Relazione Progetto Statistica e Analisi dei Dati

Margherita Maria Napolitano

Anno accademico 2023-2024

Prima Parte

Gli obiettivi della seguente analisi statistica sono:

- analizzare la **tendenza temporale**, definire quali sono le variazioni nel consumo di farmaci per il diabete nel corso degli anni nei paesi in esame e se sono presenti trend in aumento o diminuzione;
- analizzare le **differenze tra paesi**, definire se ci sono disparità significative.

Dataset e rappresentazioni grafiche

Dataset e data frame

Nel dataset oggetto di studio, estratto dal sito https://stats.oecd.org/, è analizzato il consumo farmaceutico totale secondo il *Sistema di classificazione/Defined Daily Dose (DDD) dell'ATC (Anatomical Therapeutic Chemical)*, creato dal Centro di collaborazione dell'OMS per Metodologia delle statistiche sulle droghe. In particolare, il dataset in esame analizza il **consumo di farmaci per il diabete** ed è formato da 32 righe corrispondenti ai paesi da cui sono stati prelevati i dati in esame e 13 colonne che rappresentano il periodo temporale dello studio (dal 2010 al 2022). I valori riportati sono definiti per dosaggio giornaliero per 1000 abitanti.

Come primo passo disporremo il nostro dataset in un **data frame,** un oggetto di tipo lista che si presenta in forma di tabella ed è costituito da righe (osservazioni) e colonne (variabili).

	Anno2010	Anno2011	Anno2012	Anno2013	Anno2014	Anno2015	Anno2016
Australia	57.8	59.6	60.5	57.5	57.5	57.5	56.7
Austria	45.4	46.3	45.9	46.0	46.4	46.3	46.9
Belgium	60.5	61.7	63.2	65.0	66.3	68.0	70.0
Canada	55.4	57.0	57.6	58.1	58.1	58.3	79.0
Chile	NA	34.0	30.7	45.8	52.1	59.3	46.4
Costa Rica	45.7	48.0	50.4	57.0	61.4	57.3	59.6
Czech Republic	73.7	76.2	78.2	80.2	84.7	88.0	88.9
Denmark	47.4	49.0	50.5	51.2	51.8	53.0	54.7
Estonia	45.5	47.5	53.1	54.6	57.3	59.4	61.1
Finland	83.3	84.2	84.8	85.9	88.1	89.8	92.3
France	72.9	74.5	80.2	81.6	81.3	82.4	84.5
Germany	81.4	82.6	83.2	83.2	83.6	82.6	83.8
Greece	NA	NA	NA	85.0	83.3	86.2	NA
Hungary	70.7	74.5	75.3	77.5	76.2	69.9	72.4
Iceland	31.7	38.1	38.9	41.9	42.3	44.9	46.5
Israel	NA	NA	36.8	NA	55.3	59.6	62.7
Italy	NA	61.6	62.0	62.4	61.8	62.0	62.2
Korea	65.5	64.1	64.3	65.4	60.8	61.7	60.3
Latvia	NA	NA	40.2	41.5	42.4	43.8	43.6
Lithuania	30.1	32.4	37.0	40.6	43.7	44.3	47.8

```
## Luxembourg          63.8     64.0     63.4     64.0     64.6     65.2     64.9
## Netherlands         72.3     72.9     74.1     75.1     75.0     75.8     76.2
## New Zealand           NA       NA       NA       NA       NA       NA       NA
## Norway              48.3     48.4     48.4     48.7     50.1     51.5     53.0
## Portugal            62.5     58.6     61.0     62.7     64.5     67.2     68.5
## Slovak Republic     57.8     61.3     58.0     65.6     70.4     75.2     76.3
## Slovenia            63.9     67.3     70.5     73.1     74.2     75.9     78.6
## Spain               55.8     56.0     66.4     66.6     71.1     72.9     75.1
## Sweden              51.9     53.0     53.9     55.6     56.5     58.4     60.4
## Turkiye             47.5     53.4     56.2     58.4     61.1     64.2     66.5
## United Kingdom      74.9     77.8     79.9     82.3     83.6     84.7     85.3
## Croatia             49.9     58.7     59.4     62.5     63.3     64.9     67.6
##                  Anno2017 Anno2018 Anno2019 Anno2020 Anno2021 Anno2022
## Australia           57.1     57.8     57.5     60.0     62.4       NA
## Austria             47.2     46.3     47.5     49.4     49.2       NA
## Belgium             71.0     72.8     75.8     78.0     77.7       NA
## Canada              81.9     84.3     86.4    104.9    124.5    154.7
## Chile               39.8     47.9     48.0     73.7     77.2     79.7
## Costa Rica          62.4     66.7     72.2     78.8     82.3     67.9
## Czech Republic      89.4     90.5     92.4     94.2     97.5       NA
## Denmark             55.6     57.0     58.8     62.0     67.0       NA
## Estonia             62.2     64.9     65.0     66.8     67.5     73.8
## Finland             91.7     95.8     99.9    101.8    105.6       NA
## France              84.0     84.8     86.9     85.6     87.4       NA
## Germany             83.5     85.9     88.3     91.0     93.6       NA
## Greece              80.7     81.9     83.7     88.1    100.9     98.4
## Hungary             74.2     75.4     77.0     79.2     76.4     75.3
## Iceland             47.2     48.0     48.7     51.1     54.8     57.9
## Israel              64.4     62.6     61.6     59.9     60.0     66.2
## Italy               62.5     63.0     64.0     64.8     65.1     67.2
## Korea               63.1     65.2     67.8     72.7     78.5       NA
## Latvia              45.5     46.7     48.1     50.5     50.0     51.6
## Lithuania           50.0     53.6     56.7     59.0     58.8       NA
## Luxembourg          64.2     64.0     65.1     64.5     63.8     60.0
## Netherlands         77.2     75.1     76.9     77.6     78.5       NA
## New Zealand           NA     57.0     57.2     53.0     54.9     52.3
## Norway              55.4     56.2     58.5     59.9     65.9     76.8
## Portugal            67.9     70.4     74.2     76.7     78.6     85.0
## Slovak Republic     76.0     76.8     78.6     80.4     79.5       NA
## Slovenia            80.7     82.6     83.3     85.1     87.2     91.6
## Spain               76.4     78.1     79.6     81.3     84.7     88.6
## Sweden              62.4     64.9     67.8     70.1     73.2     80.5
## Turkiye             72.6     75.4     79.1     91.9     96.3       NA
## United Kingdom      84.5     72.1     74.2     74.3     90.0       NA
## Croatia             66.9     69.2     77.8     79.5     75.5       NA
```

Si nota la presenza di numerosi valori nulli, in particolar modo nelle righe 13, 16, 19, 23 (Grecia, Israele, Latvia, Nuova Zelanda) e nelle colonne 1,2,13 (2010, 2011, 2022). Di conseguenza si è passato all'eliminazione di tali dati per avere un dataset il più coerente possibile.

```
##                  Anno2012 Anno2013 Anno2014 Anno2015 Anno2016 Anno2017 Anno2018
## Australia           60.5     57.5     57.5     57.5     56.7     57.1     57.8
## Austria             45.9     46.0     46.4     46.3     46.9     47.2     46.3
## Belgium             63.2     65.0     66.3     68.0     70.0     71.0     72.8
## Canada              57.6     58.1     58.1     58.3     79.0     81.9     84.3
```

```
## Chile               30.7     45.8     52.1     59.3     46.4     39.8     47.9
## Costa Rica          50.4     57.0     61.4     57.3     59.6     62.4     66.7
## Czech Republic      78.2     80.2     84.7     88.0     88.9     89.4     90.5
## Denmark             50.5     51.2     51.8     53.0     54.7     55.6     57.0
## Estonia             53.1     54.6     57.3     59.4     61.1     62.2     64.9
## Finland             84.8     85.9     88.1     89.8     92.3     91.7     95.8
## France              80.2     81.6     81.3     82.4     84.5     84.0     84.8
## Germany             83.2     83.2     83.6     82.6     83.8     83.5     85.9
## Hungary             75.3     77.5     76.2     69.9     72.4     74.2     75.4
## Iceland             38.9     41.9     42.3     44.9     46.5     47.2     48.0
## Italy               62.0     62.4     61.8     62.0     62.2     62.5     63.0
## Korea               64.3     65.4     60.8     61.7     60.3     63.1     65.2
## Lithuania           37.0     40.6     43.7     44.3     47.8     50.0     53.6
## Luxembourg          63.4     64.0     64.6     65.2     64.9     64.2     64.0
## Netherlands         74.1     75.1     75.0     75.8     76.2     77.2     75.1
## Norway              48.4     48.7     50.1     51.5     53.0     55.4     56.2
## Portugal            61.0     62.7     64.5     67.2     68.5     67.9     70.4
## Slovak Republic     58.0     65.6     70.4     75.2     76.3     76.0     76.8
## Slovenia            70.5     73.1     74.2     75.9     78.6     80.7     82.6
## Spain               66.4     66.6     71.1     72.9     75.1     76.4     78.1
## Sweden              53.9     55.6     56.5     58.4     60.4     62.4     64.9
## Turkiye             56.2     58.4     61.1     64.2     66.5     72.6     75.4
## United Kingdom      79.9     82.3     83.6     84.7     85.3     84.5     72.1
## Croatia             59.4     62.5     63.3     64.9     67.6     66.9     69.2
##                  Anno2019 Anno2020 Anno2021
## Australia           57.5     60.0     62.4
## Austria             47.5     49.4     49.2
## Belgium             75.8     78.0     77.7
## Canada              86.4    104.9    124.5
## Chile               48.0     73.7     77.2
## Costa Rica          72.2     78.8     82.3
## Czech Republic      92.4     94.2     97.5
## Denmark             58.8     62.0     67.0
## Estonia             65.0     66.8     67.5
## Finland             99.9    101.8    105.6
## France              86.9     85.6     87.4
## Germany             88.3     91.0     93.6
## Hungary             77.0     79.2     76.4
## Iceland             48.7     51.1     54.8
## Italy               64.0     64.8     65.1
## Korea               67.8     72.7     78.5
## Lithuania           56.7     59.0     58.8
## Luxembourg          65.1     64.5     63.8
## Netherlands         76.9     77.6     78.5
## Norway              58.5     59.9     65.9
## Portugal            74.2     76.7     78.6
## Slovak Republic     78.6     80.4     79.5
## Slovenia            83.3     85.1     87.2
## Spain               79.6     81.3     84.7
## Sweden              67.8     70.1     73.2
## Turkiye             79.1     91.9     96.3
## United Kingdom      74.2     74.3     90.0
## Croatia             77.8     79.5     75.5
```

Il sistema R è dotato di un sofisticato ambiente grafico che permette di creare grafici per illustrare i risultati di elaborazioni statistiche. Visualizzare i dati attraverso grafici è utile per diverse finalità:

1. Facilitare la comprensione: i grafici possono aiutare a rappresentare in modo visivo i dati e a renderli più facili da comprendere. Ad esempio, un grafico a barre può mostrare facilmente le frequenze di una variabile categoriale, mentre un grafico a linee può mostrare l'evoluzione di una variabile nel tempo.

2. Individuare pattern e tendenze: i grafici possono aiutare ad individuare pattern e tendenze nei dati, che altrimenti potrebbero essere difficili da rilevare dall'analisi dei dati in forma tabellare.

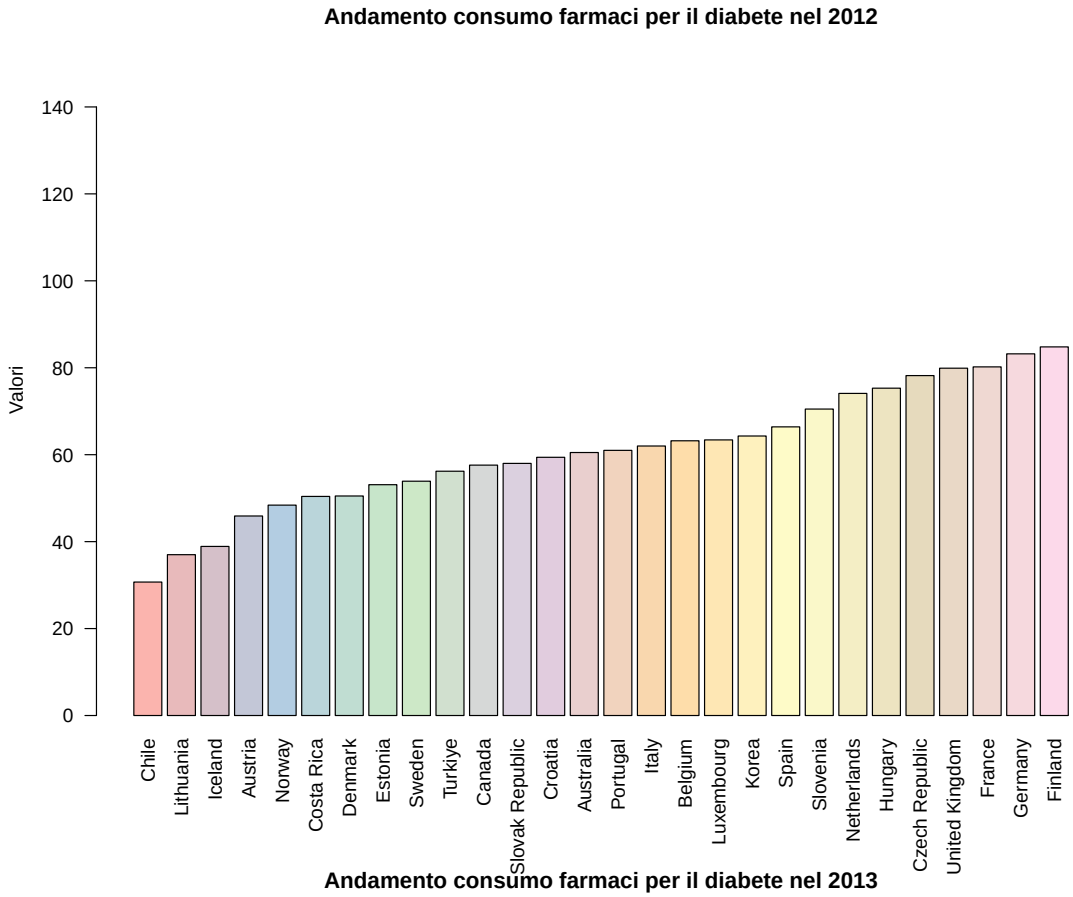
3. Fare confronti: i grafici consentono di fare confronti tra diverse categorie o gruppi di dati, ad esempio confrontando le frequenze di diverse variabili.

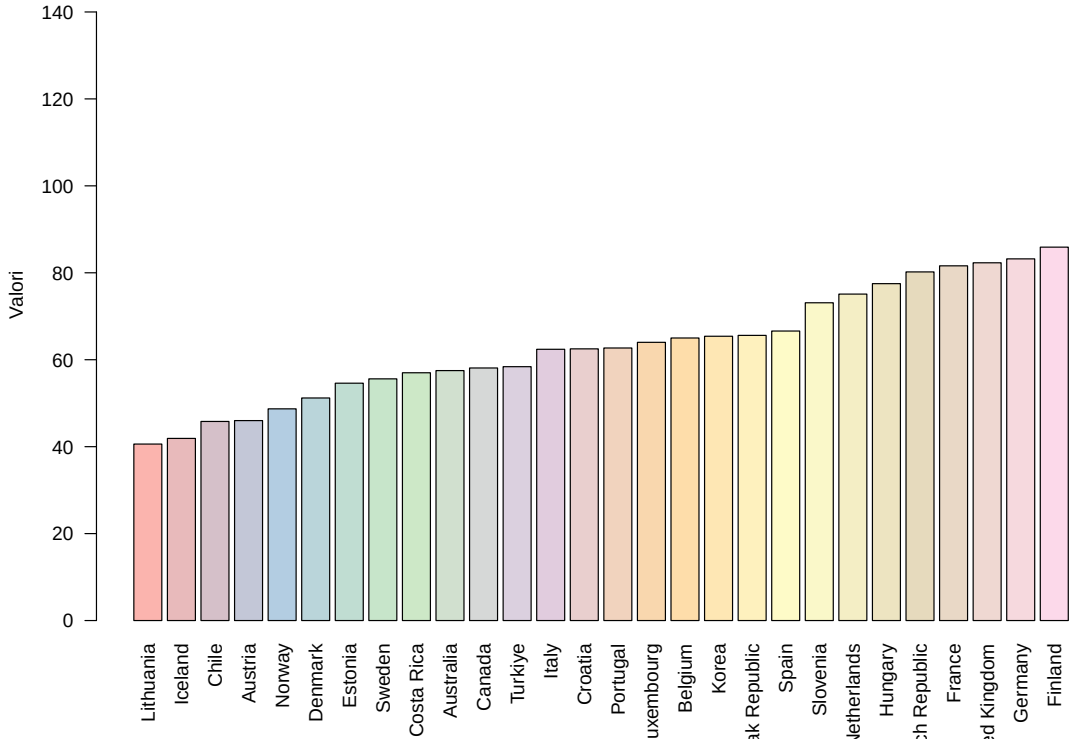
4. Comunicare i risultati: i grafici possono essere utilizzati per comunicare i risultati di un'analisi a un pubblico più ampio, anche a persone che non sono esperte di statistica o di analisi dei dati. Ad esempio, un grafico può essere incluso in un report o in una presentazione, come in questo caso, per illustrare in modo chiaro e conciso i risultati di un'analisi.

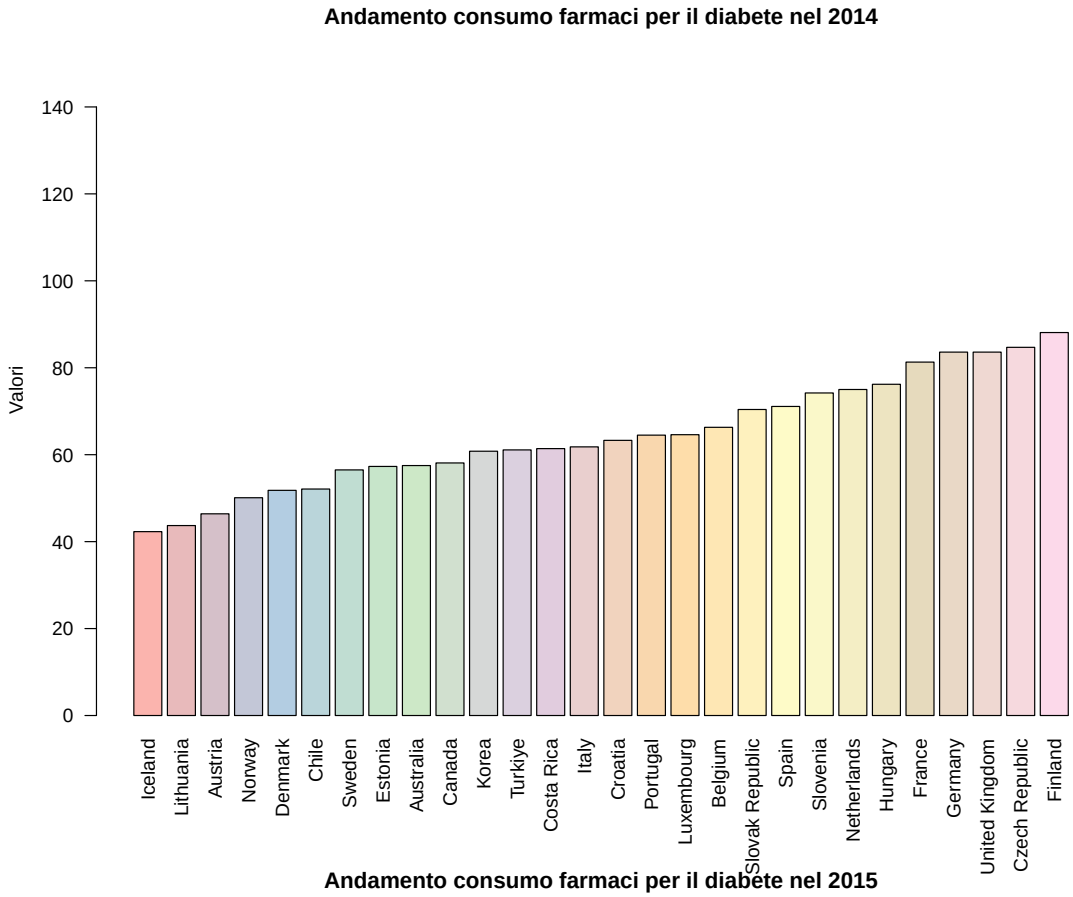
Barplot

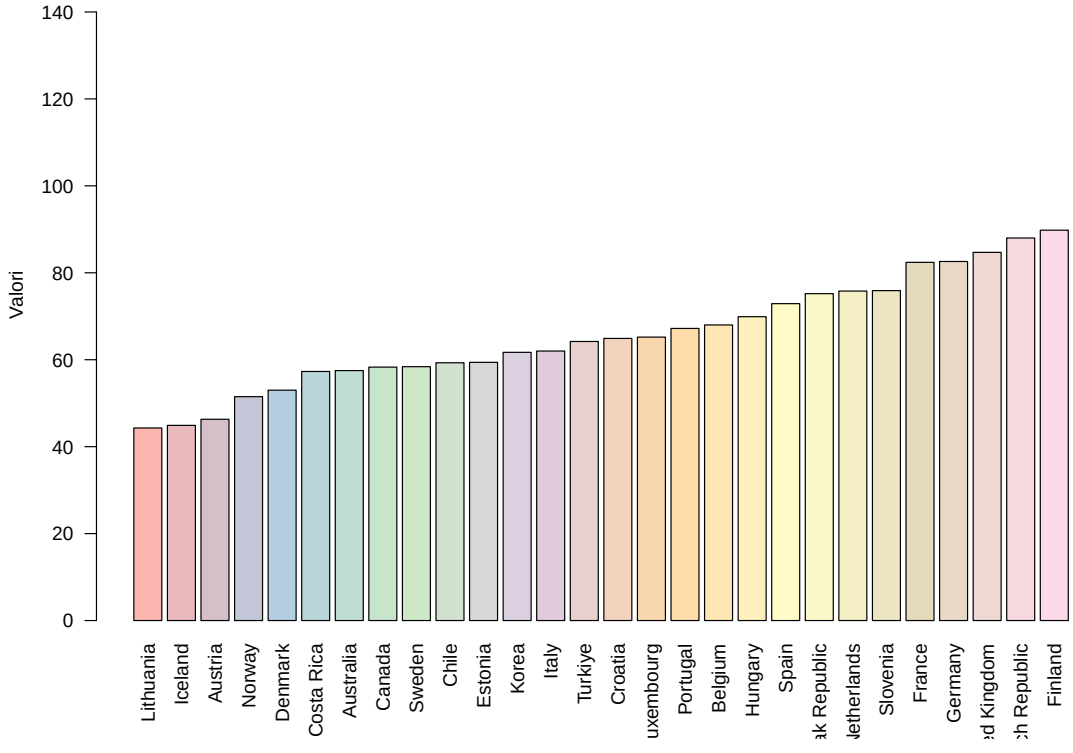
Nel nostro caso di studio disponiamo di una matrice di dati numerici e a partire da essa è possibile creare dei vettori contenenti gli elementi delle singole colonne tramite la funzione R *cbind()*. Utilizzando poi la funzione *barplot()* è possibile creare dei grafici a barre. I **grafici a barre** sono indicati per rappresentare una o più variabili categoriali. Ogni categoria viene rappresentata da una barra la cui altezza è proporzionale alla frequenza o alla percentuale della categoria stessa. Ordiniamo le barre in ordine crescente in modo da visualizzare al meglio i dati.

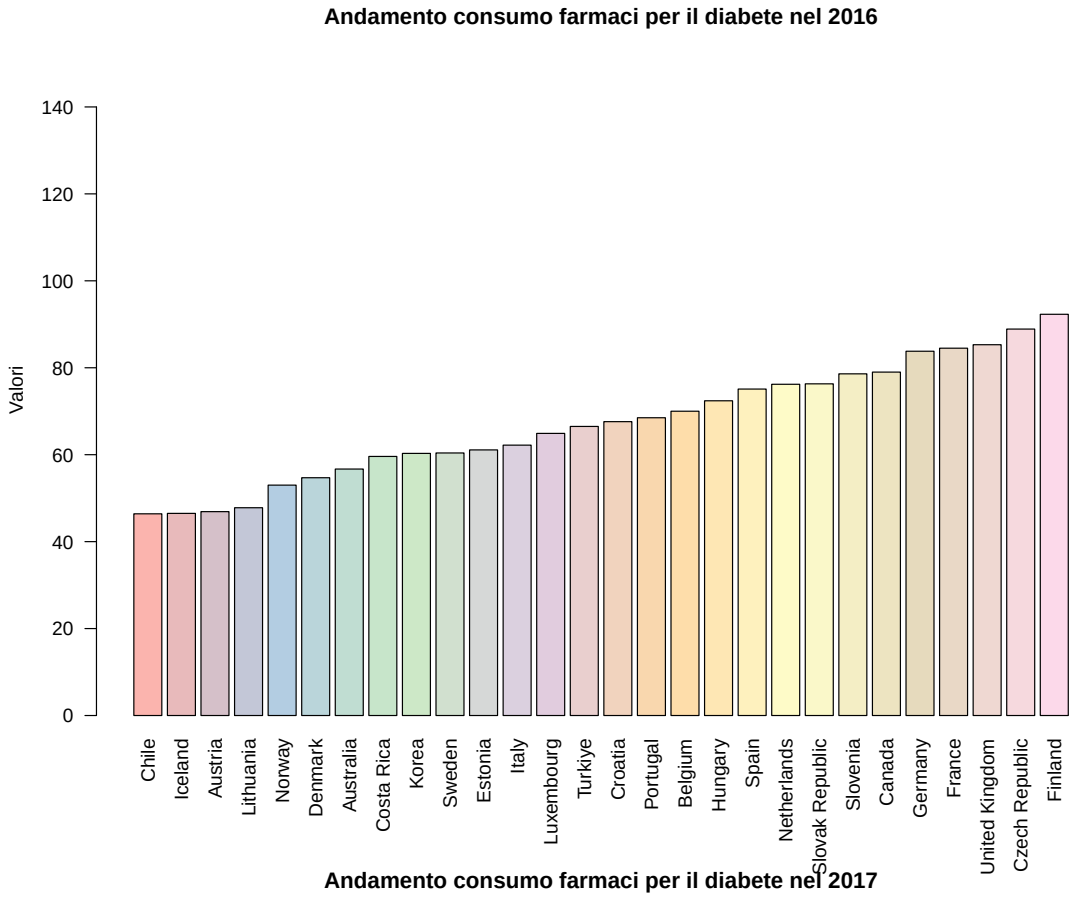
```
##                 2012 2013 2014 2015 2016 2017 2018 2019  2020  2021
## Australia       60.5 57.5 57.5 57.5 56.7 57.1 57.8 57.5  60.0  62.4
## Austria         45.9 46.0 46.4 46.3 46.9 47.2 46.3 47.5  49.4  49.2
## Belgium         63.2 65.0 66.3 68.0 70.0 71.0 72.8 75.8  78.0  77.7
## Canada          57.6 58.1 58.1 58.3 79.0 81.9 84.3 86.4 104.9 124.5
## Chile           30.7 45.8 52.1 59.3 46.4 39.8 47.9 48.0  73.7  77.2
## Costa Rica      50.4 57.0 61.4 57.3 59.6 62.4 66.7 72.2  78.8  82.3
## Czech Republic  78.2 80.2 84.7 88.0 88.9 89.4 90.5 92.4  94.2  97.5
## Denmark         50.5 51.2 51.8 53.0 54.7 55.6 57.0 58.8  62.0  67.0
## Estonia         53.1 54.6 57.3 59.4 61.1 62.2 64.9 65.0  66.8  67.5
## Finland         84.8 85.9 88.1 89.8 92.3 91.7 95.8 99.9 101.8 105.6
## France          80.2 81.6 81.3 82.4 84.5 84.0 84.8 86.9  85.6  87.4
## Germany         83.2 83.2 83.6 82.6 83.8 83.5 85.9 88.3  91.0  93.6
## Hungary         75.3 77.5 76.2 69.9 72.4 74.2 75.4 77.0  79.2  76.4
## Iceland         38.9 41.9 42.3 44.9 46.5 47.2 48.0 48.7  51.1  54.8
## Italy           62.0 62.4 61.8 62.0 62.2 62.5 63.0 64.0  64.8  65.1
## Korea           64.3 65.4 60.8 61.7 60.3 63.1 65.2 67.8  72.7  78.5
## Lithuania       37.0 40.6 43.7 44.3 47.8 50.0 53.6 56.7  59.0  58.8
## Luxembourg      63.4 64.0 64.6 65.2 64.9 64.2 64.0 65.1  64.5  63.8
## Netherlands     74.1 75.1 75.0 75.8 76.2 77.2 75.1 76.9  77.6  78.5
## Norway          48.4 48.7 50.1 51.5 53.0 55.4 56.2 58.5  59.9  65.9
## Portugal        61.0 62.7 64.5 67.2 68.5 67.9 70.4 74.2  76.7  78.6
## Slovak Republic 58.0 65.6 70.4 75.2 76.3 76.0 76.8 78.6  80.4  79.5
## Slovenia        70.5 73.1 74.2 75.9 78.6 80.7 82.6 83.3  85.1  87.2
## Spain           66.4 66.6 71.1 72.9 75.1 76.4 78.1 79.6  81.3  84.7
## Sweden          53.9 55.6 56.5 58.4 60.4 62.4 64.9 67.8  70.1  73.2
## Turkiye         56.2 58.4 61.1 64.2 66.5 72.6 75.4 79.1  91.9  96.3
## United Kingdom  79.9 82.3 83.6 84.7 85.3 84.5 72.1 74.2  74.3  90.0
## Croatia         59.4 62.5 63.3 64.9 67.6 66.9 69.2 77.8  79.5  75.5
```

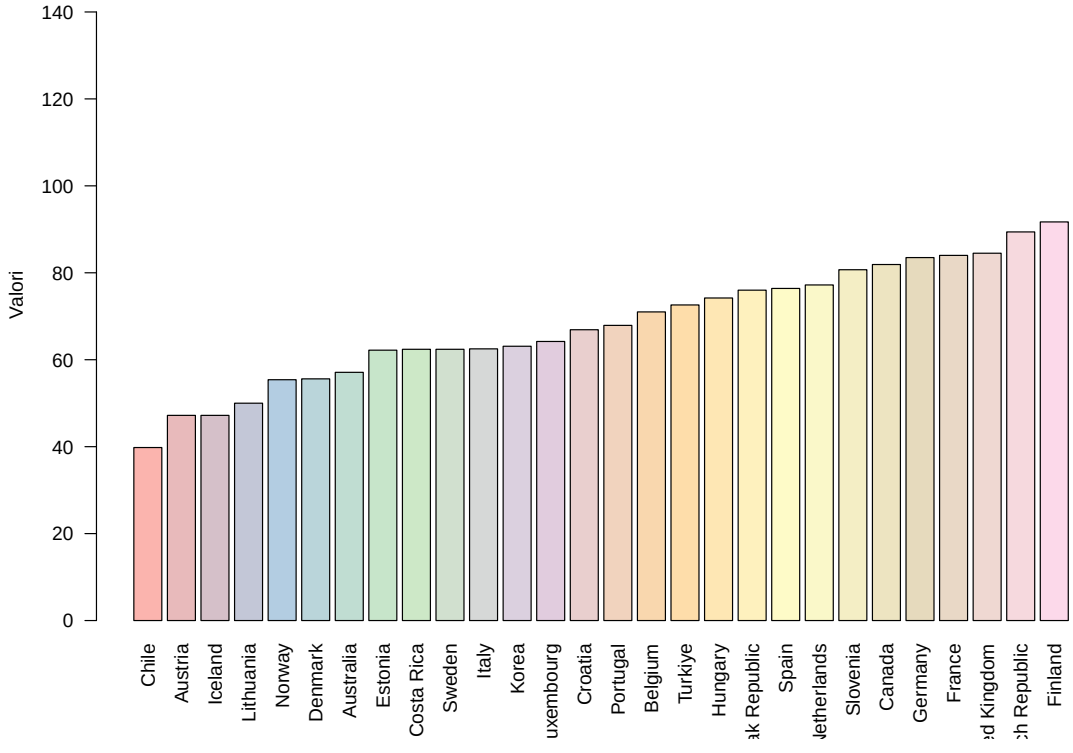

Andamento consumo farmaci per il diabete nel 2012
Valori
0
20
40
60
80
100
120
140
Chile
Lithuania
Iceland
Austria
Norway
Costa Rica
Denmark
Estonia
Sweden
Turkiye
Canada
Slovak Republic
Croatia
Australia
Portugal
Italy
Belgium
Luxembourg
Korea
Spain
Slovenia
Netherlands
Hungary
Czech Republic
United Kingdom
France
Germany
Finland

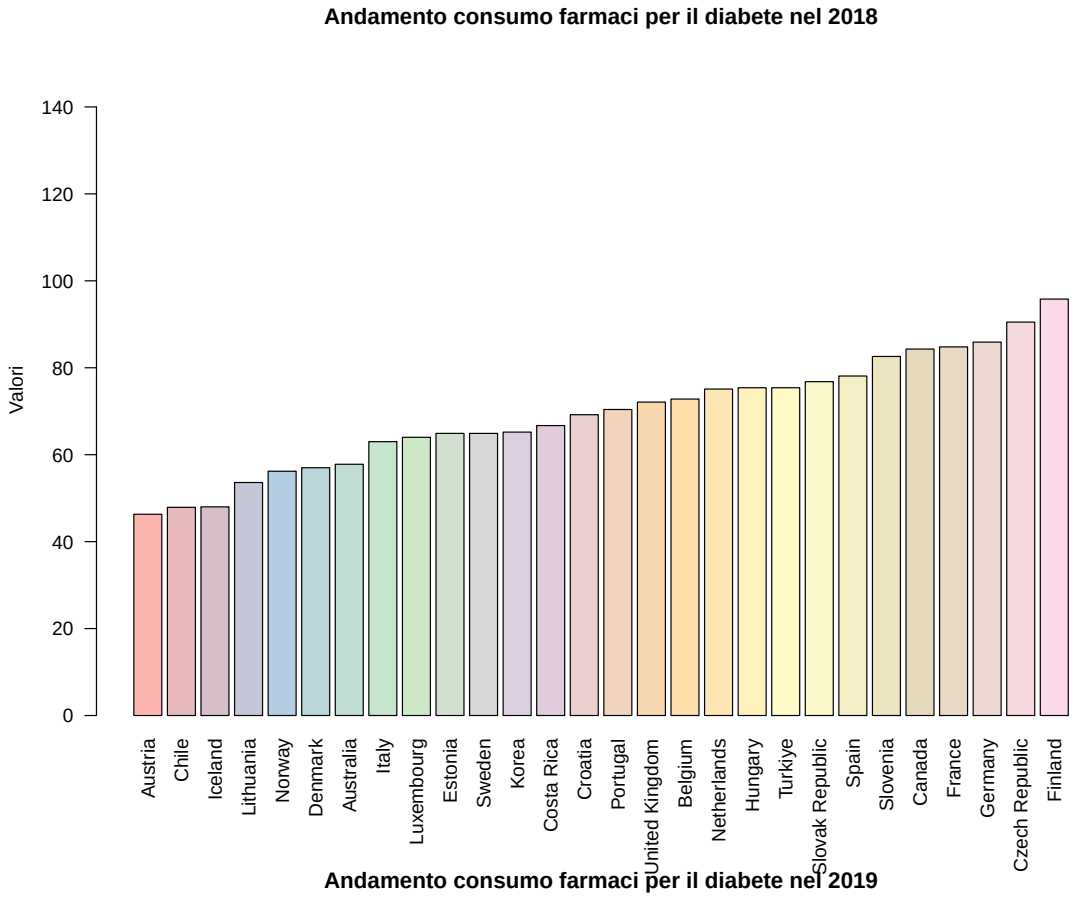

Andamento consumo farmaci per il diabete nel 2013
Valori
0
20
40
60
80
100
120
140
Lithuania
Iceland
Chile
Austria
Norway
Denmark
Estonia
Sweden
Costa Rica
Australia
Canada
Turkiye
Italy
Croatia
Portugal
Spain
Slovenia
Hungary
France
Germany
Finland

