Testing Psicologico

Filippo Gambarota

10/25/2021

Le strutture dati

Strutture dati in R

qui una slide riassuntiva delle strutture dati

- organizzare dati con dei vincoli
- metodi per accedere/modificare/aggiungere dati in modo chiaro
- metodi per far interagire la struttura dati con specifiche funzioni

Esempio - il foglio Excel

il foglio excel è una struttura dati che sicuramente conoscete

- righe e colonne formano un "rettangolo" di dati
- la prima riga sono i nomi delle colonne
- ogni colonna deve contenere lo stesso tipo di dati ma le colonne possono essere di diversa tipologia

Esempio - il foglio Excel

	A	B	C	D	E	F	G
1	id	nome	eta	facolta	altezza	regione	
2	1	Filippo	23	psicologia	155	veneto	
3	2	Andrea	22	psicologia	165	puglia	
4	3	Anna	20	filosofia	167	sicilia	
5	4	Daria	21	sociologia	170	emilia romagna	
6	5	Francesco	26	psicologia	180	piemonte	
7	6	Elettra	23	matematica	190	lombardia	
8	7	Anna	22	fisica	155	lombardia	
9	8	Carlotta	22	chimica	160	veneto	
10	9	Mattia	24	filosofia	160	sicilia	
11							
12							
13							

Vettori e matrici

Esercizi

1) Usare `seq` per creare il vettore

$$v1 = (4, 8, 10, 12)$$

2) Usare `rep` per creare

$$v2 = (a, a, a, b, b, c, c, c, c)$$

3) Costruire la matrice

$$M = \begin{pmatrix} -3 & 4 & 0.1 \\ 2 & 9 & -5 \end{pmatrix}$$

4) Controllare la struttare di M ed estrarne le dimensioni

5) Estrarre la seconda riga di M

Soluzioni

```r
# 1) Usare seq per creare il vettore v1
v1 <- seq(from=4, to=12, by=2)

# 2) Usare rep per creare il vettore v2
v2 <- rep(x=c("a","b","c"), times=c(3,2,4))

# 3) Costruire la matrice M
M <- matrix(c(-3,2,4,9,0.1, -5), ncol=3)

# 4) Controllare la struttare di M ed estrarne le dimensioni
str(M)
```

```
##  num [1:2, 1:3] -3 2 4 9 0.1 -5
```

```r
ncol(M)
```

```
## [1] 3
```

```r
nrow(M)
```

```
## [1] 2
```

```r
# 5) Estrarre la seconda riga di M
M[2,]
```

```
## [1]  2  9 -5
```

Liste

Creazione di liste

Una lista è un contenitore di oggetti di qualsiasi tipo.

Si crea con il comando `list()`.

```
# creaiamo una lista contentente i due vettori v1 e v2 e la matrice M
L <- list("v_seq"=v1, "v_rep"=v2, "matrice"=M)
L
```

```
## $v_seq
## [1]  4  6  8 10 12
##
## $v_rep
## [1] "a" "a" "a" "b" "b" "c" "c" "c" "c"
##
## $matrice
##      [,1] [,2] [,3]
## [1,]   -3    4  0.1
## [2,]    2    9 -5.0
```

```
# nomi degli elementi di L
names(L)
```

```
## [1] "v_seq"   "v_rep"   "matrice"
```

Estrazione di elementi

```r
# un elemento si può estrarre con l'indice o con il nome
# per estrarre il secondo elemento, sono equivalenti:
L[[2]]
```

```
## [1] "a" "a" "a" "b" "b" "c" "c" "c" "c"
```

```r
L$v_rep
```

```
## [1] "a" "a" "a" "b" "b" "c" "c" "c" "c"
```

```r
# salviamo tutto in un file chiamato prova.rda
save.image(file="prova.rda")

# pulizia
rm(list=ls())
```

Dataframe

Dati SWISS

Studieremo un dataframe contenuto nel pacchetto datasets.

