# 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 |

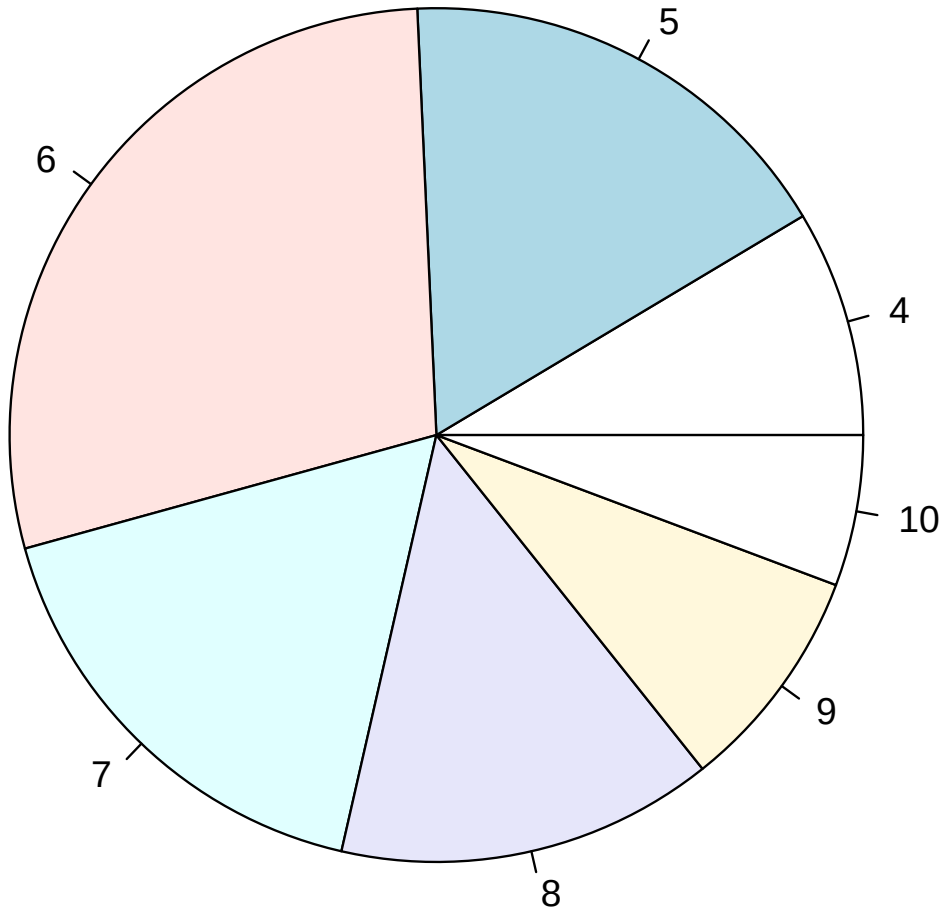
A fájl letölthető innen:
https://phon.nytud.hu/mady/courses/statistics/materials/ratings.RData

Ez nem szöveges fájl, hanem bináris, az R saját formátumában lett mentve. Beolvasás:

load("ratings.RData")
Vagyis nem kell megadni az objektum nevét, amibe mentjük, mert az R betöltéskor automatikusan létrehozza a `ratings` objektumot.

