Analiza datelor în R

Curs 5

Indicatori statistici numerici

Se analizează o caracteristică de tip cantitativ pe un eşantion de volum n dintr-o populaţie şi se obţin valorile $x_1, x_2, \ldots, x_n$.

Indicatori de poziţie

- Media de selecţie

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

În R: `mean(x)`

- Mediana
  Dacă presupunem seria de date ordonată crescător $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, atunci

$$Me = \begin{cases} x_{\frac{n+1}{2}}, & \text{dacă } x = 2k+1 \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \text{dacă } x = 2k \end{cases}$$

În R: `median(x)`

Indicatori statistici numerici

- Valoarea modală = valoarea (valorile) cu cea mai mare frecvenţă în setul de date.

  În R: pachetul `modeest`

- p-quantilele ($p \in (0,1)$)
  $y = Q(p) \iff$ o proporţie p dintre valorile $x_1, \ldots, x_n$ sunt mai mici decât y, iar restul sunt mai mari decât y

  În R: `quantile(x,p)`

- quartile: $Q(0.25), Q(0.5), Q(0.75)$

  În R: `summary(x)`

Indicatori statistici numerici

Indicatori ai împrăştierii

- Amplitudinea = $Max - Min$
- Lungimea intervalului interquartilic
  $IQR = Q(0.75) - Q(0.25)$
- Dispersia de selecţie

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

  În R: `var(x)`

- deviaţia (abaterea) standard de selecţie $s = \sqrt{s^2}$

  În R: `sd(x)`

Indicatori statistici numerici

Indicatori de formă a distribuţiei

- Asimetria ("skewness") = $\frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{ns^3}$
  - $< 0 \rightarrow$ asimetrie stânga
  - $\approx 0 \rightarrow$ distribuţie simetrică
  - $> 0 \rightarrow$ asimetrie dreapta

  În R: `skewness` din pachetul `moments`

- Aplatizarea ("kurtosis") = $\frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{ns^4}$
  - $< 3 \rightarrow$ distribuţie platikurtică
  - $\approx 3 \rightarrow$ distribuţie normokurtică
  - $> 3 \rightarrow$ distribuţie leptokurtică

  În R: `kurtosis` din pachetul `moments`

Exemplu

Setul de date *cfb* din pachetul *UsingR* conţine informaţii dintr-un studiu asupra finanţelor consumatorilor efectual de USA Federal Reserve. Vom ilustra conceptele de mai sus pentru analiza variabilelor *AGE* (vârsta participanţilor la studiu) şi *INCOME* (veniturile acestora în 2001).

Tabele şi grafice de frecvenţe

Se folosesc pentru analizarea datelor categoriale sau a celor cantitative grupate pe categorii.

Tabele de frecvenţe

- `table(x)` $\rightarrow$ valorile distincte din x şi frecvenţele lor absolute
- `table(cut(x,k))` $\rightarrow$ intervalul de valori ale lui x se împarte în k subintervale de lungime egală şi se determină frecvenţa absolută pentru fiecare subinterval

Grafice de frecvenţe

- `barplot(vectFrecv,names.arg=vectCateg)` sau `barplot(table(x))`
- `pie(vectFrecv,vectCateg)` sau `pie(table(x))`

Exemple

1. Setul de date *central*.*park*.*cloud* din pachetul *UsingR* conţine informaţii cu privire la vremea din Central Park în luna mai 2003 (cer senin / parţial noros / noros). Să se construiască tabelul de frecvenţe şi graficele de frecvenţe.

2. Setul de date *airquality* din *datasets* conţine măsurători zilnice asupra calităţii aerului în New York în perioada mai - septembrie 1973. Să se construiască un tabel de frecvenţe pentru variabila *Temp* şi graficele de frecvenţe corespunzătoare.

Histograme

Criterii de alegere a numărului de clase de grupare:

- la alegere
- $k \approx \sqrt{n}$ (Excel)
- $k = \lceil \log_2 n + 1 \rceil$ (Sturges)
- $k \approx 3.49 \cdot s \cdot n^{-1/3}$ (Scott)

În R: `hist(x)`
Argumente suplimentare:

- `breaks=m` $\rightarrow \approx m$ clase de grupare egale
- `breaks="Sturges"` (implicit) sau `"Scott"`
- `breaks=vect` unde *vect* este vectorul extremităţilor de intervale de grupare
- `prob=T` $\rightarrow$ aria histogramei = 1

Curbe de densitate

Poligonul de frecvenţe

- linie poligonală care uneşte mijloacele segmentelor orizontale superioare ale dreptunghiurilor histogramei

