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 (NEM csak átlag!),
- szóródás (NEM csak szórás!).

Ábrázolás táblázatban vagy grafikonokon.

Gyakoriság

- Abszolút érték (ha elemszámok megegyeznek).
- Arány (elemszám/összes), százalékos arány (arány*100) – jobb összehasonlíthatóság eltérő elemszám esetén.
- Kumulatív gyakoriság: előfordulás bizonyos érték ALATT – mutatja, milyen gyorsan nőnek az értékek (lapos vagy meredek növekedés).
- 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-függvények:
`table(x)`, `table(x/length(x))`, `table(x/length(x)*100)`, `prop.table(x)`, `cumsum(table(x))`

Ábrázolás

- kördiagram (pie chart),
- oszlopdiagram (barplot),
- hisztogram.

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

Példa

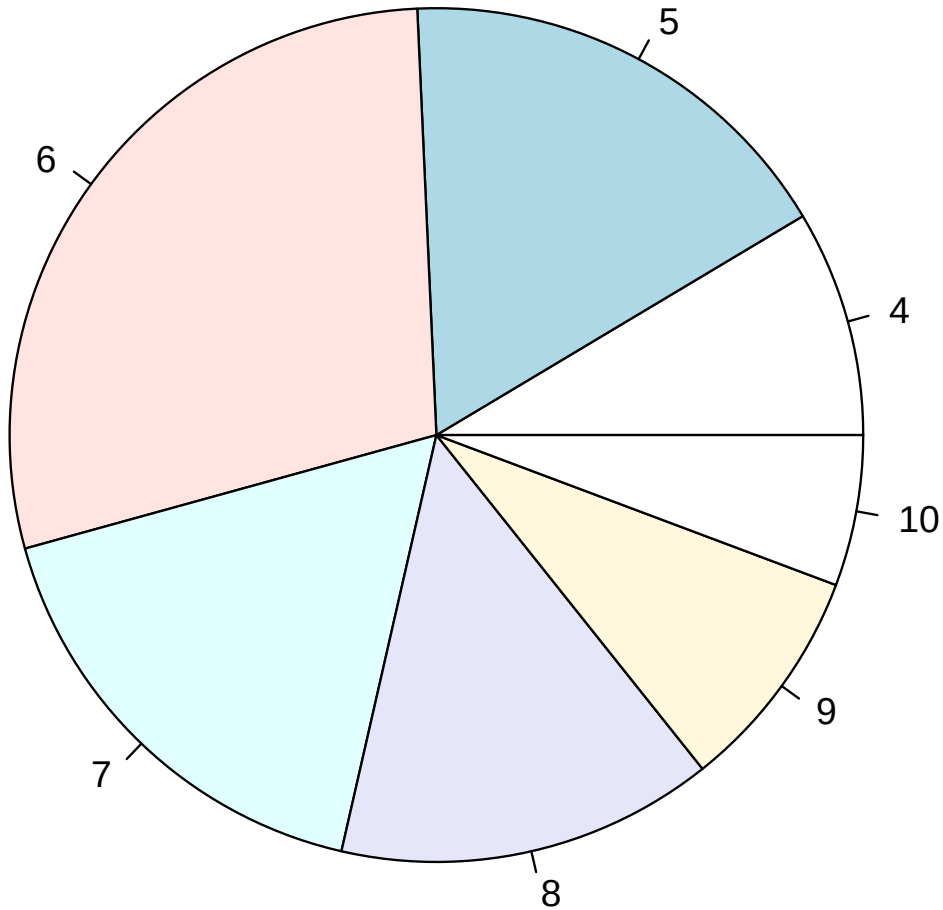
Angol növény- és állatnevek hosszúsága betűkben megadva.

típus	elemszám
növény	35
állat	46

A fájl letölthető innen:
https://phon.nytud.hu/mady/courses/statistics/materials/ratings.RData

