Valószínűség
Logikai vektorok az R-ben

Valószínűség a mindennapokban

Köznyelvi jelentés: tapasztalat alapú becslés (n megfigyelt esetből hányszor történt meg egy adott esemény). Pl. "valószínűleg mindjárt elered az eső" (mert ha ilyen borús az ég, gyakran esik), "valószínűleg idén sem lesz fizetésemelés" (mert tíz éve nem volt).

A valószínűség soha nem jelent biztos tudást! Néha mégsem esik, ha borús az ég, és néha mégis van fizetésemelés.

Az intuitív becslésnek kevés fokozata van: *nem túl valószínű, elég valószínű, nagyon valószínű, több mint valószínű*.

Valószínűség a szerencsejátékban

Fej vagy írás egy érme feldobásakor?

Megfigyelés: 10 dobás, 20, 30...

Fejek száma egyre jobban közelíti a 0,5-ös értéket.

Empirikus valószínűség, vagyis a konkrét mintán megszámolt várható valószínűség, P definíciója:

P = dobott fejek/összes dobás,

ahol a dobások száma a végtelenhez közelít.

$\Rightarrow$ A valószínűség értéke mindig 0 (egyáltalán nem valószínű) és 1 (biztos) között mozog.

További magyarázat gyanánt kötelező olvasnivaló:

Reiczigel, Harnos & Solymosi (2006): Biostatisztika nem statisztikusoknak. Nagykovácsi: Pars.

1. fejezet (13–22. oldal): a statisztika felhasználási módjai.
3.1. rész (51–56. oldal): logika és valószínűség.

Feladat 1

Mekkora a valószínűsége, hogy...

1. ötöst dobunk a kockával
2. királyt húzunk egy 32 lapos kártyapakliból,
3. kétszer feldobunk egy érmét, és mindkétszer fej lesz az eredmény,
4. egy véletlenszerűen megkérdezett magyar állampolgár katolikus, ha az összlakosság összetétele: 51% katolikus, 16% református, 3% evangélikus, 14,5% nem vallásos, 5% más vallású, 10,5% nem válaszol,
5. egy véletlenszerűen megkérdezett személy diplomás nő, ha a diplomások aránya 22,4% és a nők aránya 50%?

R

Logikai vektorok funkciója

Bizonyos változókra való szűrés: egy részhalmaz, amelynek az elemeire igaz a megadott feltétel.

Példák: a ratings mátrixban csak a növénynevek, csak bizonyos betűszűm fölötti szavak, csak a nagyon ritkák, stb.

Hasonlóan a logikához, a feltételeket itt is össze lehet kapcsolni egymással, pl. növény ÉS ritka.

Logikai vektorok jelölése

Operátorok:

==	azonos
!=	nem azonos (karakter)
<	kisebb
>	nagyobb
<>	nem egyenlő (numerikus)
|	vagy
&	és
%in%	tartalmazza a vektor valamely elemét

Logikai vektorok létrehozása

Ha csak a növényekre vagyunk kíváncsiak:
a `noveny` objektumba mentjük, hogy mely elemekre teljesül a feltétel:
`noveny = ratings$Class == "plant"`
Ha az R-be beírjuk, hogy `noveny`, látjuk, hogy az elemek vagy TRUE, vagy FALSE értéket kapnak.

Ha azokat a konkrét szavakat akarjuk látni, amikre igaz, hogy növények, akkor
`ratings$Word[noveny]`

MIÉRT?

Az R minden egyes elemszámra megjegyzi, hogy TRUE vagy FALSE értéket tartalmaz-e a feltétel alapján létrehozott logikai vektor, azaz a `noveny` objektum. A vektor 81 értéket tartalmaz. Az R a műveleteknél csak azokat a sorokat veszi figyelembe, ahol a `noveny` objektumban az adott elemszámnál TRUE érték van.

Logikai vektorok működése

Emlékeztető az első óráról:
matrixunk[1:4,]
Ekkor az R a matrixunk mátrix első négy sorát listázza ki.

Analóg módon történik a feltételt teljesítő sorok listázása:
`ratings[noveny,]`
Ekkor az R a táblázat azon sorait írja ki, amire igaz, hogy a szó osztálya növény, valamint az összes oszlopot.

Honnan tudja ezt az R? Tudja, hogy milyen elemszámokra teljesül a feltétel. Lekérdezhető így:
`which(noveny)`

Összes előfordulás, vagyis hány elemre igaz, hogy növény:
`sum(noveny)`

Ha NEM a növényekre vagyunk kíváncsiak (hanem minden másra, pl. állat, vírus, élettelen):

```
nemnoveny = ratings$Class != "plant"
```

Ha csak a nagyon ritka szavakra vagyunk kíváncsiak, pl. az egyes számú alak 10-nél kevesebbszer fordul elő:

```
ritka = ratings$FreqSingular < 10
```

Csak a ritka növények:

```
ritkanoveny = ratings$Class == "plant" &
ratings$FreqSingular < 10
```

Feladat 2

1. Hány növény és hány állat van a mátrixban?
2. Hány olyan elem van, amelyiknek az egyes számú gyakorisága 500-nál nagyobb? Melyek ezek?
3. Hány olyan elem van, amelyiknek az egyes és többes számú gyakorisága egyenlő?
4. Melyik elemnek 7 mindkét fenti mérőszáma?
5. Melyik csoport ismertebb a meanFamiliarity mutató alapján, a növények vagy az állatok? Mekkora a különbség az átlagok között?

Részhalmazok ábrázolása közös ábrában

Pontdiagram, amiben a növények és az állatok gyakoriségi (Frequency) és ismertségi (meanFamiliarity) mutatója van ábrázolva eltérő színekkel. Fontos: a tengelyhosszokat nekünk kell definiálnunk, különben az R eltérő alsó és felső értéket fog beállítani.

Növények definiálása logikai vektorral:
`noveny = ratings$Class == "plant"`

A gyakoriság és az ismertség terjedelme megkapható így:
`range(ratings$Frequency)`,
`range(ratings$meanFamiliarity)`
Ez kiváltja a `c(min(),max())` paramétert, amiben a minimum és maximum értékeket egyenként kötnénk össze.

Részhalmazok ábrázolása közös ábrában

Növények ábrázolása zöld színnel:

```
plot(ratings$Frequency[noveny],
ratings$meanFamiliarity[noveny],
col="green",xlim=range(ratings$Frequency),
ylim=range(ratings$meanFamiliarity))
```

Ha ugyanebbe az ábrába akarunk egy további függvénnyel másikat rajzolni:

```
par(new=T)
```

Itt nem fog semmi történni, az R várja a következő rajzoló parancsot.

Állatok ábrázolása kék színnel. Mivel a Class változó csak kétféle adatot tartalmaz, nem kell külön definiálni az állatokat, elég azt mondani, hogy a nem növények, vagyis `[!noveny]`.

```
plot(ratings$Frequency[!noveny],
ratings$meanFamiliarity[!noveny],
col="blue",xlim=range(ratings$Frequency),
ylim=range(ratings$meanFamiliarity))
```

Feladat 3

A fenti ábrának legyen címe és olvasható tengelyfelirata valamilyen nyelven.

Boxplot készítése: növények között az egyszerű és összetett szavak gyakorisága (ratings$Complex).

Igaz-e, hogy a ratings mátrixban szereplő állatnevek gyakorisága alacsonyabb mértékű, mint a növényeké? És az ismertségük? Szemléltessük ezt egy tetszőleges ábrán.