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

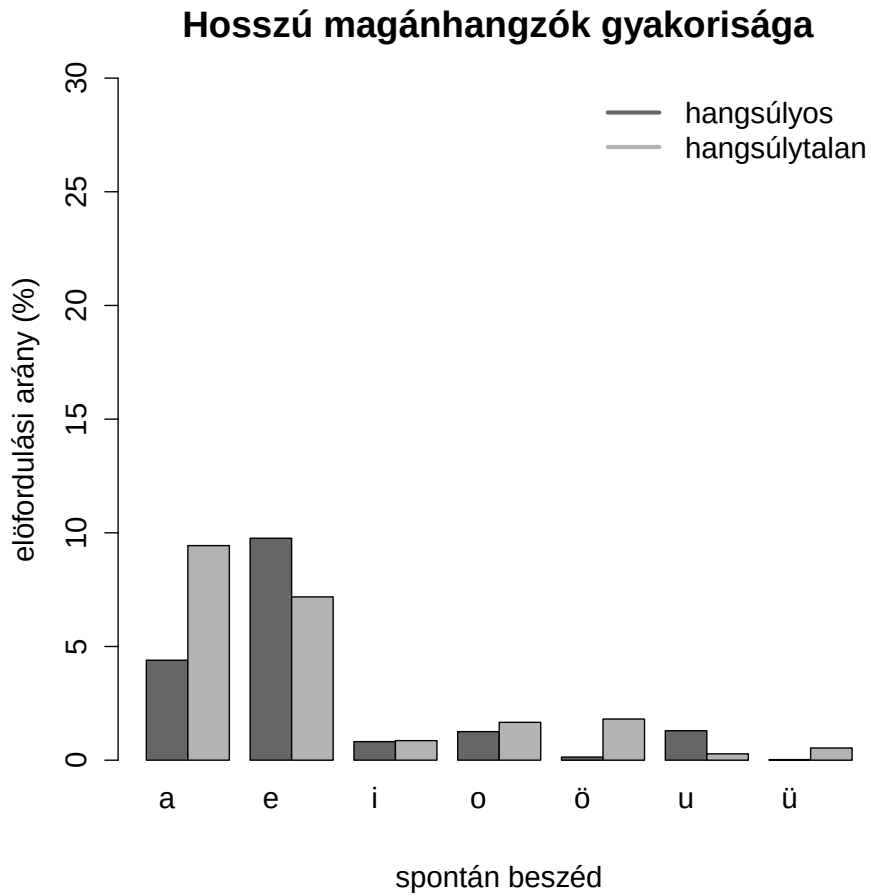
- **Módusz:** a mintában a legnagyobb gyakorisággal előforduló adatérték.
- **Medián:** a növekvő sorba rendezett adatok közül a középső. Ha az n mintaelemszám páros, a két középső érték átlaga.
- **Átlag:** mintabeli adatok számtani közepe.

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

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

Középértékek: 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.

A 9b osztályba 11 lány jár. A testmagasságuk centiméterben:

181 177 167 158 161 165 171 163 168 170 160

Most állítsuk őket tornasorba magasság szerint:
158 160 161 163 165 167 168 170 171 177 181

Középső érték: 6. elem = 167.

Ha egy 188 cm magas lány jön az osztályba, a tornasor így változik:
158 160 161 163 165 167 168 170 171 177 181 188

A medián ekkor a két középső elem átlaga, vagyis 167,5 lesz.

Középértékek: átlag (számtani közép)

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

A 9b-be járó 11 lány testmagasságának átlaga:

`átlag = (181+177+167+158+161+165+171+163+168+`
`+170+160)/11 = 167,36363 cm`

Az átlag egy statisztikai modell: olyan értéket is felvehet, ami nem szerepel a mért adatok között. Itt például egész számokban mértük a testmagasságot, de az átlag egy tört szám.

Fontos: átlagot kizárólag ekvidisztáns adatokra lehet számolni, ahol az egyes értékek egyenlő távolságra vannak egymástól. Tehát teljesülnie kell az intervallumskála vagy arányskála feltételeinek!

Iskolai osztályzatok?

Középértékek: átlag (számtani közép)

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

A 9b-be járó 11 lány testmagasságának átlaga:

`átlag = (181+177+167+158+161+165+171+163+168+`
`+170+160)/11 = 167,36363 cm`

Az átlag egy statisztikai modell: olyan értéket is felvehet, ami nem szerepel a mért adatok között. Itt például egész számokban mértük a testmagasságot, de az átlag egy tört szám.

Fontos: átlagot kizárólag ekvidisztáns adatokra lehet számolni, ahol az egyes értékek egyenlő távolságra vannak egymástól. Tehát teljesülnie kell az intervallumskála vagy arányskála feltételeinek!

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

Medián vagy átlag?

Nézzük meg, hogyan változik a 9b-be járó lányok testmagasságának átlaga az új osztálytárs megérkezése után.

`átlag = (181+177+167+158+161+165+171+163+168+ +170+160+188)/12 = 169,0833 cm`

Vagyis az új lány érkezésével az átlagos testmagasság majdnem 2 cm-vel nőtt.

Probléma: az osztályban egy kiugróan magas új érték jelentősen felfelé húzza az átlagot, holott a túlnyomó többség, 11 ember testmagassága változatlan.

