Adatelemzés az R-ben

2014. április 25.

Kísérleti adatok elemzése

Kísérlet célja: valamilyen állítás vagy megfigyelés empirikus és szisztematikus tesztelése. Pl. „a nők többet beszélnek, mint a férfiak”, „nyáron gyorsabban nő a hajunk, mint télen” stb. A kísérletek alapja az összehasonlítás.

Kísérleti adatok elemzése

Eljárás:

1. Munkahipotézis (H_1): a nők többet beszélnek, mint a férfiak.
2. Adatgyűjtés minél többféle helyzetben úgy, hogy a nők és a férfiak adatai összehasonlíthatóak legyenek.
3. Parametrizálás: számszerű mutató, pl. produkált szavak száma adott időtartamon belül, beszédidő adott időtartamon belül stb.
4. Kiindulási hipotézis (H_0) statisztikai tesztelése: feltételezzük, hogy a nők és a férfiak ugyanannyit beszélnek. Ha sikerül kimutatni, hogy a nők vagy a férfiak egységnyi idő alatt többet beszélnek, mint a másik csoport, akkor elvetjük a nullhipotézist, és feltételezzük, hogy H_1 igaz.

Az eredmények prezentálása

A kísérleti eredményeket bemutató előadások és cikkek felépítése állandó sémát követ:

1. Bevezetés: miért releváns a kérdés, mit írtak róla az irodalomban, mi az, amit még nem tudunk?
2. Anyag és módszerek: a felhasznált anyag minél pontosabb bemutatása, valamint az adatok elemzése (statisztikák, esetleges nem világos kérdések).
3. Eredmények: a konkrét kísérlet eredményeinek bemutatása szóban és diagramokon.
4. Következtetések: az eredmények értékelése a bevezetésben felvázolt összefüggések alapján, esetleges további nyitott kérdések vázolása.

Példa: három beszélő rövid és hosszú *u* – *ú* magánhangzóit hasonlítjuk össze rövid és hosszú mondatokban. Hipotézisek:

1. Feltételezzük, hogy a hosszú /uː/ tartama nagyobb, mint a rövid /u/-é.
2. Feltételezzük, hogy a hosszabb mondatokban gyorsabb a beszédtempó, ezért a magánhangzók általában rövidebbek.

A hipotéziseknek korábbi szakirodalomra kell támaszkodniuk. Feltehetünk egyéb kérdéseket is, pl.

- Ugyanúgy aránylanak-e a rövid és hosszú magánhangzó-tartamok egymáshoz a rövid és a hosszú mondatokban?
- Hosszabb-e a rövid /u/ megvalósulása a rövid mondatban, mint a hosszú mondatbeli /uː/-é?

Az elemzés menete

- Nagyobb osztású csoporttól a kisebb felé.
- Először összehasonlítjuk az összes rövid /u/ tartamát az összes hosszú /uː/ tartamával.
- Összehasonlítjuk a két magánhangzó-hosszot a kétféle hosszúságú mondaton belül.
- Megnézzük, hogy a tendencia minden beszélőre igaz-e.

Az R statisztikai szoftver

Letöltés: www.r-project.org, onnan elérhető tükrök.

Windows GUI (graphical user interface): személyre szabott telepítés: eldönthető, hogy terminál és ábrák egy ablakba kerüljenek, vagy kettőbe.

Linux: általában alapcsomag része, ha nem, repositoryból letölthető. Nincs GUI, megnyitás terminálablakban R paranccsal.

Objektumok az R-ben

Lekérdezés: `class(objektum)`

- `vector:` egydimenziós, pl. `[1,2,5,6]`, `["a","e","i","u"]`. Egy vektorban egyféle típusú adat található (csak string, csak numerikus stb.). Szám lehet string, de fordítva nem.
- `matrix:` kétdimenziós, minden sor és minden oszlop egyforma hosszú. Adatok egyféle típusúak.
- `data.frame:` kétdimenziós adattáblázat, adattípusok oszloponként változhatnak.

