R szoftver: függvények és gyakorló feladatok

Mády Katalin

MTA Nyelvtudományi Intézet

Műveletek adatmátrixokkal

`http://clara.nytud.hu/~mady/courses/statistics/materials/politeness.RData` → `pol` objektum a memóriában.

Három női és három férfi beszélő hangjának alapfrekvenciája közvetlen és udvarias beszédstílusban, három különböző helyzetben.

Számoljuk ki a női és férfi beszélők összesített átlagát a kétféle stílusban:

```
tapply(pol$frequency,
data.frame(pol$gender,pol$attitude),mean)
```

Ábrázoljuk az átlagokat:

```
interaction.plot(pol$attitude,pol$gender,pol$frequency,mean)
```

Hiányzó adatok

NA: Not Available, vagyis üres cella. Van NaN is (= Not A Number), pl. a 12/0 eredménye.

Átlag és számos más függvény számításakor ki kell zárni az üres cellákat:

```
tapply(pol$frequency,
data.frame(pol$gender,pol$attitude),mean,na.rm=T)
```

Interakció ábrázolásánál ez nem működik. Ehelyett meg kell szűrni az adatmátrixot, és kidobni az NA értékeket.

A `z = pol$frequency==NA` nem szűr, hanem NA-t ad eredményül.
Helyette `z = is.na(pol$frequency)==F`.
Utána

```
interaction.plot(pol$attitude[z],pol$gender[z],pol$frequency[z
```

Változók nevének egyszerűbb formátuma

Ha nem akarjuk a dollárjelet mindig kiírni, használhatjuk az `attach(dataframe)` függvényt, ekkor az egyes oszlopok változóira hivatkozhatunk önálló vektorként.

`attach(pol)`

Veszély: ha már van egy azonos nevű globális változónk, pl. `subject`, akkor ezzel felülírjuk a `pol` adatmátrix `subject` vektorát. Erről hibajelzést kapunk:

`The following object(s) are masked from 'package:datasets': attitude`

`search()` keresési útvonal sorrendje. Az előrébb levő csomag változóinak van elsőbbsége. Leválasztás `detach()` függvénnyel.

Alternatíva: `with(dataframe,fuggveny)`, vagy `datasets::attitude`

Hasznos függvények

Ismételt méréses próbák

```
with(pol,tapply(frequency,data.frame(subject,attitude),mean))
```

subject	inf	pol
F1	246.7857	217.28571
F2	270.0143	246.35714
...		

Nekünk viszont ilyen output kell:

subject	attitude	frequency
F1	inf	246.7857
F1	pol	217.28571
...		

Cél: `attitude` vektor előállítása 6 inf + 6 pol stringgel és 12 hozzájuk tartozó frekvenciaértékkel. Függvények: `rep(c("inf","pol"),each=6)`, `data.frame()`.

Faktorok törlése és rendezése

Állítsunk elő egy dobozdiagramot csak a három női beszélő frekvenciaértékeiből, egyenként:

```
z = pol$gender == "F"
boxplot(frequency~subject,pol[z,])
```

Fölösleges faktorszintek törlése (figyelem, ez visszafordíthatatlan művelet, készítsünk biztonsági másolatot!):

```
levels(pol$subject)[4:6] = NA
```

Faktorok nem alfanumerikus sorrendje:
Egy faktor előreállítása: `relevel(vektor, 1.faktor)`
Több faktor tetszőleges rendezése: vec = factor(pol$gender,levels=c("M","F"))

Ha fölülírtunk egy objektumot, amiről nincs másolatunk, lépjünk ki az R-ből, és ne mentsük a memóriát.

Két adatmátrix összefűzése

Feltétel: mindkettőben azonos nevű oszlop.

Példa: oz.csv és oz_terulet.csv innen:
`http://clara.nytud.hu/~mady/courses/statistics/materials`

Őzek élőhelyének kiegészítése további adatokkal, és ezek beágyazása az adatmátrixba:

```
oz.uj = merge(oz,oz_terulet,by="TERULET")
```

Faktorok kinyerése fájlnevekből

filenames.RData a szokásos linkről.

Fájlnevek szerkezete: beszélő kódjakísérlet neve_ismétlés_ mondatszám. A vektor típusa faktor.

`strsplit` függvénnyel string változóvá alakítjuk a faktorokat, majd elvágjuk a "_" mentén. Alapértelmezett output: lista, ezt vektorrá alakítjuk.

```
file.list = unlist(strsplit(as.character(filenames),"_"))
```

Utána a vektor 1., 4., 7. stb. elemét a subj változóba írjuk:

```
subj = file.list[seq(1,length(file.list),3)]
```

Ehhez hasonlóan az ismétléseket és a mondatazonosítókat is.