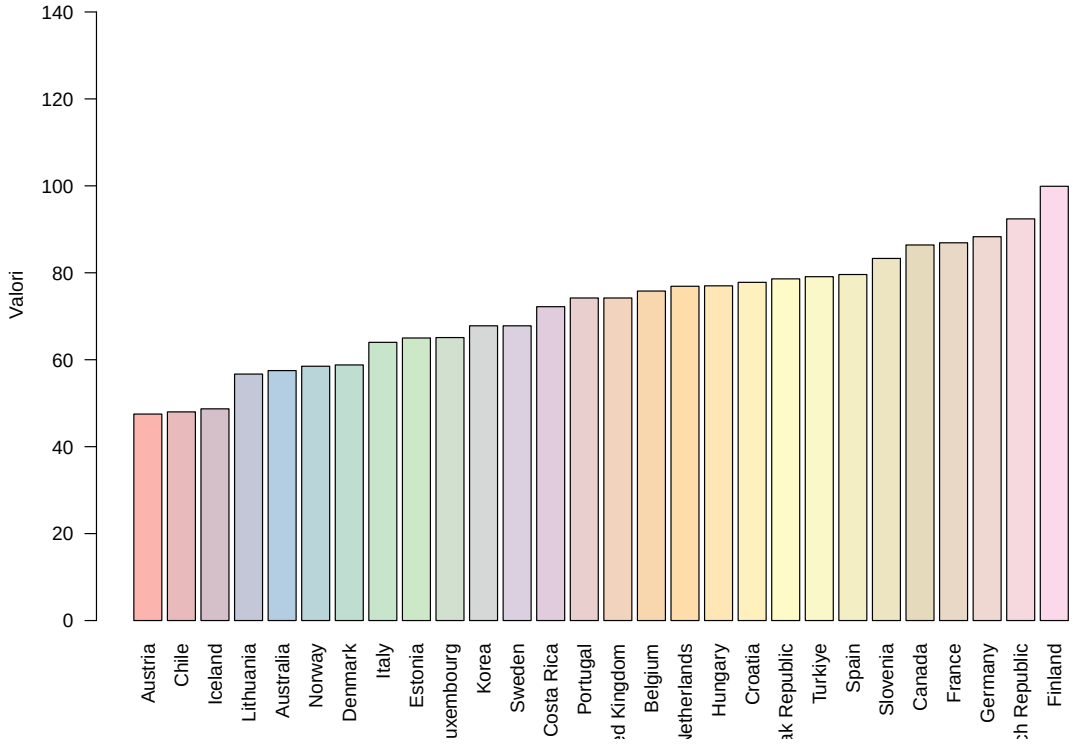

Andamento consumo farmaci per il diabete nel 2014
Valori
0
20
40
60
80
100
120
140
Iceland
Lithuania
Austria
Norway
Denmark
Chile
Sweden
Estonia
Australia
Canada
Korea
Turkiye
Costa Rica
Italy
Croatia
Portugal
Luxembourg
Belgium
Slovak Republic
Spain
Slovenia
Netherlands
Hungary
France
Germany
United Kingdom
Czech Republic
Finland

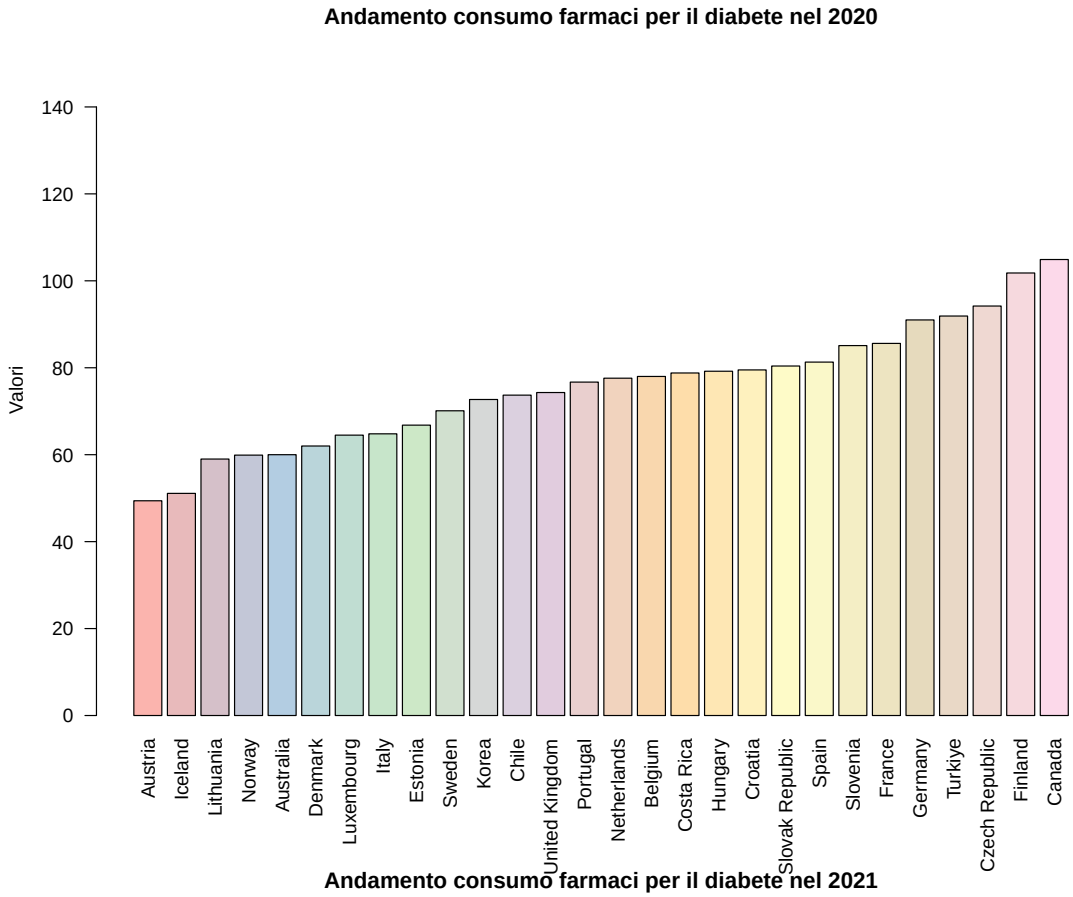

Andamento consumo farmaci per il diabete nel 2015
Valori
0
20
40
60
80
100
120
140
Lithuania
Iceland
Austria
Norway
Denmark
Costa Rica
Australia
Canada
Sweden
Chile
Estonia
Korea
Italy
Turkiye
Croatia
Portugal
Belgium
Hungary
Spain
Slovenia
France
Germany
Finland

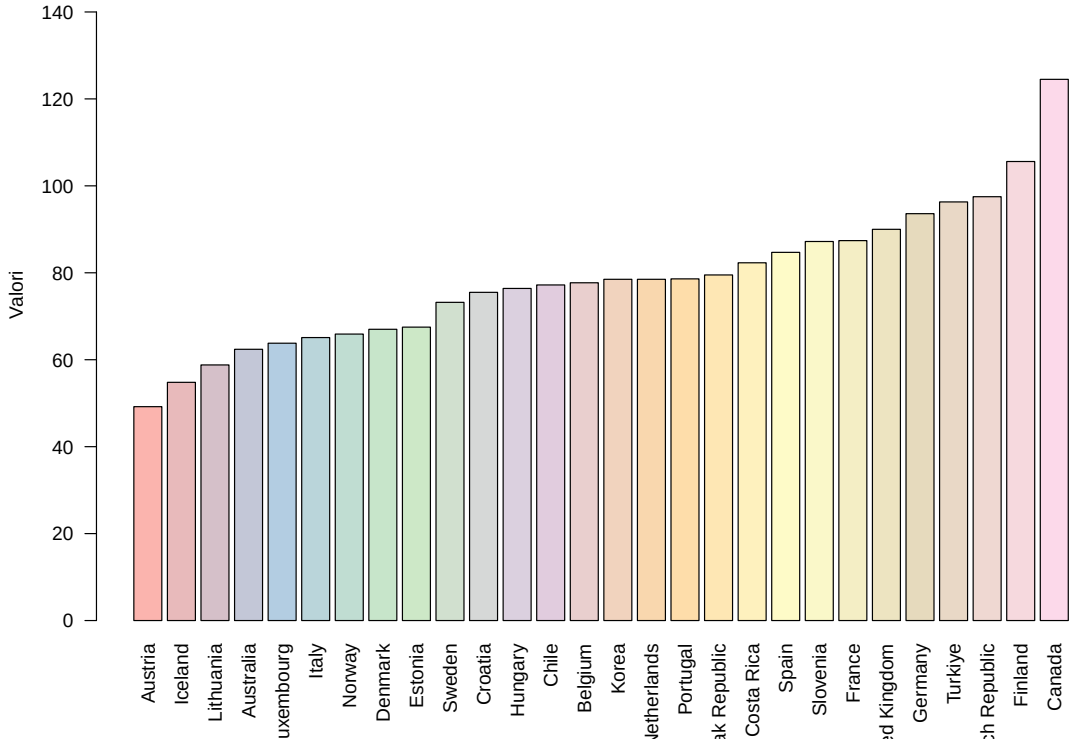

Andamento consumo farmaci per il diabete nel 2016
Valori
0
20
40
60
80
100
120
140
Chile
Iceland
Austria
Lithuania
Norway
Denmark
Australia
Costa Rica
Korea
Sweden
Estonia
Italy
Luxembourg
Turkiye
Croatia
Portugal
Belgium
Hungary
Spain
Netherlands
Slovak Republic
Slovenia
Canada
Germany
France
United Kingdom
Czech Republic
Finland


Andamento consumo farmaci per il diabete nel 2017
Valori
0
20
40
60
80
100
120
140
Chile
Austria
Iceland
Lithuania
Norway
Denmark
Australia
Estonia
Costa Rica
Sweden
Italy
Korea
Croatia
Portugal
Belgium
Turkiye
Hungary
Spain
Slovenia
Canada
Germany
France
Finland


Andamento consumo farmaci per il diabete nel 2018
Valori
0
20
40
60
80
100
120
140
Austria
Chile
Iceland
Lithuania
Norway
Denmark
Australia
Italy
Luxembourg
Estonia
Sweden
Korea
Costa Rica
Croatia
Portugal
United Kingdom
Belgium
Netherlands
Hungary
Turkiye
Slovak Republic
Spain
Slovenia
Canada
France
Germany
Czech Republic
Finland


Andamento consumo farmaci per il diabete nel 2019
Valori
0
20
40
60
80
100
120
140
Austria
Chile
Iceland
Lithuania
Australia
Norway
Denmark
Italy
Estonia
Luxembourg
Korea
Sweden
Costa Rica
Portugal
Belgium
Hungary
Croatia
Turkiye
Spain
Slovenia
Canada
France
Germany
Finland


Andamento consumo farmaci per il diabete nel 2020
Valori
0
20
40
60
80
100
120
140
Austria
Iceland
Lithuania
Norway
Australia
Denmark
Luxembourg
Italy
Estonia
Sweden
Korea
Chile
United Kingdom
Portugal
Netherlands
Belgium
Costa Rica
Hungary
Croatia
Slovak Republic
Spain
Slovenia
France
Germany
Turkiye
Czech Republic
Finland
Canada


Andamento consumo farmaci per il diabete nel 2021
Valori
0
20
40
60
80
100
120
140
Austria
Iceland
Lithuania
Australia
uxembourg
Italy
Norway
Denmark
Estonia
Sweden
Croatia
Hungary
Chile
Belgium
Korea
Netherlands
Portugal
ak Republic
Costa Rica
Spain
Slovenia
France
ed Kingdom
Germany
Turkiye
ch Republic
Finland
Canada

Dai grafici riportati possiamo notare come la Finlandia sia stato sempre il paese che ha riscontrato un maggiore consumo di farmaci per il diabete, tranne dal 2020 in cui il Canada ha avuto un notevole incremento. Lo stesso discorso può essere applicato al Cile in quanto paese con il minor tasso di consumo di farmaci per il diabete fino al 2020, anno in cui i valori sono aumentati.
E' possibile inoltre notare la differenza tra i paesi che registrano un basso tasso di consumo (ovvero quelli più a sinistra nel grafico) i cui valori si aggirano intorno ai 40/50 e i paesi con un maggiore consumo i cui valori arrivano anche a 130.

Distribuzioni di frequenza

Consideriamo una variabile X e indichiamo con $z_1, z_2, ..., z_k$ le modalità distinte da essa assunte. Prendiamo poi un campione costituito da n osservazioni di X e indichiamo con n_i il numero di volte in cui ciascuna modalità z_i è presente nel campione, ossia la frequenza assoluta con cui essa appare nel campione. L'insieme $(z_i, n_i), i = 1, 2, ...k$ si chiama **distribuzione di frequenza**.

La **frequenza assoluta** indica il numero di volte in cui ciascuna modalità è presente nel campione, dunque è pari alla numerosità del campione quando non vi sono dati mancanti, come nel nostro caso. Per le variabili quantitative è possibile calcolare le frequenze assolute distinguendo il numero di modalità, nel nostro caso è un range che va da 30 a 125. Si preferisce raccogliere le informazioni in classi e calcolare le frequenze relative a tali classi, ossia le frequenze con cui gli elementi del vettore cadono nelle diverse classi. Per fare tale operazione si utilizza la funzione *cut()* che permette di raggruppare i dati relativi ad un vettore in intervalli elencando nel parametro breaks gli estremi degli intervalli, nel nostro caso avremo quattro classi: (30,60], (60,80], (80, 100], (100, 125]

```
## [1] "frequenze assolute 2012"
```

```
##
##   (30,60]   (60,80]  (80,100] (100,125]
##        13        12         3         0
```

```
## [1] "frequenze assolute 2013"
```

```
##
##   (30,60]   (60,80]  (80,100] (100,125]
##        12        11         5         0
```

```
## [1] "frequenze assolute 2020"
```

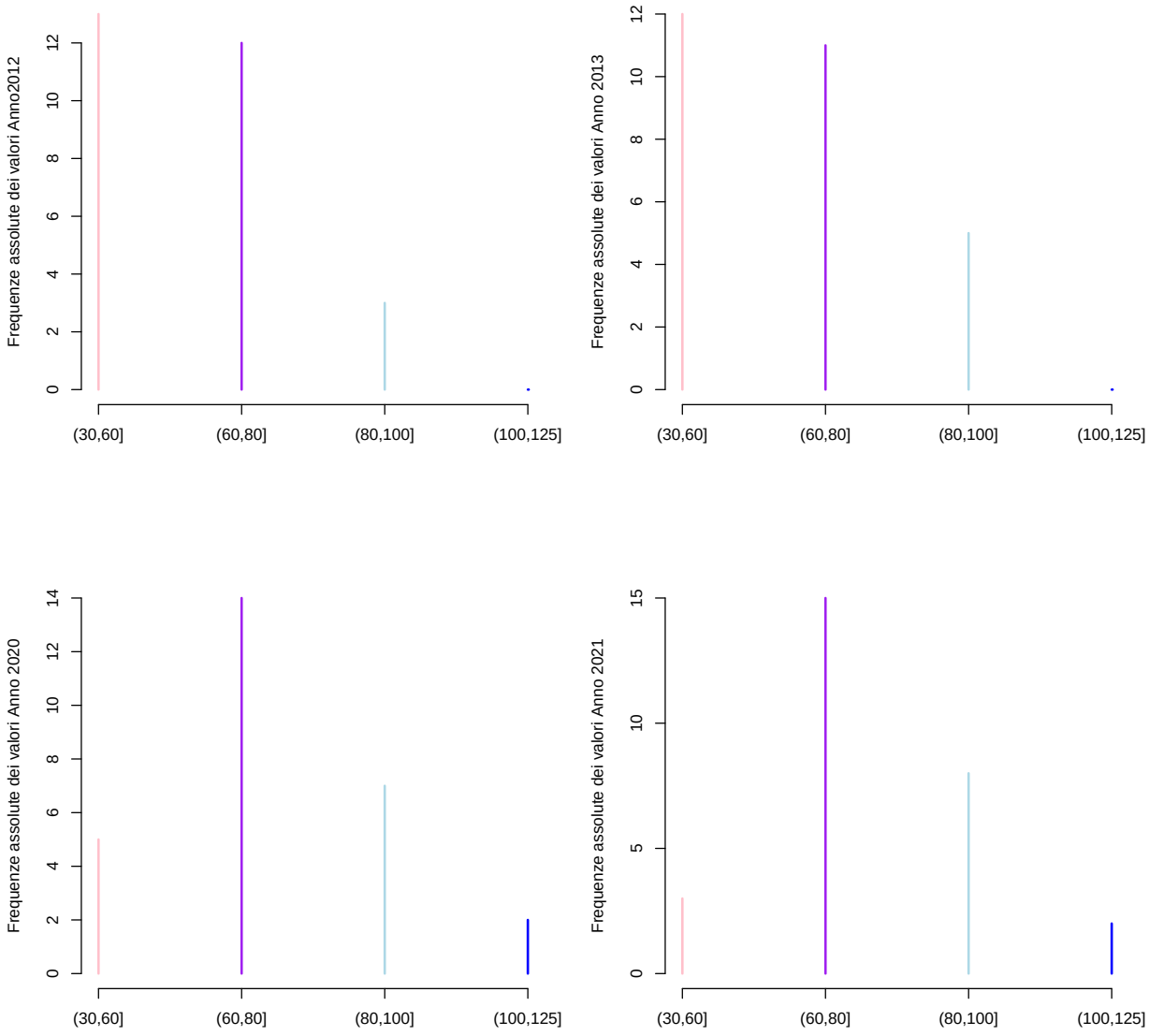
```
##
##   (30,60]   (60,80]  (80,100] (100,125]
##         5        14         7         2
```

```
## [1] "frequenze assolute 2021"
```

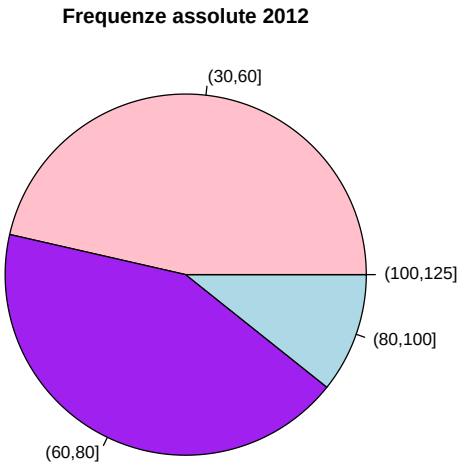
```
##
##   (30,60]   (60,80]  (80,100] (100,125]
##         3        15         8         2
```

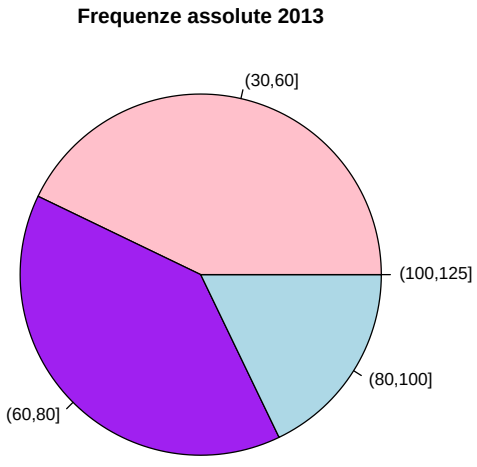
E' stata utilizzata la funzione *table()* per la costruzione della distribuzione di frequenza, quindi le frequenze assolute.

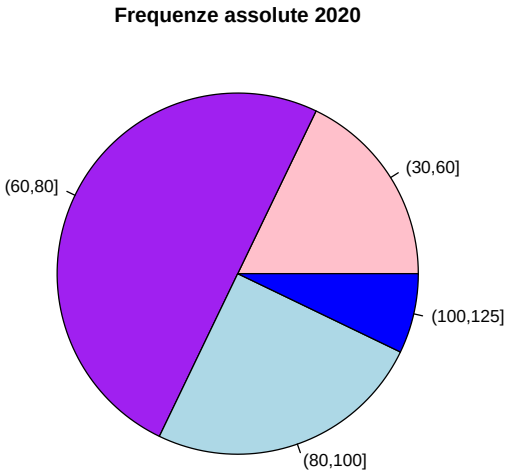
Per rappresentare correttamente la distribuzione di frequenza di una variabile quantitativa occorre utilizzare il comando *plot(table(x))* che produce un grafico a bastoncini in cui sull'asse orizzontale sono riportati i valori e sull'asse verticale le frequenze assolute dei valori distinti assunti dal vettore. Riferendosi ai valori riportati precedentemente avremo i seguenti grafici:

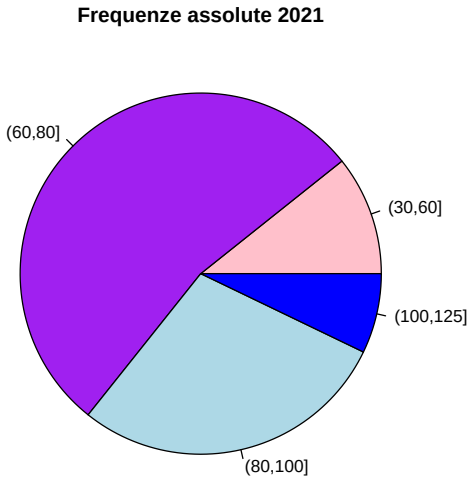


Per le variabili quantitative si può considerare anche una rappresentazione tramite diagrammi a torta generati dal comando *pie(table(x))*, dove x è un vettore o un fattore.









Calcoliamo poi le **frequenze assolute cumulate** tramite la funzione *cumsum()* che permette di calcolare le somme cumulate degli elementi di un vettore. La frequenza assoluta cumulata è definita come $N_i = n_1 + n_2 + ... + n_i (i = 1, 2, ..., k)$. E' una misura che indica il numero totale di osservazioni fino a un certo punto in una distribuzione di dati. Si calcola sommando le frequenze assolute dei valori precedenti o fino al punto desiderato nella distribuzione. La frequenza assoluta cumulata fornisce informazioni utili sulla distribuzione complessiva dei dati.

```
## [1] "frequenze assolute cumulate 2012"
```

```
##   (30,60]   (60,80]  (80,100] (100,125]
##        13        25        28        28
```

```
## [1] "frequenze assolute cumulate 2013"
```

```
##   (30,60]   (60,80]  (80,100] (100,125]
##        12        23        28        28
```

```
## [1] "frequenze assolute cumulate 2020"
```

```
##   (30,60]   (60,80]  (80,100] (100,125]
##         5        19        26        28
```

```
## [1] "frequenze assolute cumulate 2021"
```

```
##   (30,60]   (60,80]  (80,100] (100,125]
##         3        18        26        28
```

Si può calcolare la **frequenza relativa**, ovvero il rapporto tra la il numero di volte in cui ciascuna modalità è presente nel campione e la numerosità del campione stesso.

```
## [1] "frequenze relative anno 2012"
```

```
##
##   (30,60]   (60,80]  (80,100] (100,125]
## 0.4642857 0.4285714 0.1071429 0.0000000
```

```
## [1] "frequenze relative anno 2013"
```

```
##
##   (30,60]   (60,80]  (80,100] (100,125]
## 0.4285714 0.3928571 0.1785714 0.0000000
```

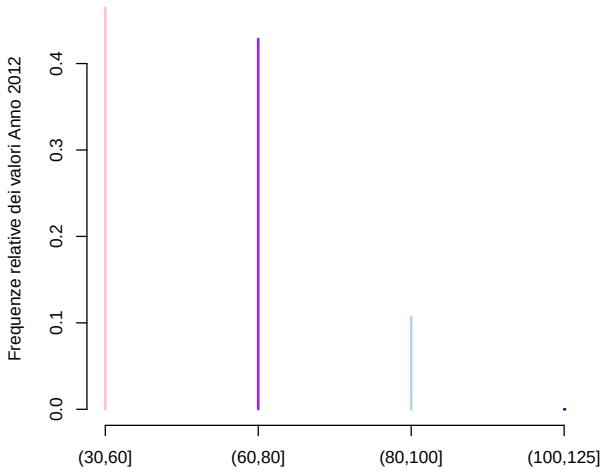
```
## [1] "frequenze relative anno 2020"
```

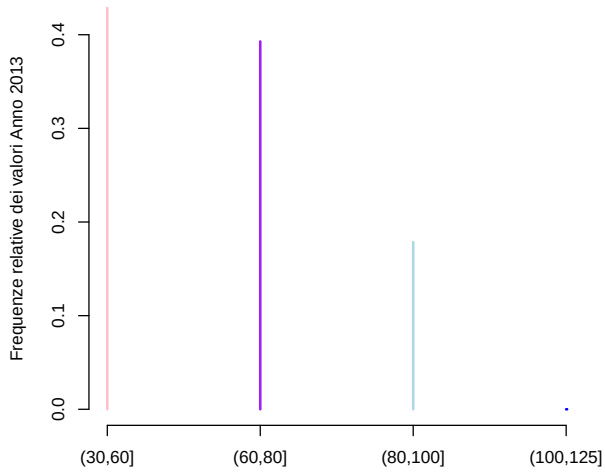
```
##
##    (30,60]    (60,80]   (80,100]  (100,125]
## 0.17857143 0.50000000 0.25000000 0.07142857
```

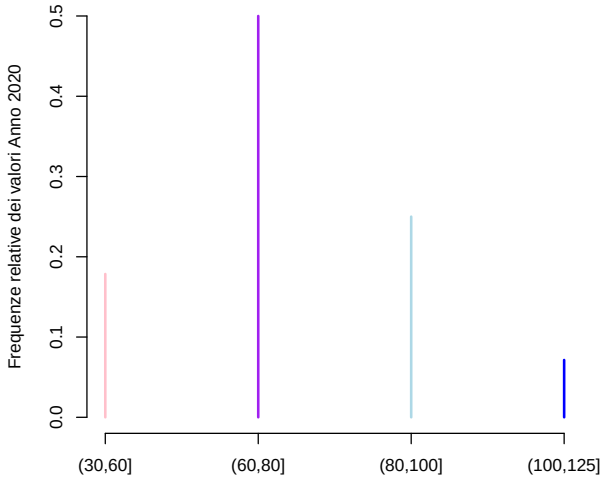
```
## [1] "frequenze relative anno 2021"
```

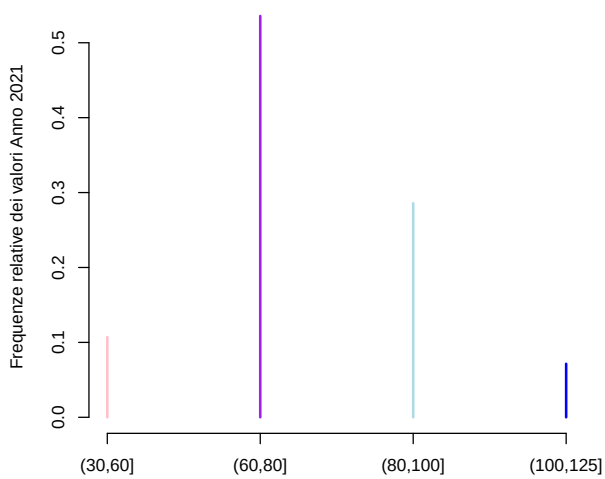
```
##
##    (30,60]    (60,80]   (80,100]  (100,125]
## 0.10714286 0.53571429 0.28571429 0.07142857
```

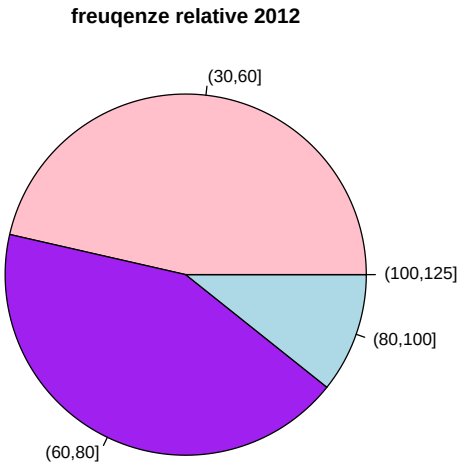
Per rappresentare le distribuzioni appena calcolate utilizziamo sia un grafico a bastoncini che un grafico a torta.

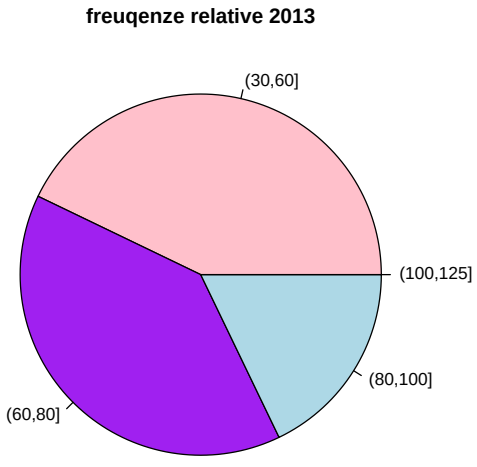

Frequenze relative dei valori Anno 2012
0.0
0.1
0.2
0.3
0.4
(30,60]
(60,80]
(80,100]
(100,125]

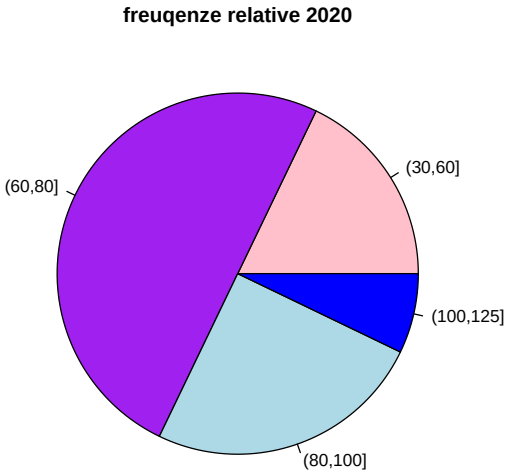

Frequenze relative dei valori Anno 2013
0.0
0.1
0.2
0.3
0.4
(30,60]
(60,80]
(80,100]
(100,125]

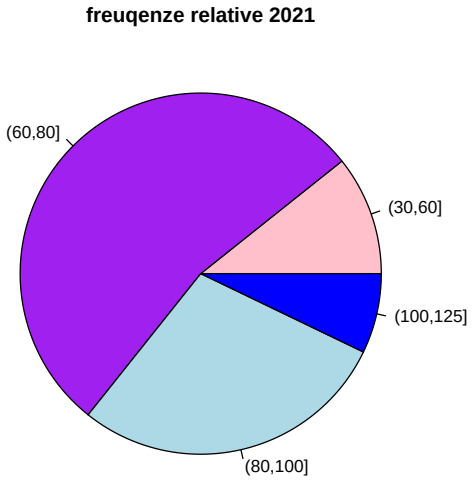

Frequenze relative dei valori Anno 2020
0.0
0.1
0.2
0.3
0.4
0.5
(30,60]
(60,80]
(80,100]
(100,125]


Frequenze relative dei valori Anno 2021
0.0
0.1
0.2
0.3
0.4
0.5
(30,60]
(60,80]
(80,100]
(100,125]









Anche in questo caso è possibile calcolare le **frequenze relative cumulate**, che serviranno per definire la funzione di distribuzione in seguito, mediante la funzione *cumsum()*. La frequenza relativa cumulata è definita come $F_i = f_1 + f_2 + ... + f_i (i = 1, 2, ..., k)$.

```
## [1] "frequenze relative cumulate anno 2012"
```

```
##   (30,60]   (60,80]  (80,100] (100,125]
## 0.4642857 0.8928571 1.0000000 1.0000000
```

```
## [1] "frequenze relative cumulate anno 2013"
```

```
##   (30,60]   (60,80]  (80,100] (100,125]
## 0.4285714 0.8214286 1.0000000 1.0000000
```

```
## [1] "frequenze relative cumulate anno 2020"
```

```
##   (30,60]   (60,80]  (80,100] (100,125]
```

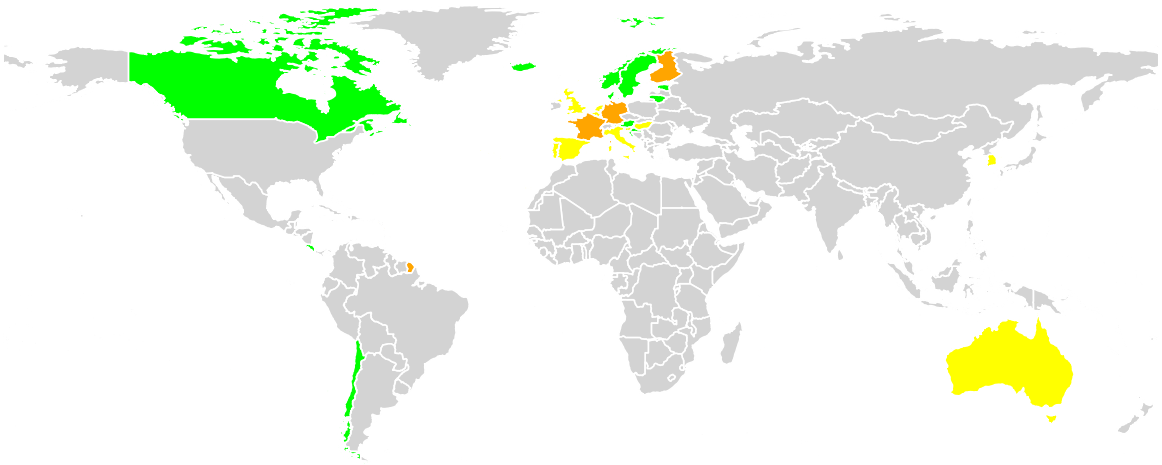
```
## 0.1785714 0.6785714 0.9285714 1.0000000
```

```
## [1] "frequenze relative cumulate anno 2021"
```

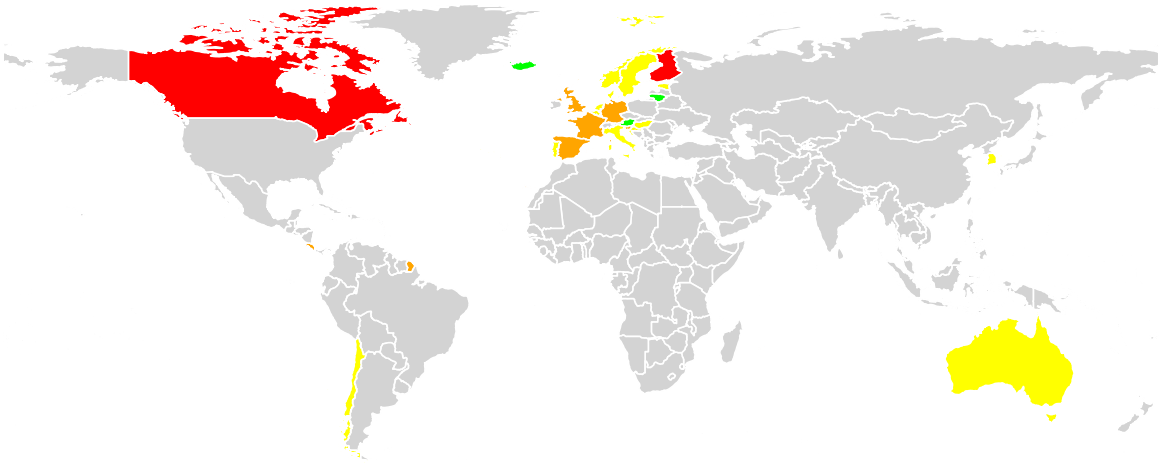
```
##   (30,60]   (60,80]  (80,100] (100,125]
## 0.1071429 0.6428571 0.9285714 1.0000000
```

Con gli stessi intervalli utilizzati per il calcolo delle distribuzioni di frequenza, visualizzo una mappa in cui coloro in verde i paesi che hanno valori compresi tra 30 e 60, in giallo tra 60 e 80, in arancione tra 80 e 100 e in rosso tra 100 e 125. In particolare effettuo questa operazione per gli anni 2012 e 2021 per visualizzare in che modo sono cambiati i valori.

Visualizziamo il grafico prodotto per l'anno 2012.



Visualizziamo il grafico prodotto per l'anno 2021.