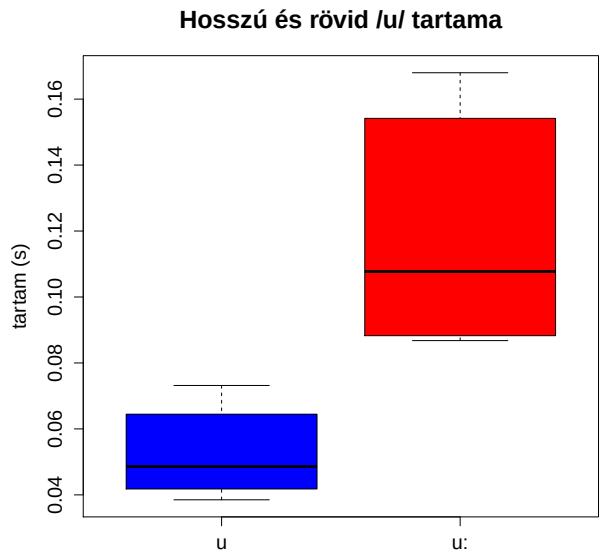
Adattípusok: numeric, integer, character, factor, logical stb.

`maganhangzo` data.frame oszlopaiban található adattípus lekérdezése: `class(maganhangzo$mondatszam)`.

Dobozdiagram (boxplot)

Adatok beolvasása:

```
objektum = read.table("file",header=T,sep=";")
```



Eljárás: összes mért adat sorrendbe állítása legkisebbtől legnagyobbig. Középső vízszintes vonal: középső adat. Doboz alsó és felső határa: 25 és 75%. Alső és felső talp: 10 és 90%. Ha az adatok szimmetrikus eloszlásúak, a dobozdiagram is szimmetrikus.

Előállítása R-ben

Függvény:

`boxplot(mertadatok~osztalyok,objektum)`

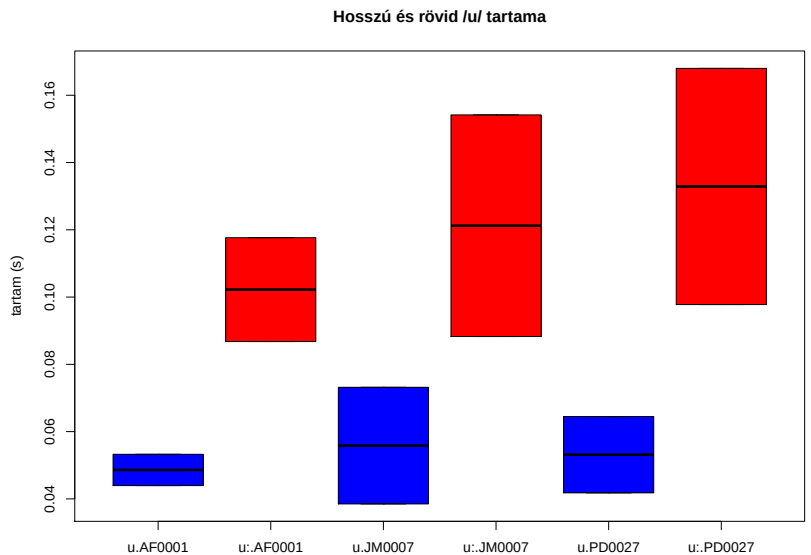
azaz

`boxplot(dur~vowel,data=u)`

Ugyanez beszélőnként

`boxplot(dur~vowel*subj,data=u)`

vagy: `boxplot(u$dur~u$vowel*u$subj)`



Részhalmaz ábrázolása

Ha az adatoknak csak egy részét akarjuk ábrázolni: logikai vektor.

Változóra igaz, hogy:

`resz = u$subj == "AF0001"`
`resz`: objektum elemeinek száma TRUE, amelyekre a feltétel teljesül. A függvények csak ezekre az elemekre lesznek érvényesek.

```
boxplot(u$dur[resz]~u$vowel[resz])
```

Diagram mentése

Windows: különböző képformátumok jobb egérgombbal.

Linux: pdf, ps.

```
dev.print("directory/file",device=postscript)
```

vagy

```
dev.print("filenev",device=pdf)
```

Adatmátrixok összekapcsolása

Fenti adatbázisban jelölni akarjuk a rövid és hosszú mondatokat.

Újabb adatmátrix létrehozása szöveges fájlként (pl. .txt):

```
sent;length
11;long
12;long
17;short
18;short
```

Beolvasás:

```
sentencelist =
read.table("sentencelist.txt",header=T,sep=";")
```

A `sent` változó adatai megegyeznek, erre építve egyesítjük a két mátrixot:

```
u = merge(u,sentencelist,by="sent")
```