Sokan csodálkoznak, ha a KSH közzéteszi az átlagfizetéseket. Most már azt is tudjuk, miért:
`https://elemzeskozpont.hu/median-es-atlagszamitas-miert-vezet-felre-az-atlag`

R

Adatok beolvasása az R-be

Az adatokat többnyire más szoftverrel állítottuk elő, vagy mástól kaptuk őket. 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.

Nyissuk meg az R-t.

Olvassunk be egy csv-fájlt. Másoljuk be ezt a parancssort egyetlen sorba, sortörés nélkül:

```
evszakok =
read.csv("https://phon.nytud.hu/mady/courses/statistics/2024/evszakok.csv")
```

Nyissuk meg az R-ben, azaz írjunk be ennyit az ablakba: `evszakok`

Fontos szempontok:

- Az egyazon típusú adatok oszlopokba vannak rendezve.
- Az oszlopokat a .csv fájlban vessző választja el egymástól (ez lehetne tabulátor, space vagy pontosvessző is).
- A tavasz és az ősz kódja egységesen nagybetűs T és O.
- Ugyanígy a nő és férfi kódja egységesen nagy N és F.

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

A kódolás ennél a fájlnál nem probléma, de ha vannak nem ASCII-kóddal írt adatok, akkor a csv-fájl és az R kódbeállítása azonos kell, hogy legyen.

Az R-ben tárolt egységek legfontosabb típusa az objektum és a függvény. Az objektum egy adatsor, amit betöltünk vagy az R-ben állítunk elő.

A függvények parancsok, amik az R-ben rendelkezésre állnak, vagy magunk írjuk meg őket. Arról ismerhetőek fel, hogy mindig kerek zárójel követi őket.

Példa: írjuk be az `ls()` sort. Ez a *list* (listázz) függvény megfelelője.

Az eredmény az R-ben tárolt objektumok listája. Itt egyelőre egyetlen objektum van, az `evszakok` nevű. Ezt a nevet mi adtuk neki, amikor a fájl betöltésekor az `evszakok=` nevű objektumba mentettük a fájlt a `read.csv("fajlnev.csv")` függvény megadásával.

Az `evszakok` objektum típusa data.frame, egy kétdimenziós táblázat, aminek több sora és több oszlopa van.

Az adattáblázatokban az oszlopoknak van egy neve, hasonlóan az Excelbe beírt adatokhoz, ahol az első sor jellemzően az oszlopnév.

Az oszlopneveket így lehet lekérdezni: `names(evszakok)`.

Az oszlopokban tárolt adatokra így lehet hivatkozni: objektum$oszlopnev.

Számoljuk ki az összes pontszám mediánját és átlagát:

```
median(evszakok$pontszam)
mean(evszakok$pontszam)
```

Mi okozhatja ezt a viszonylag nagy eltérést a medián és az átlag között?

Ábrázoljuk az adatokat dobozdiagram segítségével:
`boxplot(evszakok$pontszam)`

Az ábrán a vízszintes fekete vonal jelöli az összes adott pontszám mediánját, vagyis 15 pontot. A doboz teste a középső 50%-ot jelöli, a lábak pedig az annál alacsonyabb és magasabb értékeket.

Melyik irányba „húzza” a láb a dobozt, vagyis merre vannak erősebben kilógó értékek?

Ilyenkor segít a pontok eloszlásának ábrázolása oszlopdiagram segítségével. Ehhez a pontszámokat táblázatba kell rendeznünk így:
`table(evszakok$pontszam)`

Látszik, hogy a legalacsonyabb pontszám 4, és hogy alacsony (10 alatti) pontszámot viszonylag kevesen adtak. Ábrázoljuk az adott pontszámok gyakoriságát:
`barplot(table(evszakok$pontszam))`

Az átlag vagy a medián adja jobban vissza a pontszámok eloszlását?

A kutatási kérdésünk az, hogy az emberek jobban szeretik-e a tavaszt, mint az őszt.

Nullhipotézis?

Dobozdiagram évszakokra bontva, egyszerűsített szintaxissal:
`boxplot(pontszam~evszak, data = evszakok)`

Mi történik, ha az évszakok nevét kisbetűvel kódoljuk?

Töltsük be a módosított fájlt innen, és mentsük az evszakok2 objektumba:

```
evszakok2 =
read.csv("https://phon.nytud.hu/mady/courses/statistics/2024
/evszakok_kis.csv")
```

Mi más, mint az előbbi dobozdiagramon?

Az R-ben létrehozott objektumokat és a begépelt függvények sorát el tudjuk menteni.

Az R-hez hozzá van rendelve egy könyvtár (mappa), ahova kilépéskor menti a munkafolyamatot.

Ennek helyét így lehet lekérdezni:
`getwd()`
Azaz: *get working directory*.

Ezt érdemes megjegyezni, mert az R-et a jövőben is ebből a könyvtárból érdemes megnyitni, és az adatfájljainkat is itt érdemes tárolni.

Másoljuk be ebbe a könyvtárba a saját adatainkat tartalmazó csv-fájlt.

Az R-be így tudjuk betölteni:

```
evszakaim = read.csv("adatfajlom.csv")
```

Házi feladat: a saját gyűjtésű adatok pontszámaiból medián és átlag számítása, valamint dobozdiagram és oszlopdiagram készítése. Beküldési határidő február 26. éjfél.