I dati contengono una misura di fertilità e indicatori socio-economici per 47 province svizzere.

```
# installiamo (solo la prima volta) e richiamiamo il pacchetto
# install.packages("datasets")
library(datasets)

# richiamiamo il dataset di interesse, chiamato swiss
data("swiss")

# osserviamo il contenuto
ls()

# help per la descrizione
?swiss

# cambiamento nome da swiss a d
d <- swiss
rm(swiss)
```

Struttura dei dati

```
str(d)
```

```
## 'data.frame':    47 obs. of  6 variables:
##  $ Fertility       : num  80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9
##  $ Agriculture     : num  17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 ..
##  $ Examination     : int  15 6 5 12 17 9 16 14 12 16 ...
##  $ Education       : int  12 9 5 7 15 7 7 8 7 13 ...
##  $ Catholic        : num  9.96 84.84 93.4 33.77 5.16 ...
##  $ Infant.Mortality: num  22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9 21 24.4 ..
```

Prime righe del dataframe

```
head(d)
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           85.8        36.5          12         7    33.77
## Neuveville        76.9        43.5          17        15     5.16
## Porrentruy        76.1        35.3           9         7    90.57
##              Infant.Mortality
## Courtelary               22.2
## Delemont                 22.2
## Franches-Mnt             20.2
## Moutier                  20.3
## Neuveville               20.6
## Porrentruy               26.6
```

Caratteristiche

```r
# numero di colonne (variabili)
ncol(d)
```

```
## [1] 6
```

```r
# numero di righe (osservazioni)
nrow(d)
```

```
## [1] 47
```

```r
# nomi delle variabili
names(d)
```

```
## [1] "Fertility"        "Agriculture"      "Examination"      "Education"
## [5] "Catholic"         "Infant.Mortality"
```

```r
# nomi delle osservazioni (stampo solo i primi 10)
rownames(d)[1:10]
```

```
##  [1] "Courtelary"   "Delemont"     "Franches-Mnt" "Moutier"      "Neuveville
##  [6] "Porrentruy"   "Broye"        "Glane"        "Gruyere"      "Sarine"
```

Selezione

```r
# ultime 3 righe del dataset
sel <- d[(45:47),]

# colonna relativa all'istruzione (due modi equivalenti)
ed <- d[,4]
ed <- d$Education

# province con fertilità maggiore o uguale a 80
sel <- d[d$Fertility >= 80,]

# istruzione delle province con mortalità infantile maggiore di 22
selEd <- d[d$Infant.Mortality > 22,]$Education
```

Esercizi

Selezionare:

1) dataframe tranne le prime 10 righe
2) dataframe escludendo le colonne `Agriculture` e `Catholic`
3) solo le righe 10, 20 e 30, e solo le colonna 1 e 2
4) province con mortalità infantile minore di 18
5) province con perc. di cattolici superiore a 90% e indice di istruzione uguale a 7
6) province con fertilità inferiore a 50 oppure superiore a 90
7) fertilità delle province con `Examination` diverso da 14

Soluzioni

```r
# 1) dataframe tranne le prime 10 righe
sel <- d[-(1:10),]

# 2) dataframe escludendo le colonne Agriculture e Catholic
sel <- d[,-c(2,5)]

# 3) solo le righe 10, 20 e 30, e solo le colonna 1 e 2
sel <- d[c(10,20,30), c(1,2)]

# 4) province con mortalità infantile minore di 18
sel <- d[d$Infant.Mortality < 18,]

# 5) province con perc. di cattolici superiore a 90
# e indice di istruzione uguale a 7
sel <- d[(d$Catholic > 90) & (d$Education==7),]

# 6) province con fertilità inferiore a 50 oppure superiore a 90
sel <- d[(d$Fertility < 50) | (d$Fertility > 90),]

# 7) fertilità delle province con Examination diverso da 14
selFert <- d[d$Examination != 14,]$Fertility
```