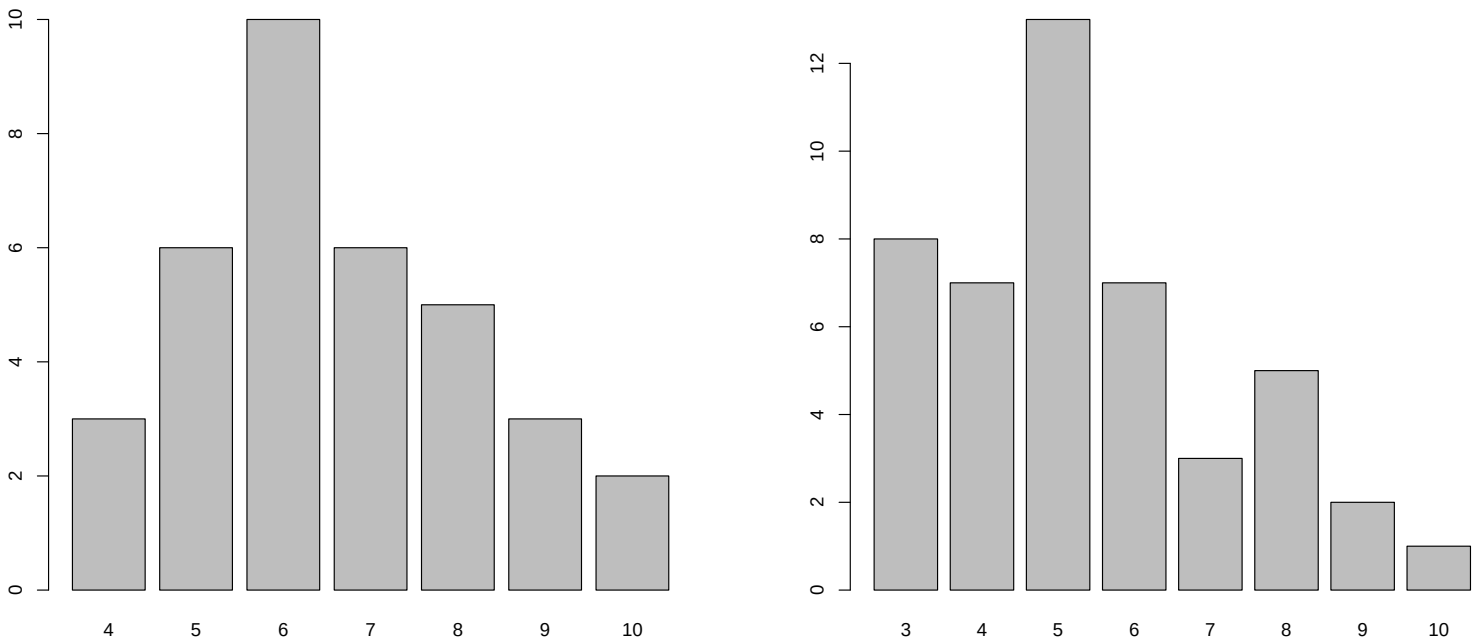
Kördiagram, R: `pie(table())`

Növénynevek hosszának gyakoriságai (= hány betűből állnak angolban)