Boxplot

Consideriamo un campione dei valori assunti dalla variabile quantitativa X. Procediamo ad ordinare i valori del campione in ordine crescente mediante i quartili. Successivamente generiamo il *boxplot* che rappresenta una scatola i cui estremi sono il quartile Q1 e il quartile Q3, tagliata da una linea orizzontale in corrispondenza del quartile Q2, ossia la mediana. Effettueremo tale analisi sui dati relativi agli anni 2012, 2013, 2020 e 2021.

```
## [1] "Quartili anno 2012"
```

```
##    0%   25%   50%   75%  100%
## 30.70 52.45 60.75 71.40 84.80
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.70   52.45   60.75   60.96   71.40   84.80
```

```
## [1] "Quartili anno 2013"
```

```
##    0%   25%   50%   75%  100%
## 40.60 55.35 62.60 73.60 85.90
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   40.60   55.35   62.60   63.16   73.60   85.90
```

```
## [1] "Quartili anno 2020"
```

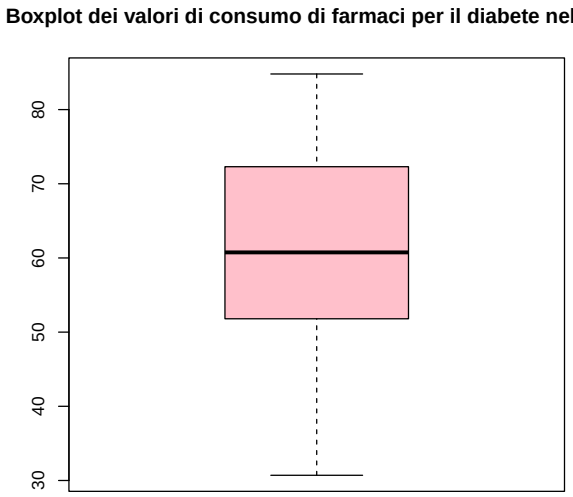
```
##      0%     25%     50%     75%    100%
##  49.400  64.725  77.150  82.250 104.900
```

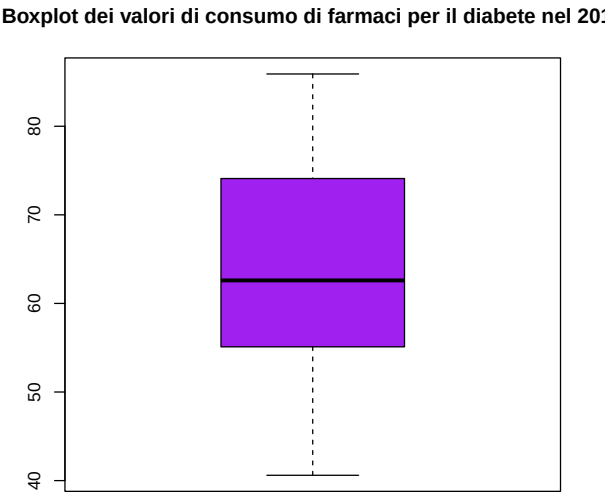
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   49.40   64.72   77.15   75.51   82.25  104.90
```

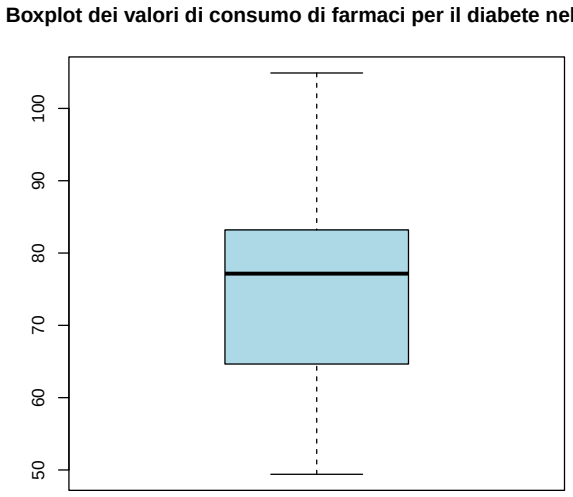
```
## [1] "Quartili anno 2021"
```

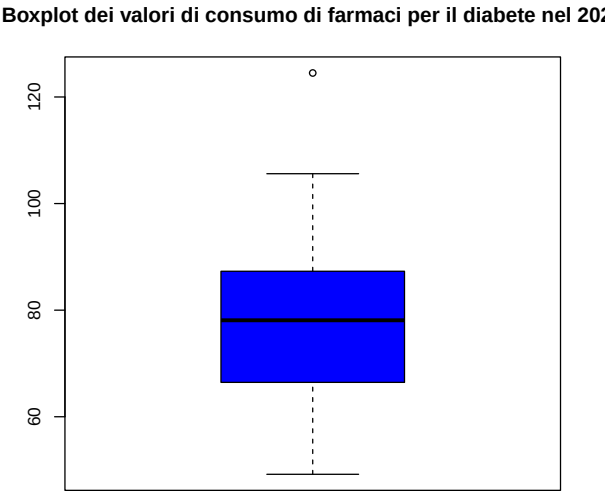
```
##      0%     25%     50%     75%    100%
##  49.200  66.725  78.100  87.250 124.500
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   49.20   66.72   78.10   78.53   87.25  124.50
```

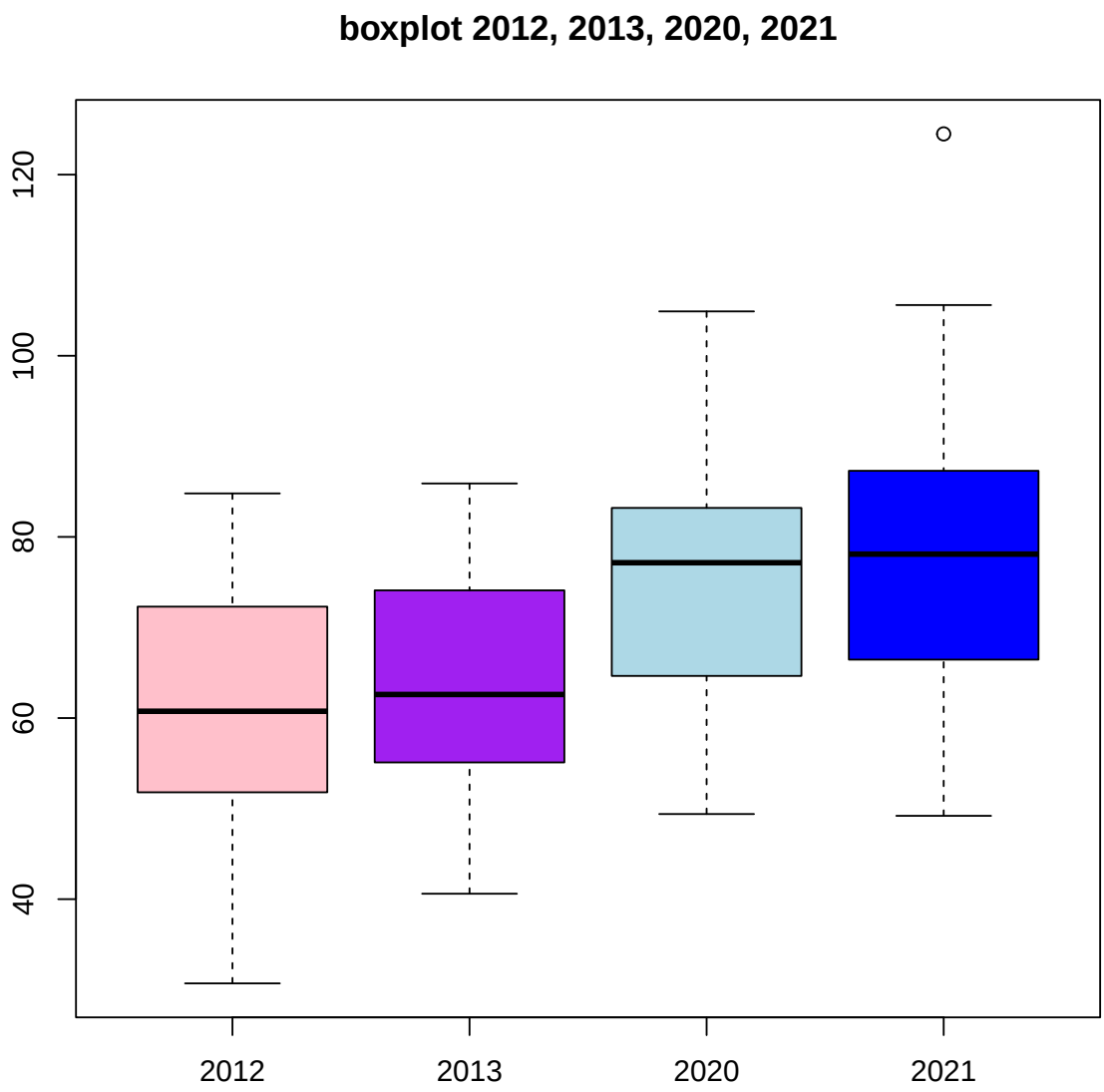




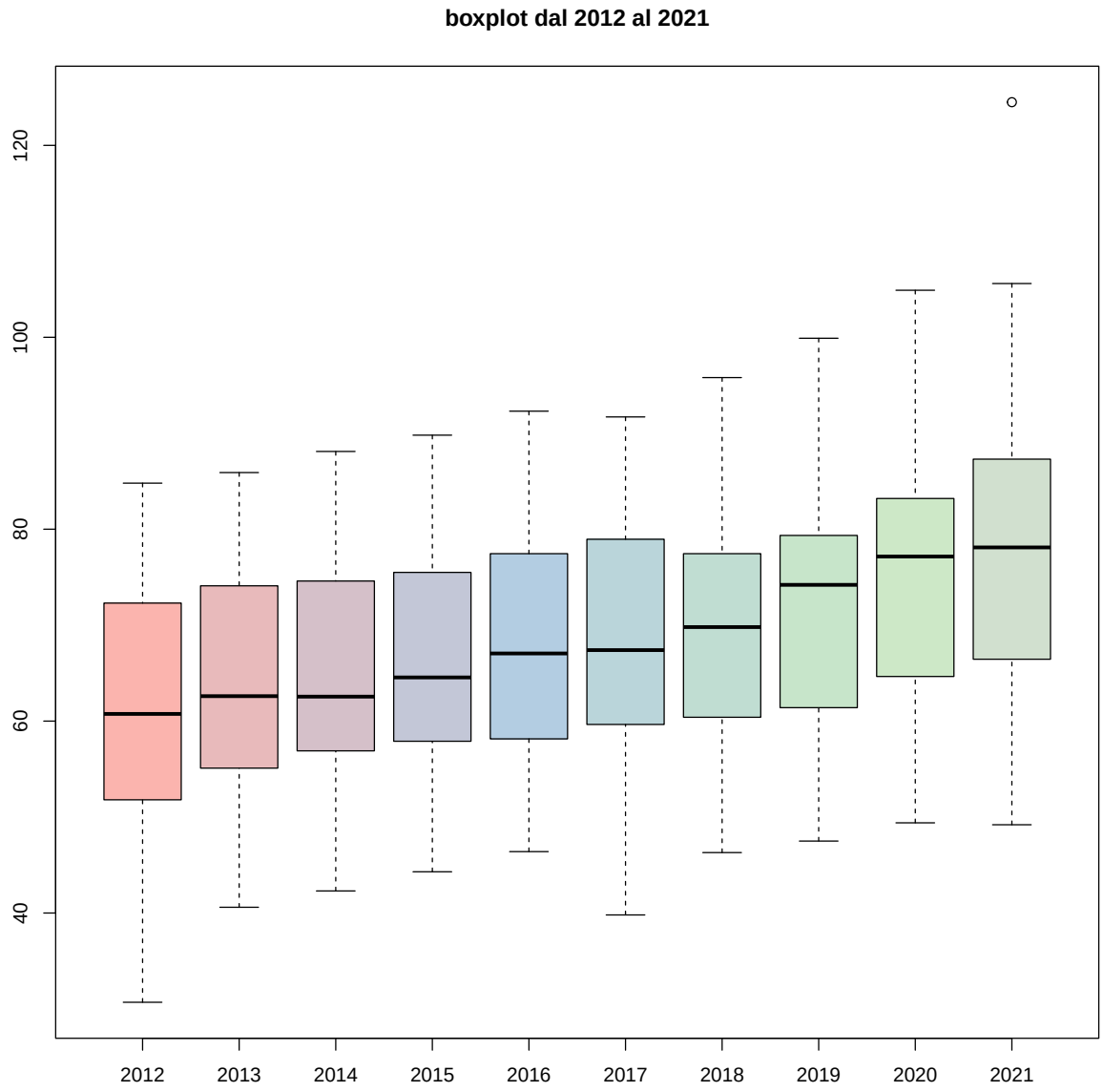




E' possibile quindi ricavare il valore dei quartili Q_0, Q_1, Q_2, Q_3, Q_4 per ciascun anno e visualizzare i singoli boxplot. Visualizziamo adesso tutti i boxplot in un unico grafico per analizzarli.



L'estremo più basso del baffo inferiore corrisponde al valore più piccolo tra le osservazioni che risulta maggiore o uguale di $Q_1 - 1,5 * (Q_3 - Q_1)$, mentre l'estremo del baffo superiore corrisponde al valore più grande delle osservazioni che risulta minore o uguale a $Q_3 + 1,5 * (Q_3 - Q_1)$. Possiamo quindi notare come entrambi i valori, sia il più piccolo che il più grande, sono aumentati nel corso degli anni implicando l'esistenza di un trend in aumento. La distanza tra il primo e il terzo quartile è detta intervallo interquartile o scarto interquartile. I dati del campione al di fuori dell'intervallo $(Q_1 - 1,5 * (Q_3 - Q_1), Q_3 + 1,5 * (Q_3 - Q_1))$ sono visualizzati nel grafico sotto forma di punti e sono detti *valori anomali* o outlier. Questi valori costituiscono un'anomalia rispetto alla maggior parte dei valori osservati. Nel nostro caso è presente solo nell'anno 2021. Tale anomalia è causata dal valore del Canada nel 2021 che è pari a 124.5. Tramite il boxplot, in particolare visualizzando la mediana, è possibile avere informazioni circa la forma simmetrica o asimmetrica della distribuzione.



Possiamo notare che in generale, visualizzando i boxplot relativi a ciascun anno, c'è stato un trend crescente del valore dei dati e sempre solo nel 2021 è presente un valore anomalo.

Serie temporali

Una *serie temporale* può essere memorizzata in un vettore, in una matrice o in un data frame. La funzione R che permette di definire una *serie temporale* è *ts()* in cui avremo i seguenti parametri:

x: valori della serie;

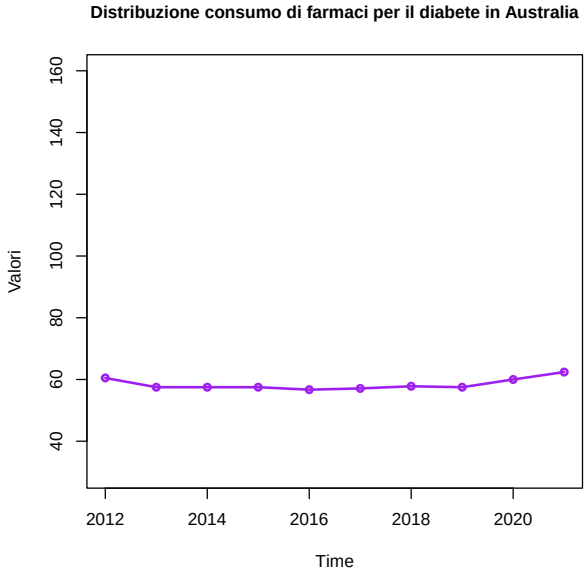
start: istante iniziale temporale;

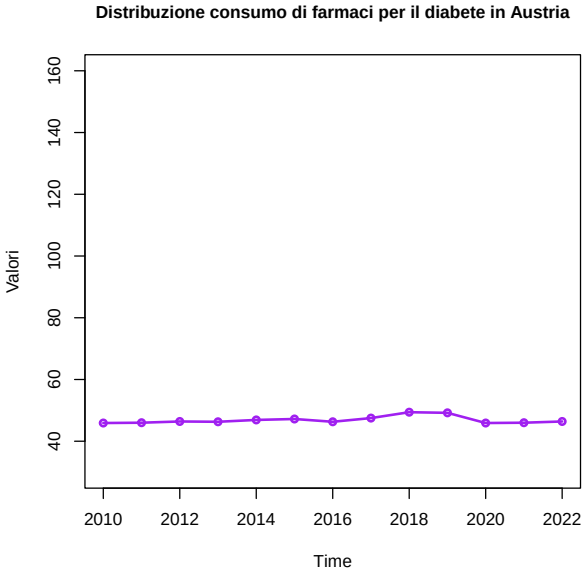
frequency: numero di osservazioni nell'unità di tempo (reciproco di deltat);

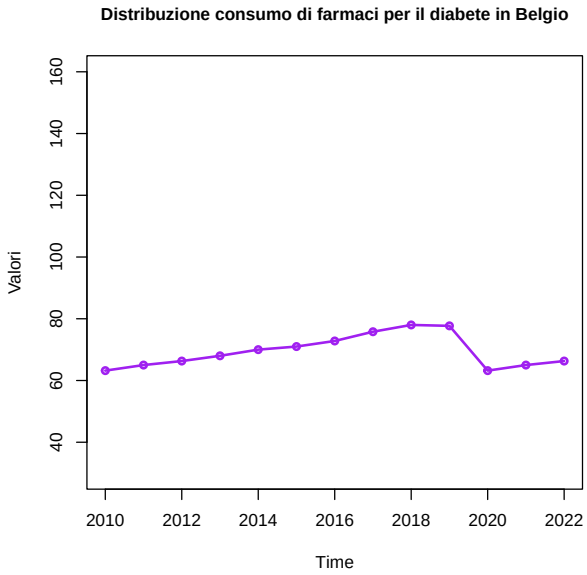
deltat: la distanza temporale tra le osservazioni;

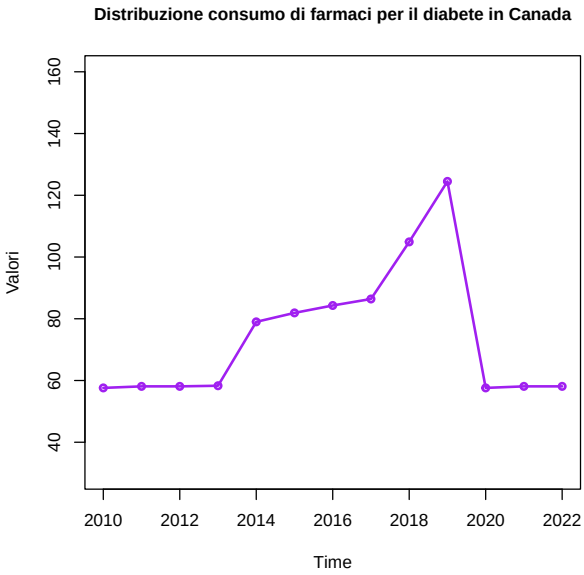
end: istante finale.

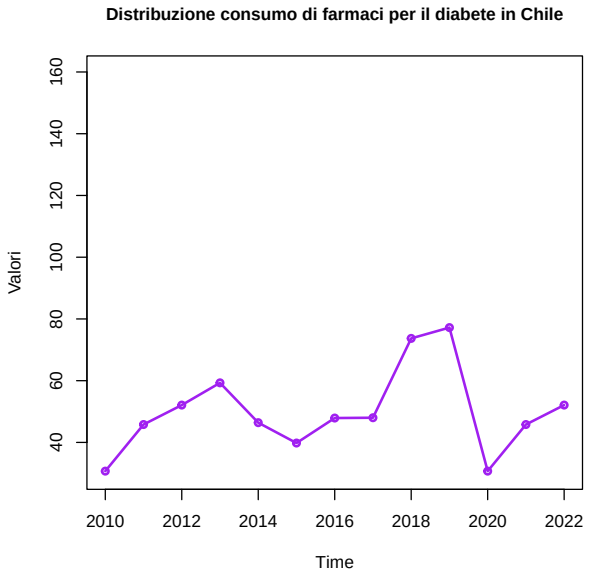
La funzione *plot()* permette successivamente di rappresentare il grafico della serie temporale creata.

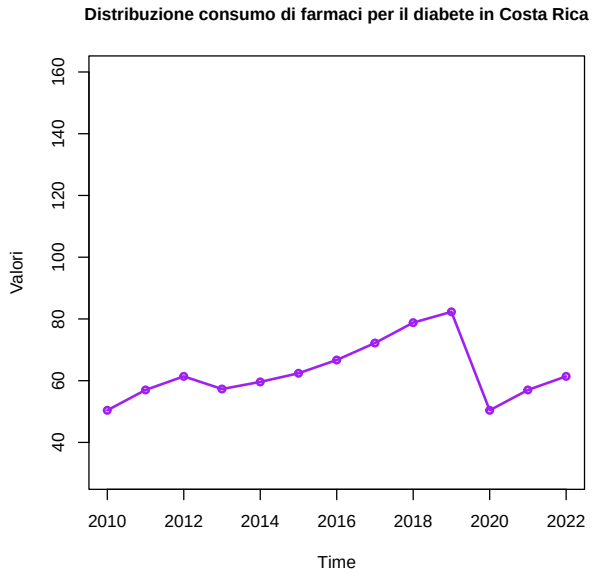


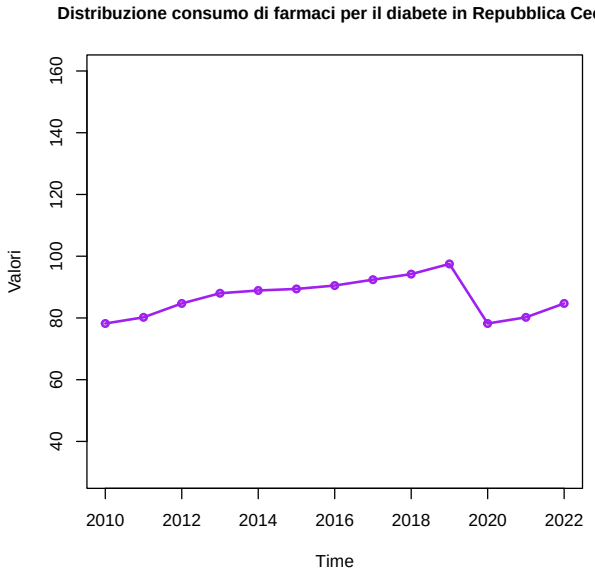


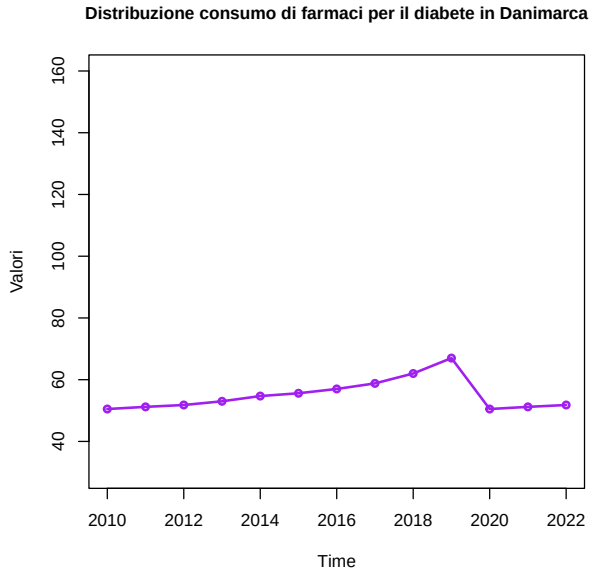


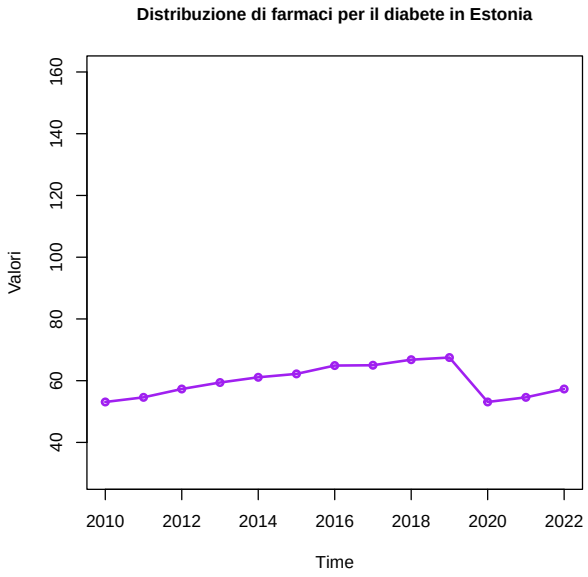


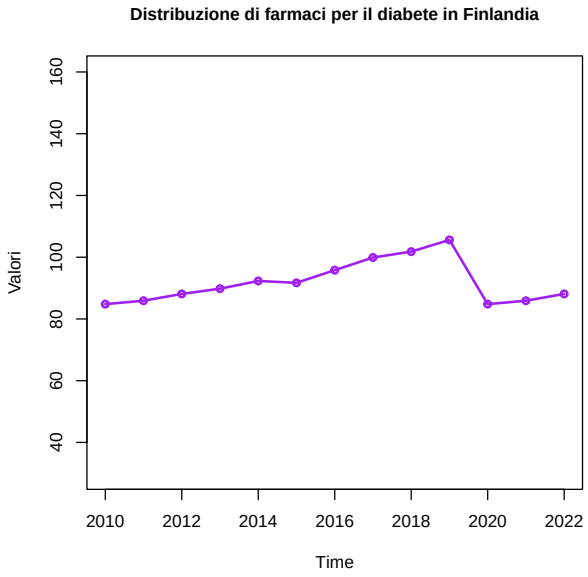

Distribuzione consumo di farmaci per il diabete in Chile
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time

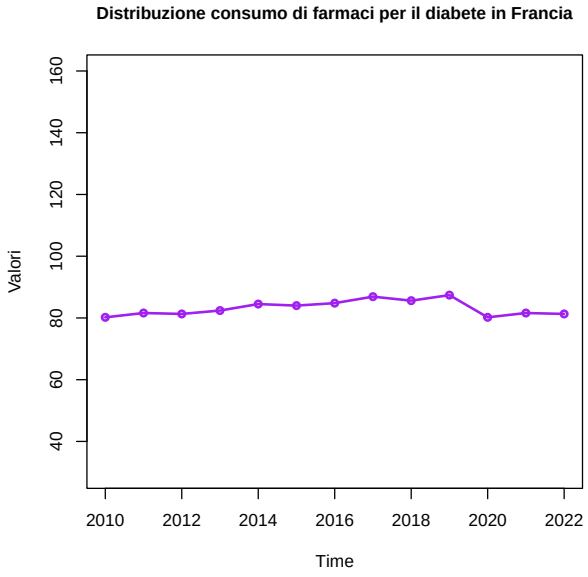

Distribuzione consumo di farmaci per il diabete in Costa Rica
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time

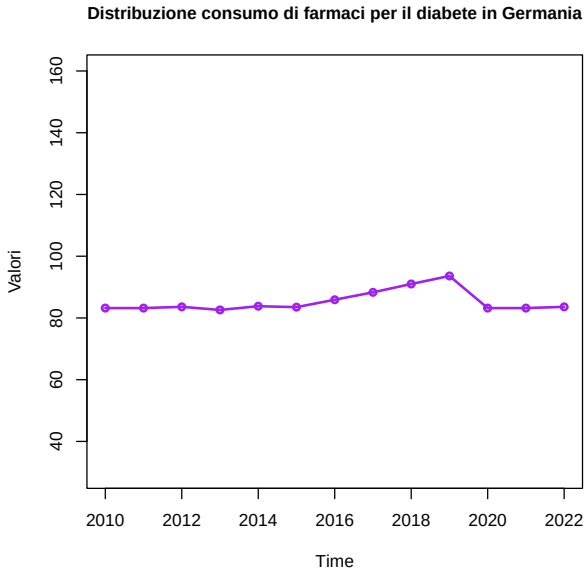

Distribuzione consumo di farmaci per il diabete in Repubblica Ceca
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time

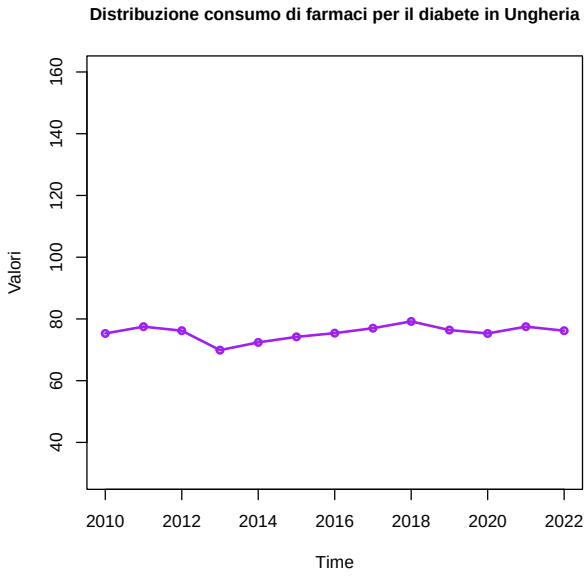

Distribuzione consumo di farmaci per il diabete in Danimarca
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time

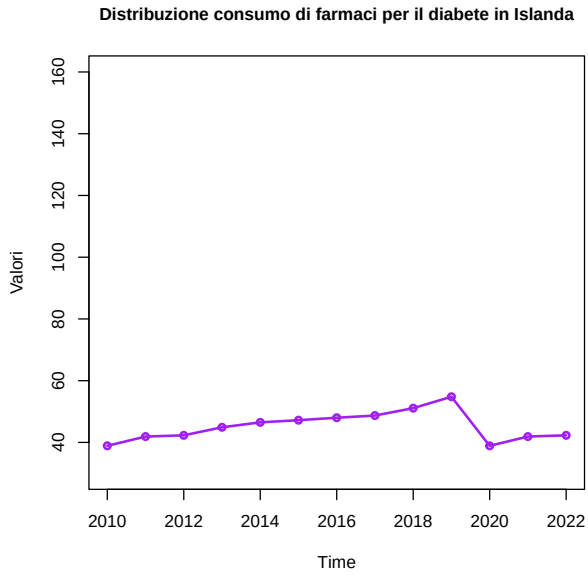

Distribuzione di farmaci per il diabete in Estonia
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time

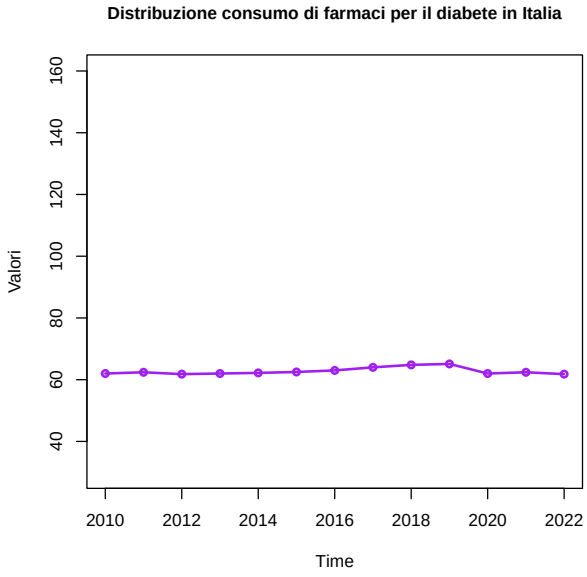

Distribuzione di farmaci per il diabete in Finlandia
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time

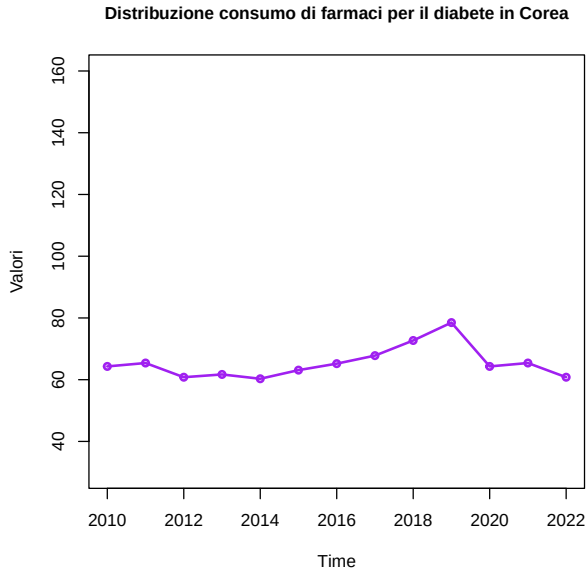

Distribuzione consumo di farmaci per il diabete in Francia
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time

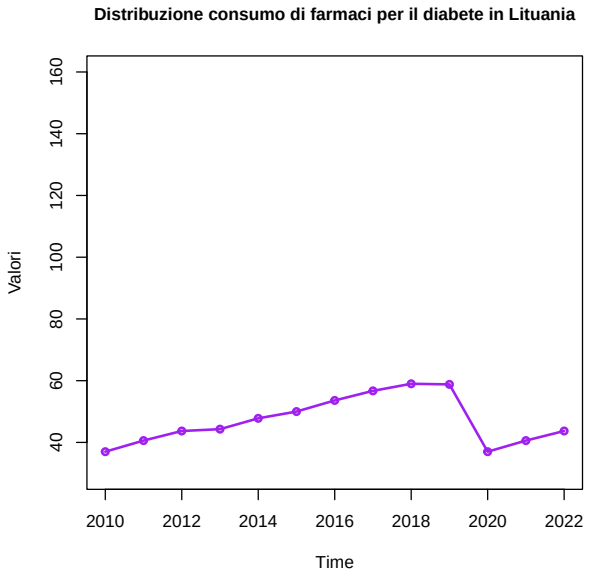

Distribuzione consumo di farmaci per il diabete in Germania
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time

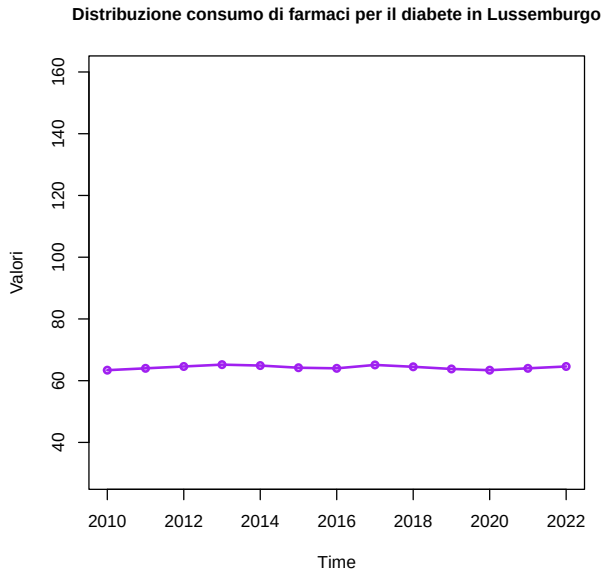

Distribuzione consumo di farmaci per il diabete in Ungheria
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time

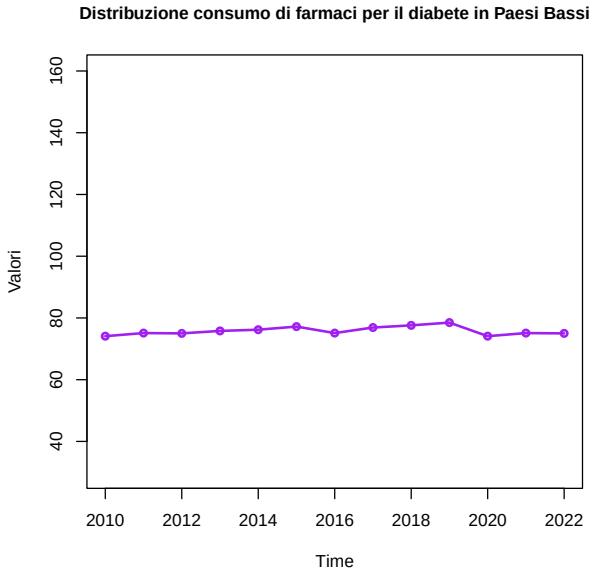

Distribuzione consumo di farmaci per il diabete in Islanda
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time

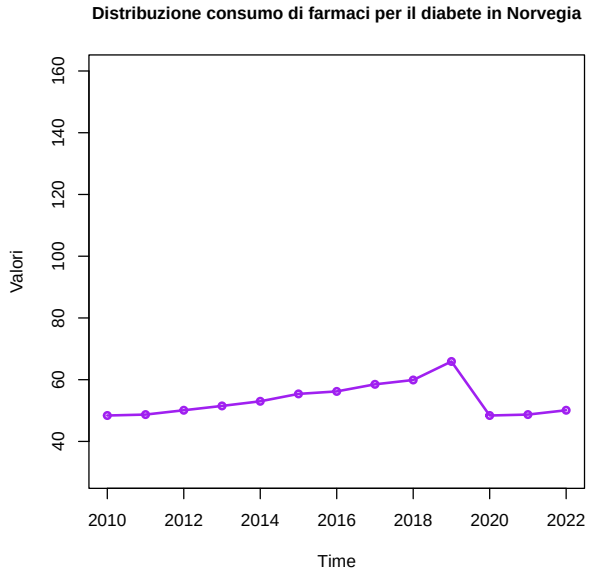

Distribuzione consumo di farmaci per il diabete in Italia
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time

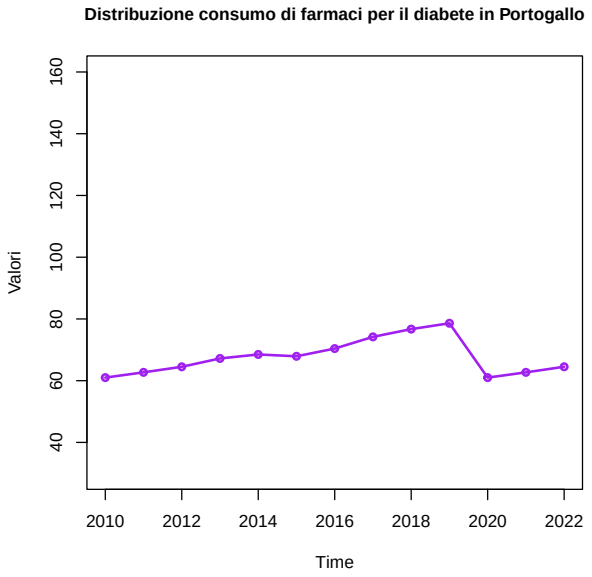

Distribuzione consumo di farmaci per il diabete in Corea
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time

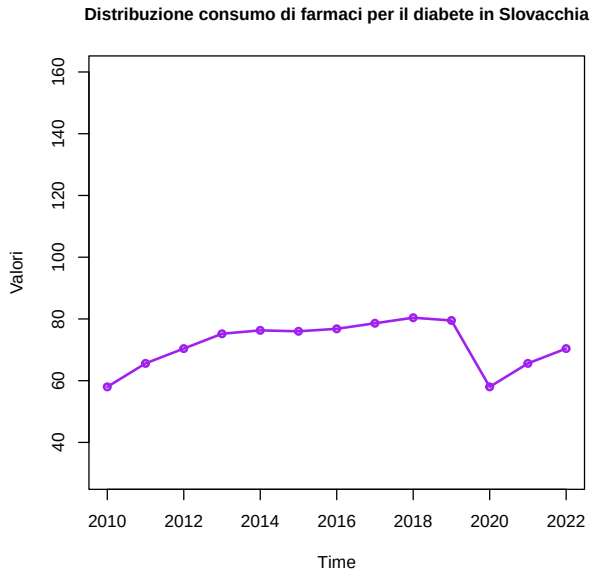

Distribuzione consumo di farmaci per il diabete in Lituania
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time

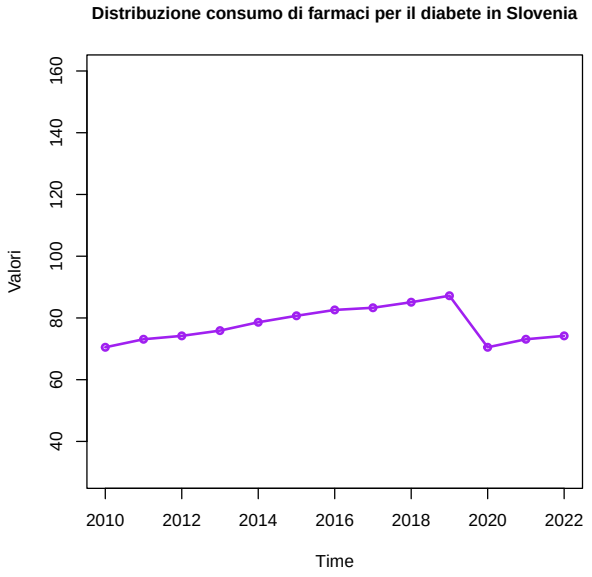

Distribuzione consumo di farmaci per il diabete in Lussemburgo
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time

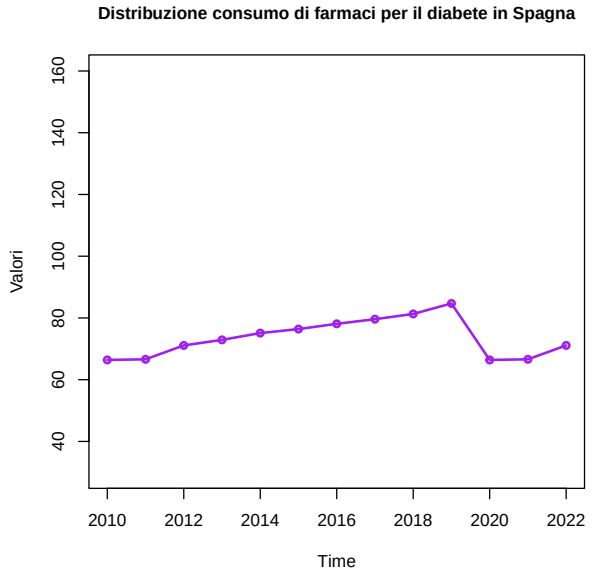

Distribuzione consumo di farmaci per il diabete in Paesi Bassi
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time

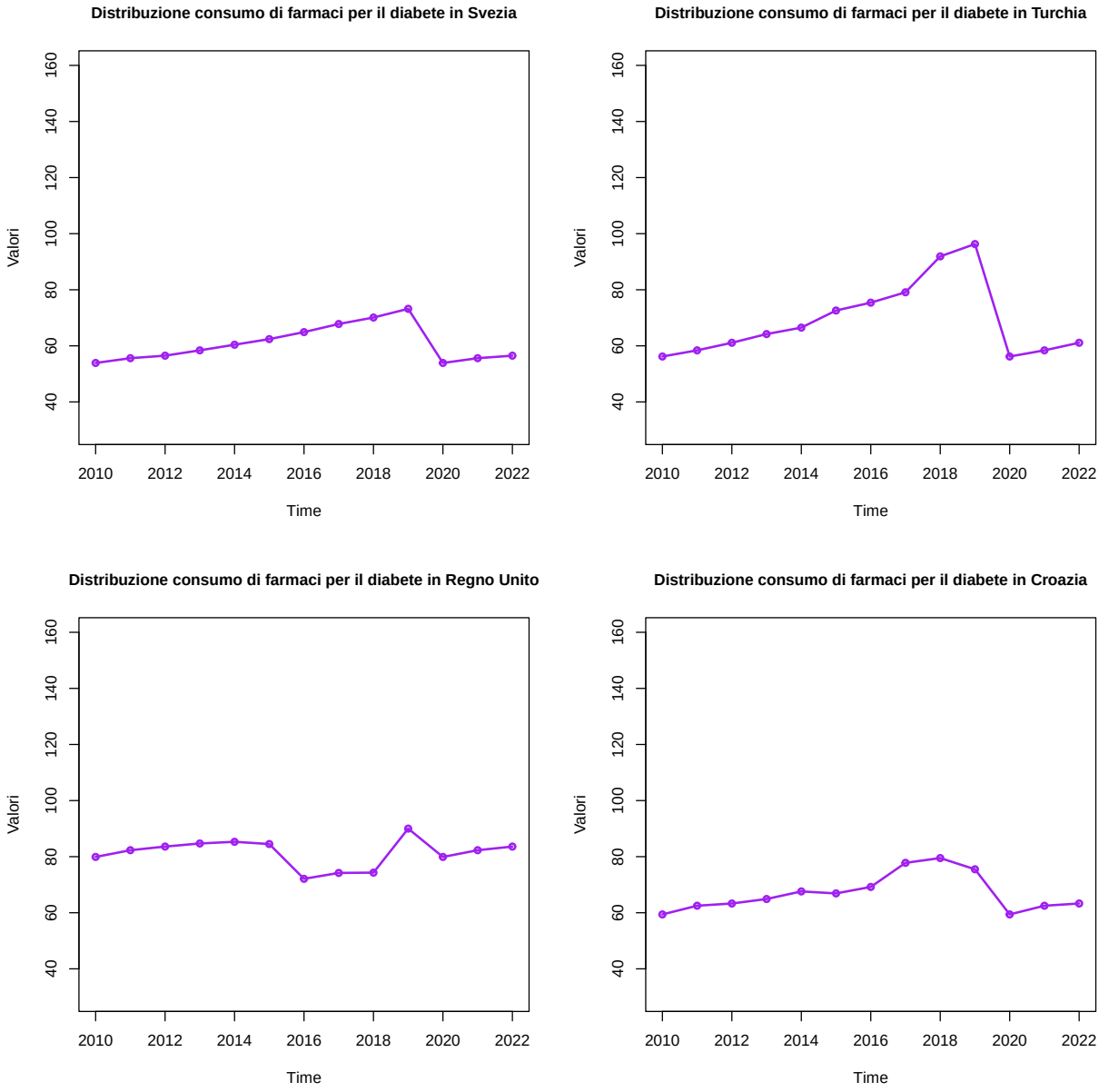

Distribuzione consumo di farmaci per il diabete in Norvegia
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time


Distribuzione consumo di farmaci per il diabete in Portogallo
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time


Distribuzione consumo di farmaci per il diabete in Slovacchia
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time

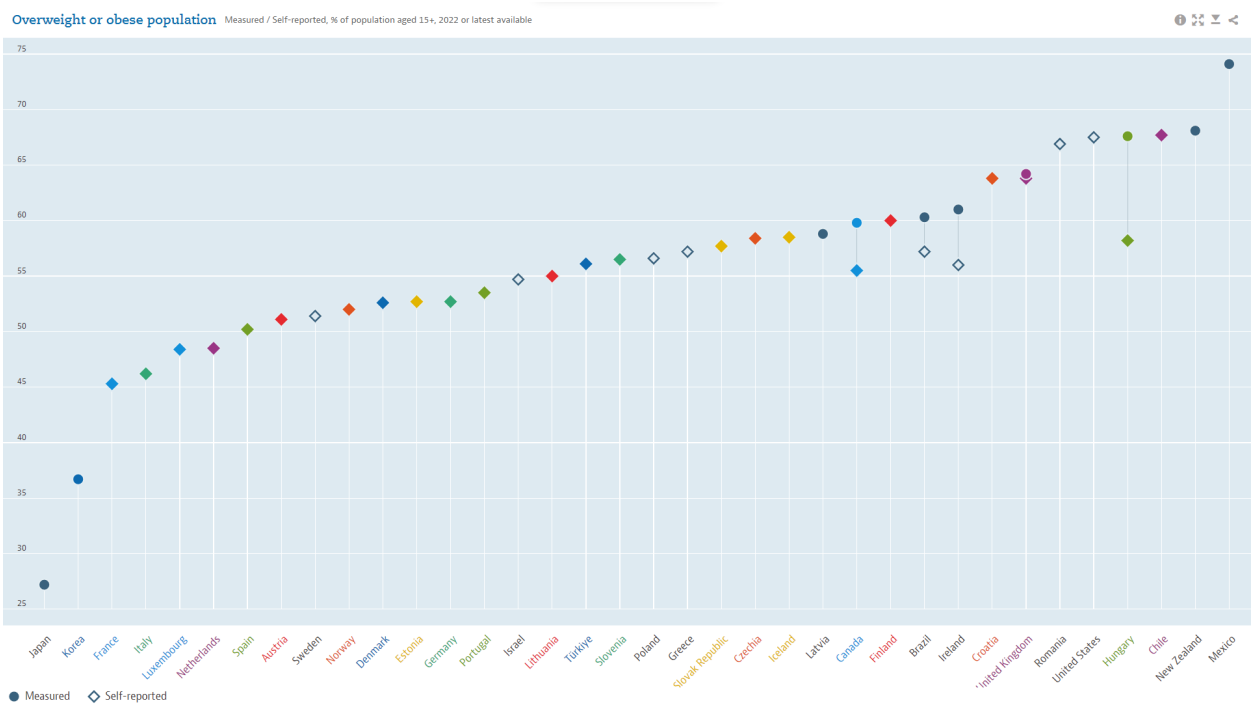

Distribuzione consumo di farmaci per il diabete in Slovenia
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time


Distribuzione consumo di farmaci per il diabete in Spagna
Valori
40
60
80
100
120
140
160
2010
2012
2014
2016
2018
2020
2022
Time



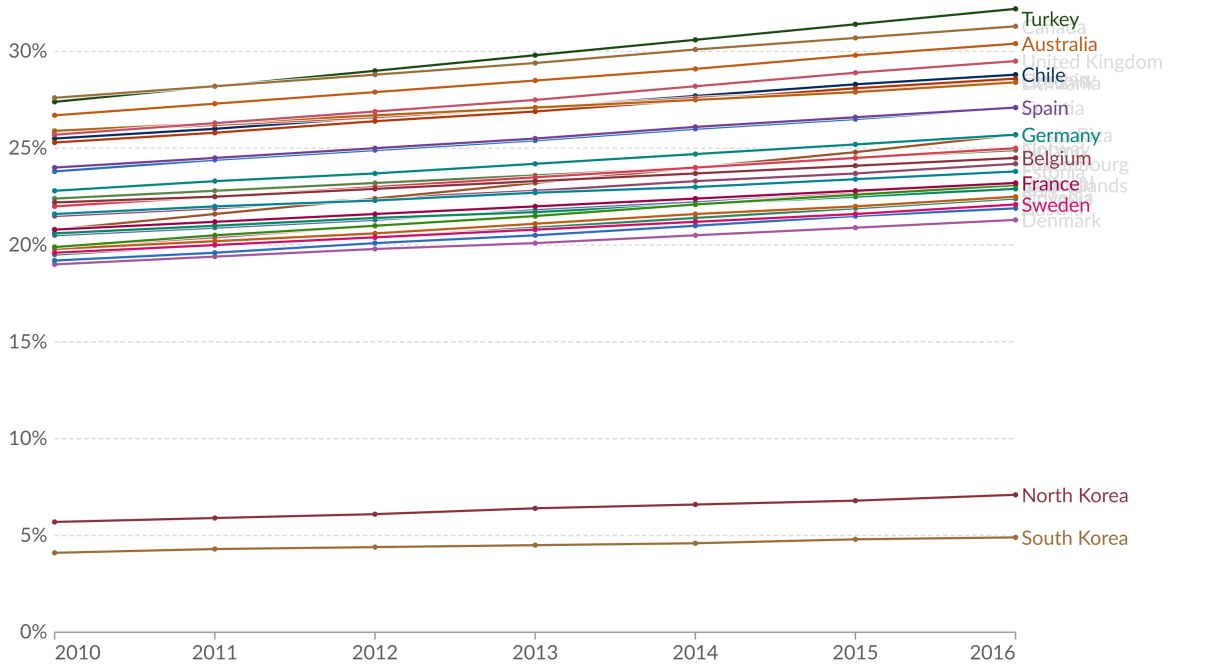
Tramite tali rappresentazioni possiamo visualizzare l'andamento del consumo di farmaci per il diabete dal 2012 al 2021 in ciascun paese. In particolare è evidente che in alcuni paesi come Austria, Australia, Italia, Lussemburgo e Paesi Bassi i valori sono quasi **costanti**, quindi non ci sono state variazioni significative. In quasi tutti i paesi e in particolare in Canada e Turchia, invece, c'è stato un **incremento** importante nel 2019. Tramite alcune ricerche ho scoperto che nel 2019 il diabete è stata la nona causa di morte nel mondo.
Una delle cause dell'aumento di malati di diabete può essere l'**obesità**, in particolare del diabete di tipo 2. Per studiare se è presente un'effettiva correlazione tra tale causa e l'aumento del diabete, ho ricercato dei dataset e analisi statistiche che riportassero il tasso di obesità nei paesi oggetti di studio. L'obesità è definita come un eccesso di grasso corporeo che può portare a una serie di problemi di salute, tra cui l'insulino-resistenza e il diabete di tipo 2. Quando una persona è obesa, il tessuto adiposo in eccesso può interferire con il modo in cui il corpo utilizza l'insulina, l'ormone che regola il livello di zucchero nel sangue. Questa condizione, chiamata resistenza all'insulina, può portare a un aumento dei livelli di zucchero nel sangue e, nel tempo, può svilupparsi il diabete di tipo 2. La misura più utilizzata si basa sull'indice di massa corporea (BMI), che è un numero unico che valuta il peso di un individuo in relazione all'altezza (peso/altezza^2, con peso

in chilogrammi e altezza in metri). In base alla classificazione dell'OMS, gli adulti con un BMI compreso tra 25 e 30 sono definiti sovrappeso, mentre quelli con un BMI pari o superiore a 30 sono obesi. Questo indicatore viene presentato sia per i dati "auto-riportati" (stime di altezza e peso provenienti da interviste sanitarie basate sulla popolazione) sia per dati "misurati" (stime precise di altezza e peso da esami sanitari) ed è misurato come percentuale del popolazione di età pari o superiore a 15 anni.
Il grafico seguente è estratto da https://data.oecd.org/healthrisk/overweight-or-obese-population.htm e i paesi colorati rappresentano i paesi oggetti di studio della nostra analisi statistica. In particolare sono riportati i valori degli ultimi dati disponibili (2019-2022).



Possiamo subito notare che non c'è corrispondenza nei valori del Cile, in quanto non è un paese con alto consumo di farmaci per il diabete ma risulta essere tra i paesi con un tasso di obesità più alto (67.7 nel 2021). Ancora la Francia nel 2019 risulta essere tra i paesi con un alto tasso di consumo di farmaci per il diabete, mentre presenta un basso indice di obesità (45.3); stesso discorso vale per la Corea (36.7). Per quanto riguarda gli altri paesi, si ha corrispondenza tra i due valori studiati.

Dal sito https://ourworldindata.org/obesity, invece, si è estratta la prevalenza stimata dell'obesità, basata su indagini sulla popolazione generale e modelli statistici dal 2010 al 2016.



1. Obesity: Obesity is defined as having a body-mass index (BMI) above 30. A person's BMI is calculated as their weight (in kilograms) divided by their height (in meters) squared. For example, someone measuring 1.60 meters and weighing 64 kilograms has a BMI of 64 / 1.6^2 = 25. Obesity increases the mortality risk of many conditions, including cardiovascular disease, gastrointestinal disorders, type 2 diabetes, joint and muscular disorders, respiratory problems, and psychological issues.

2. Risk factor: A risk factor is a condition or behavior that increases the likelihood of developing a given disease or injury, or an outcome such as death. The impact of a risk factor is estimated in different ways. For example, a common approach is to estimate the number of deaths that would occur if the risk factor was absent. Risk factors are not mutually exclusive: people can be exposed to multiple risk factors, which contribute to their disease or death. Because of this, the number of deaths caused by each risk factor is typically estimated separately. Read more: How do researchers estimate the death toll caused by each risk factor, whether it's smoking, obesity or air pollution? Read more: Why isn't it possible to sum up the death toll from different risk factors?

Possiamo notare che in questo caso c'è stato un generale trend in aumento, proprio come per il consumo di farmaci per il diabete .

Statistica descrittiva

La **statistica descrittiva** è costituita da un insieme di metodi di natura logica e matematica atti a raccogliere, elaborare, analizzare ed interpretare dati allo scopo di descrivere fenomeni collettivi e di estendere la descrizione di certi fenomeni osservati ad altri fenomeni dello stesso tipo non ancora osservati. E' utilizzata per analizzare il comportamento dei fenomeni oggetto di studio. Prima di iniziare un'elaborazione dei dati è necessario avere informazioni generali sul fenomeno tramite l'analisi di misure di sintesi. In R per calcolare le statistiche descrittive di tutte le variabili numeriche contenute nel data frame si usa la funzione *descr(dfDiabete).*