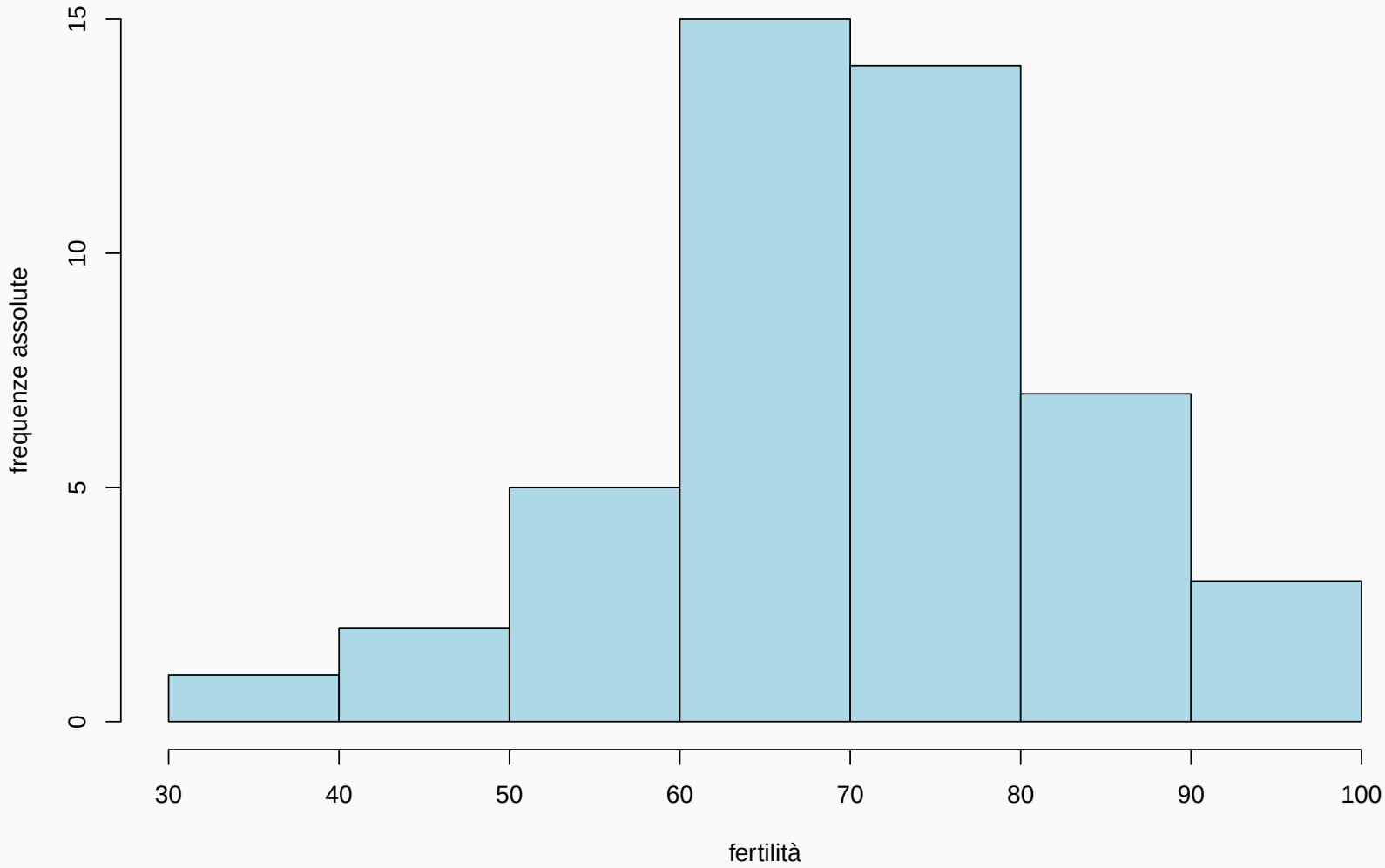
Aggiunta di variabili

```r
# nuova colonna che numera le osservazioni
d$id <- 1:nrow(d)

# nuova colonna Catholic2:
# 0 se la perc. di cattolici è bassa (perc. <= 10)
# 1 se è media (10 < perc. < 90)
# 2 se è alta (perc. >= 90)
d$Catholic2 <- 1
d[d$Catholic <= 10,]$Catholic2 <- 0
d[d$Catholic >= 90,]$Catholic2 <- 2

# variabile categoriale: la trasformo in fattore
d$Catholic2 <- factor(d$Catholic2)

summary(d$Catholic2)
```

```
##  0  1  2
## 20 12 15
```