Oszlopdiagram, R: `barplot(table())`

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

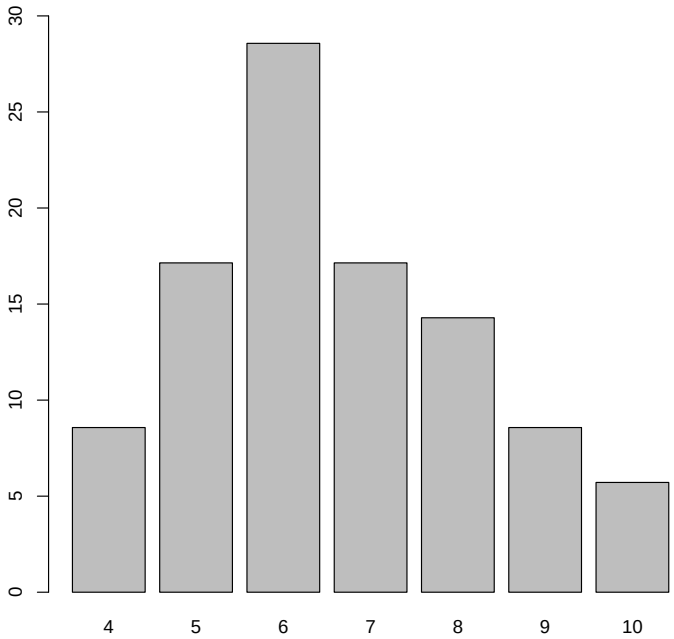


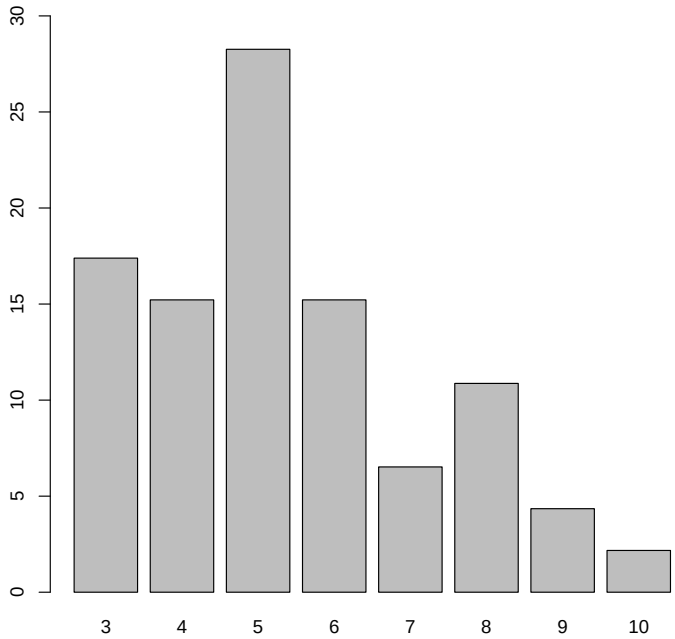
Jellemző felhasználás: nominális adatok, ordinális diszkrét adatok, kategorizált adatok.