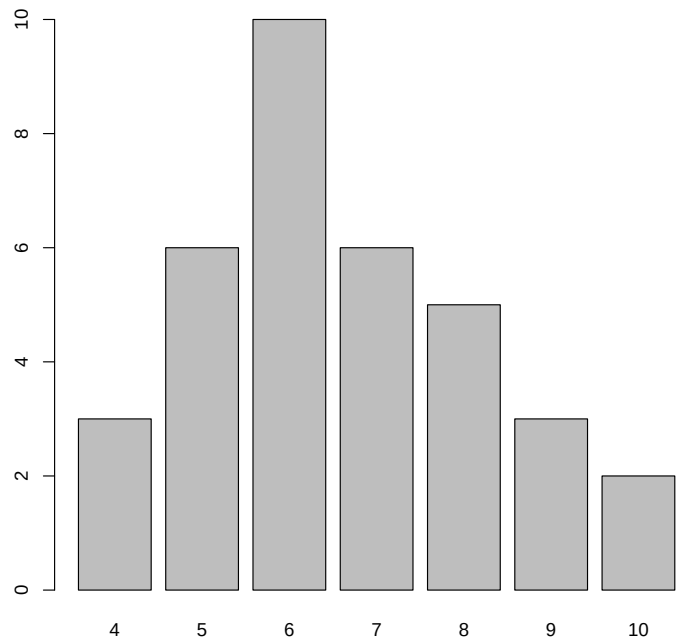
# Kördiagram, R: `pie(table())`

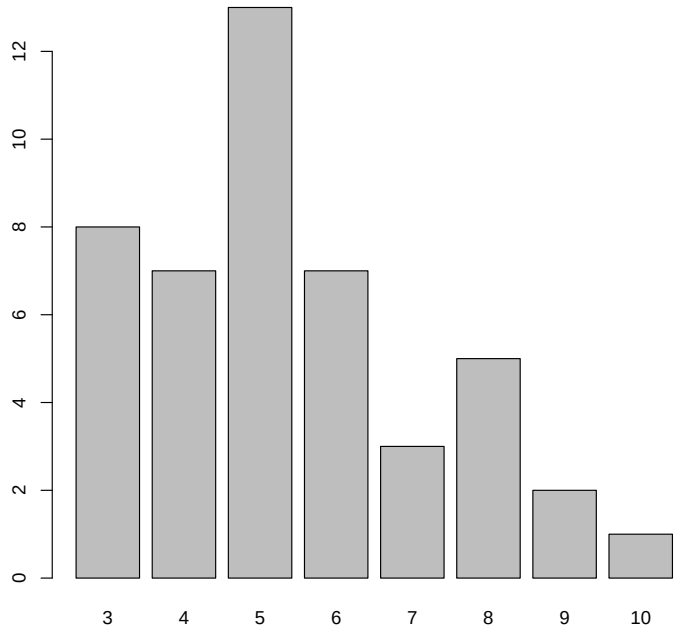
Növénynevek hosszának gyakoriságai ( = hány betűből állnak angolban)