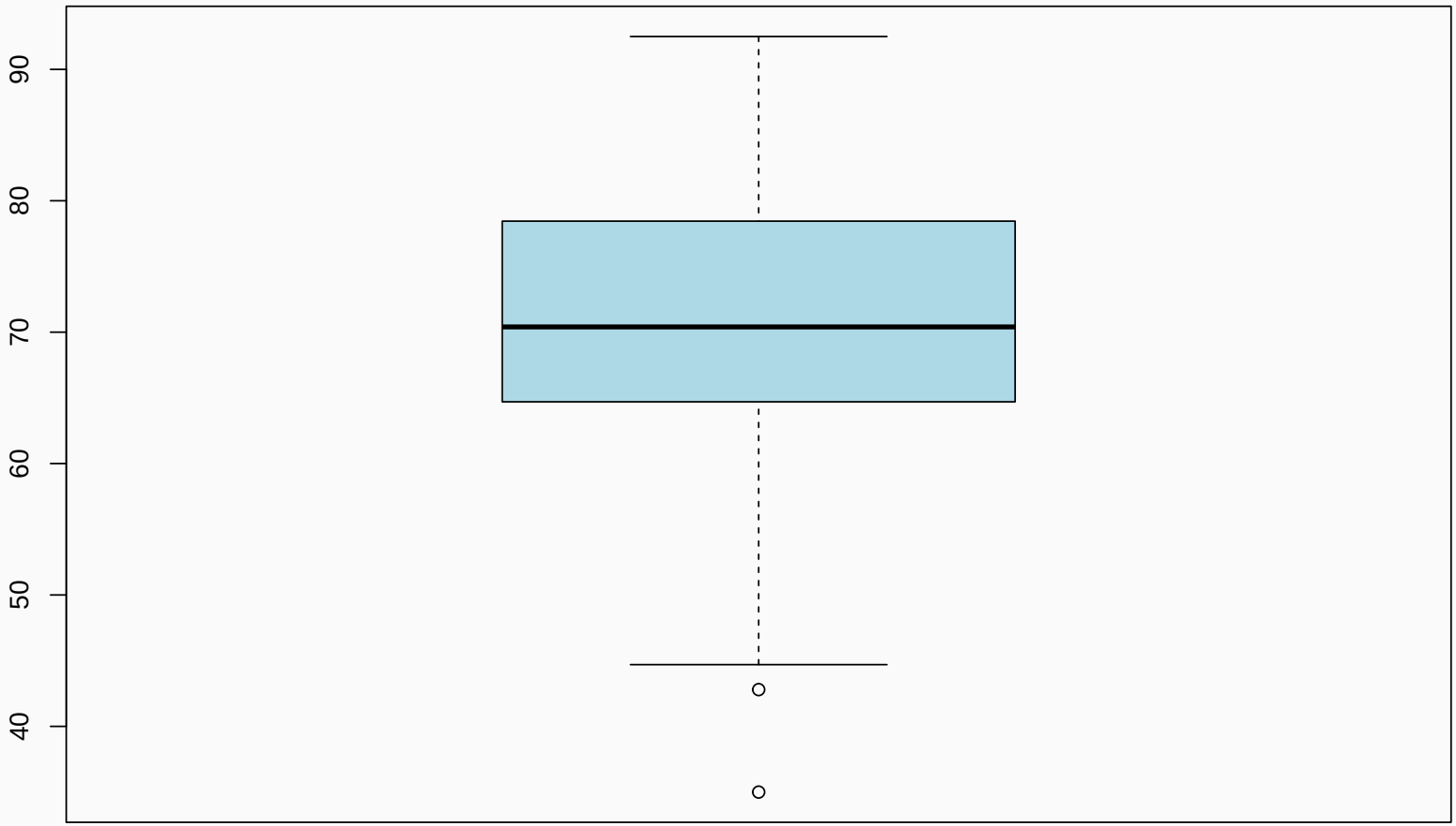
Istogramma

```
hist(d$Fertility,
     xlab="fertilità", ylab="frequenze assolute", col="lightblue", main="")
```



Boxplot

```
boxplot(d$Fertility,
        xlab="fertilità", ylab="", col="lightblue")
```