```
## Descriptive Statistics
## dfDiabete
## N: 28
##
##                     Anno2012   Anno2013   Anno2014   Anno2015   Anno2016   Anno2017   Anno2018
## ----------------- ---------- ---------- ---------- ---------- ---------- ---------- ----------
```

```
##              Mean      60.96      63.16      64.56      65.74      67.34      68.11      69.45
##           Std.Dev      14.08      12.98      12.80      12.72      13.43      13.62      13.06
##               Min      30.70      40.60      42.30      44.30      46.40      39.80      46.30
##                Q1      51.80      55.10      56.90      57.90      58.15      59.65      60.40
##            Median      60.75      62.60      62.55      64.55      67.05      67.40      69.80
##                Q3      72.30      74.10      74.60      75.50      77.45      78.95      77.45
##               Max      84.80      85.90      88.10      89.80      92.30      91.70      95.80
##               MAD      14.83      13.71      14.08      11.56      14.53      14.90      11.34
##               IQR      18.95      18.25      17.30      17.25      18.00      17.15      15.42
##                CV       0.23       0.21       0.20       0.19       0.20       0.20       0.19
##          Skewness      -0.14       0.10       0.15       0.19       0.06      -0.16       0.00
##       SE.Skewness       0.44       0.44       0.44       0.44       0.44       0.44       0.44
##          Kurtosis      -0.71      -1.05      -1.00      -0.91      -1.12      -0.94      -0.85
##           N.Valid      28.00      28.00      28.00      28.00      28.00      28.00      28.00
##         Pct.Valid     100.00     100.00     100.00     100.00     100.00     100.00     100.00
##
## Table: Table continues below
##
##
##
##                     Anno2019   Anno2020   Anno2021
## ----------------- ---------- ---------- ----------
##              Mean      71.71      75.51      78.53
##           Std.Dev      13.59      13.99      16.12
##               Min      47.50      49.40      49.20
##                Q1      61.40      64.65      66.45
##            Median      74.20      77.15      78.10
##                Q3      79.35      83.20      87.30
##               Max      99.90     104.90     124.50
##               MAD      13.57      13.94      16.09
##               IQR      16.52      17.53      20.53
##                CV       0.19       0.19       0.21
##          Skewness      -0.08       0.13       0.64
##       SE.Skewness       0.44       0.44       0.44
##          Kurtosis      -0.78      -0.60       0.56
##           N.Valid      28.00      28.00      28.00
##         Pct.Valid     100.00     100.00     100.00
```

Tale funzione mostra in output:

-**media campionaria:** la media aritmetica del campione;

-**deviazione standard:** la radice quadrata della varianza campionari

-**min:** il valore minimo assunto;

-**Q1**: è il **primo quartile**, ovvero il minimo valore osservato la cui funzione di distribuzione empirica supera 0.25;

-**mediana:** indica un punto centrale intorno al quale si dispongono lo stesso numero di valori sia a destra che a sinistra;

-**Q3: terzo quartile**, ovvero il minimo valore osservato la cui funzione di distribuzione empirica supera 0.75;

-**Max:** il valore massimo assunto;

-**deviazione mediana assoluta:** misura la dispersione statistica di un campione;

-**IQR: scarto interquartile**, la differenza tra il terzo e il primo quartile;

-**coefficiente di variazione:** quantifica quanto è grande il valore di una deviazione standard rispetto alla media;

-**skewness:** è la proprietà di una distribuzione di non poter essere suddivisa in due parti specularmente uguali;

-**curtosi:** indica la forma della distribuzione dei dati.

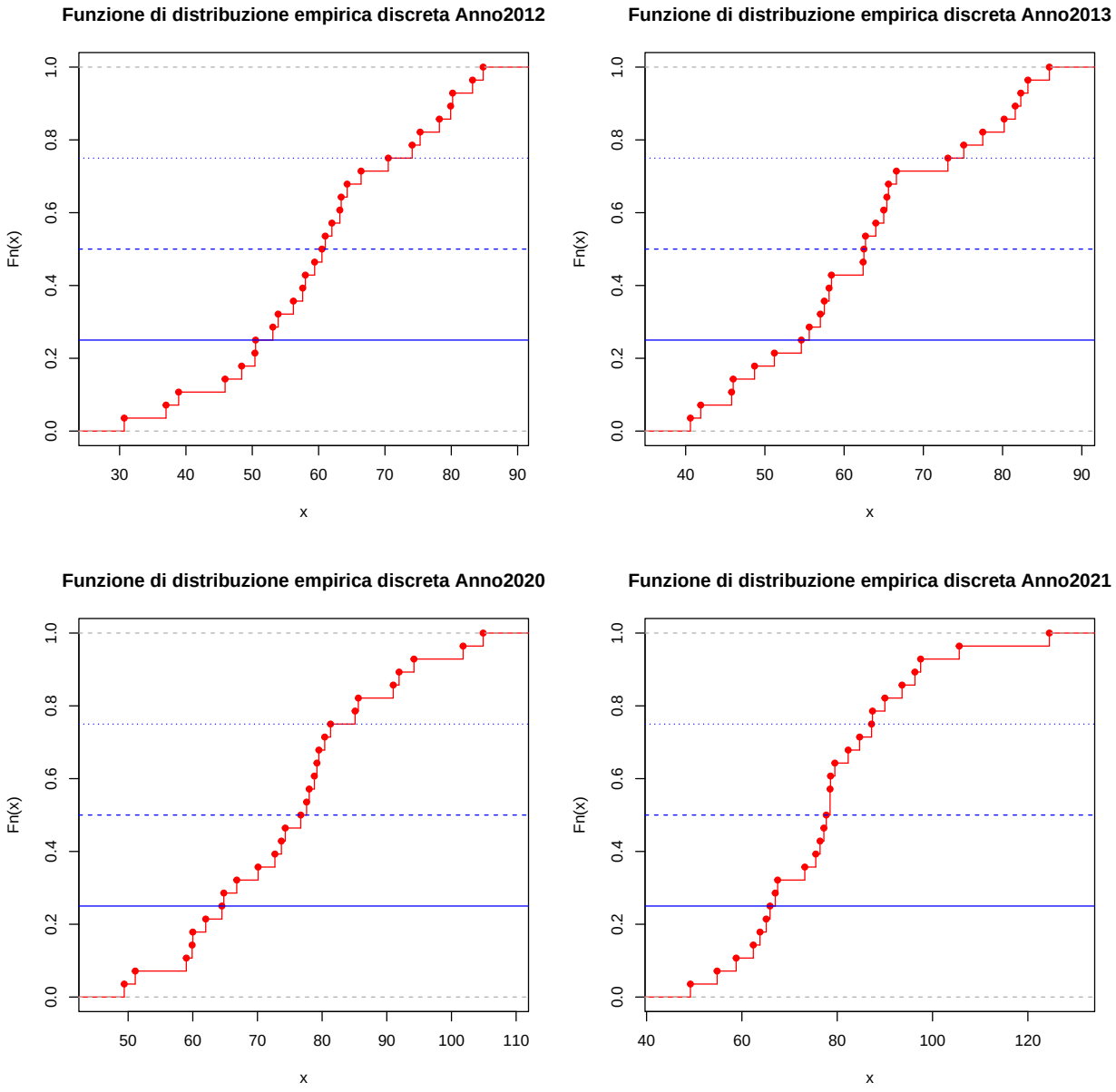
Statistica descrittiva univariata

Funzione di distribuzione

Il dataset è composto da soli fenomeni quantitativi ed è quindi utile definire la **funzione di distribuzione empirica**. Nel caso *discreto* questa funzione è definita a partire dalle frequenze relative cumulate. Consideriamo una variabile quantitativa e indichiamo con $z_1, z_2, ..., z_k$ i valori distinti da essa assunti e assumiamo che essi siano ordinati in ordine crescente. Denotiamo con n_1 il numero di volte in cui ciascun valore z_i è presente nel campione, ossia la frequenza assoluta con cui esso appare nel campione, e con $f_i = n_i/n$ le frequenze relative. Le frequenze relative cumulative sono $F_i = f_1 + f_2 + ... + f_i = (n_1 + n_2 + ...n_i)/n (i = 1, 2, ...k)$ dove F_i rappresenta la proporzione dei dati del campione minori o uguali di z_i. La funzione di distribuzione empirica $F(x)$ è così definita:

$$F(x) = \frac{\#\{x_i \leq x, i = 1, 2, ..., n\}}{n} = \begin{matrix} 0, x < z_1 \\ f_1, z_1 \leq x < z_2 \\ ... \\ f_i, z_i \leq x < z_{i+1} \\ ... \\ 1, x \geq z_k \end{matrix}$$

La funzione di distribuzione empirica è una funzione non decrescente, assume il valore a sinistra in corrispondenza ad ogni punto di salto, vale 0 per ogni valore minore dell'osservazione minima e vale 1 per ogni valore maggiore o uguale dell'osservazione massima. Tale funzione ci permette di capire *come si distribuiscono i dati osservati nel fenomeno*, in particolare come si comportano le sue caratteristiche. Viene definita a partire dalle frequenze relative cumulate. Ho deciso di studiare la funzione di distribuzione empirica relativa agli anni 2012, 2013, 2020, 2021. In R è possibile utilizzare la funzione *ecdf()* per disegnare il grafico della funzione di distribuzione empirica per le variabili quantitative discrete.



Indici di sintesi

Gli **indici di sintesi**, detti anche statistiche, sono utili a descrivere dei dati numerici. Si dividono in *misure di centralità* e *dispersione dei dati*. Le prime sono la media, mediana e moda, mentre le seconde sono la varianza e la deviazione standard. Tali indici sono riportati di seguito con i rispettivi codici in R per gli anni 2012, 2013, 2020, 2021. Possiamo visualizzare gli indici di sintesi dei dati oggetto di studio mediante la funzione *summary()*.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.70   52.45   60.75   60.96   71.40   84.80
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   40.60   55.35   62.60   63.16   73.60   85.90
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   49.40   64.72   77.15   75.51   82.25  104.90
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   49.20   66.72   78.10   78.53   87.25  124.50
```

La **media** campionaria è la media artimetica di un campione di valori numerici di ampiezza n.

```
## [1] "media anno 2012"
```

```
## [1] 60.96429
```

```
## [1] "media anno 2013"
```

```
## [1] 63.16071
```

```
## [1] "media anno 2020"
```

```
## [1] 75.51071
```

```
## [1] "media anno 2021"
```

```
## [1] 78.525
```

La **mediana** campionaria, dato un insieme di dati di ampiezza n ordinati in ordine crescente, è il valore in posizione $(n+1)/2$ se n è dispari, altrimenti è la media aritmetica dei valori che occupano le posizioni $n/2$ e $n/2+1$.

```
## [1] "mediana anno 2012"
```

```
## [1] 60.75
```

```
## [1] "mediana anno 2013"
```

```
## [1] 62.6
```

```
## [1] "mediana anno 2020"
```

```
## [1] 77.15
```

```
## [1] "mediana anno 2021"
```

```
## [1] 78.1
```

La **moda** campionaria di un insieme di dati è la modalità a cui è associata la frequenza più elevata. Se esistono più modalità con frequenza massima, ciascuna di esse è detto valore modale. Ha alcune proprietà molto importanti quali:

- è possibile identificare la moda per qualsiasi tipo di variabile;

- indica sempre un valore realmente osservato nel campione;

- non è influenzata da valori esterni;

- nel caso di distribuzioni di frequenze molto asimmetriche, la moda è il miglior indice per descrivere la tendenza centrale di un cmapione.

```
## [1] "moda anno 2012"
```

```
##
## 30.7   37 38.9 45.9 48.4 50.4 50.5 53.1 53.9 56.2 57.6   58 59.4 60.5   61   62
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 63.2 63.4 64.3 66.4 70.5 74.1 75.3 78.2 79.9 80.2 83.2 84.8
##    1    1    1    1    1    1    1    1    1    1    1    1
```

```
## [1] "moda anno 2013"
```

```
##
## 40.6 41.9 45.8   46 48.7 51.2 54.6 55.6   57 57.5 58.1 58.4 62.4 62.5 62.7   64
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
##   65 65.4 65.6 66.6 73.1 75.1 77.5 80.2 81.6 82.3 83.2 85.9
##    1    1    1    1    1    1    1    1    1    1    1    1
```

```
## [1] "moda anno 2020"
```

```
##
##  49.4  51.1    59  59.9    60    62  64.5  64.8  66.8  70.1  72.7  73.7  74.3
##     1     1     1     1     1     1     1     1     1     1     1     1     1
##  76.7  77.6    78  78.8  79.2  79.5  80.4  81.3  85.1  85.6    91  91.9  94.2
##     1     1     1     1     1     1     1     1     1     1     1     1     1
## 101.8 104.9
##     1     1
```

```
## [1] "moda anno 2021"
```

```
##
##  49.2  54.8  58.8  62.4  63.8  65.1  65.9    67  67.5  73.2  75.5  76.4  77.2
##     1     1     1     1     1     1     1     1     1     1     1     1     1
##  77.7  78.5  78.6  79.5  82.3  84.7  87.2  87.4    90  93.6  96.3  97.5 105.6
##     1     2     1     1     1     1     1     1     1     1     1     1     1
## 124.5
##     1
```

La **varianza** fornisce una misura della variabilità dei valori assunti dalla variabile stessa; nello specifico, la misura di quanto essi si discostino quadraticamente rispettivamente dalla media aritmetica o dal valore atteso.

```
## [1] "varianza anno 2012"
```

```
## [1] 198.1498
```

```
## [1] "varianza anno 2013"
```

```
## [1] 168.4499
```

```
## [1] "varianza anno 2020"
```

```
## [1] 195.8143
```

```
## [1] "varianza anno 2021"
```

```
## [1] 259.8612
```

La **deviazione standard** campionaria è la radice quadrata della varianza campionaria.

```
## [1] "deviazione standard anno 2012"
```

```
## [1] 14.07657
```

```
## [1] "deviazione standard anno 2013"
```

```
## [1] 12.97882
```

```
## [1] "deviazione standard anno 2020"
```

```
## [1] 13.99337
```

```
## [1] "deviazione standard anno 2021"
```

```
## [1] 16.12021
```

Il **coefficiente di variazione** è il rapporto tra la deviazione standard campionaria e il modulo della media campionaria. In R non esiste una funzione per calcolare tale valore ma dobbiamo definirla:

$cv <- function(x) sd(x)/abs(mean(x))$

```
## [1] "coefficiente di variazione anno 2012"
```

```
## [1] 0.2308986
```

```
## [1] "coefficiente di variazione anno 2013"
```

```
## [1] 0.2054889
```

```
## [1] "coefficiente di variazione anno 2020"
```

```
## [1] 0.1853163
```

```
## [1] "coefficiente di variazione anno 2021"
```

```
## [1] 0.2052876
```

Sulla base dei risultati ottenuti, possiamo notare che è presente la moda solo per l'anno 2021 ed è 78.5; ciò significa che negli altri casi non c'è alcun valore con frequenza superiore agli altri. Per descrivere la **forma** di una distribuzione si possono confrontare la media campionaria e la mediana campionaria. Se queste due misure sono uguali, la distribuzione di frequenze tende ad essere simmetrica; se la media campionaria è sensibilmente maggiore della mediana campionaria, la distribuzione di frequenze è più sbilanciata verso destra; se invece la media campionaria è sensibilmente minore della mediana campionaria la distribuzione di frequenze è più sbilanciata verso sinistra. Nel nostro caso i due valori sono molto simili, quindi la distribuzione di frequenze tende ad essere *simmetrica*. Inoltre, il coefficiente di variazione risulta essere minore di 0.5, ciò indica che la variabilità dei dati è contenuta e quindi la *media* può essere considerata un *buon indicatore*. E' possibile calcolare anche la **mediana per una distribuzione di frequenze** che rappresenta la modalità i-esima (i = 1,2,. . . ,k) che soddisfa la doppia disuguaglianza $F_{i-1} < 0.5, F_i \geq 0.5$. Quindi tale mediana è un valore di sintesi che indica un punto centrale intorno al quale si dispone la distribuzione di frequenze.

```
## [1] "mediana per la distribuzione di frequenze 2012"
```

```
##
##   (30,60]   (60,80]  (80,100] (100,125]
##      0.46      0.43      0.11      0.00
```

```
## [1] "mediana per la distribuzione di frequenze 2013"
```

```
##
##   (30,60]   (60,80]  (80,100] (100,125]
##      0.43      0.39      0.18      0.00
```

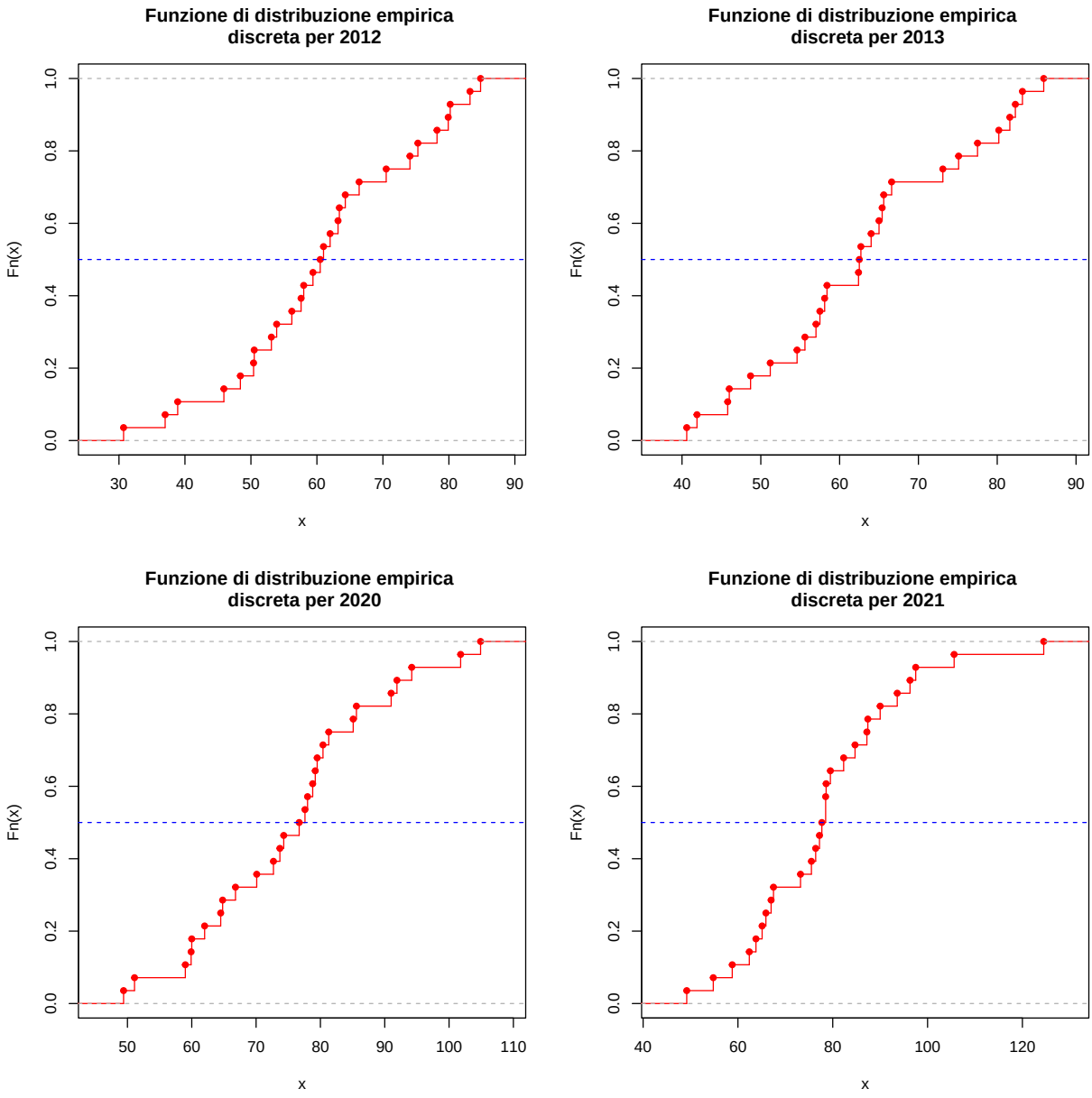
```
## [1] "mediana per la distribuzione di frequenze 2020"
```

```
##
##   (30,60]   (60,80]  (80,100] (100,125]
##      0.18      0.50      0.25      0.07
```

```
## [1] "mediana per la distribuzione di frequenze 2021"
```

```
##
##   (30,60]   (60,80]  (80,100] (100,125]
##      0.11      0.54      0.29      0.07
```

La mediana di una distirbuzione di frequenze può essere individuata graficamente a partire dalla funzione di distribuzione empirica discreta. Si traccia la funzione di distribuzione empirica e sull'asse delle ordinate si individua il punto 0.5 e da questo si traccia una linea orizzontale. Il minimo valore osservato sulle ascisse la cui funzione di distribuzione empirica supera 0.5 è proprio la mediana.



Forma di una distribuzione di frequenza

Le media e la mediana, quindi, sono indici utili per comprendere la forma delle distribuzioni di frequenze. Esistono degli indici statistici che permettono di misurare quando una distribuzione di frequenza presenta simmetria o asimmetria o se è più o meno piccata. Tali indici sono la *simmetria* e la *curtosi*.