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

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



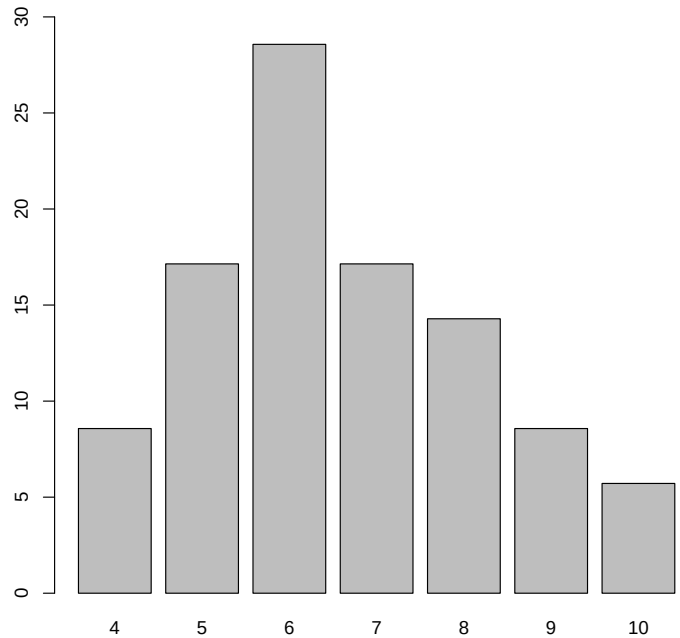



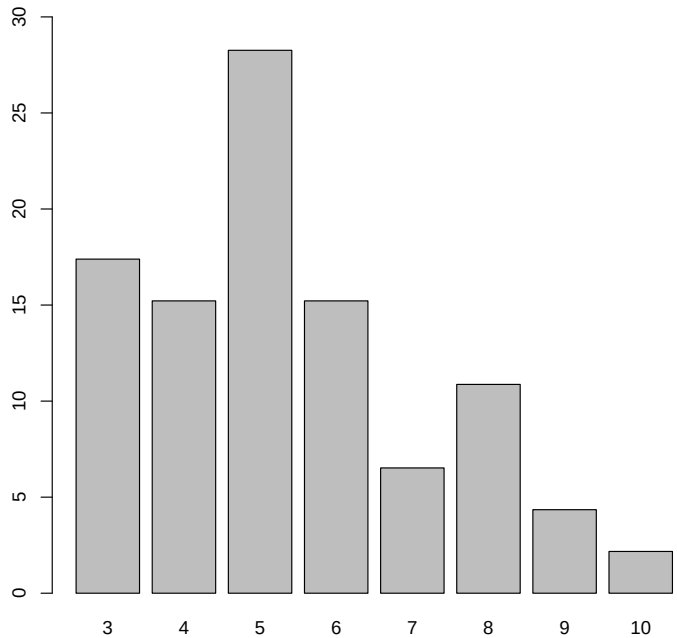

Jellemző felhasználás: nominális adatok, ordinális diszkrét adatok, kategorizált adatok.

