Skálatípusok
Objektumok az R-ben

Változók

- **Kvalitatív:** valamilyen tulajdonság (februárban születettek, nők, etnikai csoportok, szófajok stb.).
- **Diszkrét:** megszámolható, véges, gyakran egész számok (hibák száma egy tesztben, életkor években megadva).
- **Folytonos:** adott intervallumban akármilyen valós szám.
- **Kategóriák vagy csoportok:** változók összefoglalása (pl. 20 és 40 év közötti fiatal felnőttek). Előny: egyszerűbb kezelés, mert kevesebb kategória, de információvesztés.

Skálatípusok

Nominális skála: változó értékei megkülönböztethetőek, de semmilyen sorreindi viszonyban nem állnak egymással. (Nem, vallás, hajszín, szófaj.)

Ordinális skála: értékek rangsorolhatóak, de az egyes elemek távolsága nem egyenlő vagy nem értelmezhető. (Iskolai végzettség, osztályzat.)

Metrikus skálák: egy adott mértékegység többszörösei. A mértékegység részei és többszörösei is értelmezhetőek, tehát a távolság értelmezhető és összehasonlítható. Két típusa van.
Intervallumskála: nullpontja önkényes (pl. Celsius fok), mérőszámok különbsége igen, de aránya nem értelmezhető. Húsz fok nem kétszer olyan meleg, mint tíz fok.
Arányskála: nulla pont fizikailag definiált, arányok is értelmezhetőek (távolság, tömeg, energia, Kelvin fok).

Középértékek

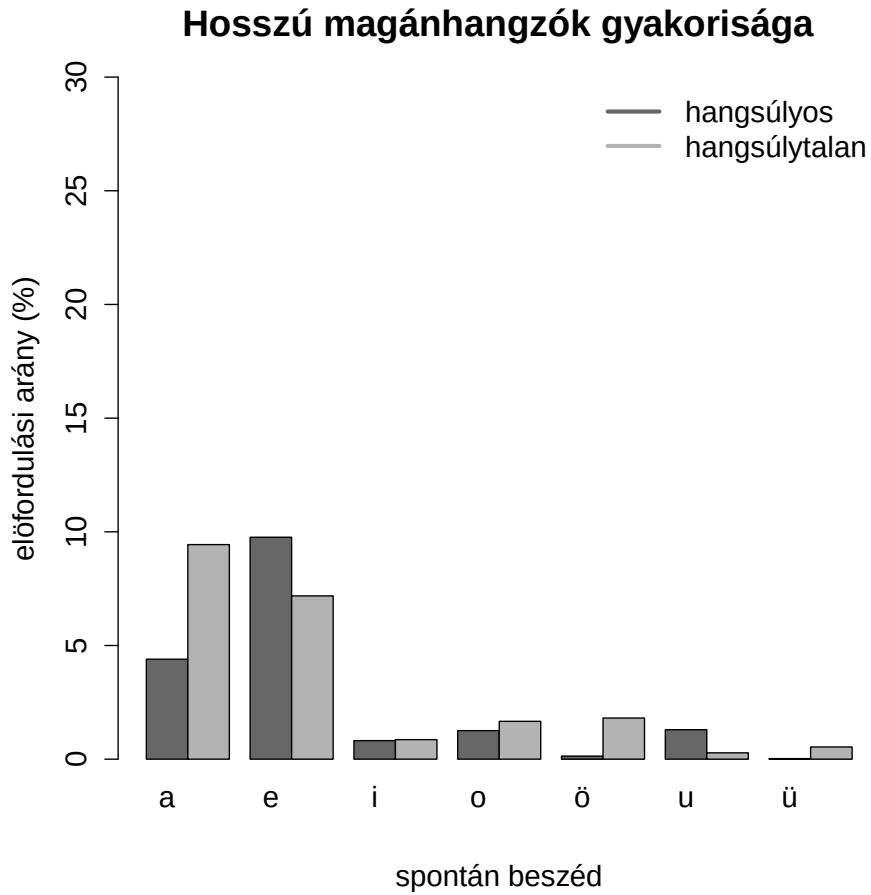
- **Módusz:** a mintában a legnagyobb gyakorisággal előforduló adatérték.
- **Medián:** a növekvő sorba rendezett adatok közül a középső. Ha az n mintaelemszám páros, a két középső érték átlaga.
- **Átlag:** mintabeli adatok számtani közepe.

