

Dataframe

Quella che i manuali di metodologia della ricerca sociale sono soliti chiamare **matrice dei dati**, in R si chiama **dataframe**.

Un dataframe è una **lista** di **vettori** (le **variabili**), che devono avere tutti la stessa lunghezza (numero di casi), ma possono essere di tipo diverso: variabili nominali (fattori), variabili cardinali (vettori numerici), ecc.

Il termine **matrice** (matrix), invece, sta ad indicare una struttura di dati in cui tutti gli elementi (i valori) sono di uno stesso tipo.

Un dataframe composto esclusivamente di valori numerici può infatti essere facilmente trasformato in una **matrice**, con la funzione `as.matrix()`.

Un dataframe può essere **⇒ costruito** inserendo i dati direttamente in R oppure — come più comunemente avviene — **importando i dati** da altre applicazioni.

Una volta caricato, un dataframe diventa un'oggetto del **Workspace (area di lavoro)**, e può essere **⇒ richiamato** nei comandi, per poter svolgere l'analisi dei dati. All'interno del workspace possono essere disponibili diversi dataframe alla volta.

Costruzione del dataframe

- **Inserimento dei dati**
- Inserimento dei dati in base a condizioni: Vedi: **Ricodifica: sostituzione di valori in base a condizioni**
- **Importare i dati in R**

Per avere informazioni sintetiche sulla struttura del dataframe:

```
str(myfile)
```

Salvataggio e caricamento da file

Per salvare il dataframe in un file separato da quello del **worspace**:

```
save(ChickWeight, file="mydata.RData")
```

Per caricare un dataframe salvato:

```
load("mydata.RData")
```

Per salvare il dataframe in un altro formato di file, vedi [Esportare la base-dati](#) e [Esportare i dati con RCommander](#)

Recupero delle informazioni

vedi: [Indicizzazione](#)

Gli elementi del dataframe (vettori-riga, vettori-colonna, celle), sono indicizzati.

Righe e colonne sono **indicizzati** con la notazione **[r,c]**: [4,] indica ad esempio la quarta riga, mentre [, 3] indica la terza colonna.

Colonne (variabili)

La funzione **colnames** consente di avere l'**elenco delle variabili** incluse nel dataframe (numero e nome):

```
> data("ChickWeight")
> colnames(ChickWeight)
[1] "weight" "Time"   "Chick"  "Diet"
```

```
> ChickWeight[1]
      weight
1         42
2         51
3         59
..        ...
```

restituisce il primo vettore colonna, ovvero la prima variabile contenuta nel dataframe, **in colonna**.

```
> ChickWeight[,1]
 [1] 42 51 59 64 76 93 106 125 149 171 199 205 40 49 58 72
84 103
[19] 122 138 162 187 209 215 43 39 55 67 84 99 115 138 163 187
198 202
[37] 42 49 56 67 74 87 102 108 136 154 160 157 41 42 48 60
79 106
....
```

restituisce il primo vettore colonna, ovvero la prima variabile contenuta nel dataframe, **in riga**, e corrisponde a

```
ChickWeight$weight
ChickWeight[[1]]
ChickWeight[["weight"]]
Chickweight[, "weight"]
```

Righe (casi)

La funzione **rownames** consente di avere l'**elenco dei casi** del dataframe:

```
rownames(ChickWeight)
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"
[10] "10" "11" "12" "13" "14" "15" "16" "17" "18"
...
[577] "577" "578"
```

```
ChickWeight[1,]
```

restituisce i dati del primo caso del dataframe.

Se le righe hanno un nome, possono essere richiamate anche facendo riferimento ad esso:

```
Chickweight["1",]
```

```
  weight Time Chick Diet
1     42    0     1    1
```

Valori (dati)

Per visualizzare singoli valori all'interno del dataframe, è sufficiente indicare sia il numero di riga del caso, che il numero o il nome fra apici della variabile (colonna):

```
> ChickWeight[310, "Diet"]
[1] 2
Levels: 1 2 3 4
```

È anche possibile recuperare le informazioni attraverso l'uso degli [Operatori](#). Ad es., per visualizzare tutti i casi che presentano un certo valore:

```
ChickWeight$Diet[ChickWeight$Diet==3]
```

Per modificare il valore (vedi anche [Ricodificare le variabili](#)):

```
mydata[310, "Genere"] <- "Uomo"
```

Numero di colonne e di righe

Dal momento che un dataframe è una [lista](#) di vettori, possiamo ottenere il numero di colonne con `length()` (= lunghezza della lista), o, più semplicemente con `ncol()`:

```
length(ChickWeight)
```

```
[1] 4
```

```
ncol(ChickWeight)
```

```
[1] 4
```

Per sapere il numero delle righe, useremo invece `nrow()`:

```
nrow(ChickWeight)
```

```
[1] 578
```

Recupero veloce (attach)

Per poter richiamare direttamente un elemento del dataframe, senza doverne digitare il nome, è possibile utilizzare il comando **attach**:

```
attach(mydata)
```

```
attach(ChickWeight)
```

```
weight
```

```
[1] 42 51 59 64 76 93 106 125 149 171 199 205 ...  
[20] 138 162 187 209 215 43 39 55 67 84 99 115 ...  
...
```

Attenzione

In questo caso, R utilizza e fa riferimento ad una immagine del dataframe: questo significa che **eventuali modifiche ai dati non verranno salvate nel dataframe originale**.

Per tornare alla situazione precedente, o per lavorare con un altro file, usare il comando `detach`.

```
detach(mydata)
```

Vedi anche

- [Valori mancanti \(in R\)](#)
- [Altre operazioni sui dataframe](#)
- [I dataset in RCommander](#)
- [Matrice dei dati](#)

Script di esempio

Esempio Lab4.R

```
# Uso il dataset di esempio
data("ChickWeight")

# Struttura del dataset
str(ChickWeight)

# Elenco delle variabili
colnames(ChickWeight)

# Richiamare le variabili
ChickWeight[1]
ChickWeight[,1]
ChickWeight$weight
ChickWeight[[1]]
ChickWeight$weight
ChickWeight[[1]]
ChickWeight[["weight"]]
ChickWeight[, "weight"]

# Elenco dei casi
rownames(ChickWeight)

# Richiamare i casi
ChickWeight[1,]
ChickWeight["1",]

# Richiamare i valori
ChickWeight[310, "Diet"]
ChickWeight$Diet[ChickWeight$Diet=="3"]
```

```
# attach: recupero veloce  
attach(ChickWeight)  
weight  
detach(ChickWeight)
```

[Concetti di base](#), [Oggetti](#), [Linguaggio](#), [Matrice dei dati](#)

From:

<https://www.agnesevardanega.eu/wiki/> - Ricerca Sociale con R

Permanent link:

https://www.agnesevardanega.eu/wiki/r/gestione_dei_dati/dataframe?rev=1632409529

Last update: **23/09/2021 15:05**

