

Ponderazione dei dati

Per una definizione generale del termine, vai alla voce **Ponderazione** del **Glossario**.

In R, non esiste una singola funzione per operare con i pesi, ma esistono diverse funzioni e diversi pacchetti che prevedono la possibilità di applicare i pesi ai dati, per produrre diversi tipi di output.

In primo luogo, è necessario creare **una variabile che contenga i pesi**: deve essere un **vettore numerico**, della stessa lunghezza degli altri vettori-colonna, che non deve avere valori negativi. In pratica: ai casi non ponderati verrà attribuito il valore 1, mentre agli altri verrà attribuito il peso previsto dal disegno di campionamento.

Altra situazione in cui è necessaria applicare una ponderazione ai dati, è rappresentata dalle tabelle di dati aggregati, come ad esempio quella che segue (questi i dati [lstat16.rda](#)):

	Età	Eta.classi	Rip.geog	Totale
1	0	Fino a 14	Nord-Ovest	32556
2	1	Fino a 14	Nord-Ovest	34407
3	2	Fino a 14	Nord-Ovest	35527
4	3	Fino a 14	Nord-Ovest	37115
...				
16	15	Giovani	Nord-Ovest	39020
17	16	Giovani	Nord-Ovest	37864
18	17	Giovani	Nord-Ovest	38073
19	18	Giovani	Nord-Ovest	38049
...				
27	26	Adulti	Nord-Ovest	41475
28	27	Adulti	Nord-Ovest	43023
29	28	Adulti	Nord-Ovest	42350
30	29	Adulti	Nord-Ovest	43724
...				
66	65	Anziani	Nord-Ovest	56414
67	66	Anziani	Nord-Ovest	56906
68	67	Anziani	Nord-Ovest	59904
69	68	Anziani	Nord-Ovest	58501
...				

Il numero dei casi corrispondenti a ciascuna riga non è "1", ma è pari al valore indicato nella colonna "Totale".

Funzioni dei pacchetti base

la funzione rep()

La funzione `rep()` replica i valori di un vettore per il numero delle volte indicato dall'argomento `times`:

```
> table(rep(Istat16$Eta.classi, times = Istat16$Totale))
```

Fino a 14	Giovani	Adulti	Anziani
8281859	6562425	32451513	13369754

In questo caso, abbiamo ottenuto la distribuzione di frequenza di `Istat16$Eta.classi`, replicando i valori un numero di volte (`times`) pari ai valori di `Istat16$Totale`.

L'argomento `times` può essere rappresentato da un valore, o da un vettore contenente i pesi dei valori o delle modalità della riga: i valori devono essere non negativi, ma possono essere inferiori a uno (0,3 ad esempio).

Possiamo applicare la funzione `rep()` anche ad altre funzioni, come ad esempio la funzione `summary()`.

```
> summary(rep(Istat16$Età, times = Istat16$Totale))
  Min. 1st Qu.  Median    Mean 3rd Qu.   Max.
  0.00   26.00   45.00   44.16   62.00  100.00
```

Questa funzione comporta però un rallentamento dei calcoli, ed è quindi preferibile utilizzare funzioni più efficienti.

Media ponderata

All'interno della distribuzione standard, esiste la funzione `weighted.mean`:

```
> weighted.mean(Istat16$Età, Istat16$Totale, na.rm = TRUE)
[1] 44.15809
```

Funzione:

```
weighted.mean(x, w, na.rm = TRUE)
```

L'argomento `na.rm = TRUE` può essere omesso in quanto è di default, e significa che i casi mancanti saranno espunti dal computo della media.

xtabs

Per produrre tabelle di frequenza e/o di contingenza ponderate, è possibile utilizzare la funzione `xtabs()`, inserendo prima della formula (`~ x + y`) il vettore dei pesi.

