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

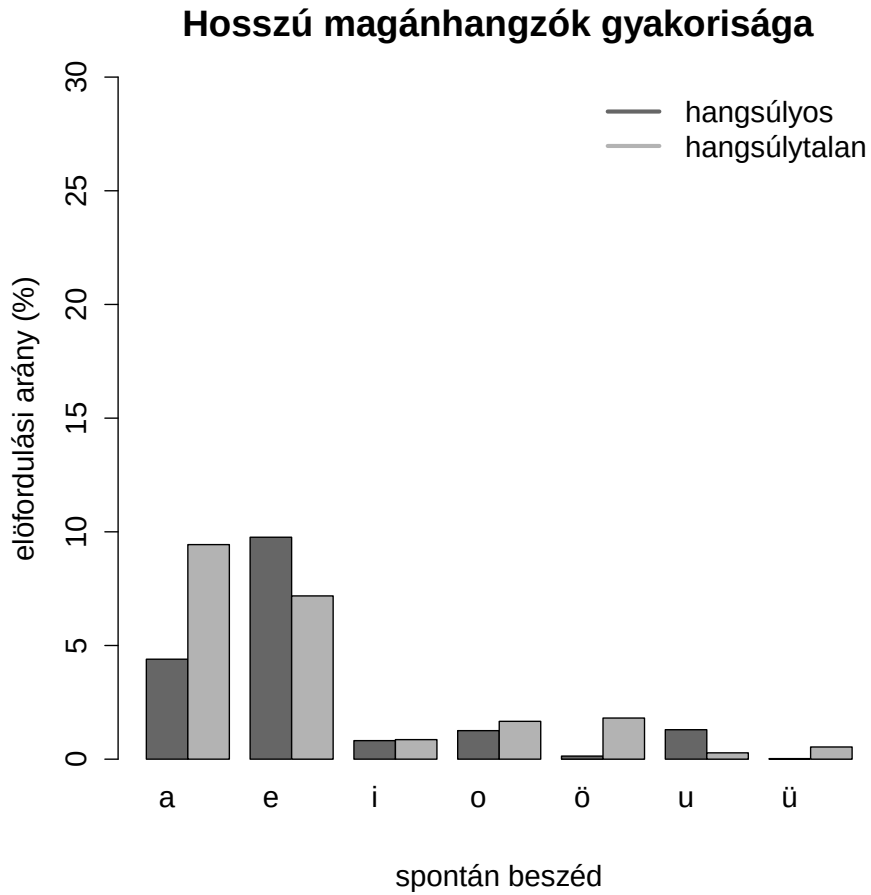
- **Módusz:** a mintában a legnagyobb gyakorisággal előforduló adatérték.
- **Medián:** a növekvő sorba rendezett adatok közül a középső. Ha az n mintaelemszám páros, a két középső érték átlaga.
- **Átlag:** mintabeli adatok számtani közepe.

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. Kezdő szinten például az Excelben.

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.

Olvassunk be egy csv-fájlt innen:
jonapot1 = read.table("http://phon.nytud.hu/mady/courses/statistics/2023/jonapot1.csv")

Nyissuk meg az R-ben.

Fontos szempontok:

- Milyen a karakterkódolás?
- Vannak-e oszlopnevek?
- A decimális pont vagy vessző?
- Az oszlopokat hogyan válasszuk el egymástól az átalakításnál?

Saját gépünkön az alapértelmezett karakterkódolás ellenőrzése:
`Sys.getlocale()`

Olvassunk be egy újabb csv-fájlt:
jonapot2 = read.table("http://phon.nytud.hu/mady/courses/statistics/2023/jonapot2.csv")

Ábrázoljuk mind a két objektum adatait egy táblázatban aszerint, hogy milyen válaszokat kaptunk a válaszadó neme szerint.

Használjuk az oszlopnevek megadásánál ezt a formátumot:
`data$oszlop`
`table(jonapot1$nem,jonapot1$koszones)`

Most készítsünk táblázatot a jonapot2 objektumból. Mi a különbség?

Feladat: a saját adatokról táblázat készítése megfelelő formában. Szempontok: ékezetek kerülése VAGY a saját laptop karakterkódolása szerinti mentés. Az oszlopok tabulátorokkal legyenek elválasztva.