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

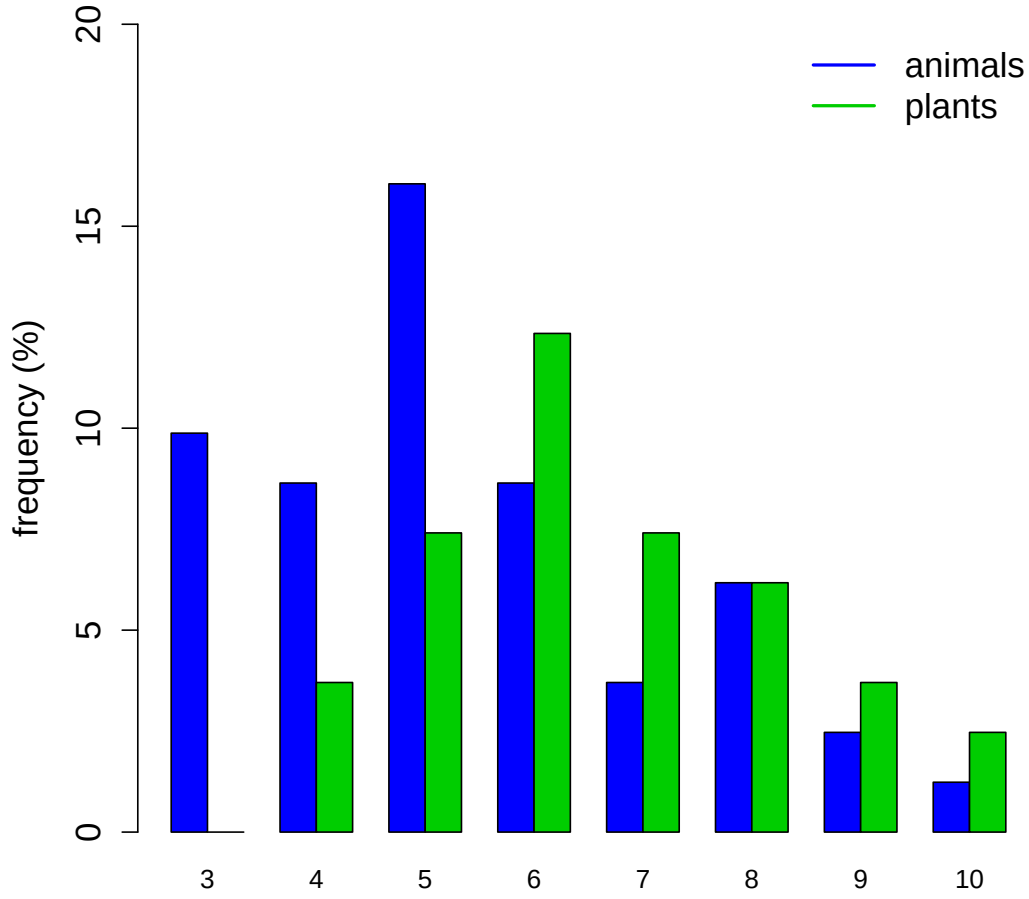
# Oszlopdiagram százalékos arányokkal

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



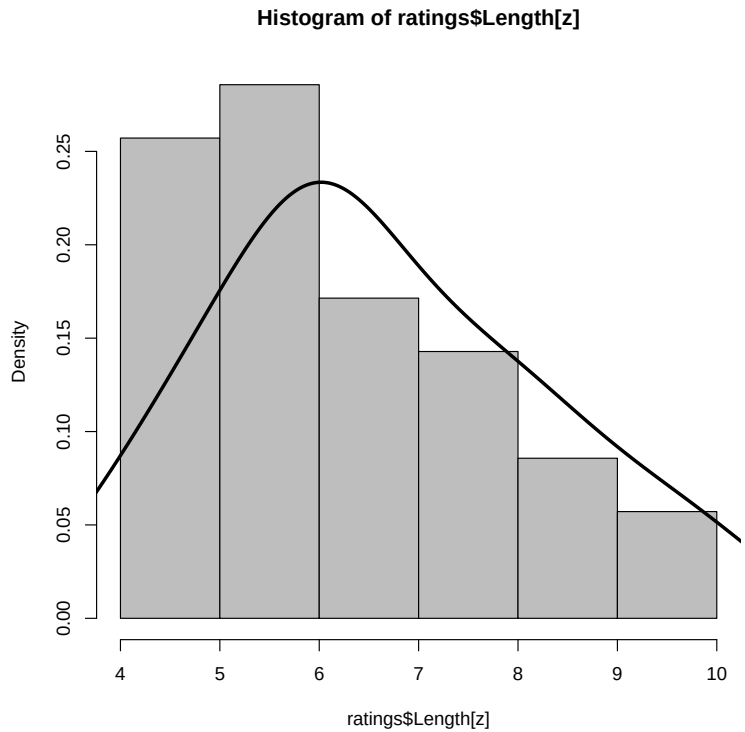
# Két minta egy oszlopdiagramban

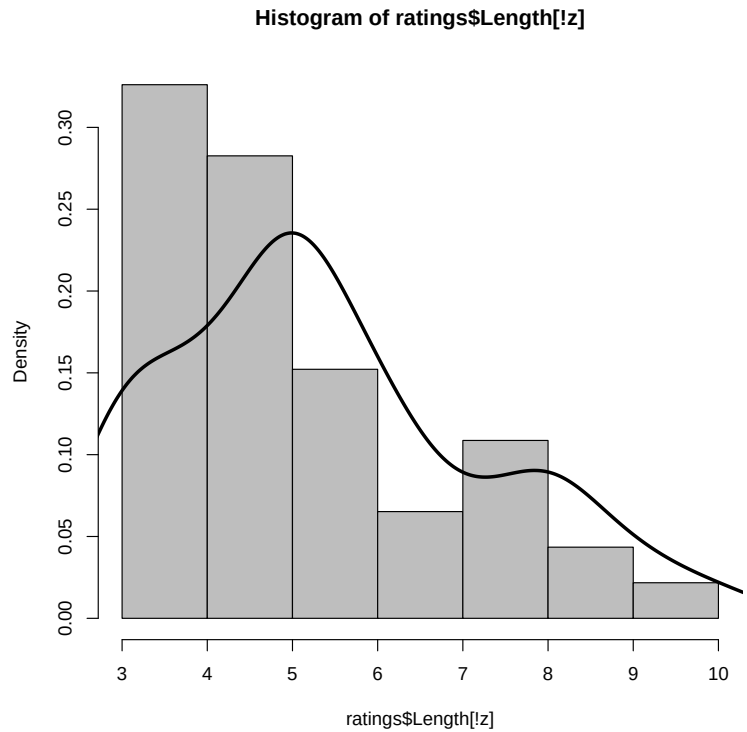



Előny: a két csoport eloszlásának jobb összehasonlíthatósága.

