Skálatípusok, leíró statisztika. Adattípusok, adatkezelés az R-ben

Változók

- **Kvalitatív:** valamilyen tulajdonság (februárban született, nő, romungró, ige stb.).
- **Diszkrét:** megszámolható, véges (hibák száma egy tesztben, életkor években megadva).
- **Folytonos:** adott intervallumban akármilyen valós szám.
- **Kategóriák vagy csoportok:** változók összefoglalása (pl. 25 és 34 év között). Egyszerűbb kezelés, de információvesztés.

Skálatípusok

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

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

Metrikus skálák: egy adott mértékegység többszörösei. A mértékegység részei és többszörösei is értelmezhetőek, tehát a távolság értelmezhető és összehasonlítható.
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

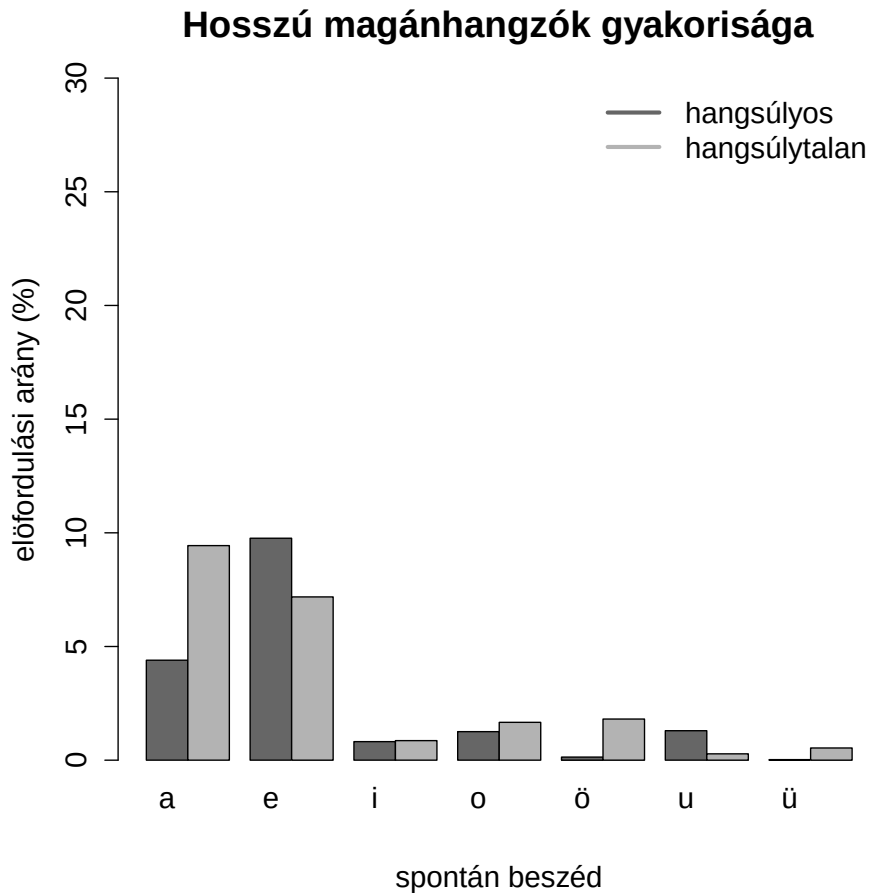
- **Módusz:** a mintában a legnagyobb gyakorisággal előforduló adatérték.
- **Medián:** a növekvő sorba rendezett adatok közül a középső. Ha az n mintaelemszám páros, a két középső érték átlaga.
- **Átlag:** mintabeli adatok számtani közepe.

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

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

Középértékek: módusz

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.

Milyen értékkel jellemezhető a 11/b osztály fiainak testmagassága?

11 fiú van, méretük cm-ben:

187 173 180 183 191 178 182 190 181 169 184

Állítsuk őket tornasorba:

169 173 178 180 181 182 183 184 187 190 191

Középső érték: 6. elem = 182.

12 fiú (= elem) 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 fiúk testmagassága:

`átlag = összeg/elemszám`
`mean()`, itt `mean(c(187 ...184)`. Eredmény: 181,6364.

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 fiúk testmagassága:

`átlag = összeg/elemszám`
`mean()`, itt `mean(c(187 ...184)`. Eredmény: 181,6364.

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 az osztályba beiratkozik Magyarország legmagasabb fiúja, aki 220 cm. Hogyan változik az átlag? És a medián?

