk!

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.

Adatok mentése

Kilépés NEM a GUI (grafikus felület, graphical user interface) bezárásával, hanem a

`q()`

függvénnyel. Save directory? yes/no/cancel

Érdemes menteni, akkor az objektumok megnyitáskor ismét betöltődnek.

Linux: automatikusan abba a könyvtárba ment, ahonnan megnyitottuk az R-t.

Mac és Windows: default: R.exe fájl könyvtára. Módosítható `setwd()` függvénnyel.

Feladat I

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

b. 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 hisztogrammal és dobozdiagrammal.