

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),
- ▶ oszlopdiaagram (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>

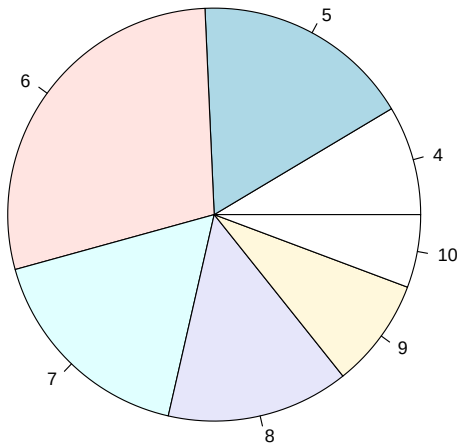
Ez nem szöveges fájl, hanem bináris, az R saját formátumában lett mentve. Beolvasás:

```
load("ratings.RData")
```

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

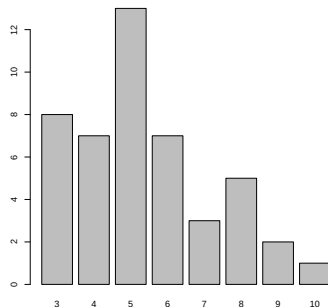
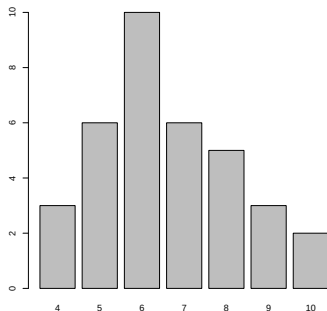
Kördiagram, R: `pie(table())`

Növény- és állatnevek hosszának gyakoriságai (= hány betűből állnak az angolban)



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

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

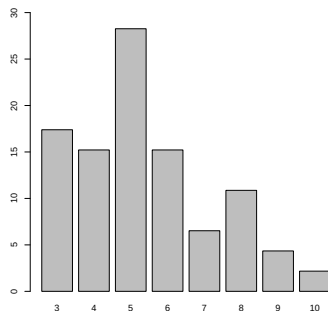
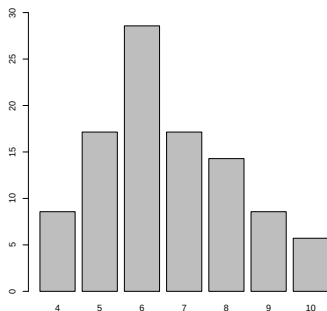


Jellemző felhasználás: nominális adatok, ordinális diszkrét adatok, kategorizált adatok.

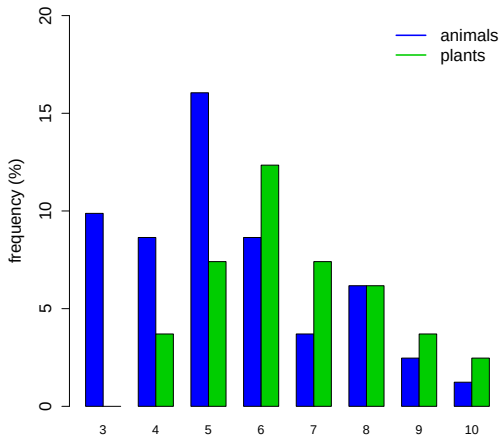
Probléma: két csoportban eltérő elemszám! ($n_n = 35$, $n_a = 46$)

Oszlopdiagram százalékos arányokkal

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



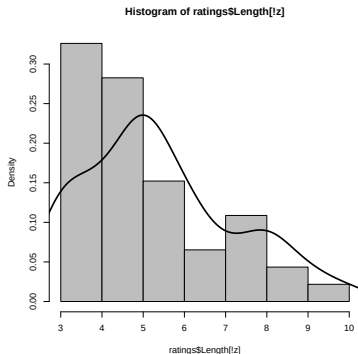
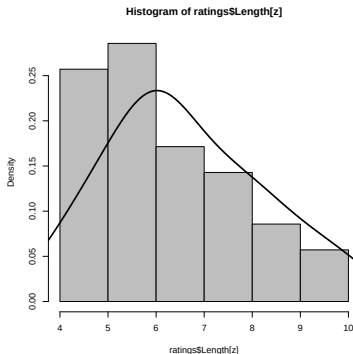
Két minta egy oszlopdiagramban



Előny: a két csoport eloszlásának jobb összehasonlíthatósága.

Hisztogram, R: hist()

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

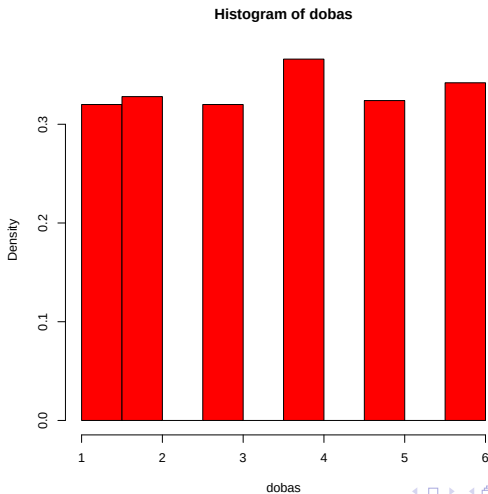
`hist(x, frequency=FALSE)`: arányos gyakoriságok,

`plot(density())`: sűrűségfüggvény.