Distribuzione di frequenze:

```
> xtabs(Totale ~ Eta.classi, data = Istat16)
Eta.classi
Fino a 14   Giovani   Adulti   Anziani
 8281859   6562425   32451513  13369754
```

Tabella di contingenza:

```
> xtabs(Totale ~ Eta.classi + Rip.geog, data = Istat16)
      Rip.geog
Eta.classi Nord-Ovest Nord-Est Centro Sud Isole
Fino a 14   2168999   1592082 1601343 1999214 920221
Giovani     1597313   1171511 1210017 1771296 812288
Adulti      8587714   6224619 6496728 7534110 3608342
Anziani     3756951   2655389 2759715 2806151 1391548
```

Con altri pacchetti

Sono diversi i pacchetti aggiuntivi di R che prevedono funzioni per la gestione e l'analisi di dati ponderati.

Fra questi, ricordiamo in particolare [Hmisc](#), installato insieme a [RCommander](#), e **weights**.

Per l'analisi dei dati di *surveys* esiste anche il pacchetto [Survey](#), che pure include alcune funzioni utili (ma più complesse da utilizzare).

Hmisc

statistiche riassuntive

`describe` è una funzione che si applica a vettori, dataframes, matrici e formule.

`describe.vector` è la funzione base per la descrizione di una singola variabile.

Con una variabile categoriale:

```
> describe(Istat16$Rip.geog, weights = Istat16$Totale)
Istat16$Rip.geog
      n missing distinct
60665551      0        5
```

Value	Nord-Ovest	Nord-Est	Centro	Sud	Isole
Frequency	16110977	11643601	12067803	14110771	6732399
Proportion	0.266	0.192	0.199	0.233	0.111

Con una variabile numerica:

```
> describe(Istat16$Età, weights = Istat16$Totale)
Istat16$Età
      n missing distinct      Info      Mean      .05      .10
60665551      0      101        1    44.16        5      11
      .25      .50      .75      .90      .95
      26      45      62      76      82

lowest :    0    1    2    3    4, highest:  96  97  98  99 100
```

varianza ponderata

```
> wtd.var(Istat16$Età, weights = Istat16$Totale, na.rm = TRUE)
[1] 545.3276
```

quantili (e mediana)

```
> wtd.quantile(Istat16$Età, weights = Istat16$Totale, na.rm = TRUE)
0%  25%  50%  75% 100%
0   26   45   62  100
```

altre funzioni

`wtd.mean`: media ponderata;

`wtd.table`: frequenze ponderate.

weights

variabili standardizzate

```
> stdz(Istat16$Età, weight = Istat16$Totale)
[1] -1.890956342 -1.848133918 -1.805311495 -1.762489071
[5] -1.719666647 -1.676844223 -1.634021799 -1.591199376
[9] -1.548376952 -1.505554528 -1.462732104 -1.419909680
[13] -1.377087257 -1.334264833 -1.291442409 -1.248619985
```

...

frequenze relative ponderate

```
> wpct(Istat16$Rip.geog, weight = Istat16$Totale, na.rm = TRUE)
Nord-Ovest  Nord-Est  Centro  Sud  Isole
0.2655704  0.1919310  0.1989235  0.2325994  0.1109757
```