L'**indice di skewness** permette di misurare la simmetria di una distribuzione di frequenze. Assegnato un insieme di dati numerisi $x_1, x_2, \ldots x_n$ si definisce skewness campionaria il valore $\gamma = \frac{m_3}{m_2^{3/2}}$ dove m_3 denota il momento centrato campionario di ordine 3.
Per calcolare tale valore si deve definire la seguente funzione in R:

```
skw<-function(x){ n<- length(x) m2 <- (n-1)*var(x)/n  m3 <- (sum((x-mean(x))^3))/n m3/(m2^1.5)}
```

```
## [1] "indice di skewness anno 2012"
```

```
## [1] -0.1500723
```

```
## [1] "indice di skewness anno 2013"
## [1] 0.1019673
## [1] "indice di skewness anno 2020"
## [1] 0.1422174
## [1] "indice di skewness anno 2021"
## [1] 0.6746057
```

Dai risultati possiamo notare che i valori riferiti all'anno 2012 presentano un'asimmetria negativa in quanto il valore calcolato γ è minore di 0, mentre i valori riferiti agli anni 2013, 2020 e 2021 presentano un'asimmetria positiva.

L'**indice di curtosi** permette di misurare la densità dei dati intorno alla media. Assegnato un insieme di dati numerici $x_1, x_2, \ldots, x_n$, si definisce curtosi campionaria il valore $\gamma_2 = \beta_2 - 3$ dove $\beta_2 = \frac{m_4}{m_2^2}$ è l'indice di Pearson.
Gli indici γ_2 e β_2 permettono di confrontare la distribuzione di frequenze dei dati con una densità di probabilità normale standard, caratterizzata da $\beta_2 = 3$ e indice di curtosi $\gamma_2 = 0$.
Anche in questo caso si deve definire la seguente funzione:

```
curt<-function(x){n<-length(x)m2<-(n-1)*var(x)/nm4<-(sum((x-mean(x))^4))/nm4/(m2^2)-3}
```

```
## [1] "indice di curtosi anno 2012"
## [1] -0.5388155
## [1] "indice di curtosi anno 2013"
## [1] -0.8982696
## [1] "indice di curtosi anno 2020"
## [1] -0.419736
## [1] "indice di curtosi anno 2021"
## [1] 0.8282838
```

Dai risultati si evince che per gli anni 2012, 2013 e 2021 avremo una distribuzione di frequenze **platicurtica** in quanto $\beta_2 < 0$, ossia la distribuzione di frequenze è più piatta di una normale; per l'anno 2022 invece avremo una distribuzione di frequenze **leptocurtica** poichè $\beta_2 > 0$, ossia la distribuzione di frequenze è più piccata di una normale.

Statistica descrittiva bivariata

La *statistica bivariata* è il ramo della statistica che si occupa dei metodi grafici e statistici per descrivere le relazioni che intercorrono tra due variabili.

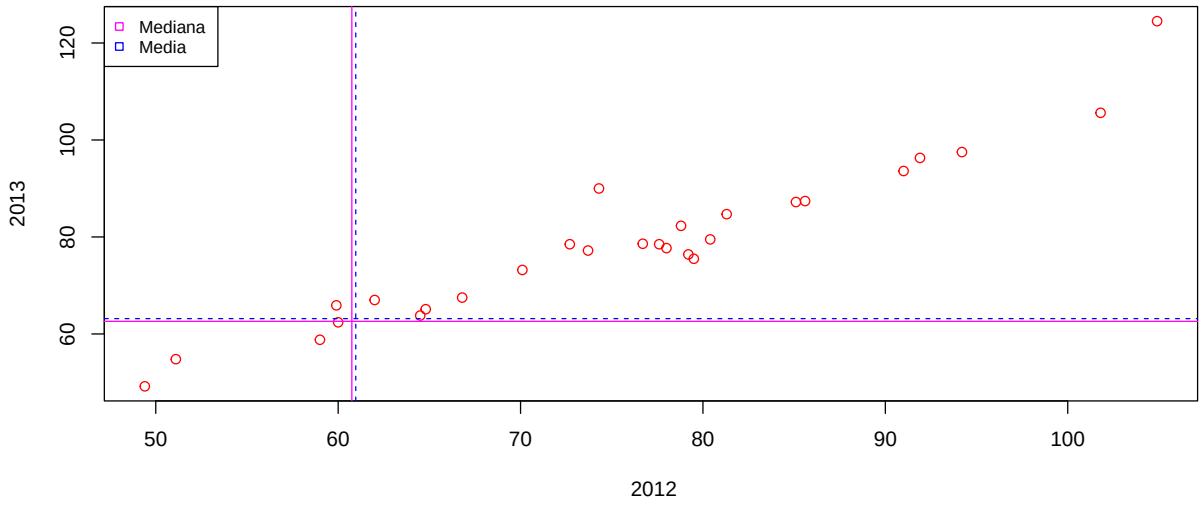
Scatterplot

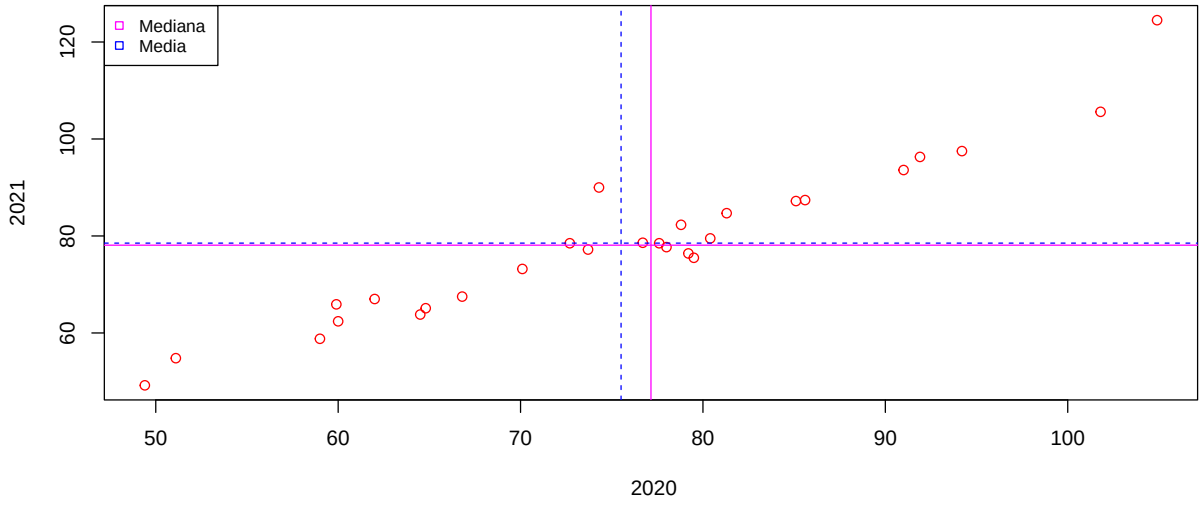
Le relazioni tra variabili quantitative possono essere rappresentate graficamente mediante **diagrammi di dispersione (scatterplot)** in cui ogni coppia di osservazioni viene rappresentata sotto forma di un punto o di un cerchietto in un piano euclideo, il risultato finale è una nuvola di punti. Il grafico ottenuto mostra se esiste una relazione tra le variabili e di quale tipo è tale relazione. Calcolo gli indici statistici di posizione e di dispersione relativi alle variabili in esame, nel mio caso alle variabili 2012, 2013, 2020 e 2021 che voglio confrontare.

	2012	2013	2020	2021
Mediana campionaria	60.75	62.6	77.15	78.1
Media campionaria	60.96	63.16	75.51	78.52

	2012	2013	2020	2021
Deviazione standard	14.07	12.97	13.99	16.12

Si nota che i valori di media e mediana aumentano con l'avanzare degli anni; la deviazione standard, invece, assume un valore alto già nel 2012.
Successivamente si realizza lo scatterplot considerando Anno2020 come variabile indipendente e Anno2021 come variabile dipendente (stesso discorso per la coppia Anno2012-Anno2013); nello scatterplot sono visualizzate le coppie del data frame. Sono tracciate anche delle linee orizzontali e verticali in corrispondenza delle mediane campionarie e delle medie campionarie dei vettori Anno 2012, Anno 2013, Anno2020 e Anno2021.





Si può notare che, in entrambi i casi, i dati (punti) sembrano posizionati intorno ad una retta ascendente e ciò induce a pensare che esista una **correlazione lineare positiva** tra le variabili, ovvero i valori delle due variabili tendono ad aumentare in parallelo.

Quando si osservano più variabili quantitative per uno stesso gruppo è necessario vedere se esiste una correlazione tra le variabili. Per ottenere una misura quantitativa della correlazione tra le variabili si considera la **covarianza campionaria** che può avere segno positivo, negativo o nullo. La covarianza campionaria tra due variabili X e Y è così definita : $C_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})(n = 2, 3, ...)$.

```
## [1] "covarianza campionaria 2012-2013"
```

```
## [1] 178.1319
```

```
## [1] "covarianza campionaria 2020-2021"
```

```
## [1] 216.256
```

Nei casi in esame entrambi i valori sono positivi e quindi le variabili sono *correlate positivamente*.

Si può usare anche il **coefficiente di correlazione campionario** che è un indice adimensionale, non fa distinzione tra variabile dipendente e indipendente, può essere calcolato solo se entrambe le variabili sono quantitative, non cambia al variare dell'unità di misura delle variabili ed è fortemente influenzato dalla presenza di eventuali valori anomali.

```
## [1] "coefficiente di correlazione campionario 2012-2013"
```

```
## [1] 0.9750109
```

```
## [1] "coefficiente di correlazione campionario 2020-2021"
```

```
## [1] 0.9586835
```

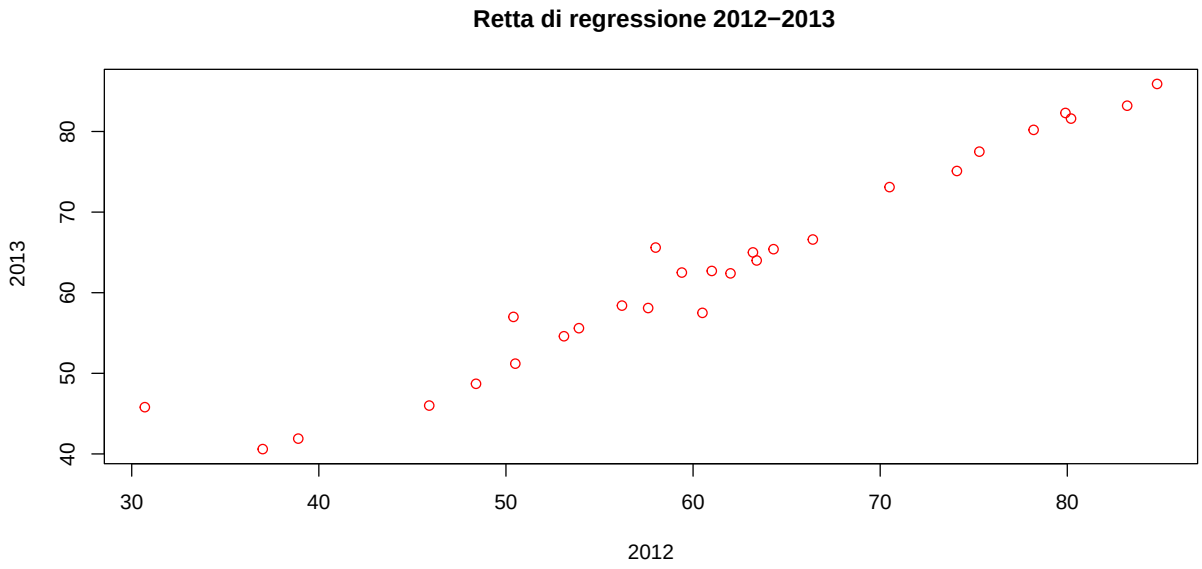
I coefficienti di correlazione sono prossimi all'unità e ciò indica che esiste una *forte correlazione* tra i dati di Anno2012-Anno2013 e Anno2020-Anno2021.

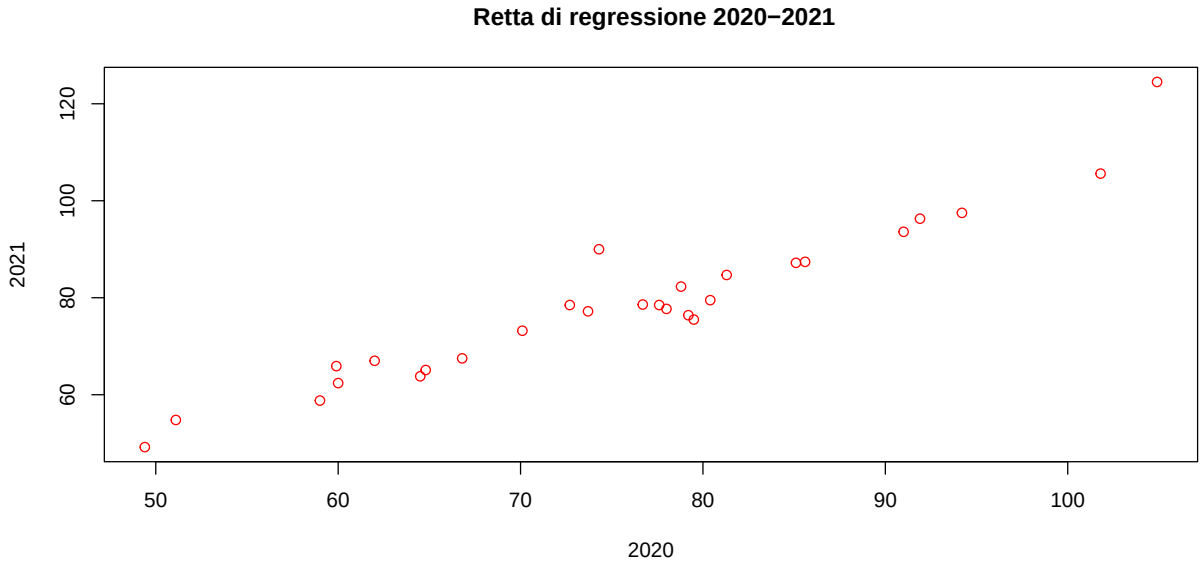
Regressione lineare semplice

Il **modello di regressione lineare semplice** è esprimibile attraverso l'equazione di una retta che riesce ad interpolare la nuvola di punti dello scatterplot meglio di tutte e altre possibili rette. Per effettuare l'analisi di regressione lineare si utilizza la funzione $lm(y \tilde{} x)$ che fornisce i valori dell'intercetta e del coefficiente angolare.

```
##
## Call:
## lm(formula = dfDiabete$Anno2013 ~ dfDiabete$Anno2012)
##
## Coefficients:
##        (Intercept)  dfDiabete$Anno2012
##              8.355               0.899
```

```
##
## Call:
## lm(formula = dfDiabete$Anno2021 ~ dfDiabete$Anno2020)
##
## Coefficients:
##        (Intercept)  dfDiabete$Anno2020
##             -4.869               1.104
```





L'equazione della retta di regressione è $Y = \alpha + \beta$ X, dove β è il **coefficiente angolare** e esprime quantitativamente la pendenza (inclinazione) della retta e α è **l'intercetta** e corrisponde all'ordinata del punto di intersezione della retta interpolante con l'asse delle ordinate. L'identificazione di questa retta viene ottenuta applicando il *metodo dei minimi quadrati*. Possiamo stimare i parametri α e β a partire dalle medie campionarie, deviazioni standard campionarie e dal coefficiente di correlazione.

```
## [1] "alpha e beta 2012-2013"
```

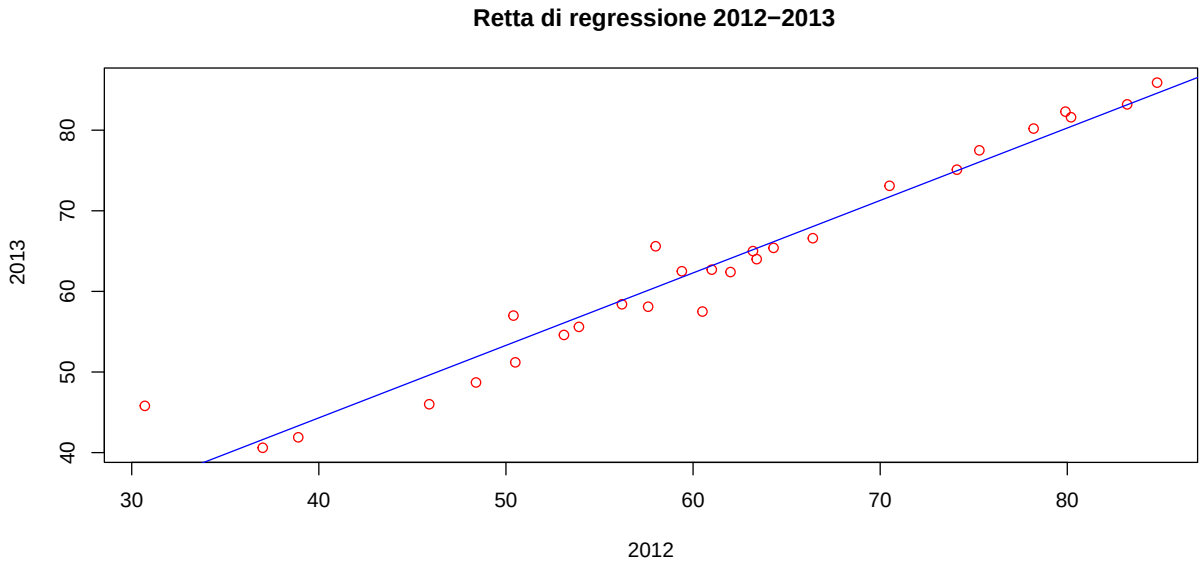
```
## [1] 8.3552925 0.8989759
```

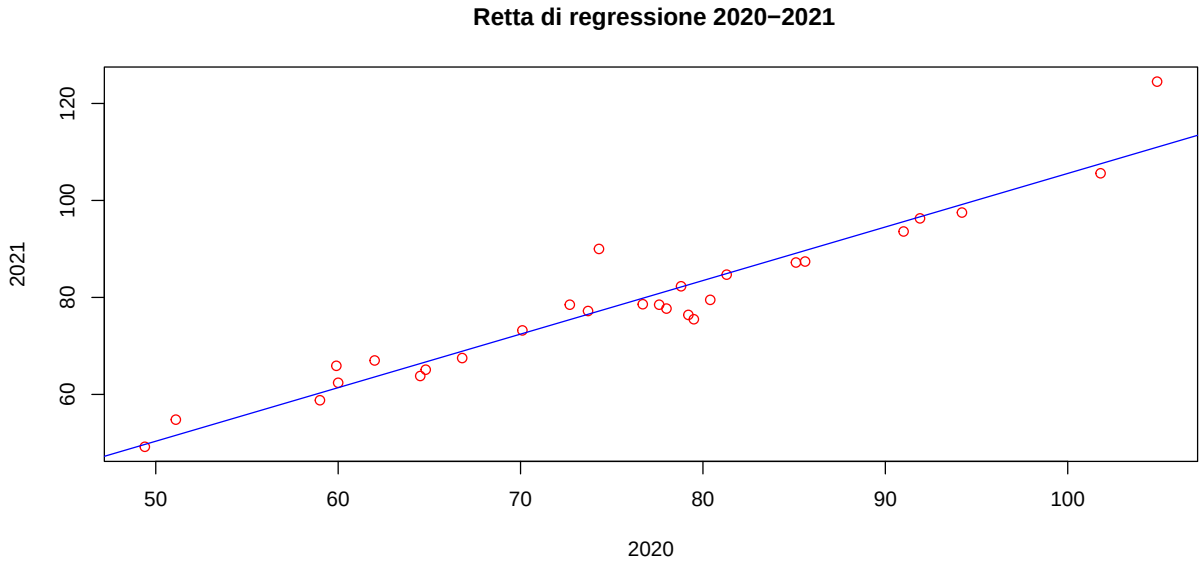
```
## [1] "alpha e beta 2020-2021"
```

```
## [1] -4.868523  1.104393
```

Le rette di regressione risulteranno dunque essere $y = 8.355 + 0.898x$ per gli anni 2012-2013 e $y = -4.868 + 1.104x$ per gli anni 2020-2021. In entrambi i casi il valore di β è positivo ed infatti le nostre rette di regressione

sono crescenti. Le rappresentazioni delle rette calcolate possono essere aggiunte allo scatterplot facendo uso della funzione *abline()*.





Residui Una volta calcolati i valori dei coefficienti α e β e disegnata la retta di regressione che interpola la nuvola dei punti nel corrispondente scatterplot, è possibile osservare quanto la retta di regressione si adatta ai punti che individuano le osservazioni. Esistono degli **scostamenti (residui)** tra le ordinate dei valori osservati e i corrispondenti valori stimati. I *residui* sono così definiti $E_i = y_i - \hat{y}_i = y_i - (\alpha + \beta x_i)(i = 1, 2, ..., n)$. La media campionaria dei residui $\bar{E}$ è nulla, ossia in media gli scostamenti positivi e negativi si compensano. Si determinano i valori stimati e i residui e la media, mediana, varianza e deviazione standard dei residui. In R per calcolare il vettore dei valori stimati $(\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)$ si utilizza la funzione *fitted()* che crea un vettore lungo n che contiene le ordinate sulla retta di regressione.; invece per calcolare il vettore dei residui $(E_1, E_2, \dots E_n)$ si utilizza la funzione *resid()*.

```
## [1] "vettore dei valori stimati 2012-2013"
```

```
##        1        2        3        4        5        6        7        8
## 62.74333 49.61828 65.17057 60.13630 35.95385 53.66368 78.65521 53.75357
##        9       10       11       12       13       14       15       16
## 56.09091 84.58845 80.45316 83.15008 76.04818 43.32545 64.09180 66.15944
##       17       18       19       20       21       22       23       24
## 41.61740 65.35036 74.96940 51.86572 63.19282 60.49589 71.73309 68.04729
##       25       26       27       28
## 56.81009 58.87774 80.18346 61.75446
```

```
## [1] "vettore dei valori stimati 2020-2021"
```

```
##        1        2        3        4        5        6        7        8
##  61.39507  49.68850  81.27415 110.98233  76.52526  82.15766  99.16532  63.60386
##        9       10       11       12       13       14       15       16
##  68.90495 107.55871  89.66754  95.63126  82.59942  51.56597  66.69616  75.42087
##       17       18       19       20       21       22       23       24
##  60.29068  66.36484  80.83239  61.28463  79.83844  83.92469  89.11534  84.91865
##       25       26       27       28
##  72.54944  96.62522  77.18790  82.93074
```

```
## [1] "vettore dei residui 2012-2013"
```

```
##          1          2          3          4          5          6
## -5.24333263 -3.61828494 -0.17056748 -2.03630261  9.84614828  3.33632365
##          7          8          9         10         11         12
##  1.54479448 -2.55357394 -1.49091120  1.31155374  1.14684274  0.04991513
##         13         14         15         16         17         18
##  1.45182450 -1.42545385 -1.69179644 -0.75944094 -1.01739970 -1.35036265
##         19         20         21         22         23         24
##  0.13059554 -3.16572461 -0.49282057  5.10410704  1.36690867 -1.44729026
##         25         26         27         28
## -1.21009189 -0.47773639  2.11653550  0.74554082
```

```
## [1] "vettore dei residui 2020-2021"
```

```
##         1         2         3         4         5         6         7
##  1.0049281 -0.4885035 -3.5741503 13.5176713  0.6747406  0.1423351 -1.6653209
##         8         9        10        11        12        13        14
##  3.3961416 -1.4049460 -1.9587096 -2.2675390 -2.0312625 -6.1994222  3.2340280
##        15        16        17        18        19        20        21
## -1.5961595  3.0791339 -1.4906786 -2.5648415 -2.3323930  4.6153674 -1.2384391
##        22        23        24        25        26        27        28
## -4.4246941 -1.9153424 -0.2186480  0.6505563 -0.3252165 12.8121047 -7.4307402
```

```
## [1] "mediana campionaria dei residui 2012-2013"
```

```
## [1] -0.4852785
```

```
## [1] "varianza campionaria dei residui 2012-2013"
```

```
## [1] 8.313621
```

```
## [1] "deviazione standard campionaria dei residui 2012-2013"
```

```
## [1] 2.883335
```

```
## [1] "mediana campionaria dei residui 2020-2021"
```

```
## [1] -1.321693
```

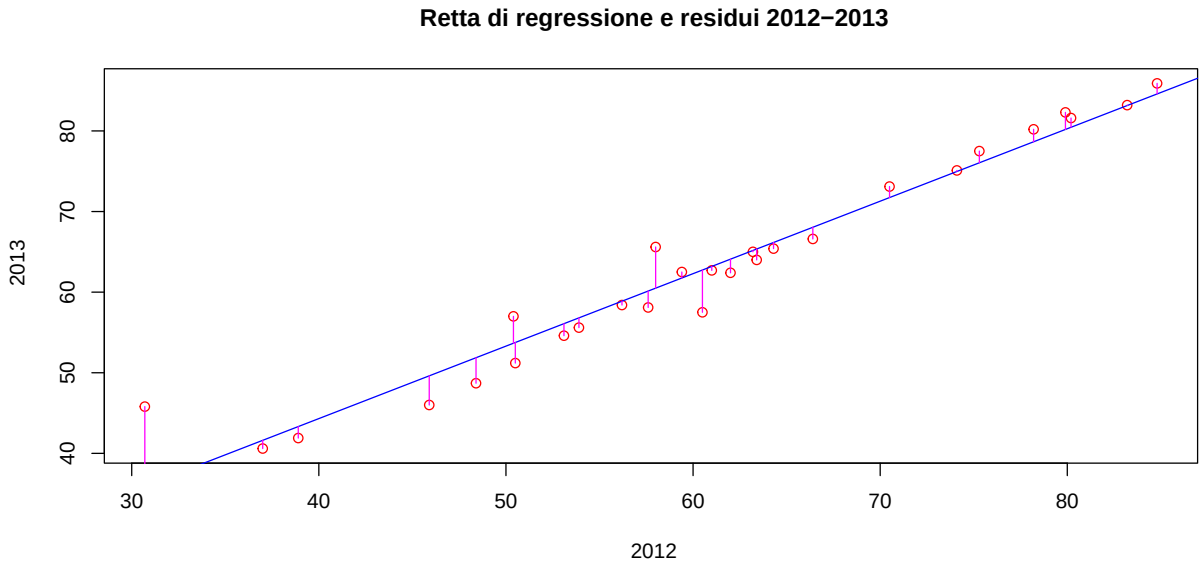
```
## [1] "varianza campionaria dei residui 2020-2021"
```

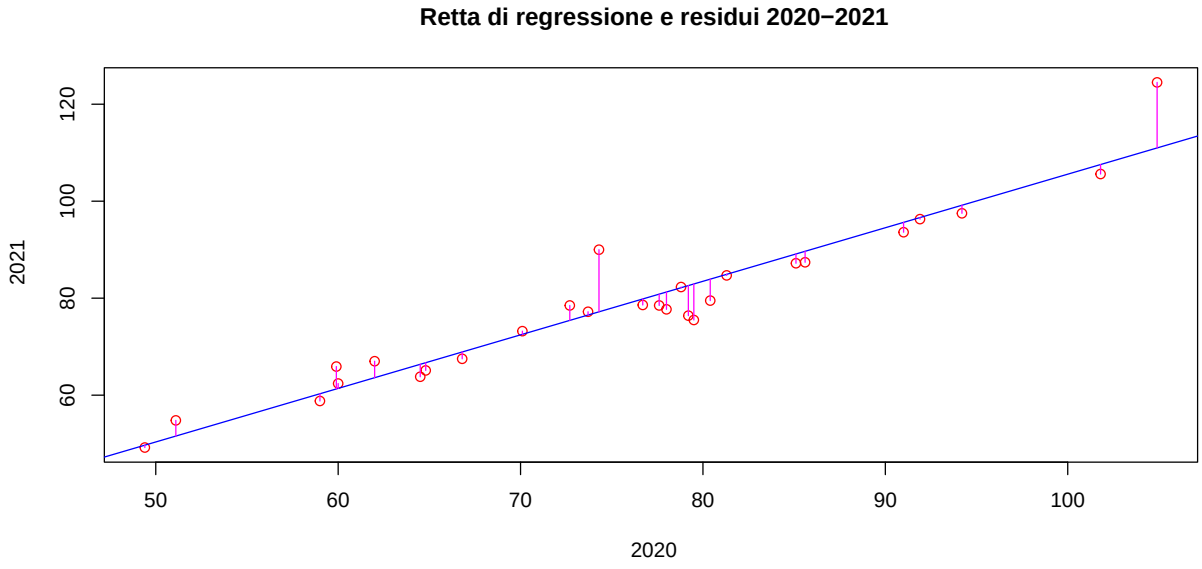
```
## [1] 21.02952
```

```
## [1] "deviazione standard campionaria dei residui 2020-2021"
```

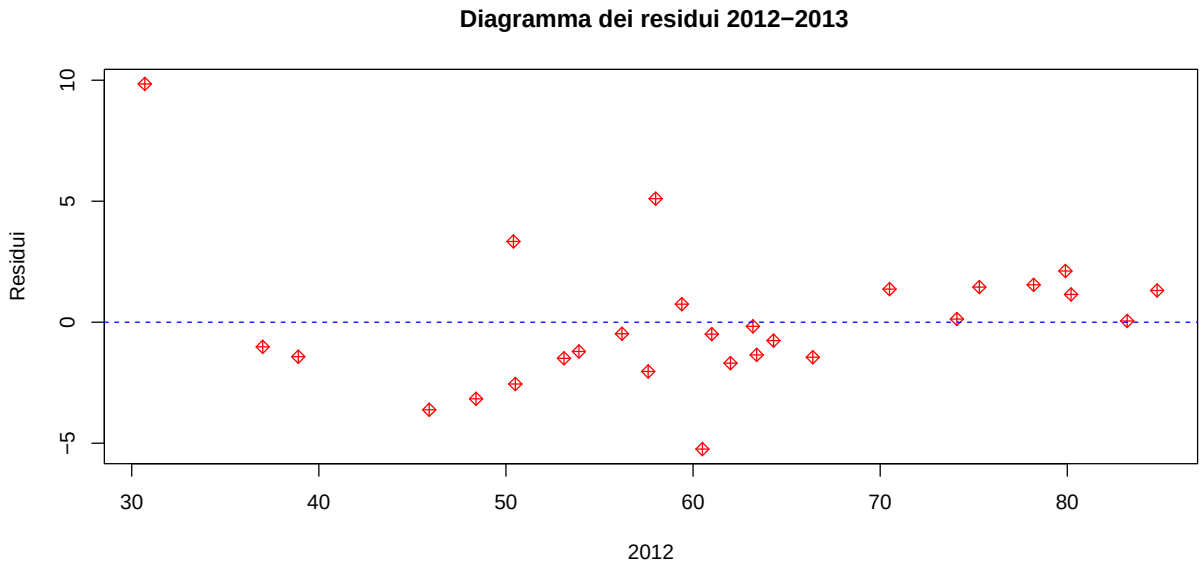
```
## [1] 4.585795
```

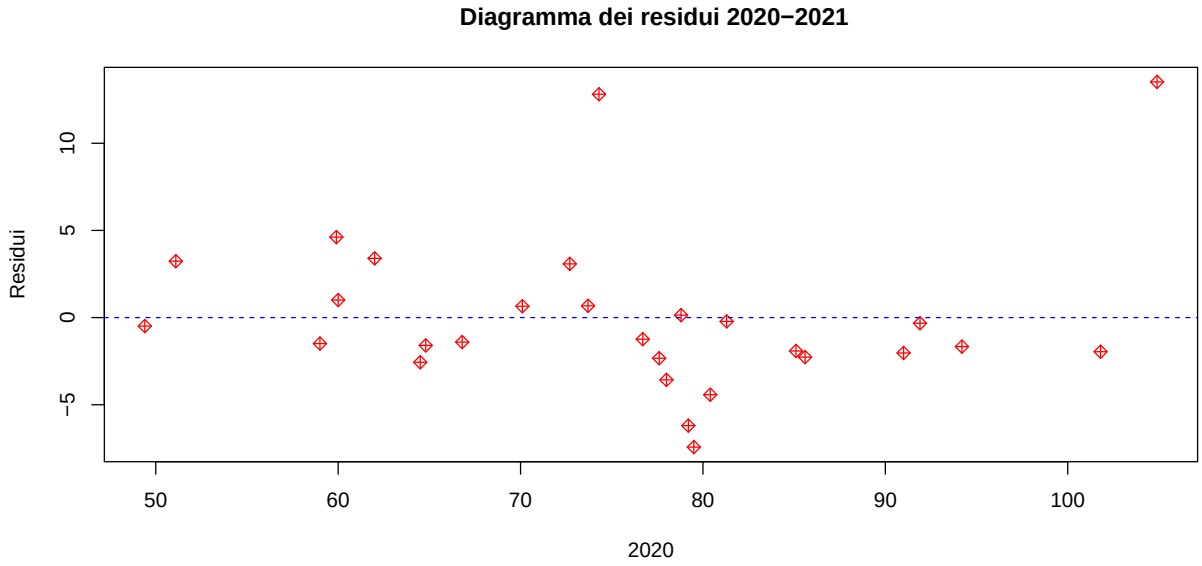
E' possibile rappresentare graficamente i residui aggiungendo allo scatterplot precedente dei segmenti verticali che congiungono i valori stimati e i valori osservati, rappresentando i valori dei residui rispetto alle osservazioni e rappresentando i residui standardizzati rispetto ai valori stimati. Realizziamo i grafici dei residui ottenuti aggiungendo dei segmenti verticali che visualizzano i residui.





Si può studiare in modo più accurato il modo con cui la retta di regressione interpola i dati e come i residui si dispongono intorno alla retta interpolante influenzandone la posizione attraverso il diagramma dei residui. Tale diagramma è un grafico in cui i valori dei residui sono posti sull'asse delle ordinate e quelli della variabile indipendente sull'asse delle ascisse.





I punti indicano la posizione dove si collocano i residui rispetto ai valori attesi. La retta orizzontale è posta nello zero e corrisponde alla media campionaria dei residui. Si nota che i punti sono disposti in modo casuale attorno alla linea orizzontale e non si evidenzia alcun comportamento particolare nella distribuzione dei punti.

E' possibile calcolare il valore dei **residui standardizzati** $E_i^{(s)} = \frac{E_i}{s_E}$ per capire se esiste qualche *relazione tra i residui e i residui standardizzati*. E' possibile realizzare un grafico in cui tali residui standardizzati (ordinate) vengono disegnati in funzione dei valori stimati (ascisse) mediante la retta di regressione. I residui standardizzati sono caratterizzati da media campionaria nulla e varianza unitaria.

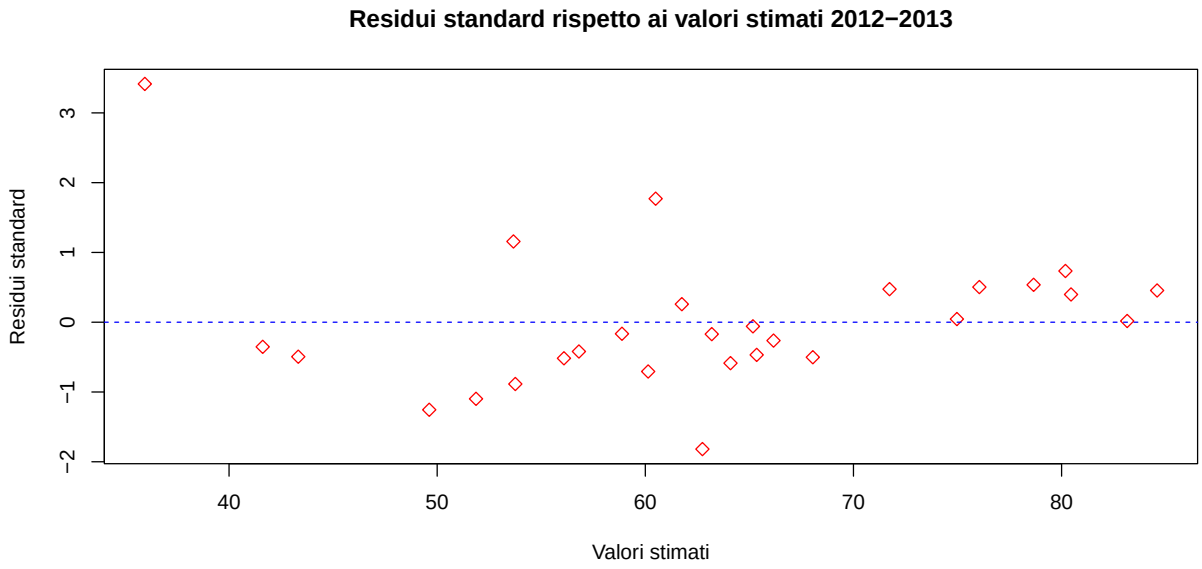
```
## [1] "residui standard 2012-2013"
```

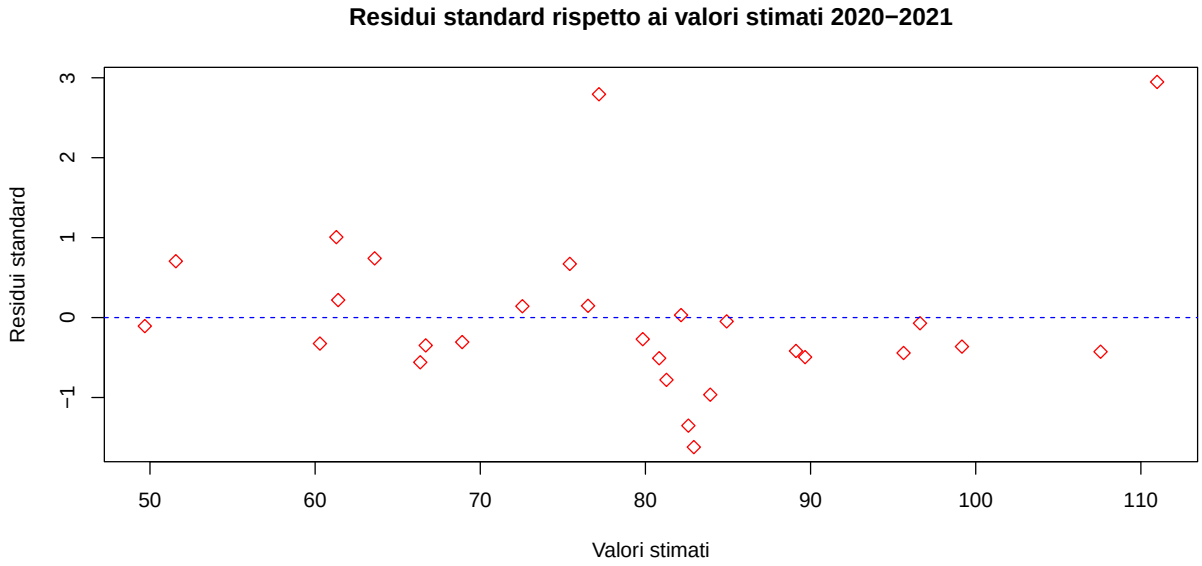
```
##           1           2           3           4           5           6
## -1.81849581 -1.25489578 -0.05915632 -0.70623171  3.41484714  1.15710580
##           7           8           9          10          11          12
##  0.53576656 -0.88563207 -0.51707873  0.45487386  0.39774870  0.01731159
```

```
##          13          14          15          16          17          18
##  0.50352266 -0.49437677 -0.58674987 -0.26338977 -0.35285518 -0.46833360
##          19          20          21          22          23          24
##  0.04529323 -1.09793853 -0.17092033  1.77020951  0.47407210 -0.50195009
##          25          26          27          28
## -0.41968481 -0.16568883  0.73405813  0.25856892
```

```
## [1] "residui standard 2020-2021"
```

```
##           1           2           3           4           5           6
##  0.21913933 -0.10652536 -0.77939598  2.94772680  0.14713711  0.03103825
##           7           8           9          10          11          12
## -0.36314769  0.74057857 -0.30636911 -0.42712540 -0.49447019 -0.44294663
##          13          14          15          16          17          18
## -1.35187508  0.70522731 -0.34806602  0.67145037 -0.32506438 -0.55930137
##          19          20          21          22          23          24
## -0.50861256  1.00644867 -0.27005984 -0.96486955 -0.41766854 -0.04767942
##          25          26          27          28
##  0.14186336 -0.07091823  2.79386763 -1.62038205
```





Anche in questo caso i punti sono disposti casualmente attorno alla linea orizzonatale (media campionaria dei residui standardizzati) e non si evidenzia alcuna tendenza particolare nella distribuzione dei punti.