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

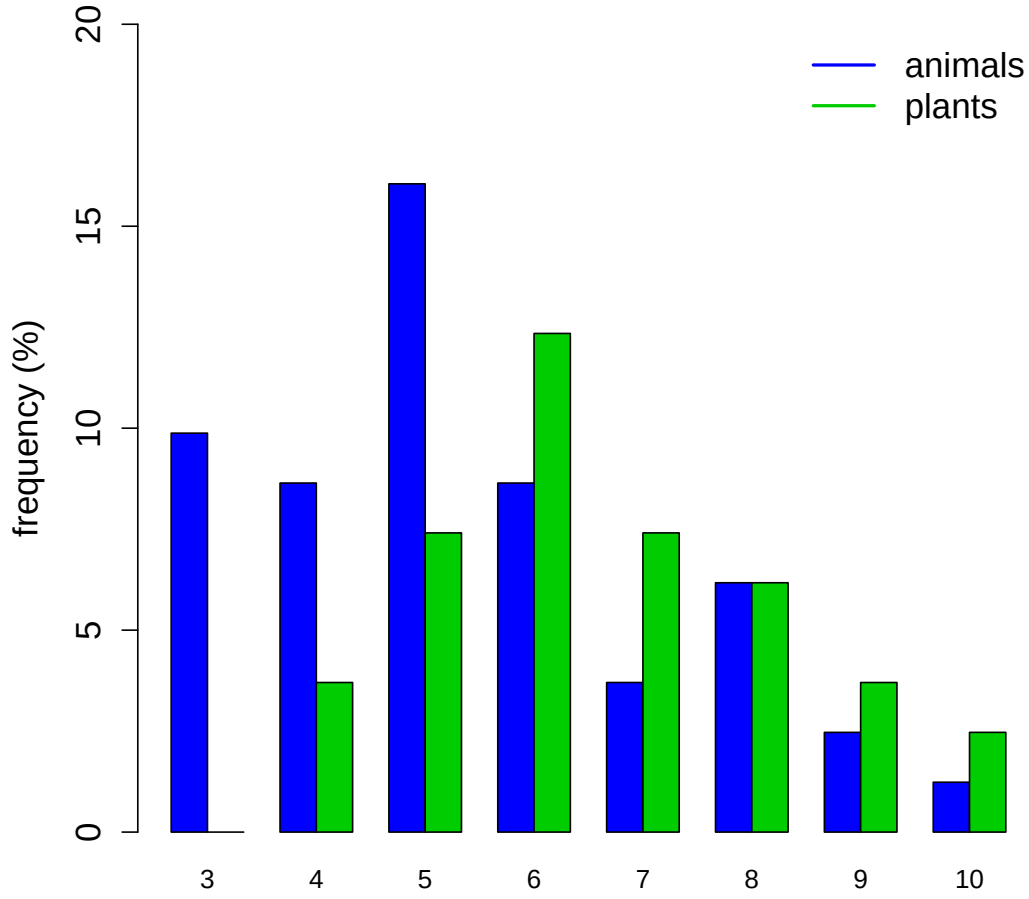
Oszlopdiagram százalékos arányokkal

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





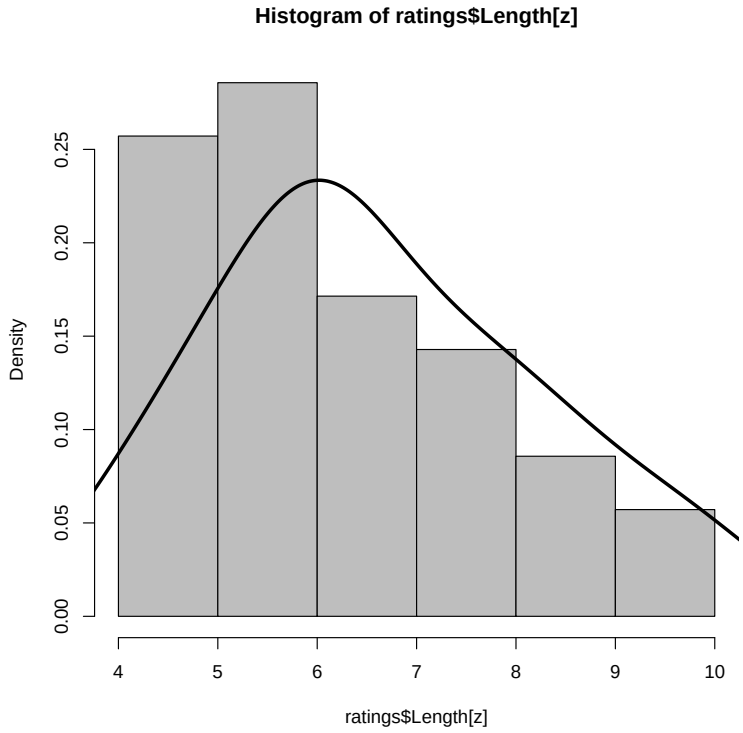
Két minta egy oszlopdiagramban

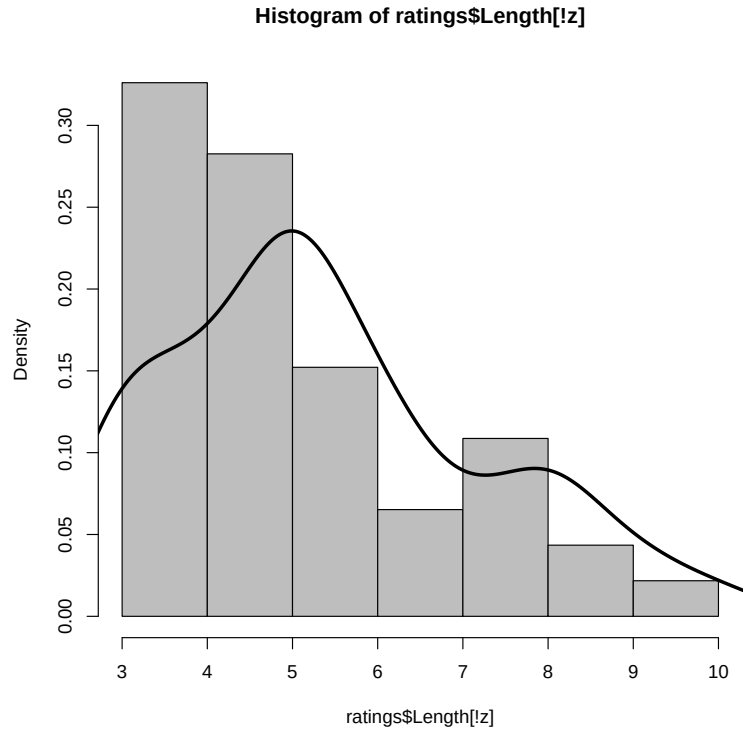


Előny: a két csoport eloszlásának jobb összehasonlíthatósága.

Hisztogram, R: `hist()`

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





