Analiza datelor în R

Curs 6

Analiză exploratorie multivariată

Scop: descrierea relaţiilor dintre variabile

- cantitative vs. calitative
- calitative vs. calitative
- cantitative vs. cantitative

Relaţia dintre o variabilă cantitativă şi una calitativă

1. Vom considera setul de date *iris* din *datasets*.
Vom analiza lungimea petalelor, lăţimea petalelor, lungimea sepalelor şi lăţimea sepalelor, comparativ pe specii:
   - calcul de indicatori numerici
   - reprezentare grafică (boxplot grupat pe categorii):
     `boxplot(varCantitativa~varCalitativa)`

Relaţia dintre o variabilă cantitativă şi una calitativă

2. Vom utiliza setul de date *energy* din pachetul *ISwR* pentru a compara consumul de energie (*expend*) pentru 22 femei, în funcţie de clasa de greutate corporală (*stature* $=$ *lean* sau *obese*).

- boxplot grupat
- Atunci când grupurile analizate sunt mici, este preferabil să se reprezinte grafic datele sub formă de puncte (stripchart):
  `stripchart(expend~stature,method="jitter")`

Relaţia dintre două sau mai multe variabile calitative

Se utilizează pentru reprezentare tabele de contingenţă.

Introducerea unui tabel preexistent

Se analizează preferinţa pentru un anumit tip de băutură (cola, cafea sau ceai) pe un eşantion de 105 persoane şi se obţin răspunsurile din tabelul de mai jos:

		băutura		
		cola	cafea	ceai
F/M	F	10	12	28
	M	20	31	4

Datele se preiau într-o matrice, care se converteşte în tabel cu funcţia `as.table`.

Relaţia dintre două sau mai multe variabile calitative

Cod R:

```
bauturi=matrix(c(10,12,28,20,31,4),byrow=T,nrow=2)

# denumirea liniilor si coloanelor
colnames(bauturi)=c("cola","cafea","ceai")
rownames(bauturi)=c("F","M")
bauturi

# denumirea variabilelor reprezentate pe linii
# si coloane
names(dimnames(bauturi))=c("F/M","bautura")
bauturi

# conversie in tabel
bauturi=as.table(bauturi)
bauturi
```

Relaţia dintre două sau mai multe variabile calitative

Frecvenţe marginale=frecvenţe pentru fiecare variabilă calitativă

`margin.table(bauturi,1)` $\rightarrow$ frecvenţe marginale pentru atributul de pe linii
`margin.table(bauturi,2)` $\rightarrow$ frecvenţe marginale pentru atributul de pe coloane

Tabele de frecvenţe relative

`prop.table(bauturi,1)` $\rightarrow$ proporţii pentru fiecare celulă din totalul de pe linie (frecvenţe relative pentru fiecare nivel al atributului de pe linii)
`prop.table(bauturi,2)` $\rightarrow$ proporţii pentru fiecare celulă din totalul de pe coloană
`prop.table(bauturi)` $\rightarrow$ proporţii pentru fiecare celulă din totalul pe tabel

Relaţia dintre două sau mai multe variabile calitative

Tabularea datelor dintr-un dataframe: funcţiile `table` şi `xtabs`

Setul de date *UCBAdmissions* conţine informaţii despre situaţia admiterilor la Universitatea Berkeley în 1973.

```
xtabs(Freq~Gender+Admit)
prop.table(xtabs(Freq~Gender+Admit),1)
```
$\rightarrow$ proporţia de admişi pe sexe

Se observă că, dintre aplicanţii bărbaţi, 44.5% au fost admişi, în timp ce dintre femei, doar 30.4%. Analiza situaţiei pe fiecare departament arata însă că procentajele de femei şi bărbaţi admişi sunt comparabile.
$\longrightarrow$ **paradoxul lui Simpson**: un trend care apare într-o combinaţie de grupuri dispare sau se inversează în interiorul fiecărui grup.

Relaţia dintre două variabile cantitative

Se analizează pe un eşantion de volum n două caracteristici, X şi Y, şi se obţin perechile de valori $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

- grafic de împrăştiere: `plot(x,y)`
- coeficientul de corelaţie liniară (Pearson)

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

  - $r \in [-1, 1]$
  - $r < 0$ - corelaţie liniară negativă
  - $r \approx 0$ - nu există asociere liniară între variabile
  - $r > 0$ - corelaţie liniară pozitivă
  - r este invariant la translaţie sau scalare, deci datele poti fi standardizate.