# 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

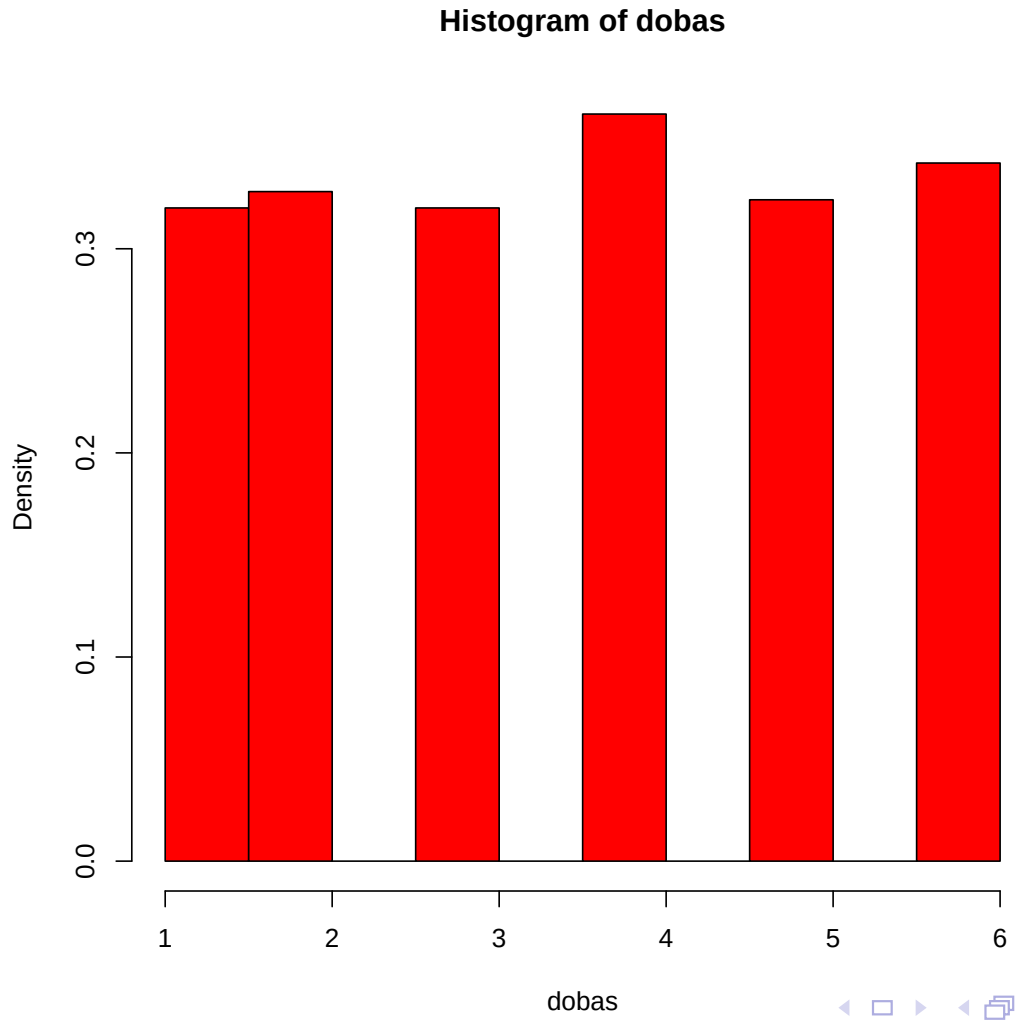
- **Definíció:** sorrendbe állított elemek milyen gyakorisággal fordulnak elő.
- **Felhasználás:** ordinális skálától felfelé.
- **Előállítás:** folytonos vagy diszkrét értékek közötti interpoláció.
- **Jelentőség:** valószínűségi statisztikai elemzés alapja.