Felhasználás: legalább ordinális skála, de a sűrűségfüggvénynek csak folytonos metrikus adatok esetén van igazán értelme.

Eloszlás

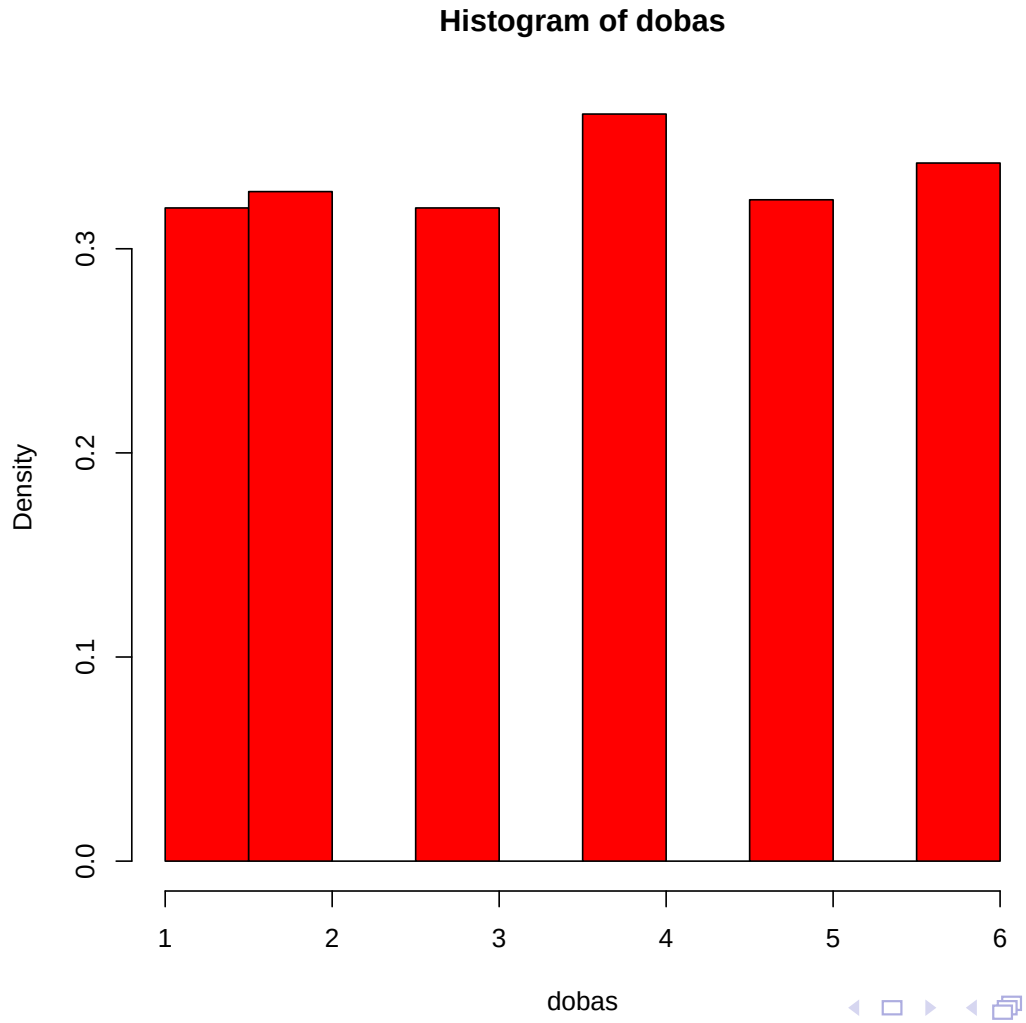
- **Definíció:** sorrendbe állított elemek milyen gyakorisággal fordulnak elő.
- **Felhasználás:** ordinális skálától felfelé.
- **Előállítás:** folytonos vagy diszkrét értékek közötti interpoláció.
- **Jelentőség:** valószínűségi statisztikai elemzés alapja.