Eloszlás típusai

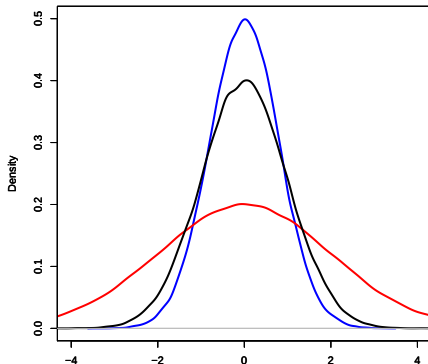
Egyenletes eloszlás

pl. dobott számok gyakorisága



Eloszlás típusai

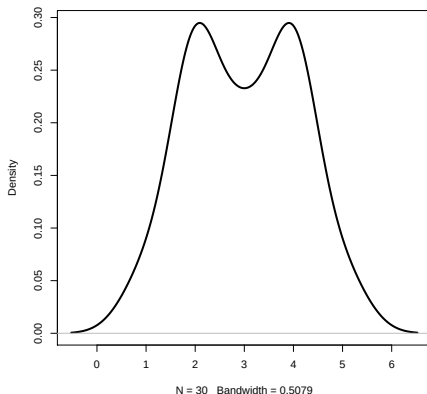
Unimodális: egy módusza van.



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, mert a minta valószínűleg két külön populációhoz tartozik. Pl.: nők és férfiak testmagassága.

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 az extrém szélső értékekre.

A 9b-be járó lányok testmagasságának terjedelem:

$$\text{terjedelem} = 181 - 158 = 23 \text{ cm}$$

Az új lány érkezése utáni terjedelem:

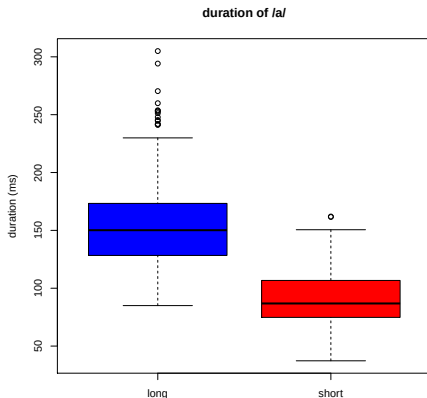
$$\text{terjedelem} = 188 - 158 = 30 \text{ cm}$$

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.

1. és 3. kvartilis: doboz alsó és felső szélei.



Pontok: extrém (kieső, kiugró, kilógó) értékek, a jellemző tartományon kívül esnek.

Interkvartilis tartomány

- ▶ Jelentőség: ha ordinális skála vagy nem szimmetrikus eloszlású parametrikus adatok.
- ▶ Interkvartilis tartomány/terjedelem: az X változó értékskálájának középén elterülő övezete, ahol a populáció 50%-a található. Függvénye `quantile()`.
- ▶ Folytonos változó esetén: alsó 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 különbségének fele.

A dobozdiagramon ponttal ábrázolt extrém értékek a doboz aljától és tetejétől 1,5 kvartilissal távolabb esnek, vagyis nem tartoznak a jellemző értékek közé.

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

Házi feladat:

Hány személy adatait tartalmazza az evszakok táblázat, hány férfi és hány nő szerepel benne? A nemek szerinti gyakoriság ábrázolása kördiagrammal és oszlopdiagrammal.

Szükséges függvények:

```
table(evszakok$nem)
pie(table(evszakok$nem))
barplot(table(evszakok$nem))
```

Hogy kerül ide a harmadik kategória? Van más elírás is a táblázatban?

Ellenőrzés a `unique()` függvénnyel. Ez az adott oszlophoz tartozó összes előfordulást kilistázza.

Töltsünk le két további fájlt az R-be a következő két fájlt az evszakok.szavak és evszakok.kisbetu objektumba:
evszakok_szavakkal.csv, evszakok_kisbetu.csv

Rajzoljunk mindegyikből dobozdiagramot évszakok szerint bontva.

Mi itt a probléma? Orvoslás az Excelen keresztül vagy az R-ben (ld. később). Egyelőre töltsük le újra a javított evszakok.csv fájlt a kurzus honlapjáról.

R munkamemóriája

Ha az R-et az asztalon lévő ikonra kattintva nyitjuk meg, a következő függvénnyel tudhatjuk meg, hova menti a fájlokat bezáráskor:

get working directory, azaz `getwd()`
például: "C:/Users/mady/Documents"

Az R Unix-alapú szoftver, hasonlóan a Machez és a Linuxhoz, ezért nem `\`, hanem `/` jel választja el a könyvtárakat (mappákat).