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

# Eloszlás típusai

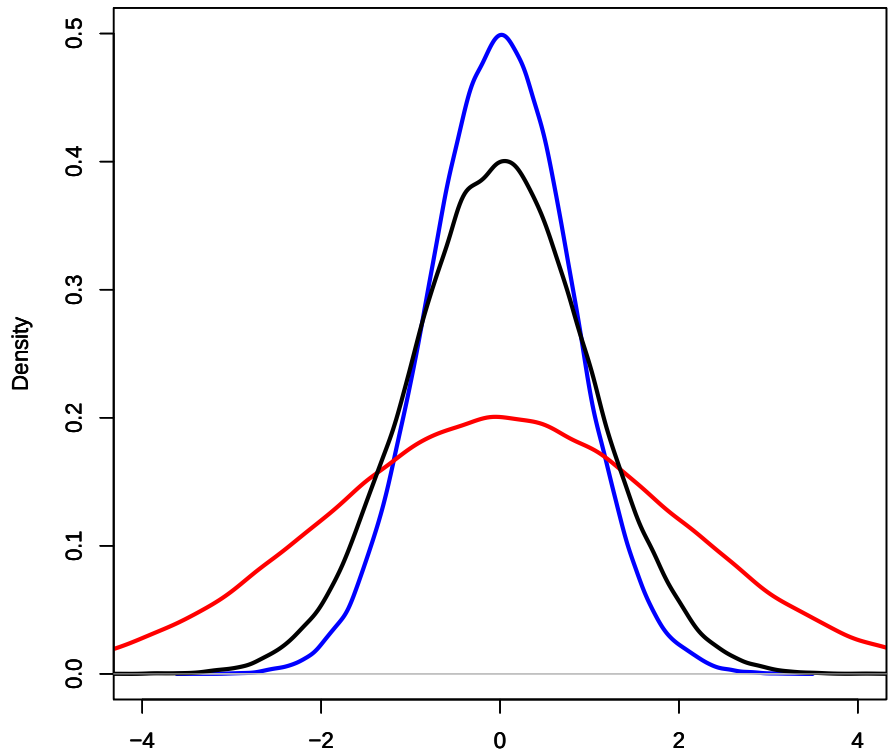
## Egyenletes eloszlás

pl. dobott számok gyakorisága 



# Eloszlás típusai

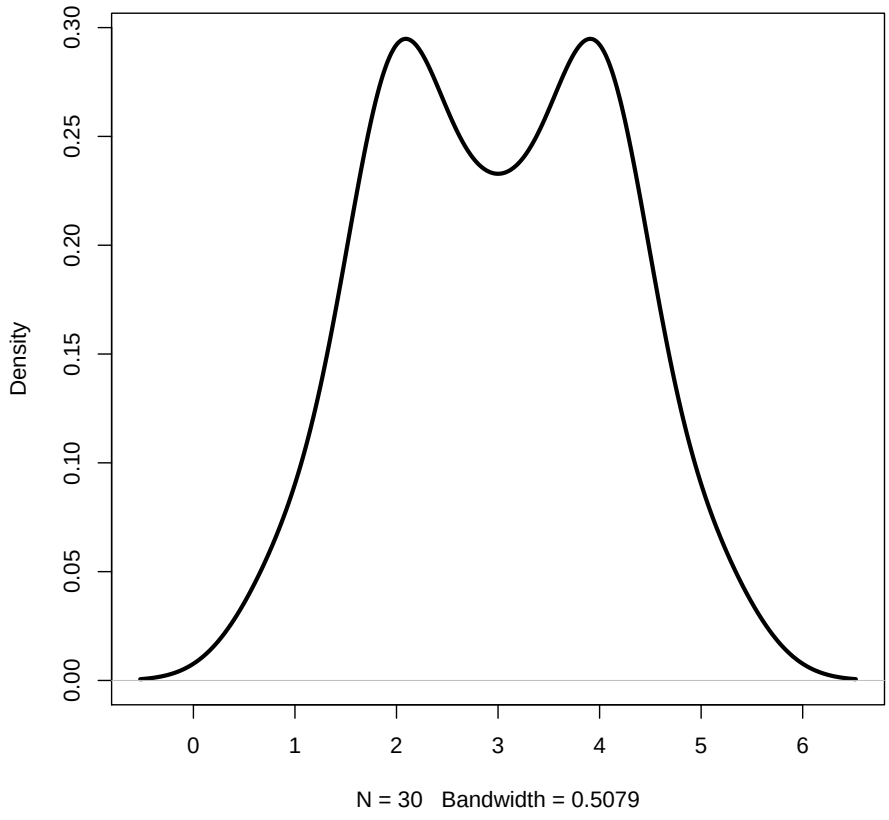
Unimodális: egy módusza van.