Coefficiente di determinazione Poichè si è interessati a vedere quanto la retta si adatta ai dati, l'accento può essere posto sul quadrato del coefficiente di correlazione e su quanto esso si avvicini ad uno e quindi che tutti i punti tenderanno ad allinearsi lungo la retta di regressione. Il *coefficiente di determinazione (r-square)* per la regressione lineare semplice è il rapporto tra la varianza dei valori stimati tramite la retta di regressione e la varianza dei valori osservati. Nel caso di regressione lineare semplice, il coefficiente di determinazione coincide con il quadrato del coefficiente di correlazione.

```
## [1] "coefficiente di determinazione 2012-2013"
```

```
## [1] 0.9506463
```

```
## [1] "coefficiente di determinazione 2020-2021"
```

```
## [1] 0.919074
```

Dal momento che i coefficienti di determinazione sono vicini all'unità, allora le rette interpolano bene i nostri dati e quindi tutti i punti tendono ad allinearsi lungo la retta di regressione.

Regressione lineare multipla

Il *modello di regressione lineare multipla* viene utilizzato per spiegare la relazione tra una variabile quantitativa Y, detta **variabile dipendente**, e le **variabili quantitative indipendenti** $X_1, X_2, ..., X_p$. Tale modello è esprimibile attraverso l'equazione: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$ dove α è l'**intercetta** e $\beta_1, \beta_2, ...\beta_p$ sono i **regressori**. In particolare, β_1 rappresenta l'inclinazione di Y rispetto alla variabile X_1 tenendo costanti le variabili $X_1, X_2, ..., X_p$, β_p rappresenta l'inclinazione di Y rispetto alla variabile X_p tenendo costanti le variabili $X_1, X_2, ..., X_(p-1)$. Si calcolano due matrici *cov(dfDiabete)* e *cor(dfDiabete)* i cui elementi sono le covarianze e le correlazioni tra coppie di variabili e tali matrici sono simmetriche. La **matrice delle covarianze** contiene sulla diagonale principale la varianza delle singole colonne del data frame, mentre la **matrice delle correlazioni** contiene il numero 1 sulla diagonale principale. La matrice di correlazione misura la forza del legame di natura lineare esistente tra tutte le coppie di variabili quantitative, ossia misura la forza del legame di natura lineare esistente tra tutte le coppie di variabili quantitative.

```
## [1] "matrice delle covarianze"
```

```
##          Anno2012 Anno2013 Anno2014 Anno2015 Anno2016 Anno2017 Anno2018
## Anno2012 198.1498 178.1319 169.0620 159.2406 171.0078 172.3847 154.1079
## Anno2013 178.1319 168.4499 163.7115 157.1870 161.8364 159.8514 145.4603
## Anno2014 169.0620 163.7115 163.8698 159.8428 161.9574 158.6132 146.3190
## Anno2015 159.2406 157.1870 159.8428 161.7076 159.7960 154.3597 143.3591
## Anno2016 171.0078 161.8364 161.9574 159.7960 180.4551 180.8219 168.2497
## Anno2017 172.3847 159.8514 158.6132 154.3597 180.8219 185.6392 172.7366
## Anno2018 154.1079 145.4603 146.3190 143.3591 168.2497 172.7366 170.5241
## Anno2019 156.4542 148.1447 149.1431 145.5254 172.2690 177.4784 176.0481
## Anno2020 119.5293 122.3375 127.4297 128.3396 154.3733 157.4944 165.2468
## Anno2021 123.7435 125.3006 129.6980 131.4091 168.1464 172.8513 175.3501
##          Anno2019 Anno2020 Anno2021
## Anno2012 156.4542 119.5293 123.7435
## Anno2013 148.1447 122.3375 125.3006
## Anno2014 149.1431 127.4297 129.6980
## Anno2015 145.5254 128.3396 131.4091
## Anno2016 172.2690 154.3733 168.1464
## Anno2017 177.4784 157.4944 172.8513
## Anno2018 176.0481 165.2468 175.3501
## Anno2019 184.7939 173.3087 182.3959
## Anno2020 173.3087 195.8143 216.2560
## Anno2021 182.3959 216.2560 259.8612
```

```
## [1] "matrice delle correlazioni"
```

```
##           Anno2012  Anno2013  Anno2014  Anno2015  Anno2016  Anno2017  Anno2018
## Anno2012 1.0000000 0.9750109 0.9382094 0.8895939 0.9043456 0.8988083 0.8383692
## Anno2013 0.9750109 1.0000000 0.9853588 0.9523929 0.9282316 0.9039538 0.8582551
## Anno2014 0.9382094 0.9853588 1.0000000 0.9819254 0.9418175 0.9094003 0.8753032
## Anno2015 0.8895939 0.9523929 0.9819254 1.0000000 0.9354403 0.8909103 0.8633111
## Anno2016 0.9043456 0.9282316 0.9418175 0.9354403 1.0000000 0.9879426 0.9591286
## Anno2017 0.8988083 0.9039538 0.9094003 0.8909103 0.9879426 1.0000000 0.9708604
## Anno2018 0.8383692 0.8582551 0.8753032 0.8633111 0.9591286 0.9708604 1.0000000
## Anno2019 0.8176110 0.8396666 0.8570576 0.8418406 0.9433632 0.9582237 0.9917329
## Anno2020 0.6068136 0.6735998 0.7113754 0.7212292 0.8212310 0.8260530 0.9043114
## Anno2021 0.5453244 0.5988902 0.6285111 0.6410464 0.7764836 0.7869861 0.8329951
```

```
##            Anno2019  Anno2020  Anno2021
## Anno2012 0.8176110 0.6068136 0.5453244
## Anno2013 0.8396666 0.6735998 0.5988902
## Anno2014 0.8570576 0.7113754 0.6285111
## Anno2015 0.8418406 0.7212292 0.6410464
## Anno2016 0.9433632 0.8212310 0.7764836
## Anno2017 0.9582237 0.8260530 0.7869861
## Anno2018 0.9917329 0.9043114 0.8329951
## Anno2019 1.0000000 0.9110756 0.8323398
## Anno2020 0.9110756 1.0000000 0.9586835
## Anno2021 0.8323398 0.9586835 1.0000000
```

La funzione *pairs()* permette di visualizzare in un'unica finestra grafica più scatterplot ottenuti mettendo in relazione tutte le coppie di variabili quantitative definite nel data frame.

Nel modello di regressione lineare multipla, a differenza del modello semplice, avremo diversi coefficienti angolari, uno per ciascuna variabile indipendente.

Per effettuare l'analisi di regressioni lineari multiple si utilizza la funzione $lm(y \tilde{} x_1 + x_2 + ... + x_p)$ ed è possibile visualizzare tutti i coefficienti.

```
##
## Call:
## lm(formula = dfDiabete$Anno2021 ~ dfDiabete$Anno2020 + dfDiabete$Anno2019 +
##     dfDiabete$Anno2018)
##
## Coefficients:
##        (Intercept)  dfDiabete$Anno2020  dfDiabete$Anno2019  dfDiabete$Anno2018
##            -4.0148              1.3561             -0.7585              0.4973
```

```
##        (Intercept) dfDiabete$Anno2020 dfDiabete$Anno2019 dfDiabete$Anno2018
##         -4.0147854          1.3560882         -0.7585025          0.4972540
```

Da cui ricaviamo che l'intercetta $\alpha = -4.0147$ e i tre regressori sono $\beta_1 = 1.3560, \beta_2 = -0.7585, \beta_3 = 0.4972$. L'equazione della retta sarà quindi $y = -4.01 + 1.35x_1. - 0.75x_2 + 0.49x_3$ I regressori $\beta_1 e \beta_3$ sono positivi e quindi avranno un effetto positivo sui valori del 2020, a differenza del regressore β_2 che è negativo.

Anche in questo caso i **residui** mostrano di quanto si discostano i valori osservati dai valori stimati con la regressione lineare multipla. Calcoliamo dunque i valori stimati e i residui e le relative mediana, varianza e deviazione standard.

```
## [1] "vettore delle stime"
```

```
##         1         2         3         4         5         6         7         8
##  62.47789  49.96996  80.46569 114.62276  83.33926  81.24792  98.64457  63.80621
##         9        10        11        12        13        14        15        16
##  69.54102 105.89752  88.31963  95.12758  82.47565  52.21044  66.64257  75.56731
##        17        18        19        20        21        22        23        24
##  59.64014  65.89864  80.23259  60.78817  78.72297  83.58551  89.27824  84.69392
##        25        26        27        28
##  71.89231  98.10512  76.31369  79.19270
```

```
## [1] "vettore dei residui"
```

```
##            1            2            3            4            5            6
## -0.077890449 -0.769960194 -2.765691273  9.877242592 -6.139258150  1.052078350
##            7            8            9           10           11           12
## -1.144573370  3.193789634 -2.041024422 -0.297520739 -0.919631194 -1.527583231
##           13           14           15           16           17           18
## -6.075654400  2.589561161 -1.542568025  2.932686152 -0.840137596 -2.098642773
##           19           20           21           22           23           24
## -1.732587367  5.111827242 -0.122971128 -4.085511754 -2.078237425  0.006081234
##           25           26           27           28
##  1.307691621 -1.805119044 13.686308759 -3.692704213
```

```
## [1] "mediana campionaria dei residui"
```

```
## [1] -0.8798844
```

```
## [1] "varianza campionaria dei residui"
```

```
## [1] 17.75321
```

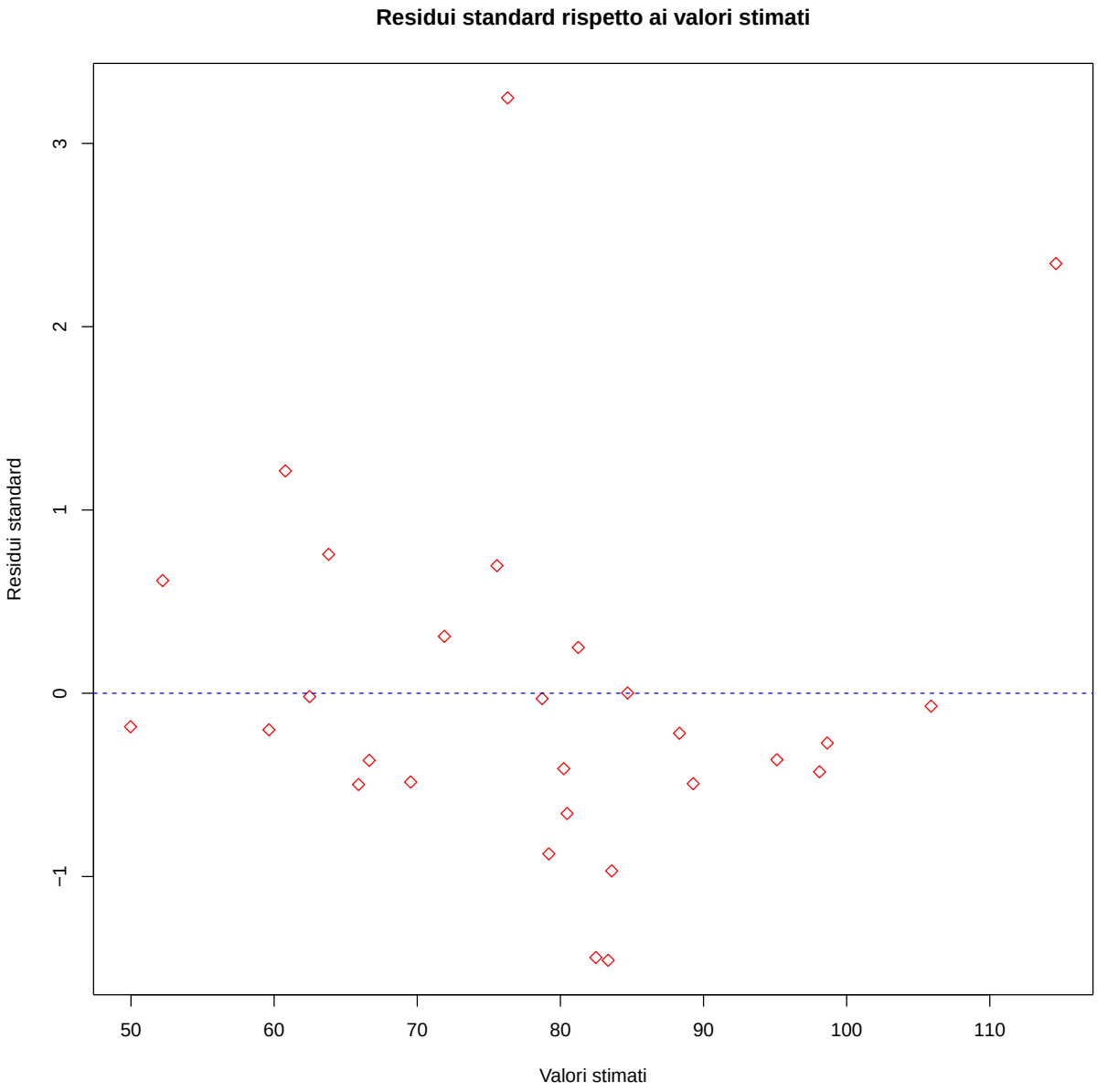
```
## [1] "deviazione standard campionaria dei residui"
```

```
## [1] 4.213456
```

Anche nel modello multivariato è importante calcolare i **residui standardizzati** e rappresentarli in un grafico in cui tali residui standardizzati (ordinate) vengono disegnati in funzione dei valori stimati (ascisse) con il metodo dei minimi quadrati.

```
## [1] "vettore dei residui standardizzati"
```

```
##           1           2           3           4           5           6
## -0.01848611 -0.18273840 -0.65639496  2.34421403 -1.45705999  0.24969487
##           7           8           9          10          11          12
## -0.27164716  0.75799763 -0.48440625 -0.07061204 -0.21826054 -0.36254876
##          13          14          15          16          17          18
## -1.44196461  0.61459315 -0.36610517  0.69602867 -0.19939394 -0.49808110
##          19          20          21          22          23          24
## -0.41120339  1.21321482 -0.02918534 -0.96963438 -0.49323820  0.00144329
##          25          26          27          28
##  0.31036081 -0.42841768  3.24823823 -0.87640745
```



I punti indicano la posizione dove si collocano i residui standardizzati rispetto ai valori stimati con la retta di regressione. La retta orizzontale è posizionata nello zero, che corrisponde alla media campionaria dei residui standardizzati. Anche in questo caso i punti sono disposti quasi casualmente attorno alla linea orizzontale e non si evidenzia alcuna tendenza particolare nella distribuzione dei punti.

Il **coefficiente di determinazione** in un modello di regressione lineare multipla è il rapporto tra la varianza dei valori stimati tramite la funzione di regressione multipla e la varianza dei valori osservati dalla variabile dipendente. Tale coefficiente D^2 risulta adimensionale e $0 \leq D^2 \leq 1$. Quando $D^2 = 0$ il modello non spiega per nulla i dati; invece quando $D^2 = 1$ il modello spiega perfettamente i dati. In R per calcolare l'indice D^2 per la regressione multipla si usa la funzione $summary(lm(y \sim x_1 + x_2 + ... + x_p))\$r.square$.

```
## [1] 0.9316819
```

Il coefficiente di determinazione è 0.9316, ossia il modello di regressione multipla utilizzato è prossimo all'unità e quindi **può spiegare significativamente i dati**. Non si è ottenuto un significativo miglioramento del coefficiente di determinazione in quanto nel modello di regressione semplice era 0.9190 mentre adesso è 0.9316.

Analisi dei cluster

L'*analisi dei cluster* è una metodologia che permette di raggruppare in sottoinsiemi, detti *cluster*, entità appartenenti ad un insieme più ampio. I metodi di analisi dei cluster permettono di raggiungere i seguenti obiettivi: individuazione di una reale tipologia, previsioni basate su gruppi, esplorazione dei dati, generazione di ipotesi di ricerca, verifica di ipotesi di ricerca, riduzione della complessità dei dati. Sia $I = I_1, I_2, ..., I_n$ un insieme di n individui appartenenti ad una popolazione. Assumiamo che esista un insieme di caratteristiche $C = C_1, C_2, ..., C_p$ che sono osservabili e sono possedute da ogni individuo in I. Il termine *osservabile* denota caratteristiche che danno origine a misure. Il problema dell'analisi dei cluster consiste nel determinare m sottoinsiemi, detti cluster, di individui in I, con m intero minore di n, tali che I_i appartenga soltanto ad un unico sottoinsieme. Gli individui che sono assegnati allo stesso cluster sono detti *simili* mentre gli individui che sono assegnati a differenti cluster sono detti *dissimili*. L'obiettivo della seguente analisi dei cluster è applicare tutti i metodi studiati, calcolare la misura di non omogeneità e analizzare qual è il metodo più efficiente. E' infatti buona norma applicare una pluralità di metodi per verificare la stabilità dei gruppi e scegliere la partizione (e il metodo) che, a parità di numero di cluster, fornisce le migliori misure di non omogeneità. Per analizzare i nostri dati e creare dei cluster possiamo utilizzare le misure di distanza o di similarità, ovvero la distanza tra due individui differenti. Le *misure metriche di somiglianza* sono soprattutto basate sulle *funzioni distanza* tra i vettori delle caratteristiche. Occorre dunque definire tale funzione. Una funzione a valori reali $d(X_i, X_j)$ è detta **funzione distanza** se e soltanto se essa soddisfa le seguenti condizioni:

(i) $d(X_i, X_j) = 0$ se e solo se $X_i = X_j$, con X_i e X_j in E_p;

(ii) $d(X_i, X_j) \geq 0$ per ogni X_i e X_j in E_p;

(iii) $d(X_i, X_j) = d(X_j, X_i)$ per ogni X_i e X_j in E_p;

(iv) $d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)$ per ogni X_i, X_j e X_k in E_p.

Le distanze tra tutte le possibili coppie di unità sono inserite in una matrice D di cardinalità n x n. Calcoliamo dunque le distanze per il nostro data frame mediante la funzione *dist(X)* e avremo una matrice delle distanze, in questo caso euclidea.

La *metrica Euclidea* è così definita: $d_2(X_i, X_j) = \left[\sum_{k=1}^{p} (x_{ij} - x_{jk})^2\right]^2$

```
##                 Australia    Austria    Belgium     Canada      Chile
## Australia       0.0000000  2.6475715  3.0386269  6.3161097  3.2491483
## Austria         2.6475715  0.0000000  5.5448618  8.3479881  2.9588638
## Belgium         3.0386269  5.5448618  0.0000000  4.0036753  5.0459048
## Canada          6.3161097  8.3479881  4.0036753  0.0000000  7.0386779
## Chile           3.2491483  2.9588638  5.0459048  7.0386779  0.0000000
## Costa Rica      2.4020931  4.4104912  1.8526005  4.2089151  3.5280924
## Czech Republic  7.0604020  9.6371788  4.1264082  4.4124952  8.8484226
## Denmark         1.1012818  2.1982988  3.4596983  6.2101417  2.5180107
## Estonia         1.2379382  3.3670927  2.2366983  5.3529204  3.3047337
## Finland         8.2284257 10.8157590  5.2899976  4.9631650  9.9664693
## France          5.9651029  8.5722136  3.1612393  4.4861496  8.0213556
## Germany         6.3887674  9.0056441  3.5657395  4.4163498  8.3357371
## Hungary         4.0021213  6.6108861  1.5605302  4.5195566  6.2854416
## Iceland         2.9840433  0.7892070  5.6893439  8.1970118  2.6711488
## Italy           1.1178458  3.7050437  1.9842874  5.5796658  3.8792819
## Korea           1.8976468  4.4082454  1.4579638  4.6850738  4.0446168
## Lithuania       2.7558262  1.5114998  5.1038980  7.4150863  2.4001316
## Luxembourg      1.5035414  4.0681359  1.7925968  5.5770656  4.2262136
## Netherlands     4.1673494  6.7768255  1.5824786  4.4522412  6.3940865
## Norway          1.3836350  1.9390169  3.7847952  6.4463935  2.4593696
## Portugal        2.7049887  5.1765380  0.4507485  4.0862726  4.6206074
## Slovak Republic 3.8991974  6.3314478  1.0461890  3.8430737  5.6500815
```

```
## Slovenia         4.9189383  7.4759939  1.9530126  3.6202982  6.8346642
## Spain            4.0540683  6.5796050  1.0696785  3.5821980  5.9490381
## Sweden           1.5228524  3.6491272  2.0334976  4.9475753  3.3453181
## Turkiye          4.0300972  6.2034860  1.8051206  2.5077039  5.0660975
## United Kingdom   5.4354197  8.0136906  3.1117592  5.0558128  7.5564546
## Croatia          2.7328702  5.1201044  0.6888838  4.0950602  4.5681184
##                 Costa Rica Czech Republic    Denmark    Estonia    Finland
## Australia        2.4020931      7.0604020  1.1012818  1.2379382  8.2284257
## Austria          4.4104912      9.6371788  2.1982988  3.3670927 10.8157590
## Belgium          1.8526005      4.1264082  3.4596983  2.2366983  5.2899976
## Canada           4.2089151      4.4124952  6.2101417  5.3529204  4.9631650
## Chile            3.5280924      8.8484226  2.5180107  3.3047337  9.9664693
## Costa Rica       0.0000000      5.7053697  2.2636853  1.4461509  6.8022968
## Czech Republic   5.7053697      0.0000000  7.5661452  6.3388586  1.2698116
## Denmark          2.2636853      7.5661452  0.0000000  1.3018816  8.7262948
## Estonia          1.4461509      6.3388586  1.3018816  0.0000000  7.5170451
## Finland          6.8022968      1.2698116  8.7262948  7.5170451  0.0000000
## France           4.8746940      1.2981318  6.5658564  5.3536249  2.3943548
## Germany          5.1894018      1.0123412  6.9695911  5.7804181  1.9276053
## Hungary          3.2099144      3.2846016  4.6731069  3.5300310  4.3817184
## Iceland          4.3887846      9.7951895  2.2601441  3.4689672 10.9693744
## Italy            1.9002197      5.9739224  1.8715737  1.0013322  7.1485996
## Korea            1.3728721      5.3921819  2.3502187  1.4585002  6.5073191
## Lithuania        3.6733359      9.1999609  1.8122808  2.9182201 10.3543269
## Luxembourg       2.1076577      5.6553210  2.2980645  1.3338840  6.8423258
## Netherlands      3.3656941      2.9860902  4.8188843  3.6343853  4.1582340
## Norway           2.5383441      7.8901494  0.3722328  1.5995558  9.0522830
## Portugal         1.4844823      4.5136523  3.0660597  1.8669743  5.6729011
## Slovak Republic  2.5712261      3.4326900  4.2722132  3.0110999  4.6454096
## Slovenia         3.6008058      2.1948947  5.3959816  4.1689418  3.3631020
## Spain            2.7235777      3.0941485  4.4835674  3.2552331  4.2713487
## Sweden           1.0295537      6.1341778  1.4848673  0.4952318  7.2849230
## Turkiye          1.9059896      4.3351248  4.0234658  3.0398475  5.3280131
## United Kingdom   4.7599682      2.5166202  6.1323745  5.0279840  3.5637052
## Croatia          1.3483758      4.6646968  3.0197561  1.8427398  5.7970893
##                     France    Germany    Hungary    Iceland      Italy
## Australia        5.9651029  6.3887674  4.0021213  2.9840433  1.1178458
## Austria          8.5722136  9.0056441  6.6108861  0.7892070  3.7050437
## Belgium          3.1612393  3.5657395  1.5605302  5.6893439  1.9842874
## Canada           4.4861496  4.4163498  4.5195566  8.1970118  5.5796658
## Chile            8.0213556  8.3357371  6.2854416  2.6711488  3.8792819
## Costa Rica       4.8746940  5.1894018  3.2099144  4.3887846  1.9002197
## Czech Republic   1.2981318  1.0123412  3.2846016  9.7951895  5.9739224
## Denmark          6.5658564  6.9695911  4.6731069  2.2601441  1.8715737
## Estonia          5.3536249  5.7804181  3.5300310  3.4689672  1.0013322
## Finland          2.3943548  1.9276053  4.3817184 10.9693744  7.1485996
## France           0.0000000  0.6416795  2.0971351  8.7889667  4.8731642
## Germany          0.6416795  0.0000000  2.4925021  9.2124327  5.3126161
## Hungary          2.0971351  2.4925021  0.0000000  6.8772854  2.9288443
## Iceland          8.7889667  9.2124327  6.8772854  0.0000000  3.9870284
## Italy            4.8731642  5.3126161  2.9288443  3.9870284  0.0000000
## Korea            4.3621230  4.7152783  2.4736193  4.5867887  1.1067779
## Lithuania        8.2521337  8.6585345  6.3863555  1.0022789  3.6102471
## Luxembourg       4.5344347  4.9957054  2.6168001  4.3737949  0.4599966
```

```
## Netherlands      1.8181297  2.2935285  0.6516449  7.0228555  3.0780386
## Norway           6.8986985  7.3067227  5.0156143  1.9366447  2.2057475
## Portugal         3.5690855  3.9595298  1.9366024  5.3003894  1.7038076
## Slovak Republic  2.6260783  3.0649674  1.7033804  6.4416281  2.8398763
## Slovenia         1.3506145  1.7213327  1.4367526  7.6268208  3.8346716
## Spain            2.2281596  2.6150092  1.3239590  6.7120743  2.9917567
## Sweden           5.1817179  5.5752187  3.3851737  3.7037775  1.1706538
## Turkiye          3.8067041  3.9935565  2.8595929  6.1490887  3.2529586
## United Kingdom   1.6031288  1.9420772  2.0896406  8.2858901  4.4174948
## Croatia          3.7400820  4.1184219  2.1044870  5.2295861  1.7724381
##                      Korea  Lithuania Luxembourg Netherlands     Norway
## Australia        1.8976468  2.7558262  1.5035414   4.1673494  1.3836350
## Austria          4.4082454  1.5114998  4.0681359   6.7768255  1.9390169
## Belgium          1.4579638  5.1038980  1.7925968   1.5824786  3.7847952
## Canada           4.6850738  7.4150863  5.5770656   4.4522412  6.4463935
## Chile            4.0446168  2.4001316  4.2262136   6.3940865  2.4593696
## Costa Rica       1.3728721  3.6733359  2.1076577   3.3656941  2.5383441
## Czech Republic   5.3921819  9.1999609  5.6553210   2.9860902  7.8901494
## Denmark          2.3502187  1.8122808  2.2980645   4.8188843  0.3722328
## Estonia          1.4585002  2.9182201  1.3338840   3.6343853  1.5995558
## Finland          6.5073191 10.3543269  6.8423258   4.1582340  9.0522830
## France           4.3621230  8.2521337  4.5344347   1.8181297  6.8986985
## Germany          4.7152783  8.6585345  4.9957054   2.2935285  7.3067227
## Hungary          2.4736193  6.3863555  2.6168001   0.6516449  5.0156143
## Iceland          4.5867887  1.0022789  4.3737949   7.0228555  1.9366447
## Italy            1.1067779  3.6102471  0.4599966   3.0780386  2.2057475
## Korea            0.0000000  4.0947758  1.2353999   2.6684414  2.6979460
## Lithuania        4.0947758  0.0000000  4.0134950   6.5391725  1.4758442
## Luxembourg       1.2353999  4.0134950  0.0000000   2.7324064  2.6245592
## Netherlands      2.6684414  6.5391725  2.7324064   0.0000000  5.1561975
## Norway           2.6979460  1.4758442  2.6245592   5.1561975  0.0000000
## Portugal         1.1521890  4.7128418  1.5813928   1.9689193  3.3920312
## Slovak Republic  2.4355587  5.8371917  2.5773886   1.4351160  4.5799577
## Slovenia         3.2504077  7.0330139  3.5522702   1.1493809  5.7192101
## Spain            2.4239662  6.1137732  2.7412172   1.0923246  4.8024152
## Sweden           1.2008365  3.0914620  1.4948971   3.5063269  1.7875608
## Turkiye          2.4006971  5.4007831  3.2625876   2.8652190  4.2952294
## United Kingdom   4.0398694  7.8885637  4.0506028   1.6797058  6.4674408
## Croatia          1.2455533  4.5950782  1.7005442   2.2013679  3.3364341
##                   Portugal Slovak Republic   Slovenia      Spain     Sweden
## Australia        2.7049887       3.8991974  4.9189383  4.0540683  1.5228524
## Austria          5.1765380       6.3314478  7.4759939  6.5796050  3.6491272
## Belgium          0.4507485       1.0461890  1.9530126  1.0696785  2.0334976
## Canada           4.0862726       3.8430737  3.6202982  3.5821980  4.9475753
## Chile            4.6206074       5.6500815  6.8346642  5.9490381  3.3453181
## Costa Rica       1.4844823       2.5712261  3.6008058  2.7235777  1.0295537
## Czech Republic   4.5136523       3.4326900  2.1948947  3.0941485  6.1341778
## Denmark          3.0660597       4.2722132  5.3959816  4.4835674  1.4848673
## Estonia          1.8669743       3.0110999  4.1689418  3.2552331  0.4952318
## Finland          5.6729011       4.6454096  3.3631020  4.2713487  7.2849230
## France           3.5690855       2.6260783  1.3506145  2.2281596  5.1817179
## Germany          3.9595298       3.0649674  1.7213327  2.6150092  5.5752187
## Hungary          1.9366024       1.7033804  1.4367526  1.3239590  3.3851737
## Iceland          5.3003894       6.4416281  7.6268208  6.7120743  3.7037775
```

```
## Italy            1.7038076       2.8398763  3.8346716  2.9917567  1.1706538
## Korea            1.1521890       2.4355587  3.2504077  2.4239662  1.2008365
## Lithuania        4.7128418       5.8371917  7.0330139  6.1137732  3.0914620
## Luxembourg       1.5813928       2.5773886  3.5522702  2.7412172  1.4948971
## Netherlands      1.9689193       1.4351160  1.1493809  1.0923246  3.5063269
## Norway           3.3920312       4.5799577  5.7192101  4.8024152  1.7875608
## Portugal         0.0000000       1.3461495  2.3566026  1.4462101  1.6380941
## Slovak Republic  1.3461495       0.0000000  1.4211664  0.7280729  2.8559334
## Slovenia         2.3566026       1.4211664  0.0000000  0.9529249  3.9616366
## Spain            1.4462101       0.7280729  0.9529249  0.0000000  3.0472825
## Sweden           1.6380941       2.8559334  3.9616366  3.0472825  0.0000000
## Turkiye          1.7763545       1.9922327  2.5154747  1.9041795  2.6448056
## United Kingdom   3.4348600       2.7173732  2.0033786  2.4196732  4.9164356
## Croatia          0.4691793       1.5179051  2.5223249  1.6541758  1.5799616
##                    Turkiye United Kingdom    Croatia
## Australia        4.0300972      5.4354197  2.7328702
## Austria          6.2034860      8.0136906  5.1201044
## Belgium          1.8051206      3.1117592  0.6888838
## Canada           2.5077039      5.0558128  4.0950602
## Chile            5.0660975      7.5564546  4.5681184
## Costa Rica       1.9059896      4.7599682  1.3483758
## Czech Republic   4.3351248      2.5166202  4.6646968
## Denmark          4.0234658      6.1323745  3.0197561
## Estonia          3.0398475      5.0279840  1.8427398
## Finland          5.3280131      3.5637052  5.7970893
## France           3.8067041      1.6031288  3.7400820
## Germany          3.9935565      1.9420772  4.1184219
## Hungary          2.8595929      2.0896406  2.1044870
## Iceland          6.1490887      8.2858901  5.2295861
## Italy            3.2529586      4.4174948  1.7724381
## Korea            2.4006971      4.0398694  1.2455533
## Lithuania        5.4007831      7.8885637  4.5950782
## Luxembourg       3.2625876      4.0506028  1.7005442
## Netherlands      2.8652190      1.6797058  2.2013679
## Norway           4.2952294      6.4674408  3.3364341
## Portugal         1.7763545      3.4348600  0.4691793
## Slovak Republic  1.9922327      2.7173732  1.5179051
## Slovenia         2.5154747      2.0033786  2.5223249
## Spain            1.9041795      2.4196732  1.6541758
## Sweden           2.6448056      4.9164356  1.5799616
## Turkiye          0.0000000      4.0698505  1.7465534
## United Kingdom   4.0698505      0.0000000  3.7227244
## Croatia          1.7465534      3.7227244  0.0000000
```

E' sufficiente considerare la matrice triangolare al di sopra o al di sotto della diagonale principale. Infatti, i termini sulla diagonale principale sono tutti uguali a zero mentre i termini simmetrici sono uguali a due a due, pertanto il numero di distanze che è necessario conoscere affinchè sia definita la posizione di ciascuna delle n variabili rispetto alle rimanenti $n-1$ è $n(n-1)/2$. In questo caso è stata applicata la metrica euclidea che però è fortemente influenzata dall'unità di misura in base alla quale è valutata ciascuna delle p caratteristiche. Per ovviare a tale problema, si sono scalate e standardizzate le misure riportate nel data frame iniziale prima di calcolare la matrice delle distanze mediante la funzione *scale(X)*. Mediante lo scalamento e la standardizzazione si ottengono dei nuovi dati le cui medie campionarie sono nulle e le varianze campionarie unitarie. Calcoliamo dunque la media campionaria, la varianza campionaria e la deviazione standard campionaria delle colonne della matrice mediante la funzione *apply()* di R. Alla matrice di partenza

si può associare una matrice W_x detta **matrice delle varianze e covarianze** che calcoliamo con la funzione *cov(matscal)*.

```
## [1] "media campionaria delle colonne di matscal"
```

```
##      Anno2012      Anno2013      Anno2014      Anno2015      Anno2016
## -6.609108e-17  3.816392e-17 -2.852497e-16  4.693894e-16  4.406198e-16
##      Anno2017      Anno2018      Anno2019      Anno2020      Anno2021
## -1.964574e-16  5.230191e-16  4.411212e-16 -5.403780e-17 -3.593351e-16
```

```
## [1] "varianza campionaria delle colonne di matscal"
```

```
## Anno2012 Anno2013 Anno2014 Anno2015 Anno2016 Anno2017 Anno2018 Anno2019
##        1        1        1        1        1        1        1        1
## Anno2020 Anno2021
##        1        1
```

```
## [1] "deviazione standard campionaria delle colonne di matscal"
```

```
## Anno2012 Anno2013 Anno2014 Anno2015 Anno2016 Anno2017 Anno2018 Anno2019
##        1        1        1        1        1        1        1        1
## Anno2020 Anno2021
##        1        1
```

```
## [1] "matrice delle varianze e covarianze"
```

```
##           Anno2012  Anno2013  Anno2014  Anno2015  Anno2016  Anno2017  Anno2018
## Anno2012 1.0000000 0.9750109 0.9382094 0.8895939 0.9043456 0.8988083 0.8383692
## Anno2013 0.9750109 1.0000000 0.9853588 0.9523929 0.9282316 0.9039538 0.8582551
## Anno2014 0.9382094 0.9853588 1.0000000 0.9819254 0.9418175 0.9094003 0.8753032
## Anno2015 0.8895939 0.9523929 0.9819254 1.0000000 0.9354403 0.8909103 0.8633111
## Anno2016 0.9043456 0.9282316 0.9418175 0.9354403 1.0000000 0.9879426 0.9591286
## Anno2017 0.8988083 0.9039538 0.9094003 0.8909103 0.9879426 1.0000000 0.9708604
## Anno2018 0.8383692 0.8582551 0.8753032 0.8633111 0.9591286 0.9708604 1.0000000
## Anno2019 0.8176110 0.8396666 0.8570576 0.8418406 0.9433632 0.9582237 0.9917329
## Anno2020 0.6068136 0.6735998 0.7113754 0.7212292 0.8212310 0.8260530 0.9043114
## Anno2021 0.5453244 0.5988902 0.6285111 0.6410464 0.7764836 0.7869861 0.8329951
##           Anno2019  Anno2020  Anno2021
## Anno2012 0.8176110 0.6068136 0.5453244
## Anno2013 0.8396666 0.6735998 0.5988902
## Anno2014 0.8570576 0.7113754 0.6285111
## Anno2015 0.8418406 0.7212292 0.6410464
## Anno2016 0.9433632 0.8212310 0.7764836
## Anno2017 0.9582237 0.8260530 0.7869861
## Anno2018 0.9917329 0.9043114 0.8329951
## Anno2019 1.0000000 0.9110756 0.8323398
## Anno2020 0.9110756 1.0000000 0.9586835
## Anno2021 0.8323398 0.9586835 1.0000000
```

Misura di non omogeneità totale

Si vuole adesso definire la **matrice statistica di non omogeneità** per l'insieme I di individui, di cardinalità *pxp* definita come: $H_I = (n-1)W_I$ dove W_I è la matrice delle varianze e covarianze calcolata in precedenza. La traccia di una matrice di non omogeneità di un insieme di individui fornisce una misura della dispersione dei dati intorno al valore medio dell'insieme dal quale è stata ricavata. Per l'analisi dei cluster è necessario calcolare il valore della misura di non omogeneità totale, calcolabile mediante diversi metodi. Si definisce *misura di non omogeneità statistica* dell'insieme I di individui la traccia della matrice H_I: $trHI = \sum_{r=1}^{p} h_{rr} = (n-1)\sum_{r=1}^{p} s_r^2$.