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

Eloszlás típusai

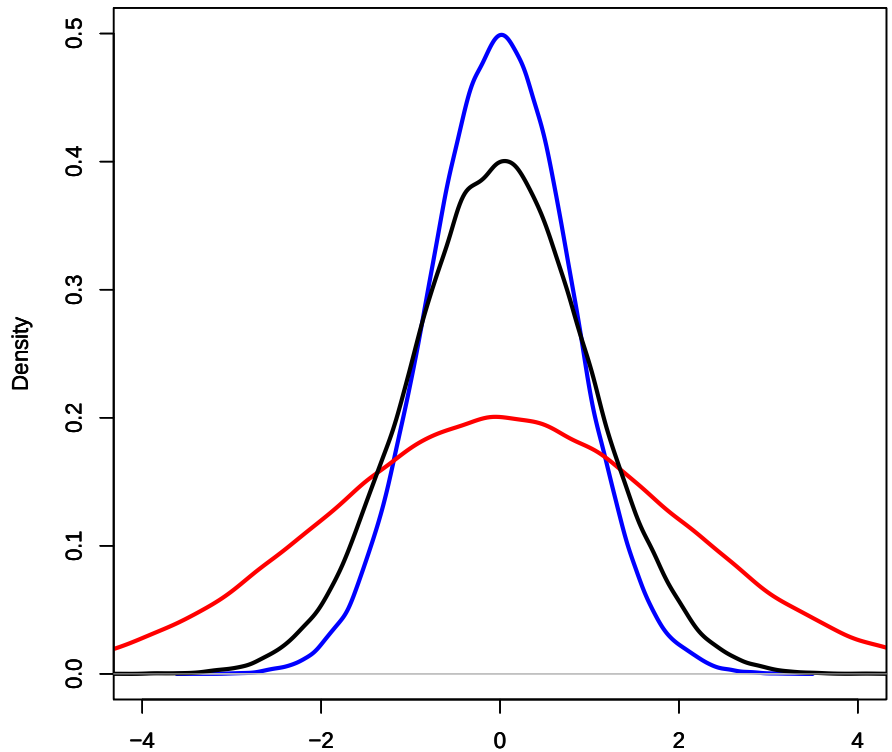
Egyenletes eloszlás

pl. dobott számok gyakorisága



Eloszlás típusai

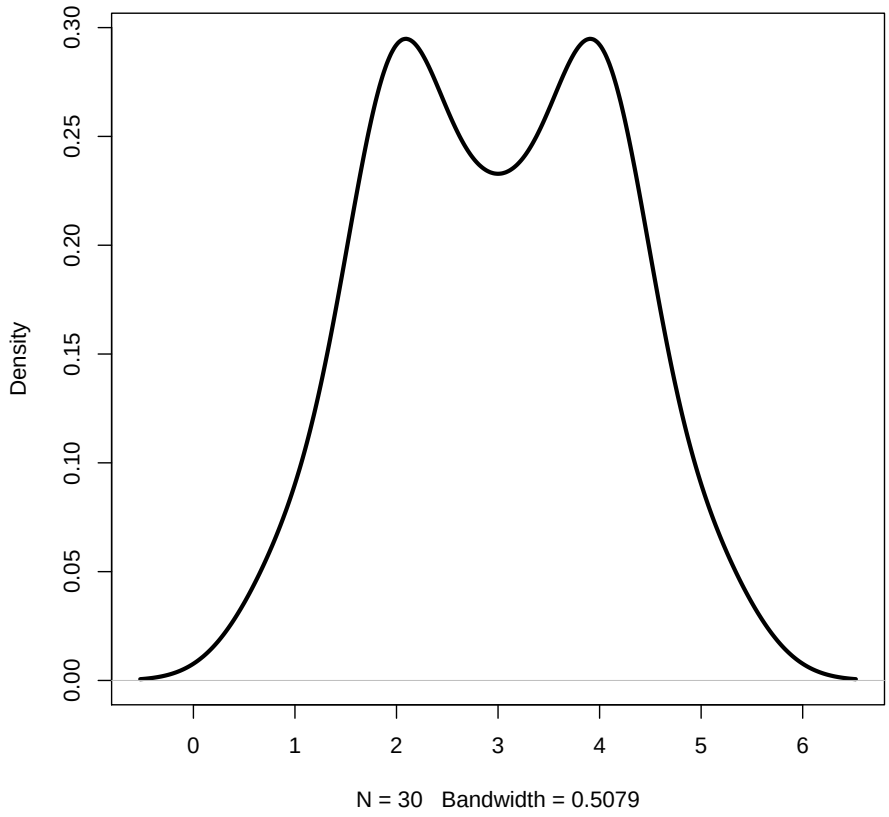