Nominális skála: módusz, ordinális skála: medián, metrikus skála: átlag.

Alacsonyabb skálára érvényes statisztikai módszerek mindig alkalmazhatóak a magasabbakra, de információvesztéssel jár(hat)nak.

Középértékek: modusz

A mintában előforduló leggyakoribb kategória. Minden skálatípusra alkalmazható.



Középértékek: medián

Egy sorozat középső eleme. Ha az n elemből álló sorozat elemszáma páros, akkor a medián a két középső elem átlaga. Nominális adatokra NEM számolható medián.

Hány ismerőse van a Facebook-os ismerőseimnek?

11 véletlenszerűen kiválasztott ismerős ismerőseinek száma:

546 388 724 269 113 467 682 178 149 382 196

Sorba rendezett értékek:
113 149 178 196 269 382 388 467 546 682 724

Középső érték: 6. elem = 382.

12 ismerős esetén a 6. és 7. elem átlaga a medián.

Középértékek: átlag (számtani közép)

Az összes érték átlaga, azaz az értékek összege osztva az elemszámmal.

A Facebook-os ismerőseim ismerőseinek száma:

```
átlag = (546+388+724+269+113+467+682+
+178+149+382+196)/11 = 382,1818
```

Az átlag egy statisztikai modell, nem feltétlenül tükröz reális adatokat. Senkinek nincs 0,1818-ad ismerőse.

Fontos: átlagot csak parametrikus adatokra lehet számolni, amelyek ekvidisztánsak, azaz egyenlő távolságra vannak egymástól.

Iskolai osztályzatok?

Középértékek: átlag (számtani közép)

Az összes érték átlaga, azaz az értékek összege osztva az elemszámmal.

A Facebook-os ismerőseim ismerőseinek száma:

```
átlag = (546+388+724+269+113+467+682+
+178+149+382+196)/11 = 382,1818
```

Az átlag egy statisztikai modell, nem feltétlenül tükröz reális adatokat. Senkinek nincs 0,1818-ad ismerőse.

Fontos: átlagot csak parametrikus adatokra lehet számolni, amelyek ekvidisztánsak, azaz egyenlő távolságra vannak egymástól.

Iskolai osztályzatok? Az 1-es és 2-es különbsége nagyobb, mint a 4-esé és 5-ösé, ezért nem ekvidisztáns skála.

Medián vagy átlag?

Képzeljük el, hogy a 11 ismerősünk közül valaki csak tegnap iratkozott fel a Facebook-ra, ezért még csak 11 barátja van. Egy másik ismerős híres színésznő, és 5439 ismerőse van.

`átlag = (11+149+178+196+269+382+388+467+546+682+5429)/11 = 791,5455`

Ha 11 helyett 111 ismerős ismerőseit vizsgáljuk, kiderül, hogy kevés embernek van 11 vagy 5429 ismerőse - ezek szélső vagy extrém értékek (ld. később).

Az adatok eloszlását érdemes ábrázolni, illetve az átlag mellett a mediánt is kiszámolni, ami robusztusabb, mert kevésbé érzékeny a szélső értékekre.

A fenti adatok mediánja szintén 382, ami reálisabban tükrözi az adatok középértékét.

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.csv
Érdemes a felhasználói név alatt létrehozni egy R könyvtárat erre a célra. Ide töltjük le a fájlt, és ide mentjük majd az R munkamemóriáját is.

Fájl beolvasása Linuxban

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

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

Ezzel a `soc` változóba (objektumba) írtuk a soc.csv 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.csv-ot ide mentettük, elég a `read.table("soc.csv",...)` 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.csv-t mentettük.

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

VAGY

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

Fontos: Windows-ban is / jelet használunk!

Ha a decimálisunk vessző, a cellákat pedig pontosvesszővel választottuk el, akkor a `read.csv2()` függvény alapbeállításai pont megfelelőek.

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

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?

Feladat II

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