Il **primo metodo** calcola la misura di non omogeneità statistica dell'insieme I di individui utilizzando tale definizione.

```
n<-nrow(dfDiabete)
if(n>1)
  trHI<-(n-1)*sum(apply(matscal,2,var)) else trHI<-0
trHI
```

```
## [1] 270
```

Tale misura è esprimibile anche in termine della somma dei quadrati delle distanze euclidee tra ogni vettore $X_1, X_2, ..., X_n$ e il vettore $\bar{X}$ delle medie campionarie.

Il **secondo metodo** determina la matrice di non omogeneità statistica dell'insieme I di individui e somma gli elementi sulla diagonale.

```
n<-nrow(matscal)
WI<-cov(matscal)
HI<-(n-1)*WI
trHI<-sum(diag(HI))
trHI
```

```
## [1] 270
```

Il **terzo metodo** calcola la misura di non omogeneità statistica dell'insieme I di individui utilizzando i quadrati delle distanze euclidee.

```
trHI<-sum(dist_E^2)/n
trHI
```

```
## [1] 270
```

Metodi non gerarchici

L'obiettivo dei metodi non gerarchici è quello di ottenere un'**unica partizione** degli n individui di partenza in cluster. Tali metodi consentono di riallocare gli individui già classificati ad un livello precedente. Il numero di cluster in cui suddividere l'insieme totale degli n individui può essere fissato a priori dal ricercatore oppure può essere determinato nel corso dell'analisi. Gli algoritmi di tipo non gerarchico procedono, data una prima partizione, a riallocare gli individui nel gruppo con centroide più vicino; fin quando per nessun individuo si verifica che sia minima la distanza rispetto al centroide di un gruppo diverso da quello a cui esso appartiene. Il centroide è un punto nello spazio che rappresenta un cluster e che corrisponde al punto medio dei punti del cluster stesso. Il metodo più utilizzato prende il nome di **k-means** e richiede che il numero di cluster sia specificato a priori e fornisce in output un'unica partizione.

Tale metodo applica i seguenti passi:

1) fissare a priori il numero k di cluster specificando i punti di riferimento iniziali che inducono una prima partizione provvisoria;

2) considerare tutti gli individui e attribuire ciascuno di essi al cluster individuato dal punto di riferimento da cui ha distanza minore;

3) calcolare il baricentro (centroide) di ognuno dei k gruppi così ottenuti;

4) valutare la distanza di ogni unità da ogni centroide ottenuto al passo precedente.

5) Ricalcolare i centroidi dei k gruppi così ottentui;

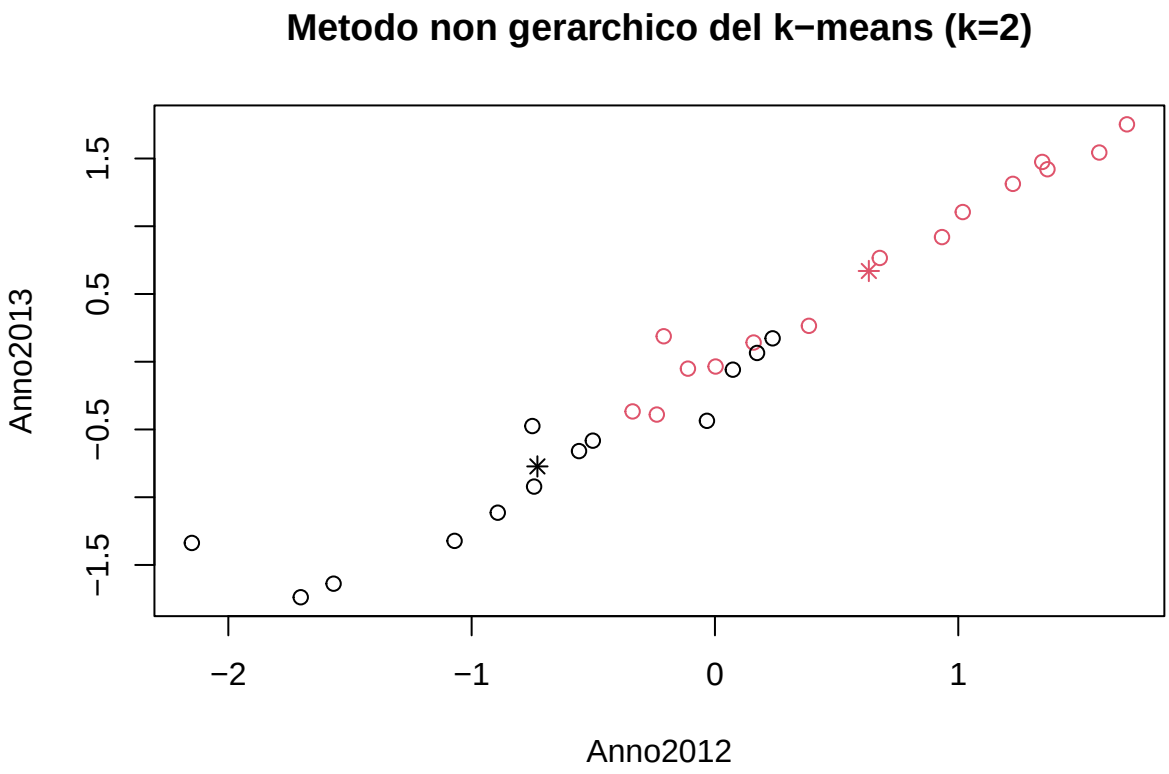
6) ripetere il procedimento a partire dal punto 4) fino a che i centroidi non subiscono ulteriori modifiche rispetto all'iterazione precedente. Si procede iterativamente fino a raggiungere una configurazione stabile.

Per garantire la convergenza della procedura iterativa, come misura di distanza tra i vettori delle caratterisiche e i centroidi viene utilizzata la distanza euclidea e si considera la matrice contenente i quadrati delle distanze euclidee.

Applichiamo il **metodo k-means** tramite la funzione *kmeans(x, center, iter.max = N, nstart = M)* effettuando un'unica scelta casuale dei punti di riferimento con un numero massimo di iterazioni pari a 10. Attueremo una scelta casuale dei punti di riferimento.

Due cluster Analizziamo il caso in cui il numero di cluster è pari a 2 e visualizziamo i cluster generati.

```
## K-means clustering with 2 clusters of sizes 13, 15
##
## Cluster means:
##     Anno2012  Anno2013   Anno2014   Anno2015   Anno2016   Anno2017   Anno2018
## 1 -0.7297217 -0.772793 -0.7994208 -0.8091549 -0.8870612 -0.8823907 -0.8682620
## 2  0.6324255  0.669754  0.6928314  0.7012676  0.7687863  0.7647386  0.7524937
##     Anno2019   Anno2020   Anno2021
## 1 -0.8753124 -0.8181859 -0.7402318
## 2  0.7586041  0.7090945  0.6415342
##
## Clustering vector:
##       Australia         Austria         Belgium          Canada           Chile
##               1               1               2               2               1
##      Costa Rica  Czech Republic         Denmark         Estonia         Finland
##               1               2               1               1               2
##          France         Germany         Hungary         Iceland           Italy
##               2               2               2               1               1
##           Korea       Lithuania      Luxembourg     Netherlands          Norway
##               1               1               1               2               1
##        Portugal Slovak Republic        Slovenia           Spain          Sweden
##               2               2               2               2               1
##         Turkiye  United Kingdom         Croatia
##               2               2               2
##
## Within cluster sum of squares by cluster:
## [1] 42.45709 64.31510
##  (between_SS / total_SS =  60.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

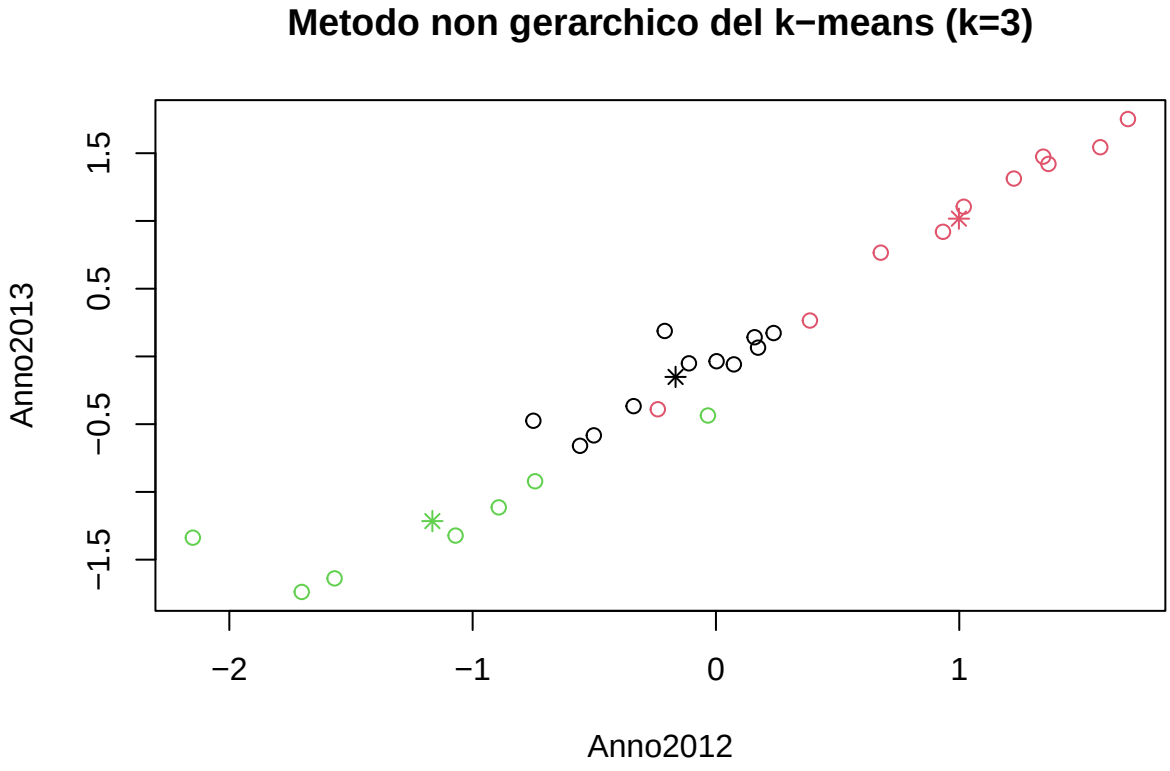


Tale metodo individua la seguente partizione in due cluster $G_1 = \{$Repubblica Ceca, Finlandia, Francia, Germania, Ungheria, Paesi Bassi, Portogallo, Repubblica Slovacca, Slovenia, Spagna, Turchia, Regno Unito, Croazia$\}$ e $G_2 = \{$Cile, Costa Rica, Danimarca, Estonia, Islanda, Italia, Corea, Lituania, Lussemburgo, Norvegia, Svezia.$\}$ Possiamo notare che il *rapporto tra between e total* è minore del 70% (60.5%), quindi dobbiamo scegliere un numero più alto di cluster.

Tre cluster Analizziamo il caso in cui il numero di cluster è pari a 3.

```
## K-means clustering with 3 clusters of sizes 11, 10, 7
##
## Cluster means:
##     Anno2012   Anno2013   Anno2014   Anno2015   Anno2016   Anno2017    Anno2018
## 1 -0.1660215 -0.1510703 -0.1577068 -0.1400683 -0.1578983 -0.1199576 -0.07441546
## 2  0.9985185  1.0169862  1.0175414  0.9668033  1.0623342  1.0453513  0.99601350
## 3 -1.1655641 -1.2154414 -1.2058056 -1.1610402 -1.2694944 -1.3048543 -1.30593786
##       Anno2019    Anno2020   Anno2021
## 1 -0.009744637 -0.04169415 -0.1453568
## 2  0.939812938  0.85678347  0.8694055
## 3 -1.327276910 -1.15845701 -1.0135901
##
## Clustering vector:
##       Australia         Austria         Belgium          Canada           Chile
##               3               3               1               2               3
##      Costa Rica  Czech Republic         Denmark         Estonia         Finland
##               1               2               3               1               2
##          France         Germany         Hungary         Iceland           Italy
##               2               2               2               3               1
```

```
##           Korea       Lithuania      Luxembourg     Netherlands          Norway
##               1               3               1               2               3
##        Portugal Slovak Republic        Slovenia           Spain          Sweden
##               1               1               2               2               1
##         Turkiye  United Kingdom         Croatia
##               1               2               1
##
## Within cluster sum of squares by cluster:
## [1] 17.21571 37.56683 14.02809
##  (between_SS / total_SS =  74.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```



Sono individuati adesso tre partizioni: $G_1 = \{$Costa Rica, Estonia, Italia, Corea, Lussemburgo, Portogallo, Repubblica Slovacca, Svezia, Turchia, Croazia$\}$, $G_2 = \{$Repubblica Ceca, Finlandia, Francia, Germania, Ungheria, Paesi Bassi, Slovenia, Spagna, Regno Unito$\}$ e $G_3 = \{$Cile, Danimarca, Islanda, Lituania, Norvegia$\}$. In questo caso il valore del rapporto tra between e total è maggiore del 70% (74.5%), dunque possiamo considerare tre cluster.

I **vantaggi** del metodo k-means sono la velocità di esecuzione dei calcoli e la libertà che viene lasciata agli individui di raggrupparsi e allontanarsi. Uno **svantaggio**, invece, è che la classificazione finale può essere influenzata dalla scelta iniziale dei k vettori delle caratteristiche come punti di riferimento.

Metodi gerarchici

I metodi di clustering gerarchico possono essere di due tipi: **agglomerativi** e **divisivi**. I metodi gerarchici di tipo agglomerativo partono da una situazione in cui si hanno n cluster distinti ognuno contenente un solo individuo per giungere, attraverso successive unioni dei cluster meno distanti tra loro, ad una situazione in cui si ha un solo cluster che contiene tutti gli n individui. Invece, i metodi gerarchici di tipo divisivo partono da una situazione in cui si ha un solo cluster che contiene tutti gli n individui per giungere, attraverso successive divisioni dei cluster più distanti tra loro, ad una situazione in cui si hanno n cluster distinti ognuno contenente un solo individuo.
L'obiettivo finale dei metodi gerarchici non è quello di ottenere una singola partizione degli n individui di partenza, ma di ottenere una sequenza di partizioni che possono essere rappresentate graficamente mediante una struttura ad albero, detta **dendrogramma**, nella quale sull'insieme delle ordinate sono riportati i livelli di distanza mentre sull'asse delle ascisse sono riportati i singoli individui. Ad ogni livello di distanza corrisponde una partizione, mentre ad ogni partizione corrispondono infiniti livelli di distanza compresi tra quelli che individuano due successive unioni o divisioni. Il dendrogramma fornisce un quadro completo della struttura dell'insieme in termini delle misure di distanza tra gli individui.
Sia $I = I_1, I_2, ..., I_n$ un insieme di n individui o entità appartenenti ad una popolazione.

Saranno utilizzati tutti i metodi studiati per effettuare un confronto tra il rapporto between e total per valutare quale metodo è più opportuno da utilizzare, si considera un *buon risultato il rapporto con valore maggiore del 70%.*

Metodo del legame singolo Applico il metodo gerarchico del legame singolo. In questo metodo *la distanza tra i gruppi è definita come la minima tra tutte le distanze che si possono calcolare tra ogni individuo del primo e del secondo gruppo.*
Nella procedura gerarchica si considera inizialmente, ossia al livello 0, un insieme di n cluster; al passo successivo si cerca nella matrice delle distanze il coefficiente di distanza minima e si raggruppano nello stesso cluster i due individui associati secondo tale coefficiente.

```
##             Length Class  Mode
## merge       54     -none- numeric
## height      27     -none- numeric
## order       28     -none- numeric
## labels      28     -none- character
## method       1     -none- character
## call         3     -none- call
## dist.method  1     -none- character
```

```
## List of 7
##  $ merge      : int [1:27, 1:2] -8 -3 -15 -28 -9 -11 -13 -22 -2 -23 ...
##  $ height     : num [1:27] 0.372 0.451 0.46 0.469 0.495 ...
##  $ order      : int [1:28] 4 5 26 27 17 2 14 13 19 28 ...
##  $ labels     : chr [1:28] "Australia" "Austria" "Belgium" "Canada" ...
##  $ method     : chr "single"
##  $ call       : language hclust(d = dist_E, method = "single")
##  $ dist.method: chr "euclidean"
##  - attr(*, "class")= chr "hclust"
```

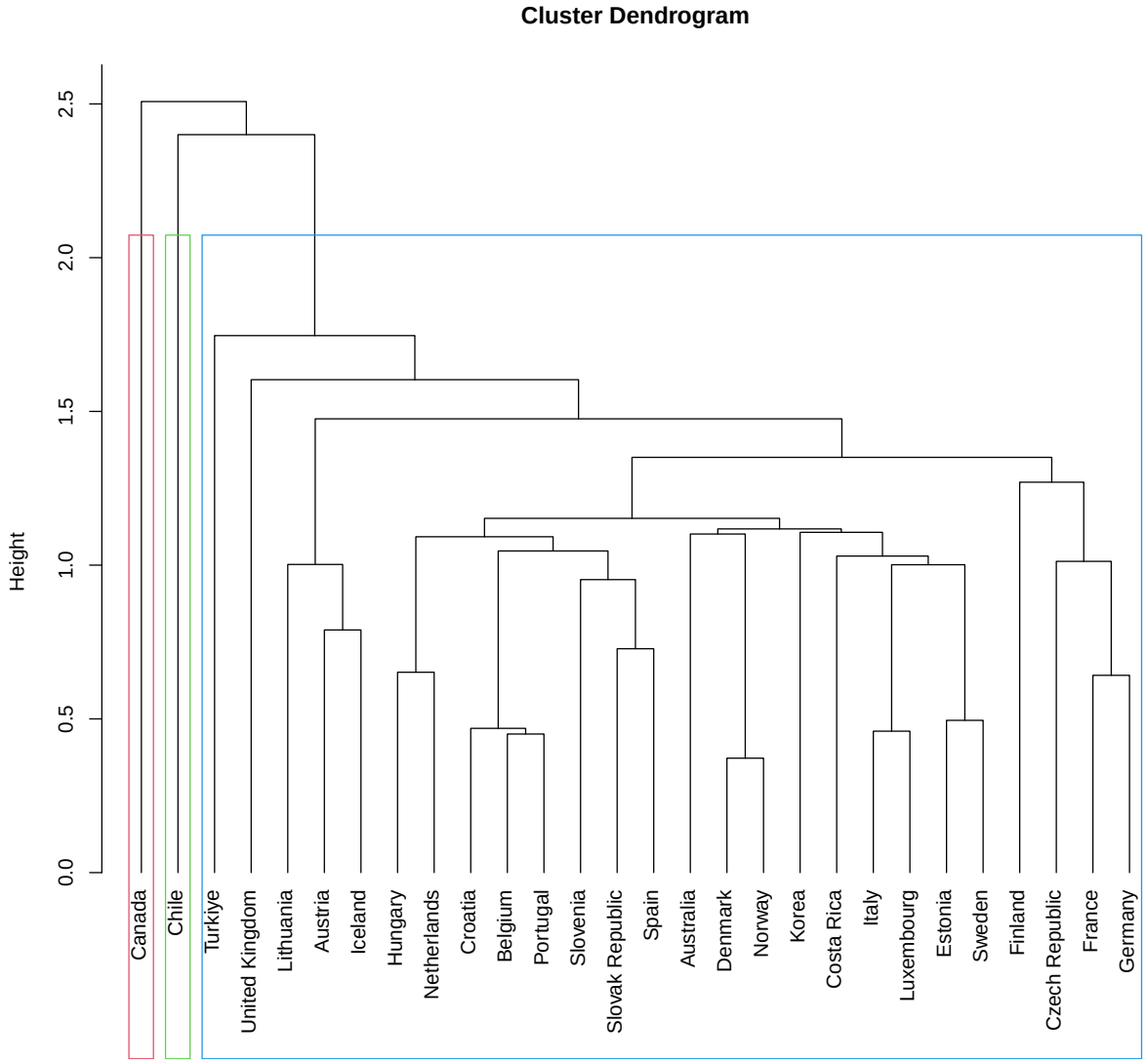
```
##       [,1] [,2]
##  [1,]   -8  -20
##  [2,]   -3  -21
##  [3,]  -15  -18
##  [4,]  -28    2
##  [5,]   -9  -25
##  [6,]  -11  -12
##  [7,]  -13  -19
```

```
##  [8,]  -22  -24
##  [9,]   -2  -14
## [10,]  -23    8
## [11,]    3    5
## [12,]  -17    9
## [13,]   -7    6
## [14,]   -6   11
## [15,]    4   10
## [16,]    7   15
## [17,]   -1    1
## [18,]  -16   14
## [19,]   17   18
## [20,]   16   19
## [21,]  -10   13
## [22,]   20   21
## [23,]   12   22
## [24,]  -27   23
## [25,]  -26   24
## [26,]   -5   25
## [27,]   -4   26
```

```
##  [1] 0.3722328 0.4507485 0.4599966 0.4691793 0.4952318 0.6416795 0.6516449
##  [8] 0.7280729 0.7892070 0.9529249 1.0013322 1.0022789 1.0123412 1.0295537
## [15] 1.0461890 1.0923246 1.1012818 1.1067779 1.1178458 1.1521890 1.2698116
## [22] 1.3506145 1.4758442 1.6031288 1.7465534 2.4001316 2.5077039
```

I risultati *merge* sono stati disposti su due colonne: i numeri con il segno negativo indicano i singoli individui, mentre i numeri positivi indicano i cluster che si formano; *height* indica invece la distanza in cui è avvenuta l'agglomerazione tra i cluster.

Adesso possiamo costruire il dendrogramma per visualizzare i risultati ottenuti.



Attuo adesso il calcolo e l'analisi delle misure di non omogeneità statistiche che saranno utilizzate per valutare la bontà di suddivisione in cluster ottenuta con i vari metodi.

```
##       Australia         Austria         Belgium          Canada           Chile
##               1               1               1               2               3
##      Costa Rica  Czech Republic         Denmark         Estonia         Finland
##               1               1               1               1               1
##          France         Germany         Hungary         Iceland           Italy
##               1               1               1               1               1
##           Korea       Lithuania      Luxembourg     Netherlands          Norway
##               1               1               1               1               1
##        Portugal Slovak Republic        Slovenia           Spain          Sweden
##               1               1               1               1               1
##         Turkiye  United Kingdom         Croatia
##               1               1               1
```

```
## [[1]]
##       Australia         Austria         Belgium          Canada           Chile
##               1               1               1               2               3
##      Costa Rica  Czech Republic         Denmark         Estonia         Finland
##               1               1               1               1               1
##          France         Germany         Hungary         Iceland           Italy
##               1               1               1               1               1
##           Korea       Lithuania      Luxembourg     Netherlands          Norway
##               1               1               1               1               1
##        Portugal Slovak Republic        Slovenia           Spain          Sweden
##               1               1               1               1               1
##         Turkiye  United Kingdom         Croatia
##               1               1               1
```

```
##   Anno2012  Anno2013 Anno2014 Anno2015  Anno2016  Anno2017  Anno2018  Anno2019
## 1 0.884039 0.9977582 1.028514 1.054248 0.9519377 0.8646059 0.9189194 0.9109094
## 2       NA        NA       NA       NA        NA        NA        NA        NA
## 3       NA        NA       NA       NA        NA        NA        NA        NA
##   Anno2020  Anno2021
## 1 0.896916 0.7425685
## 2       NA        NA
## 3       NA        NA
```

```
## [1] "misura di non omogeneità del primo gruppo"
```

```
## [1] 231.2604
```

```
## [1] "misura di non omogeneità del secondo gruppo"
```

```
## [1] 0
```

```
## [1] "misura di non omogeneità del terzo gruppo"
```

```
## [1] 0
```

```
## [1] "misura di non omogeneità interna"
```

```
## [1] 231.2604
```

```
## [1] "misura di non omogeneità tra i cluster"
```

```
## [1] 38.73961
```

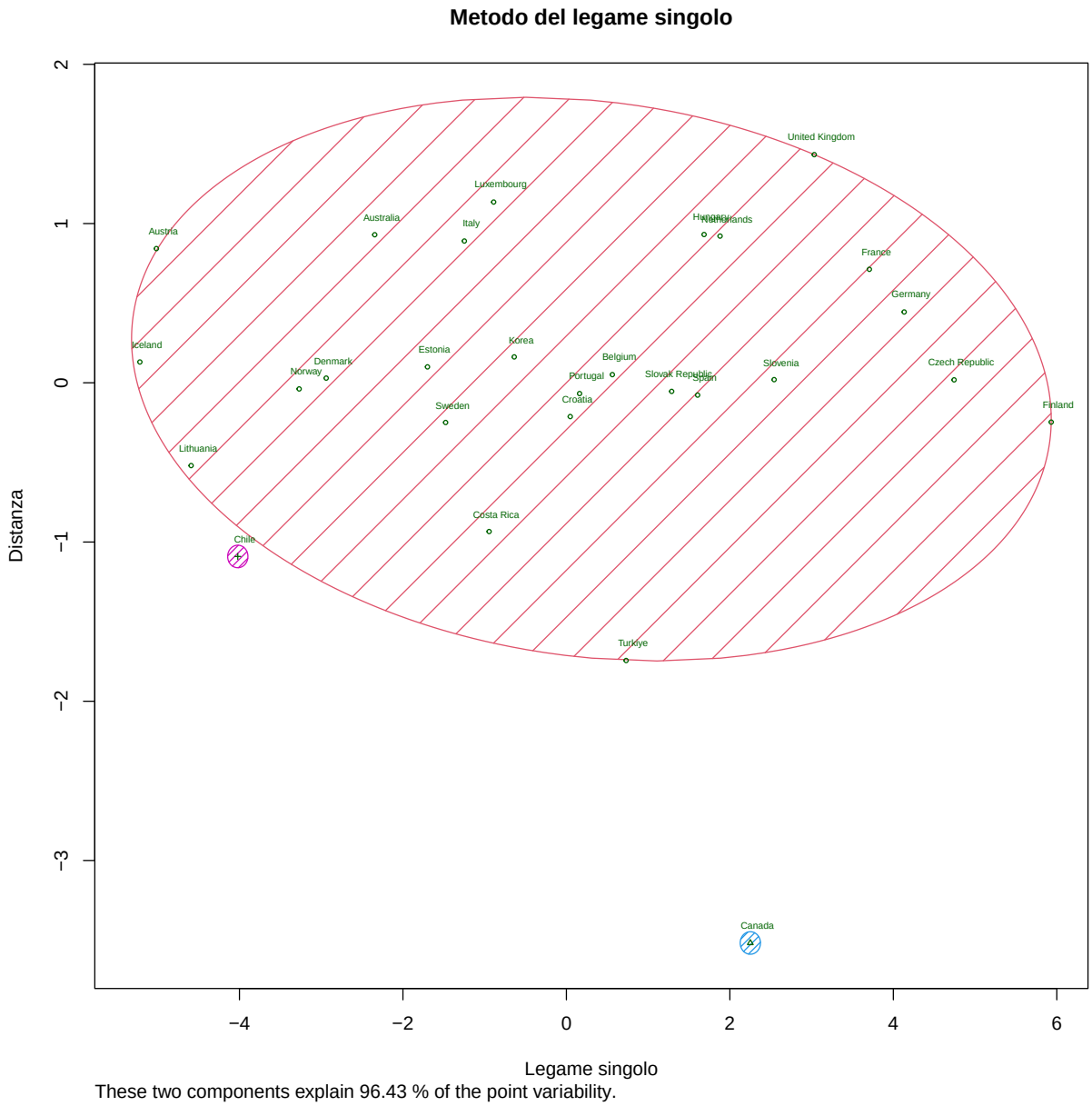
```
## [1] "rapporto tra between e total"
```

```
## [1] 0.1434801
```

I risultati suggeriscono che il rapporto tra between e total risulta minore del 70% (14%), dunque i risultati di questo metodo per la divisione in tre cluster **non è soddisfacente**. Possiamo visualizzare un diagramma di Ven rappresentante i tre cluster e i punti appartenenti agli insiemi.

```
##
## Caricamento pacchetto: 'cluster'
```

```
## Il seguente oggetto è mascherato da 'package:maps':
##
##     votes.repub
```



These two components explain 96.43 % of the point variability.

Dal diagramma possiamo notare che vi è un singolo grande gruppo e due più piccoli, questo potrebbe essere un effetto collaterale del metodo del legame singolo, proprio perchè viene a crearsi un effetto a catena che lega elementi dissimili poichè si guarda solo alla distanza minima.

Metodo del legame completo Nel metodo del legame completo *la distanza tra due gruppi è definita come la massima tra tutte le distanze che si possono calcolare tra ogni individuo del primo gruppo e ogni individuo del secondo.* La massima distanza esistente tra gli individui dei due cluster rappresenta il diametro della sfera che contiene tutti i punti appartenenti ai due gruppi. Tale metodo identifica soprattutto gruppi di forma ellissoidale, ossia una serie di punti che si addensano intorno ad un nucleo centrale.

```
##             Length Class  Mode
## merge       54     -none- numeric
## height      27     -none- numeric
## order       28     -none- numeric
```

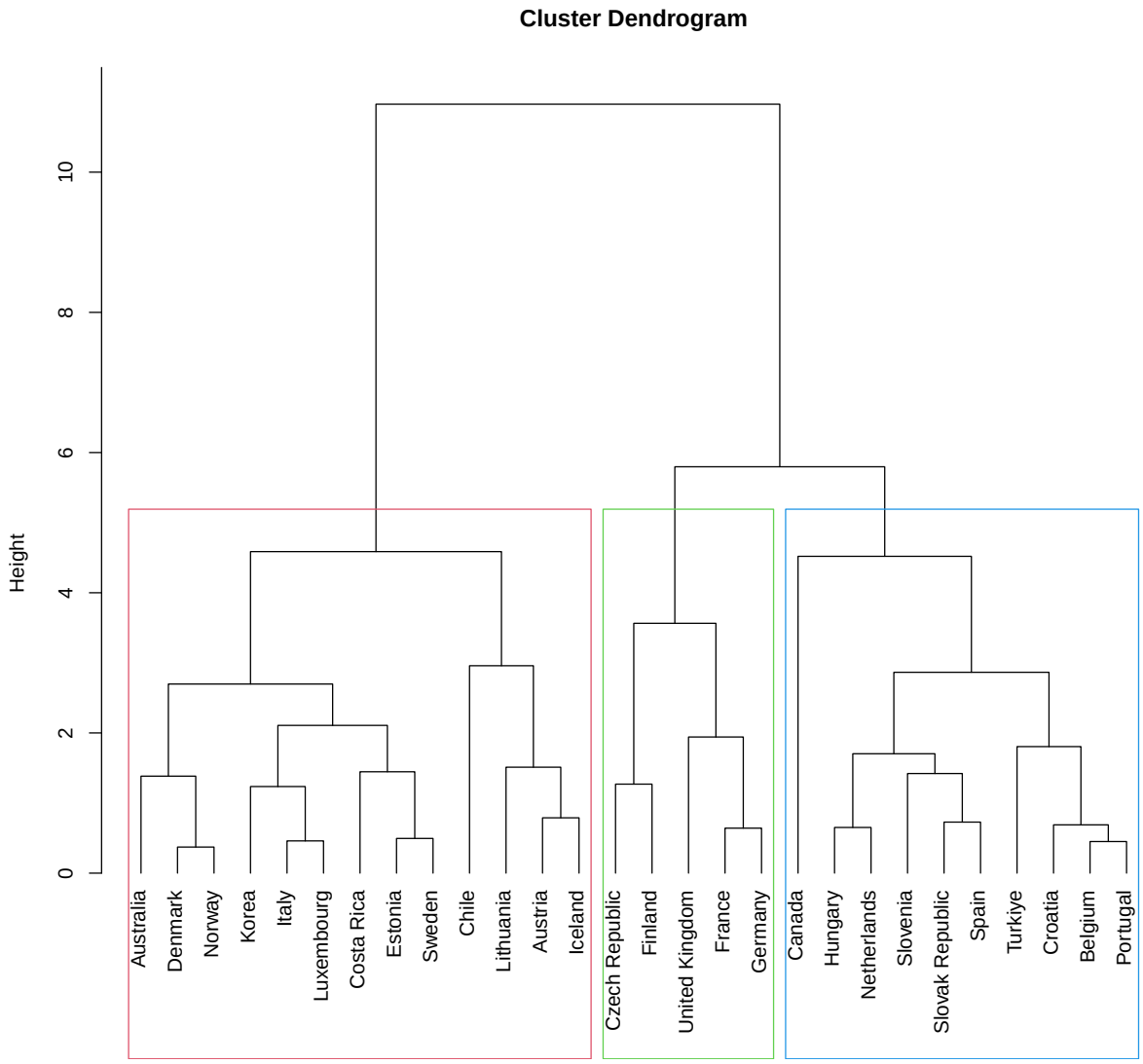
```
## labels      28     -none- character
## method       1     -none- character
## call         3     -none- call
## dist.method  1     -none- character
```

```
## List of 7
##  $ merge      : int [1:27, 1:2] -8 -3 -15 -9 -11 -13 -28 -22 -2 -16 ...
##  $ height     : num [1:27] 0.372 0.451 0.46 0.495 0.642 ...
##  $ order      : int [1:28] 1 8 20 16 15 18 6 9 25 5 ...
##  $ labels     : chr [1:28] "Australia" "Austria" "Belgium" "Canada" ...
##  $ method     : chr "complete"
##  $ call       : language hclust(d = dist_E, method = "complete")
##  $ dist.method: chr "euclidean"
##  - attr(*, "class")= chr "hclust"
```

```
##  [1]  0.3722328  0.4507485  0.4599966  0.4952318  0.6416795  0.6516449
##  [7]  0.6888838  0.7280729  0.7892070  1.2353999  1.2698116  1.3836350
## [13]  1.4211664  1.4461509  1.5114998  1.7033804  1.8051206  1.9420772
## [19]  2.1076577  2.6979460  2.8652190  2.9588638  3.5637052  4.5195566
## [25]  4.5867887  5.7970893 10.9693744
```

```
##       [,1] [,2]
##  [1,]   -8  -20
##  [2,]   -3  -21
##  [3,]  -15  -18
##  [4,]   -9  -25
##  [5,]  -11  -12
##  [6,]  -13  -19
##  [7,]  -28    2
##  [8,]  -22  -24
##  [9,]   -2  -14
## [10,]  -16    3
## [11,]   -7  -10
## [12,]   -1    1
## [13,]  -23    8
## [14,]   -6    4
## [15,]  -17    9
## [16,]    6   13
## [17,]  -26    7
## [18,]  -27    5
## [19,]   10   14
## [20,]   12   19
## [21,]   16   17
## [22,]   -5   15
## [23,]   11   18
## [24,]   -4   21
## [25,]   20   22
## [26,]   23   24
## [27,]   25   26
```

```
##  [1]  0.3722328  0.4507485  0.4599966  0.4952318  0.6416795  0.6516449
##  [7]  0.6888838  0.7280729  0.7892070  1.2353999  1.2698116  1.3836350
## [13]  1.4211664  1.4461509  1.5114998  1.7033804  1.8051206  1.9420772
## [19]  2.1076577  2.6979460  2.8652190  2.9588638  3.5637052  4.5195566
## [25]  4.5867887  5.7970893 10.9693744
```