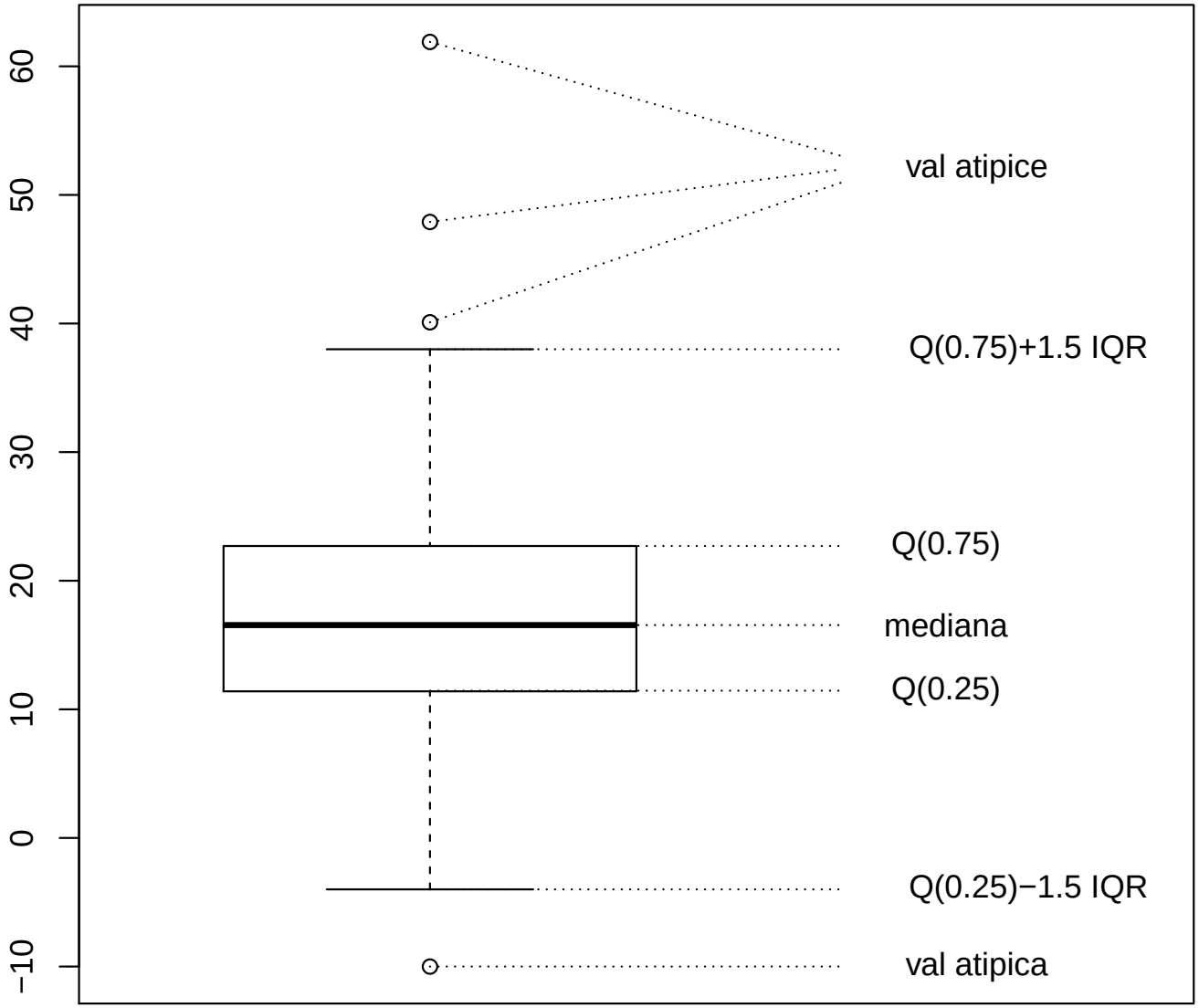
Curba de densitate

- versiune "continuă" a histogramei
- înălţimea curbei de densitate într-un punct estimează proporţia de valori din eşantion care se găsesc într-un interval de lăţime specificată ("bandwidth"), centrat în acel punct

În R: `plot(density(x))`

Exemplu: Histograma şi curba de densitate pentru variabila *Temp* din *airquality*

Grafice de tip boxplot



În R: `boxplot(x)`

Grafice de tip quantilă - quantilă

- Se utilizează în cazul unui singur eşantion pentru a compara distribuţia datelor cu o distribuţie de referinţă (e.g. normală).
- se reprezintă grafic puncte având ca şi coordonate:
  - pe Oy, valorile din eşantion, sortate crescător, fiecare reprezentând o anumită *p*-quantilă a respectivului eşantion
  - pe Ox, *p*-quantilele corespunzătoare ale distribuţiei teoretice de referinţă
- Dacă există concordanţă, punctele se vor afla aproximativ pe o dreaptă.

În R: `qqnorm` (comparare cu distribuţia normală); `qqplot`

Grafice de tip quantilă - quantilă

Exemplu:

```
x=rnorm(1000); y=rchisq(1000,df=5)
qqnorm(x)
qqnorm(y)
```