Hozzunk létre a Dokumentumok (vagy más) mappában egy újat, **statisztika** néven. Itt tároljuk az órán használt fájlokat.

Állítsuk be, hogy az R ebbe a könyvtárba mentse a munka során létrehozott fájlokat:

```
setwd("C:/Users/mady/Documents/statisztika")
```

Mentés bezárás előtt: `q()` vagy File - Exit, a válasz Igen/yes.

Az R az objektumokat az itt beállított munkamemóriába írja ki egy .RData nevű bináris fájlba, a beírt parancsokat pedig egy .Rhistory sima szöveges fájlba.

Ha a Windows fájlkezelő úgy van beállítva, hogy a rendszerfájlokat rejtse el (alapértelmezett beállítás), akkor lehetséges, hogy az .RData és .Rhistory fájlokat nem fogja megjeleníteni.

Ezt a fájlkezelőben lehet megváltoztatni. Továbbiak (...) - Beállítások - Nézet - Rejtett fájlok, mappák és meghajtók MEGJELENÍTÉSE - kiválasztás az elrejtés helyett.

Ugyanebben az ablakban feljebb az Ismert fájltypusok kiterjesztéseinek elrejtése menüpont előtt vegyük ki a pipát. Így látni fogjuk a .csv, .txt stb. kiterjesztést a fájlnev végén, így mindig tudjuk, milyen típusú fájl látunk.

Adatok beolvasása az R-be

Az R-be csak szöveges fájlokat tudunk beolvasni, MS-Office kódolású fájlokat nem. Ezért az Excel-ből .csv fájlként (comma-separated values) mentve alakítjuk szöveges fájlá a táblázatot.

Fontos szempontok:

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

A betöltéskor alkalmazott függvényt ezen szempontok szerint kell kiválasztani vagy pontosítani.

Töltsük le a 3_statisztika_R.txt fájlt, ami a további parancsokat tartalmazza:

`https://phon.nytud.hu/mady/courses/statistics/2025/`

Töltsük le az evszakok.csv fájlt. Betöltés közvetlenül a böngészőből (a pdf-ből kimásolt függvény nem fog működni az idézőjelek miatt):

```
evszakok =  
read.csv2(url("https://phon.nytud.hu/mady/courses/statistics/  
2025/evszakok.csv"))
```

Ha már letöltöttük a fájlt, mert a saját gépünkön tároljuk:

```
evszakok =  
read.csv2("C:/Users/mady/Documents/statisztika/evszakok.csv")
```

Töltsük le az `evszakok_vesszo.csv` fájlt az `evszakok.vesszo` objektumba. Rajzoljunk boxplotot. Miért nem sikerül?

Így nézzük meg, mi van az objektumban tárolva:
`head(evszakok.vesszo)`

Mi más itt, mint az eddig mentett fájloknál?

Megadjuk, hogy az oszlopokat elválasztó karakter vessző, nem pontosvessző (ami az alapértelmezett):
`read.csv2("fájlnév", sep=",")`

Miért kell figyelni, ha az oszlopok vesszővel vannak elválasztva?
Nézzük meg, mi történik, ha betöltjük `evszakok_dec_vesszo.csv` fájlt a `read.csv2()` függvénnyel a fenti módon.

Szöveges táblázatok beolvasása

A függvényekben számos paramétert lehet állítani attól függően, milyen formában hoztuk létre a csv-fájlt. Ezeket a paramétereket a súgón keresztül lehet lekérdezni:

`help(read.csv2)` vagy `?read.csv2`

`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`: pontosvessző, vessző vagy más. Választhatunk szóközt vagy tabulátort is, de csak ha biztosak vagyunk benne, hogy az Excel-táblázatban minden cellában van adat.

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:

```
header = TRUE, sep = ";", dec=","
```

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

Néhány hasznos függvény

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

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

`head(evszakok)`: első hat adatsor.

`data.frame` változóira (oszlopaire) 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`

`class(objektum$oszlop)`: megmutatja, hogy milyen típusú értékek vannak az oszlopban: integer, numeric, character, factor stb.

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()
```


Feladatok

Adjuk meg az evszakok objektum alapján:

1. Életkor átlaga és mediánja. Ábrázolás dobozdiagrammal. Mekkora az 1. és 3. kvartilis, a medián és az interkvartilis félterjedelem?
2. 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.)

Házi feladat

Növény- és állatnevekre vonatkozó adatok a ratings objektumból.

1. A gyakoriságok adatai (Frequency) növényekre és állatokra vonatkozóan külön-külön, dobozdiagramon.
2. A szavak ismertsége (meanFamiliarity) külön-külön, dobozdiagramon.
3. A szóhosszok ábrázolása oszlopdiagramon, a 81 növény- és állatnév százalékos arányában.