Unimodális: egy módusza van.



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.

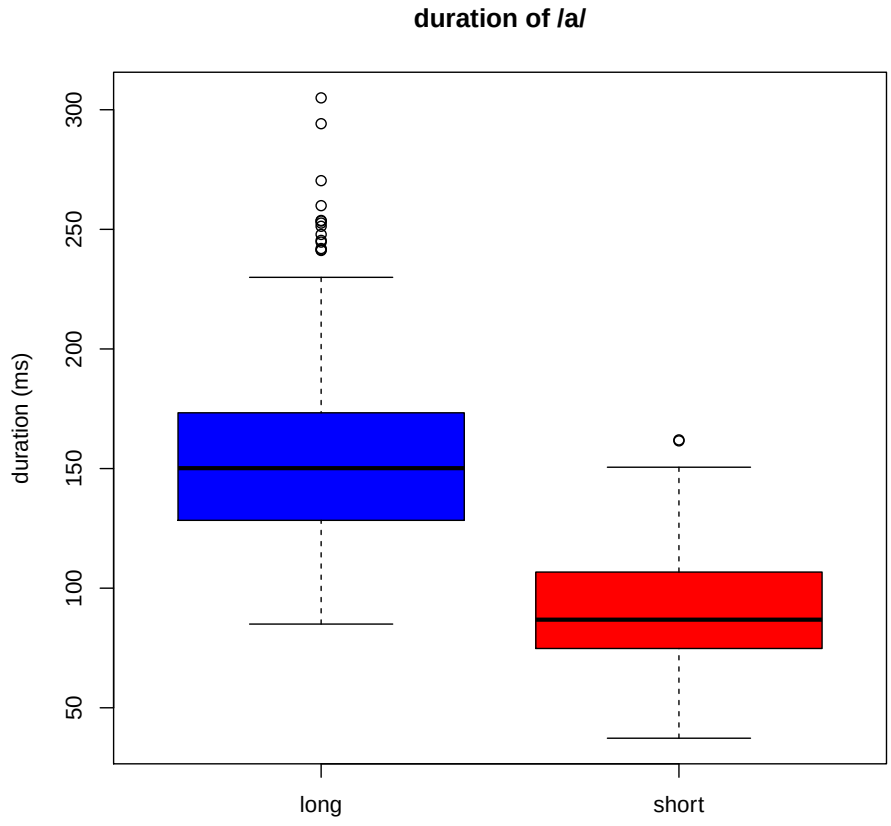
Átlagos Facebook-felhasználó ismerőseinek száma:
$\texttt{terjedelem} = 724 - 113 = 611$