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.

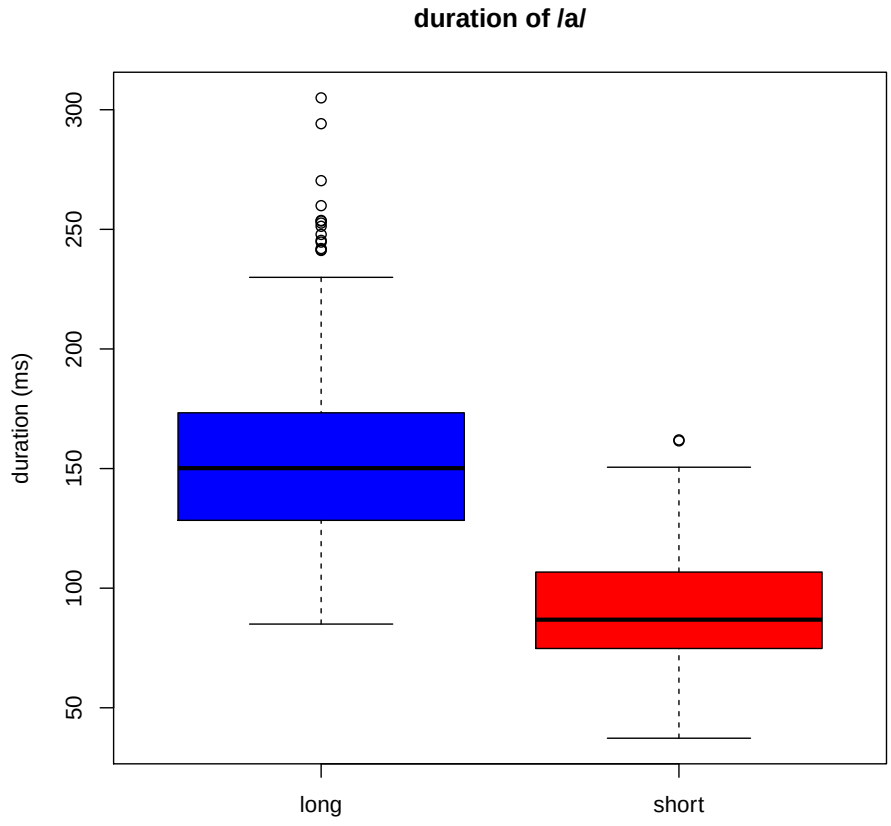
A 9b-be járó lányok testmagasságának terjedelme:
$\texttt{terjedelem} = 181 - 158 = 23 \ \texttt{cm}$

Az új lány érkezése utáni terjedelem:
$\texttt{terjedelem} = 188 - 158 = 30 \ \texttt{cm}$

Probléma: az első érték valószínűleg jobb becslése a populációra jellemző terjedelemnek, mert a 188 cm-es testmagasság Magyarországon ritka.