Boxplot condizionato

```
boxplot(d$Fertility ~ d$Catholic2,
        xlab="perc. cattolici", ylab="fertilità", col="lightblue")
```

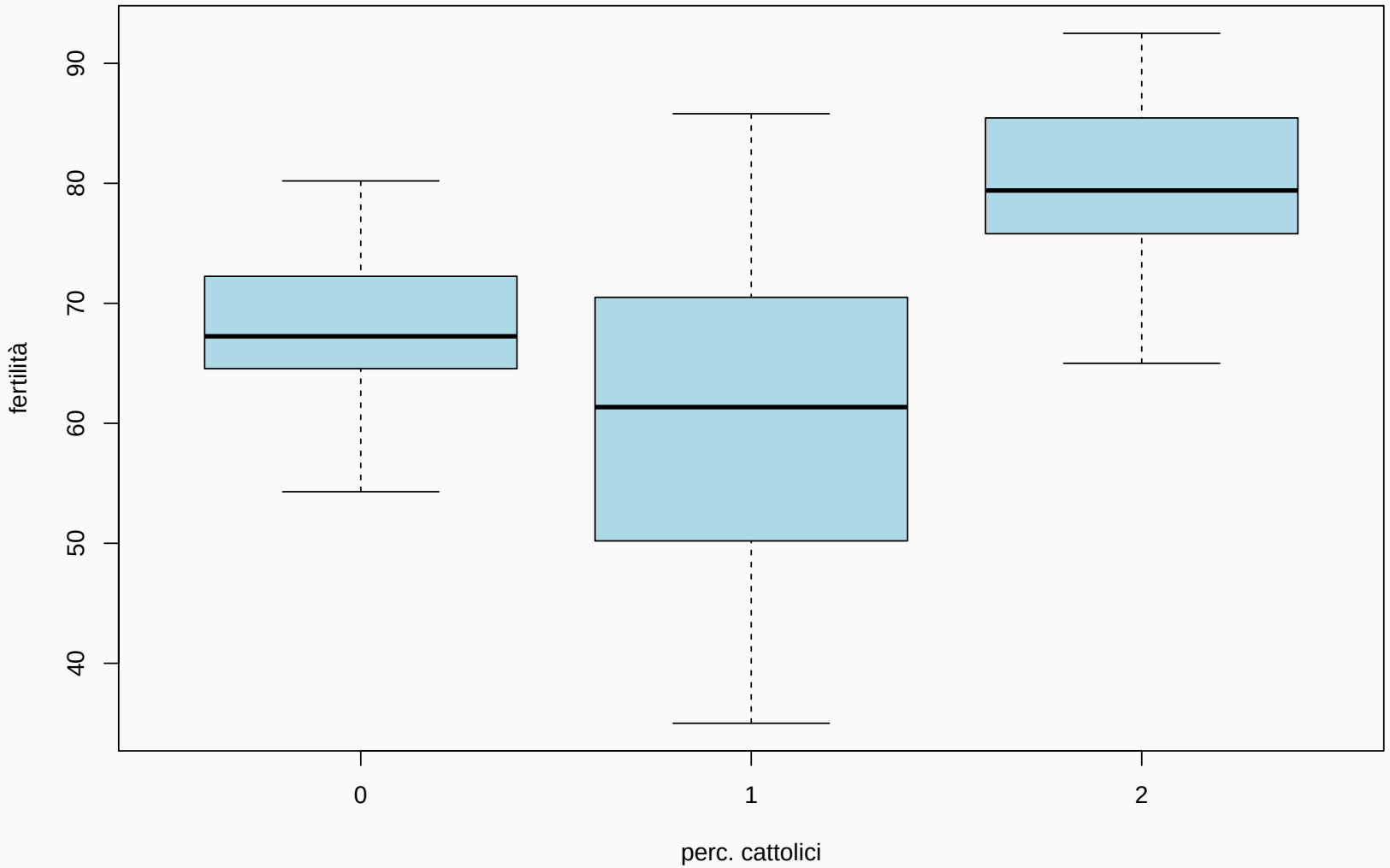


Grafico a dispersione

```
plot(d$Catholic, d$Fertility,
     xlab="perc. cattolici", ylab="fertilità", pch=16)
```