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.

Dobozdiagram, R: `boxplot()`

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



Pontok: szélső értékek, a jellemző tartományon kívül esnek.

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 – azonos a doboz alsó és felső értékével.
- 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.

R

R munkamemóriája

Ha megnyitás után látjuk a mentett objektumokat, így tudjuk meg, hova történik az automatikus mentés: get working directory, azaz

```
getwd()
```

Ha nem látunk semmit, lehet, hogy nem sikerült a mentés. Ilyenkor a jövőben az R-et MINDIG rendszergazdaként kell megnyitni. *munkamem* lehet C:/Dokumentumok/Felhasználó/én/R vagy C:/Programok/R/lib vagy egyéb.

Bezárás előtti mentéskor (`q()`, `yes`) az R az objektumokat az aktuális munkamemóriába, egy .RData nevű fájlba írja ki, a parancsokat egy .Rhistory fájlba.

Figyelem! Ha a Windows fájlkezelő úgy van beállítva, hogy a rendszerfájlokat rejtse el (alapértelmezett beállítás), akkor az .RData és .Rhistory fájlokat nem fogja megjeleníteni. Ezt a Beállítások menüpontban meg kell változtatni. Linux alatt a megjelenítés l -a vagy ls -a paranccsal történik.

Munkamemória beállítása

`setwd("munkamem")`
Ide átmásolhatjuk az alapértelmezett helyről ezt a két fájlt:
`.RData .Rhistory`
Betöltés:
`load(".RData")`
`load(".Rhistory")`

Itt tárolhatjuk az adatfájlokat (pl. jonapot1.csv), amikkel dolgozunk. Így tudjuk betölteni:

`jonapot1 = read.csv("jonapot1.csv")`
(alternatíva: `read.table, read.csv2`)

Vagy ha eleve R-formátumú:
`load("jonapot1.RData")`

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.

`dim(soc)`: a soc objektum sorainak és oszlopainak száma (dimenziói).

`table(soc$oszlop)`: táblázat létrehozása.
%-os arány: `table(soc$oszlop)/osszesadat*100`.

Feladat I

Adjuk meg a `soc` objektum alapján:

- Hány személy adatait tartalmazza a táblázat?
- Hány nő és hány férfi szerepel benne?

Házi feladatok:

1. Származási hely (`loc` oszlop) szerinti gyakoriságok ábrázolása kördiagrammal és oszlopdiagrammal. Melyik ábratípus informatívabb? Miért?
2. Életkor átlaga és mediánja. Ábrázolás dobozdiagrammal (`boxplot(soc$age)`). Mekkora az 1. és 3. interkvartilis, a medián és az interkvartilis félterjedelem?
3. Életkor ábrázolása hisztogrammal, oszlopdiagrammal, majd a kumulatív eloszlások oszlopdiagramjával. Mit mutatnak az eloszlások? (Megjegyzés: a hisztogram függvényénél egy, az oszlopdiagramnál kettő, a kumulatív eloszlásnál három egymásba ágyazott zárójelre lesz szükség.)

Feladat II

Két dobókockával való 10, 100, 1000 dobálás összege. Mi a minták módusza, mediánja és átlaga? A három minta eloszlásának ábrázolása oszlopdiagrammal, hisztogrammal és dobozdiagrammal.