# 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 – szimmetrikus eloszlás esetén egyezik a mediánnal.

Eredeti tornasorunk kvartilisei:
158 160 161 163 165 167 168 170 171 177 181

Ha egy 188 cm magas lány jön az osztályba, az eddigi medián 167 cm-s pedig elmegy, a tornasor kvartilisei így változnak:
158 160 161 163 165 168 170 171 177 181 188

Az interkvartilis tartomány kevésbé érzékeny a kiugró szélső értékekre, mint a terjedelem (minimum és maximum érték).

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. evszakok.csv), amikkel dolgozunk. Így tudjuk betölteni:

```
evszakok = read.csv("evszakok.csv")
```

Vagy ha eleve R-formátumban van mentve:
`load("evszakok.RData")`

# Adatok beolvasása az R-be

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 választottuk el egymástól az átalakításnál?

Töltsünk le négy evszakok kezdetű csv-fájlt innen:
https://phon.nytud.hu/mady/courses/statistics/2024/

# Szöveges táblázatok beolvasása

Az alapértelmezett függvény:
`read.table()`
A függvényben számos paramétert lehet állítani attól függően, milyen formában hoztuk létre a csv-fájlt. Ezeket a paramétereket a súgón keresztül lehet lekérdezni:
`help(read.table)` vagy `?read.table`

`header`
Azt szabályozza, hogy a táblázat első sorában oszlopnevek vannak-e. Ha igen, `header=T` a megfelelő beállítás. Mi történik, ha a paramétert nem állítjuk át?
`sep`
Ha az Excel-fájlunkban minden cellában van adat, akkor választhatunk szóközt vagy tabulátort is, így a szöveges fájlban jobban elkülönülnek az oszlopok.

Most olvassuk be az evszakok_tab.csv fájlt így:
`evszakok=read.table=("evszakok_tab.csv")`. Mi történik?

Megoldás: szóköz vagy tab helyett válasszunk vesszőt vagy pontosvesszőt.

`dec`

A tizedes jelölése a csv-fájlunkban. Mi történik, ha nem állítunk ezen a paraméteren?

Egy magyar nyelvű Excelből kiírt csv-fájl esetén tehát a következő paramétereket érdemes használni, ha eltérnek az alapértelmezett beállításoktól:

`read.table(file, header = TRUE, sep = ";", dec=","`

Nézzük meg a `read.csv2()` függvény alapértelmezett beállításait a súgóban. Mit látunk?

Hozzunk létre egy evszakok nevű objektumot az R-ben a megfelelő paraméterbeállításokkal.

# Néhány hasznos függvény

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

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

`head(evszakok)`: első hat adatsor.

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

`dim(evszakok)`: az evszakok objektum sorainak és oszlopainak száma (dimenziói).

`table(evszakok$evszak)`: táblázat létrehozása.
%-os arány: `table(evszakok$evszak)/osszessorszama*100`.

# Ábrák mentése

Az R-et használók korábban gyakran programozó kedvű kutatók voltak, akik Linux operációs rendszert használtak, és LaTeX szövegszerkesztőben készítették az írásaikat. Ezért az R-be alapvetően beillesztett képformátum a .pdf (Portable Document Format) és a .ps (PostScript).

```
pie(file)
dev.print("célfájl",device=pdf)
```

A Wordben viszont képformátumú ábrákra van szükség. Mentés ábraként, pl. png (jpg és tif nem jó, mert képtömörítéssel készül!):

```
png(file = "fajlnev", bg = "white")
plot(blablabla)
dev.off()
```

# Feladat I

Adjuk meg az `evszakok` objektum alapján:

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

Házi feladatok:

1. Válaszadók neme szerinti gyakoriságok ábrázolása kördiagrammal és oszlopdiagrammal. Melyik ábratípus informatívabb? Miért?
2. Életkor átlaga és mediánja. Ábrázolás dobozdiagrammal. 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.