Medián vagy átlag?

Képzeljük el, hogy az osztályba beiratkozik Magyarország legmagasabb fiúja, aki 220 cm. Hogyan változik az átlag? És a medián?

`átlag = 184,8333, medián = 182,5`

Egyetlen új elem 3 cm-mel növeli az átlagot, a medián viszont csak 0,5-tel.

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

Ábrázolás például a `barplot()` függvénnyel.

R

Változók típusa az R-ben

Honnan tudjuk, milyen változó?
`class()`
pl. numeric, character, matrix

Változók típusa az R-ben

Honnan tudjuk, milyen változó?
`class()`
pl. numeric, character, matrix

Ha bizonyos változókkal bizonyos műveleteket végzünk, az R néha átalakítja őket!
`e = cbind(a,d)`
Ok: mátrix csak egyféle változótípusból állhat, ezért a számból betű lesz.

Változók típusa az R-ben

Honnan tudjuk, milyen változó?
`class()`
pl. numeric, character, matrix

Ha bizonyos változókkal bizonyos műveleteket végzünk, az R néha átalakítja őket!
`e = cbind(a,d)`
Ok: mátrix csak egyféle változótípusból állhat, ezért a számból betű lesz.

`e = data.frame(a,b,d)`
Oszlopok: önmagukban is változók, osztályuk lekérdezhető így: class(e[,3]), azaz az e dataframe 3. oszlopa.

Adatok beolvasása az R-be

Adatokat többnyire más szoftverrel állítottuk elő (E-Prime, Praat, manuális lejegyzés, Excel-táblázat).

Praktikus formátum: .csv (comma-separated file)
`read.csv("fájlnév")`

Feladat: a *visz(ont)lát(ásra)* köszönésre kapott válaszok táblázatba rendezése Excelben vagy hasonló táblázatkezelő szoftverben. Mentés viszlat.csv fájlként.

Érdemes kerülni az ékezetes betűket és bármi nem-ASCII karaktert.

Grafikus felület (Mac, Windows)

Betöltés nem lehetséges közvetlen elérési útvonallal. Ehelyett:

(1) R-konzolban (ablak) File > Change directory... megkeressük a könyvtárat, ahova viszlat.csv-t mentettük.

```
viszlat=read.csv("C:/Users/en/Documents/viszlat.csv")
```

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(viszlat)`: oszlopban tárolt változók neve.

`head(viszlat)`: első hat adatsor.

data.frame változóira (oszlopaira) hivatkozás: viszlat$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

Dobjunk fel egy dobókockát tizenkétszer, és írjuk le, milyen számokat kaptunk.

Ha nincs dobókockánk...

Egész számok véletlenszerű generálása:
`sample()`
Argumentumok és jelentés: `?sample` vagy `help(sample)`.
Két argumentum megadása kötelező: `x` = miből húzzon, `size` = hányszor.

x lehet `c(1,2,3,4,5,6)` vagy `1:6` vagy `seq(1,6)`.

Ha nincs dobókockánk...

Egész számok véletlenszerű generálása:
`sample()`
Argumentumok és jelentés: `?sample` vagy `help(sample)`.
Két argumentum megadása kötelező: `x` = miből húzzon, `size` = hányszor.

x lehet `c(1,2,3,4,5,6)` vagy `1:6` vagy `seq(1,6)`.

```
kocka = sample(1:6,12)
```

Ha nincs dobókockánk...

Egész számok véletlenszerű generálása:
`sample()`
Argumentumok és jelentés: `?sample` vagy `help(sample)`.
Két argumentum megadása kötelező: `x` = miből húzzon, `size` = hányszor.

x lehet `c(1,2,3,4,5,6)` vagy `1:6` vagy `seq(1,6)`.

```
kocka = sample(1:6,12)
```

$\Rightarrow$ hibajelzés.

Figyelem! Default argumentum szerint egy számot csak egyszer lehet kihúzni: `replace=FALSE`. Megoldás: replace=TRUE.

Feladatok

Foglaljuk össze egy mátrixban, hányadikra hányat dobtunk. Így nézzen ki:

1	6
2	6
3	5
4	6
.	.
.	.
.	.

Adjuk meg a móduszt, mediánt, átlagot, és a dobások összegét.

Hasznos parancsok: `cbind()`, `rbind()`, `table()`, `median()`, `mean()`, `sum()`.

Feladatok

Dobjunk 100-at, majd 1000-et. Hasonlítsuk össze a középértékeket.