altre funzioni

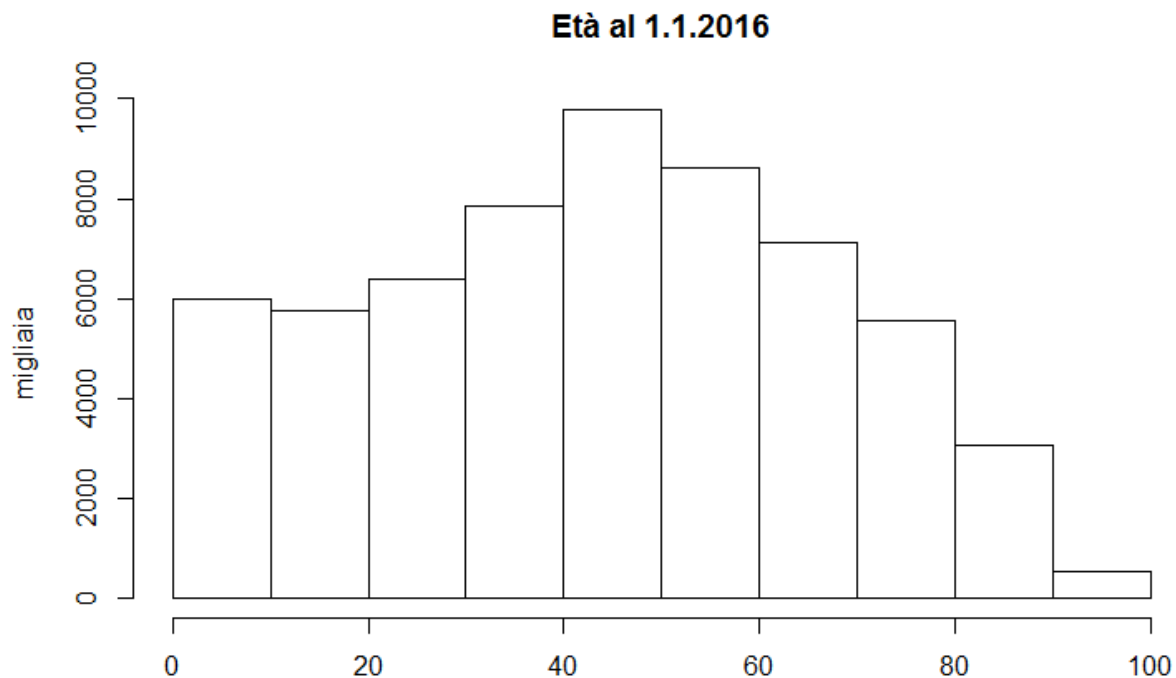
`wtd.chi.sq`: chi quadrato ponderato; `onecor.wtd`: coefficiente di correlazione;
`wtd.t.test`: t-test; `wtd.hist`: istogramma (vedi oltre).

Rappresentazioni grafiche

Le rappresentazioni grafiche di variabili categoriali ponderate è possibile attraverso le funzioni grafiche standard, a partire dalle tabelle di frequenza, comunque prodotte.

Per produrre un istogramma di una variabile continua ponderata, è disponibile la funzione `wtd.hist` nel pacchetto `weights`. Es.:

```
wtd.hist(Istat16$Età,
         weight = Istat16$Totale/1000,
         main = "Età al 1.1.2016",
         xlab = "",
         ylab = "migliaia")
```



Script

Ponderazione.R

```
# rep
table(rep(Istat16$Eta.classi, times = Istat16$Totale))
summary(rep(Istat16$Età, times = Istat16$Totale))

# weighted.mean
weighted.mean(Istat16$Età, Istat16$Totale, na.rm = TRUE)

# xtabs
xtabs(Totale ~ Eta.classi, data = Istat16)
xtabs(Totale ~ Eta.classi + Rip.geog, data = Istat16)

#Hmisc
require(Hmisc)
describe(Istat16$Rip.geog, weights = Istat16$Totale)
describe(Istat16$Età, weights = Istat16$Totale)
wtd.var(Istat16$Età, weights = Istat16$Totale, na.rm = TRUE)
wtd.quantile(Istat16$Età, weights = Istat16$Totale, na.rm = TRUE)

# weights
require(weights)
stdz(Istat16$Età, weight = Istat16$Totale)
wpct(Istat16$Rip.geog, weight = Istat16$Totale, na.rm=TRUE)
```

```
wtd.hist(Istat16$Età,  
         weight = Istat16$Totale/1000,  
         main = "Età al 1.1.2016",  
         xlab = "",  
         ylab = "migliaia")
```

Gestione dei dati

From:

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

Permanent link:

https://www.agnesevardanega.eu/wiki/r/gestione_dei_dati/ponderazione

Last update: **05/09/2025 16:42**