În R: `cor(x,y)`

Relaţia dintre două variabile cantitative

1. Analizăm relaţia dintre înălţime şi greutate pentru 15 femei cu vârstele între 30 şi 39 de ani (setul de date *women* din *datasets*).

2. Setul de date *mtcars* din *datasets* conţine caracteristicile tehnice a 32 modele de maşini din perioada 1973-1974. Analizăm dacă există o relaţie între greutatea unei maşini (*wt*) şi consumul de combustibil (*mpg*).

Dacă relaţia dintre variabile nu este liniară dar este monotonă, se poate calcula coeficientul de corelaţie Spearman ρ = coeficientul Pearson calculat pentru rangurile observaţiilor. Coeficientul Spearman poate fi calculat şi pentru date ordinale.

În R: `cor(x,y,method="spearman")`

Relaţia dintre două variabile cantitative

Cantităţile care se modifică multiplicativ (i.e., rata de schimbare este proporţională cu cantitatea) pot fi logaritmate înainte de a fi analizate.

Exemplu:
Setul de date *Animals* din *datasets* conţine informaţii despre greutatea corporală (kg) şi greutatea creierului (g) la mai multe specii de animale. Vom analiza relaţia dintre cele două cantităţi în forma dată, apoi logaritmate.

Grafice quantilă - quantilă

În cazul a două serii de date (i.e. măsurători ale aceleiaşi caracteristici în două grupuri distincte), se utilizează pentru a determina dacă acestea urmează aceeaşi distribuţie.

- grupuri de aceeaşi mărime: se ordonează crescător valorile în cele două grupuri, i.e. $x[1] \leq x[2] \leq \ldots x[n]$, $y[1] \leq y[2] \leq \ldots y[n]$, apoi se reprezintă grafic punctele $(x[i], y[i])$
- grupuri de mărimi diferite: se reduce grupul mai mare la dimensiunea celui mai mic păstrând *Min*, *Max* şi alegând quantile echidistante între acestea.

Dacă distribuţiile sunt similare, punctele se vor găsi aproximativ pe prima bisectoare.

În R: `qqplot(x,y)`

Exerciţii

1. Pentru setul de date *airquality* din *datasets*, să se reprezinte temperaturile înregistrate sub formă de boxplot-uri grupate pe luni.

2. Setul de date *reaction.time* din *UsingR* conţine timpii de reacţie la un stimul extern pentru participanţii la un studiu, precum şi informaţii despre vârsta şi sexul acestora şi dacă foloseau sau nu telefonul mobil la momentul respectiv. Să se reprezinte timpul de reacţie în funcţie de:

a) categoria de vârstă

b) sex

c) utilizarea sau nu a telefonului mobil.

Exerciţii

3. În setul de date *twins* din *UsingR* se găsesc IQ-urile a 27 perechi de gemeni separaţi la naştere, dintre care unul a fost crescut de familia biologică (*Biological*), iar celălalt de o familie adoptivă (*Foster*). Să se determine dacă există o asociere între IQ-urile celor două categorii de gemeni.

4. Setul de date *kid.weights* din *UsingR* conţine înălţimile şi greutăţile a 250 copii cu vârstele cuprinse între 0 şi 12 ani. Să se reprezinte greutatea în funcţie de înălţime şi să se determine tipul relaţiei între aceste două mărimi şi coeficientul de corelaţie adecvat.

Exerciţii

5. Setul de date *UNLifeExpectancy*, disponibil la adresa `http://instruction.bus.wisc.edu/jfrees/jfreesbooks/RegressionModeling/BookWebDec2010/CSVData/UNLifeExp.csv`, conţine informaţii cu privire la speranţa de viaţă şi alţi parametri socio-economici ai ţărilor lumii.

a) Să se reprezinte histograma şi curba de densitate pentru speranţa de viaţă (*LIFEEXP*).

b) Să se reprezinte histograma pentru cheltuielile de sănătate (*HEALTHEXPEND*).

c) Să se determine dacă există o asociere între speranţa de viaţă într-o ţară şi cheltuielile cu sănătatea, respectiv speranţa de viaţă şi fertilitatea (*FERTILITY*) în ţara respectivă.