Metodo gerarchico agglomerativo
del legame completo

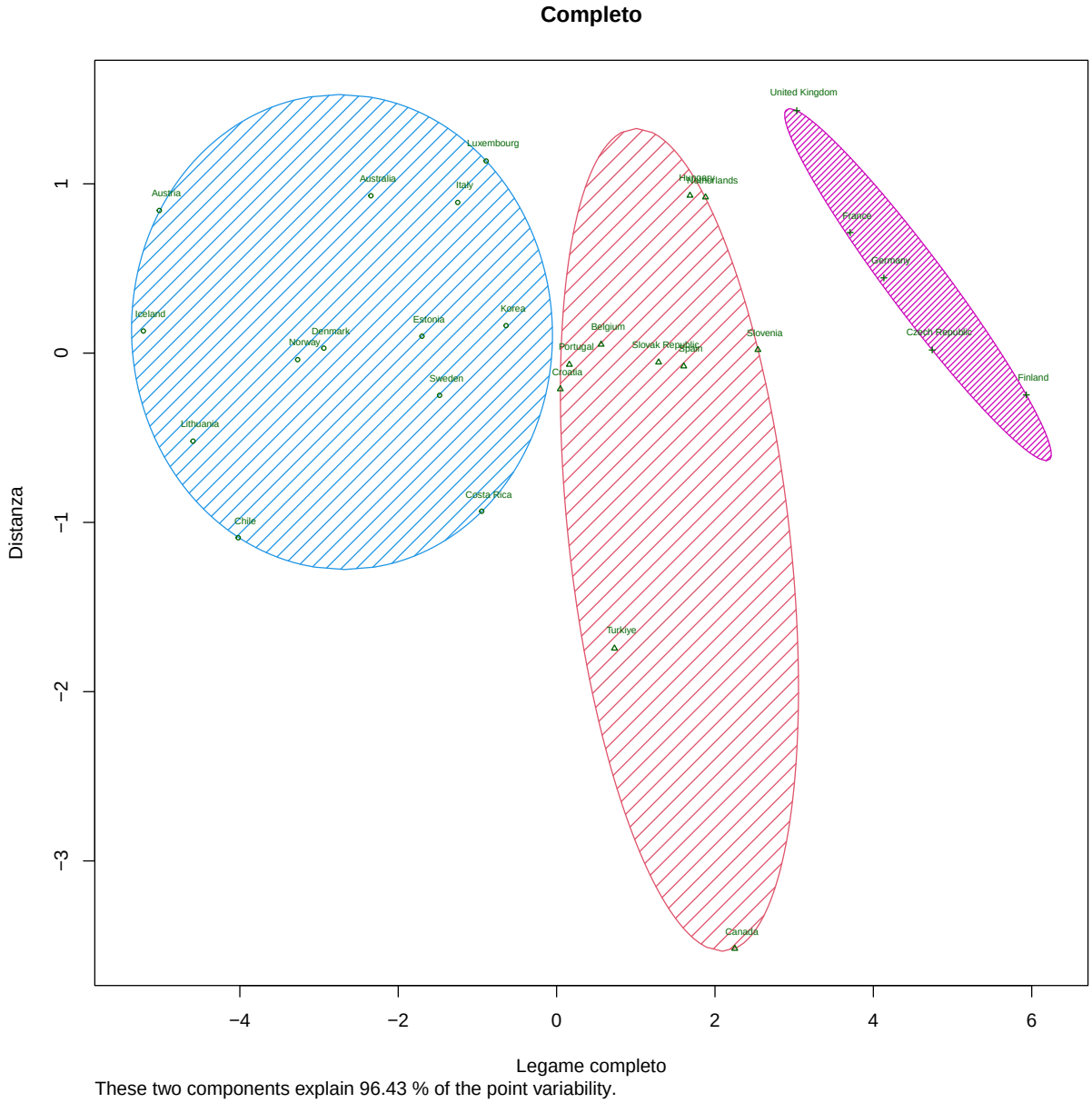
A differenza del dendrogramma ottenuto dall'analisi del metodo a legame singolo, in questo caso abbiamo rami molto più lunghi poichè i gruppi si formano a livelli di distanza maggiori. Attuo adesso il calcolo e l'analisi del rapporto tra between e total.

```
##       Australia         Austria         Belgium          Canada           Chile
##               1               1               2               2               1
##      Costa Rica  Czech Republic         Denmark         Estonia         Finland
##               1               3               1               1               3
##          France         Germany         Hungary         Iceland           Italy
##               3               3               2               1               1
##           Korea       Lithuania      Luxembourg     Netherlands          Norway
##               1               1               1               2               1
##        Portugal Slovak Republic        Slovenia           Spain          Sweden
##               2               2               2               2               1
##         Turkiye  United Kingdom         Croatia
```

```
##               2               3               2
## [[1]]
##       Australia         Austria         Belgium          Canada           Chile
##               1               1               2               2               1
##      Costa Rica  Czech Republic         Denmark         Estonia         Finland
##               1               3               1               1               3
##          France         Germany         Hungary         Iceland           Italy
##               3               3               2               1               1
##           Korea       Lithuania      Luxembourg     Netherlands          Norway
##               1               1               1               2               1
##        Portugal Slovak Republic        Slovenia           Spain          Sweden
##               2               2               2               2               1
##         Turkiye  United Kingdom         Croatia
##               2               3               2
## [1] "misura di non omogeneità del primo gruppo"
## [1] 42.45709
## [1] "misura di non omogeneità del secondo gruppo"
## [1] 24.13424
## [1] "misura di non omogeneità del terzo gruppo"
## [1] 7.911562
## [1] "misura di non omogeneità interna"
## [1] 74.50289
## [1] "misura di non omogeneità tra i cluster"
## [1] 195.4971
## [1] "rapporto tra between e total"
## [1] 0.7240634
```

I risultati suggeriscono che il rapporto tra between e total risulta maggiore del 70% (72%) e quindi i risultati di tale metodo sono **soddisfacenti** per la suddivisione in tre cluster.

Possiamo visualizzare un diagramma di Ven rappresentante i tre cluster e i punti appartenenti agli insiemi.



Notiamo che nel metodo del legame completo i gruppi hanno una forma ellissoidale e i primi due cluster sono molto distanti dal terzo.

Metodo del legame medio Nel metodo del legame medio *la distanza tra due gruppi è definita come la media aritmetica delle distanze tra tutte le coppie di unità che compongono i due gruppi.* Uno svantaggio di tale metodo è che se le misure dei due cluster da unire sono molto differenti la distanza sarà molto vicina a quella del cluster più numeroso.

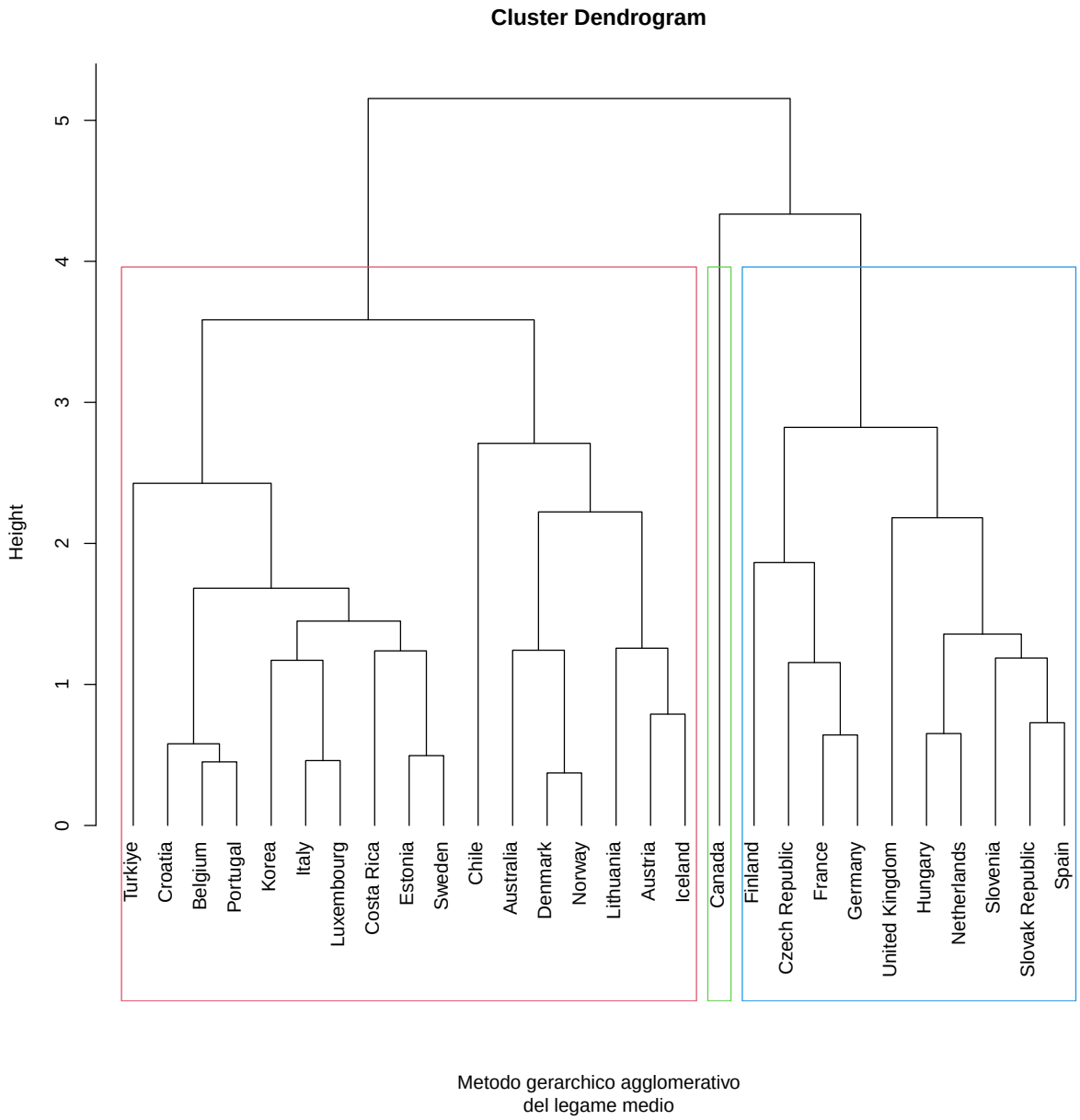
```
##             Length Class  Mode
## merge       54     -none- numeric
## height      27     -none- numeric
## order       28     -none- numeric
## labels      28     -none- character
## method       1     -none- character
```

```
## call         3     -none- call
## dist.method  1     -none- character
```

```
## List of 7
##  $ merge      : int [1:27, 1:2] -8 -3 -15 -9 -28 -11 -13 -22 -2 -7 ...
##  $ height     : num [1:27] 0.372 0.451 0.46 0.495 0.579 ...
##  $ order      : int [1:28] 26 28 3 21 16 15 18 6 9 25 ...
##  $ labels     : chr [1:28] "Australia" "Austria" "Belgium" "Canada" ...
##  $ method     : chr "average"
##  $ call       : language hclust(d = dist_E, method = "average")
##  $ dist.method: chr "euclidean"
##  - attr(*, "class")= chr "hclust"
```

```
##       [,1] [,2]
##  [1,]   -8  -20
##  [2,]   -3  -21
##  [3,]  -15  -18
##  [4,]   -9  -25
##  [5,]  -28    2
##  [6,]  -11  -12
##  [7,]  -13  -19
##  [8,]  -22  -24
##  [9,]   -2  -14
## [10,]   -7    6
## [11,]  -16    3
## [12,]  -23    8
## [13,]   -6    4
## [14,]   -1    1
## [15,]  -17    9
## [16,]    7   12
## [17,]   11   13
## [18,]    5   17
## [19,]  -10   10
## [20,]  -27   16
## [21,]   14   15
## [22,]  -26   18
## [23,]   -5   21
## [24,]   19   20
## [25,]   22   23
## [26,]   -4   24
## [27,]   25   26
```

```
##  [1] 0.3722328 0.4507485 0.4599966 0.4952318 0.5790316 0.6416795 0.6516449
##  [8] 0.7280729 0.7892070 1.1552365 1.1710889 1.1870457 1.2378523 1.2424584
## [15] 1.2568893 1.3568189 1.4489837 1.6818998 1.8639239 2.1819543 2.2232967
## [22] 2.4261016 2.7094455 2.8227178 3.5851967 4.3351340 5.1545709
```

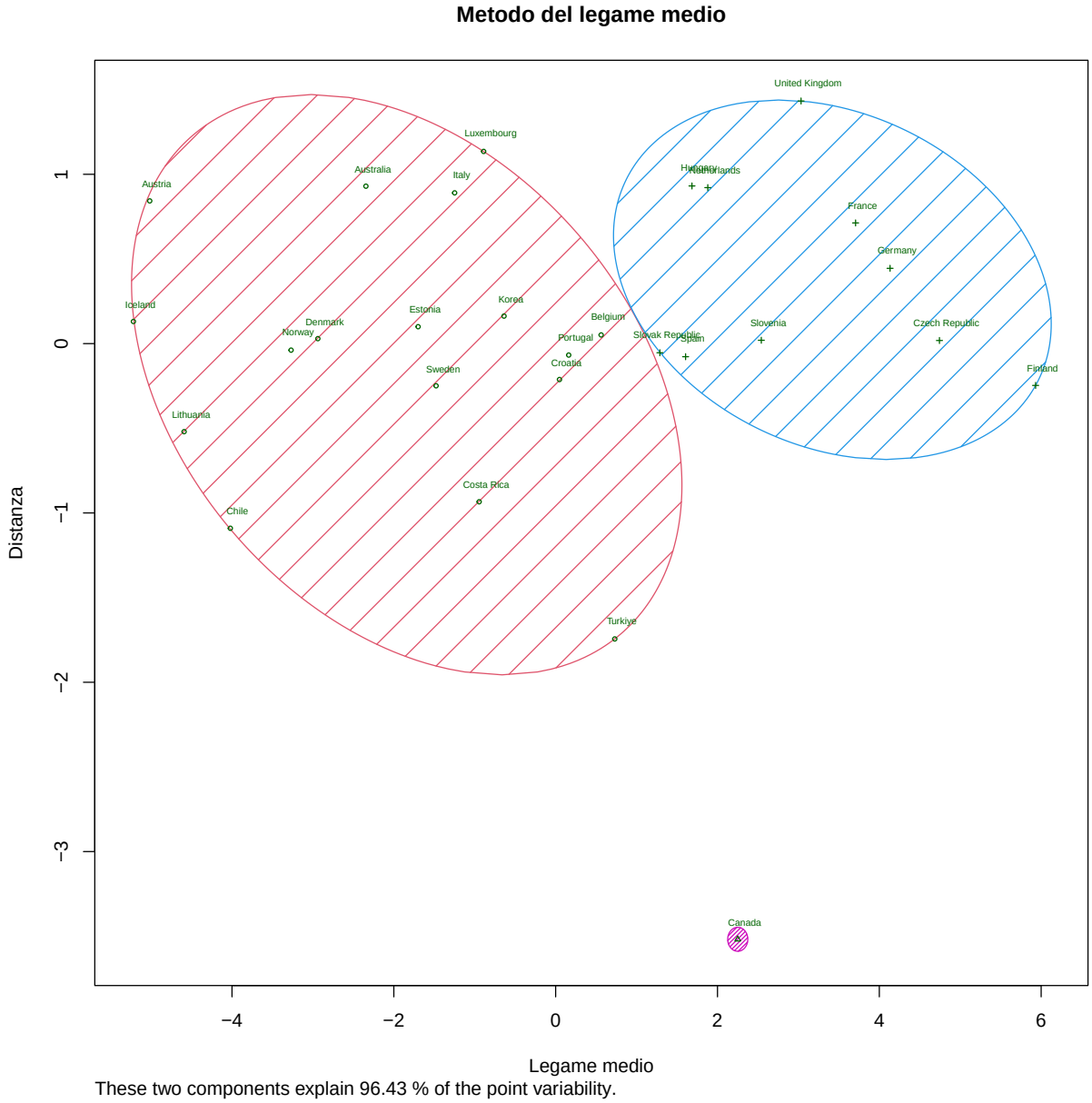


Attuo adesso il calcolo e l'analisi del rapporto tra between e total.

```
##       Australia         Austria         Belgium          Canada           Chile
##               1               1               1               2               1
##      Costa Rica  Czech Republic         Denmark         Estonia         Finland
##               1               3               1               1               3
##          France         Germany         Hungary         Iceland           Italy
##               3               3               3               1               1
##           Korea       Lithuania      Luxembourg     Netherlands          Norway
##               1               1               1               3               1
##        Portugal Slovak Republic        Slovenia           Spain          Sweden
##               1               3               3               3               1
##         Turkiye  United Kingdom         Croatia
##               1               3               1
```

```
## [[1]]
##       Australia         Austria         Belgium          Canada           Chile
##               1               1               1               2               1
##      Costa Rica  Czech Republic         Denmark         Estonia         Finland
##               1               3               1               1               3
##          France         Germany         Hungary         Iceland           Italy
##               3               3               3               1               1
##           Korea       Lithuania      Luxembourg     Netherlands          Norway
##               1               1               1               3               1
##        Portugal Slovak Republic        Slovenia           Spain          Sweden
##               1               3               3               3               1
##         Turkiye  United Kingdom         Croatia
##               1               3               1

## [1] "misura di non omogeneità del primo gruppo"

## [1] 73.14755

## [1] "misura di non omogeneità del secondo gruppo"

## [1] 0

## [1] "misura di non omogeneità del terzo gruppo"

## [1] 26.47352

## [1] "misura di non omogeneità interna"

## [1] 99.62107

## [1] "misura di non omogeneità tra i cluster"

## [1] 170.3789

## [1] "rapporto tra between e total"

## [1] 0.6310331
```

I risultati suggeriscono che il rapporto tra between e total risulta minore del 70% (63%) e quindi i risultati di tale metodo **sono soddisfacenti** per la suddivisione in tre cluster. Possiamo visualizzare un diagramma di Ven rappresentante i tre cluster e i punti appartenenti agli insiemi.



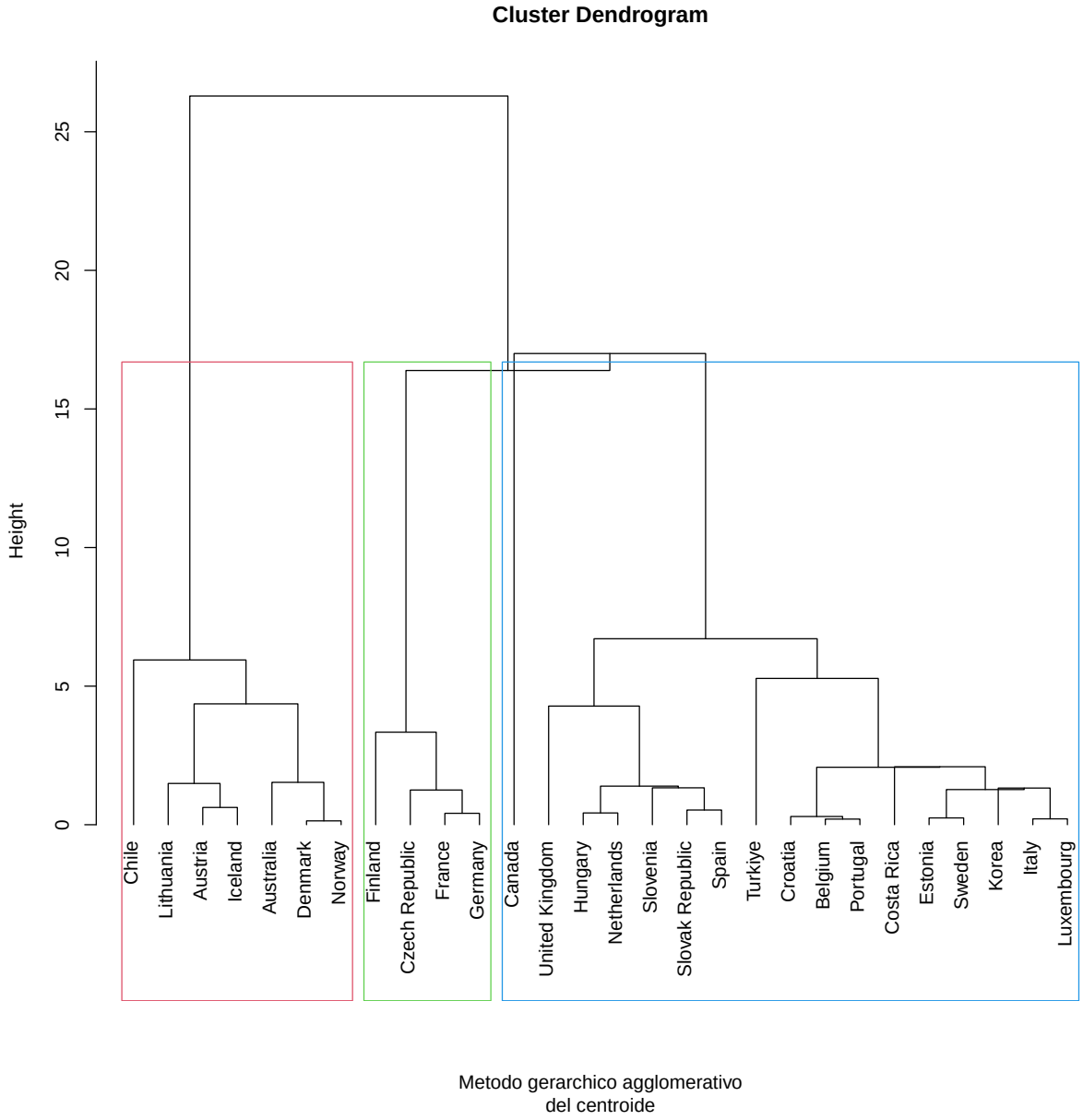
Metodo del centroide Nel metodo del centroide la distanza tra i due gruppi è definita come *la distanza tra i centroidi, ossia tra le medie campionarie calcolate sugli individui appartenenti ai due gruppi.* Il metodo del centroide può dare origine a fenomeni gravitazionali, per cui i gruppi grandi tendono ad attrarre al loro interno i piccoli gruppi.

```
##             Length Class  Mode
## merge       54     -none- numeric
## height      27     -none- numeric
## order       28     -none- numeric
## labels      28     -none- character
## method       1     -none- character
## call         3     -none- call
## dist.method  1     -none- character
```

```
## List of 7
##  $ merge      : int [1:27, 1:2] -8 -3 -15 -9 -28 -11 -13 -22 -2 -7 ...
##  $ height     : num [1:27] 0.139 0.203 0.212 0.245 0.297 ...
##  $ order      : int [1:28] 5 17 2 14 1 8 20 10 7 11 ...
##  $ labels     : chr [1:28] "Australia" "Austria" "Belgium" "Canada" ...
##  $ method     : chr "centroid"
##  $ call       : language hclust(d = (dist_E)^2, method = "centroid")
##  $ dist.method: chr "euclidean"
##  - attr(*, "class")= chr "hclust"
```

```
##       [,1] [,2]
##  [1,]   -8  -20
##  [2,]   -3  -21
##  [3,]  -15  -18
##  [4,]   -9  -25
##  [5,]  -28    2
##  [6,]  -11  -12
##  [7,]  -13  -19
##  [8,]  -22  -24
##  [9,]   -2  -14
## [10,]   -7    6
## [11,]  -16    3
## [12,]    4   11
## [13,]  -23    8
## [14,]    7   13
## [15,]  -17    9
## [16,]   -1    1
## [17,]   -6   12
## [18,]    5   17
## [19,]  -10   10
## [20,]  -27   14
## [21,]   15   16
## [22,]  -26   18
## [23,]   -5   21
## [24,]   20   22
## [25,]   -4   24
## [26,]   19   25
## [27,]   23   26
```

```
##  [1]  0.1385572  0.2031743  0.2115969  0.2452546  0.2965515  0.4117526
##  [7]  0.4246411  0.5300902  0.6228476  1.2520524  1.3226859  1.2688719
## [13]  1.3313674  1.3917049  1.4888853  1.5289944  2.0912609  2.0734923
## [19]  3.3401468  4.2810963  4.3604984  5.2778357  5.9438358  6.7094288
## [25] 17.0001112 16.3875076 26.2908529
```

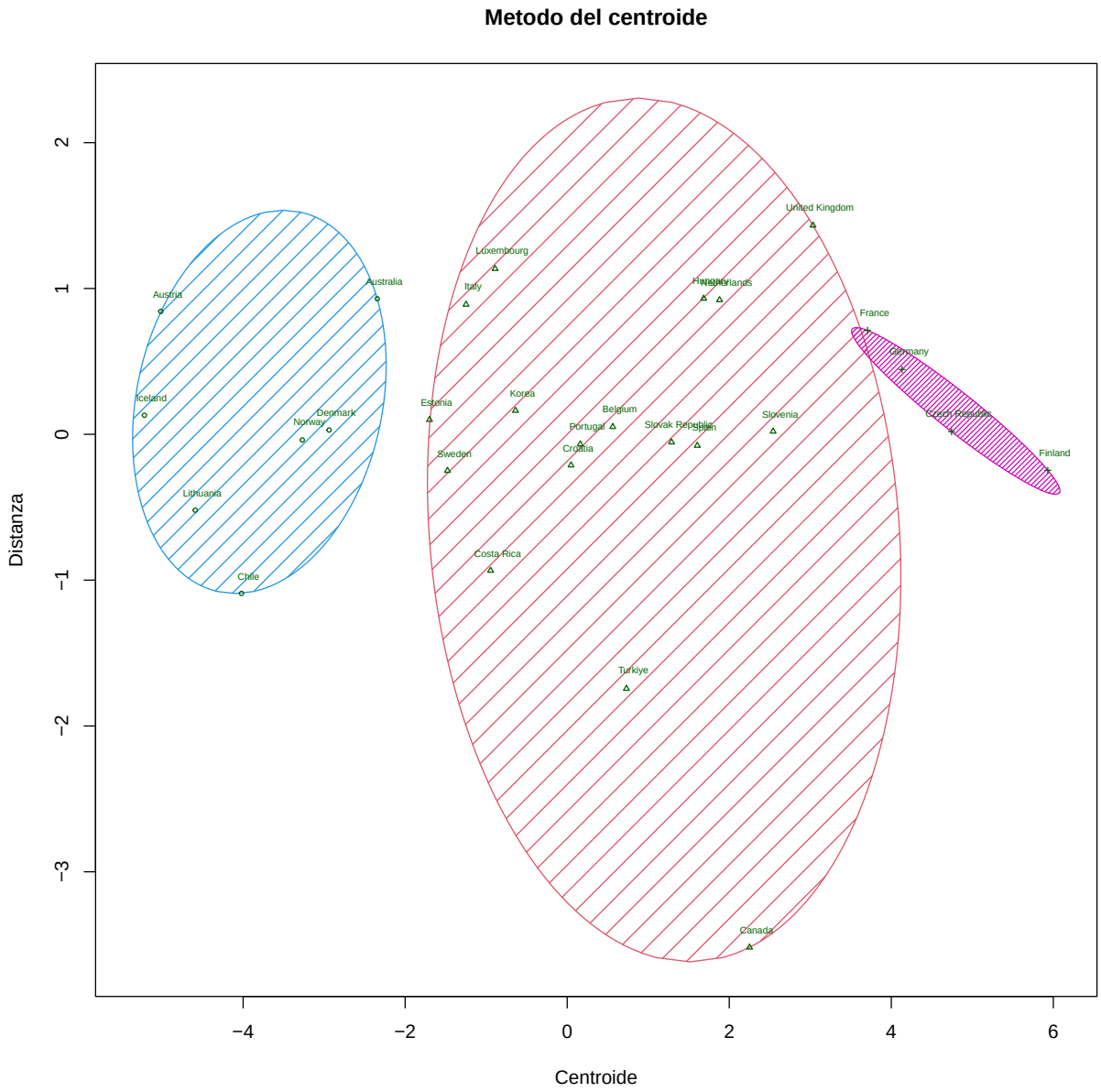


Attuo adesso il calcolo e l'analisi del rapporto tra between e total.

```
##       Australia         Austria         Belgium          Canada           Chile
##               1               1               2               2               1
##      Costa Rica  Czech Republic         Denmark         Estonia         Finland
##               2               3               1               2               3
##          France         Germany         Hungary         Iceland           Italy
##               3               3               2               1               2
##           Korea       Lithuania      Luxembourg     Netherlands          Norway
##               2               1               2               2               1
##        Portugal Slovak Republic        Slovenia           Spain          Sweden
##               2               2               2               2               2
##         Turkiye  United Kingdom         Croatia
##               2               2               2
```

```
## [[1]]
##       Australia         Austria         Belgium          Canada           Chile
##               1               1               2               2               1
##      Costa Rica  Czech Republic         Denmark         Estonia         Finland
##               2               3               1               2               3
##          France         Germany         Hungary         Iceland           Italy
##               3               3               2               1               2
##           Korea       Lithuania      Luxembourg     Netherlands          Norway
##               2               1               2               2               1
##        Portugal Slovak Republic        Slovenia           Spain          Sweden
##               2               2               2               2               2
##         Turkiye  United Kingdom         Croatia
##               2               2               2

## [1] "misura di non omogeneità del primo gruppo"

## [1] 14.02809

## [1] "misura di non omogeneità del secondo gruppo"

## [1] 61.33494

## [1] "misura di non omogeneità del secondo gruppo"

## [1] 3.545688

## [1] "misura di non omogeneità interna"

## [1] 78.90871

## [1] "misura di non omogeneità tra i cluster"

## [1] 191.0913

## [1] "rapporto tra between e total"

## [1] 0.7077455
```

I risultati suggeriscono che il rapporto tra between e total risulta maggiore del 70% (70.7%) quindi i risultati di tale metodo sono **soddisfacenti** per la suddivisione in tre cluster. Possiamo visualizzare un diagramma di Ven rappresentante i tre cluster e i punti appartenenti agli insiemi.



These two components explain 96.43 % of the point variability.

Metodo della mediana Il metodo della mediana è simile a quello del centroide, con la differenza che *la procedura è indipendente dalla numerosità dei cluster.* Infatti, quando due gruppi si aggregano, il nuovo centroide è calcolato come la semisomma dei due centroidi precedenti.

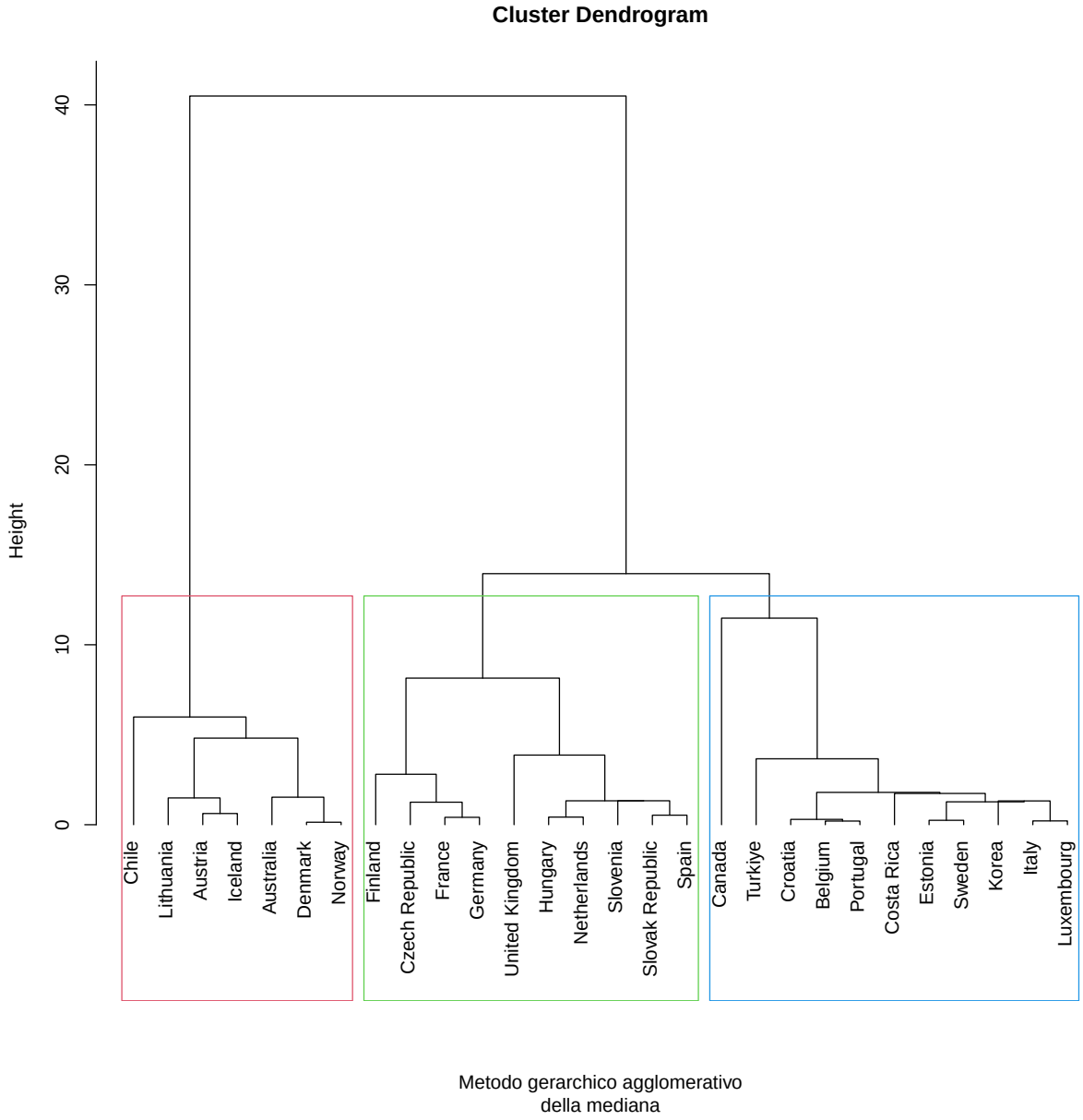
```
##             Length Class  Mode
## merge       54     -none- numeric
## height      27     -none- numeric
## order       28     -none- numeric
## labels      28     -none- character
## method       1     -none- character
## call         3     -none- call
## dist.method  1     -none- character
```

```
## List of 7
```

```
##  $ merge      : int [1:27, 1:2] -8 -3 -15 -9 -28 -11 -13 -22 -2 -7 ...
##  $ height     : num [1:27] 0.139 0.203 0.212 0.245 0.297 ...
##  $ order      : int [1:28] 5 17 2 14 1 8 20 10 7 11 ...
##  $ labels     : chr [1:28] "Australia" "Austria" "Belgium" "Canada" ...
##  $ method     : chr "median"
##  $ call       : language hclust(d = (dist_E)^2, method = "median")
##  $ dist.method: chr "euclidean"
##  - attr(*, "class")= chr "hclust"
```

```
##       [,1] [,2]
##  [1,]   -8  -20
##  [2,]   -3  -21
##  [3,]  -15  -18
##  [4,]   -9  -25
##  [5,]  -28    2
##  [6,]  -11  -12
##  [7,]  -13  -19
##  [8,]  -22  -24
##  [9,]   -2  -14
## [10,]   -7    6
## [11,]  -16    3
## [12,]    4   11
## [13,]  -23    8
## [14,]    7   13
## [15,]  -17    9
## [16,]   -1    1
## [17,]   -6   12
## [18,]    5   17
## [19,]  -10   10
## [20,]  -26   18
## [21,]  -27   14
## [22,]   15   16
## [23,]   -5   22
## [24,]   19   21
## [25,]   -4   20
## [26,]   24   25
## [27,]   23   26
```

```
##  [1]  0.1385572  0.2031743  0.2115969  0.2452546  0.2965515  0.4117526
##  [7]  0.4246411  0.5300902  0.6228476  1.2520524  1.3226859  1.2722555
## [13]  1.3313674  1.3294581  1.4888853  1.5289944  1.7383785  1.7974079
## [19]  2.8038778  3.6625691  3.8702546  4.8117740  5.9864814  8.1457681
## [25] 11.4792899 13.9512091 40.4952073
```



Attuo adesso il calcolo e l'analisi del rapporto tra between e total.

```
##       Australia         Austria         Belgium          Canada           Chile
##               1               1               2               2               1
##      Costa Rica  Czech Republic         Denmark         Estonia         Finland
##               2               3               1               2               3
##          France         Germany         Hungary         Iceland           Italy
##               3               3               3               1               2
##           Korea       Lithuania      Luxembourg     Netherlands          Norway
##               2               1               2               3               1
##        Portugal Slovak Republic        Slovenia           Spain          Sweden
##               2               3               3               3               2
##         Turkiye  United Kingdom         Croatia
##               2               3               2
```

```
## [[1]]
##       Australia         Austria         Belgium          Canada           Chile
##               1               1               2               2               1
##      Costa Rica  Czech Republic         Denmark         Estonia         Finland
##               2               3               1               2               3
##          France         Germany         Hungary         Iceland           Italy
##               3               3               3               1               2
##           Korea       Lithuania      Luxembourg     Netherlands          Norway
##               2               1               2               3               1
##        Portugal Slovak Republic        Slovenia           Spain          Sweden
##               2               3               3               3               2
##         Turkiye  United Kingdom         Croatia
##               2               3               2
```

```
##    Anno2012   Anno2013   Anno2014   Anno2015  Anno2016  Anno2017  Anno2018
## 1 0.4939001 0.19704038 0.17344491 0.22630799 0.1062763 0.2063988 0.1397848
## 2 0.1115263 0.08916383 0.06089092 0.08669747 0.1858443 0.2040782 0.2378230
## 3 0.3412929 0.28640830 0.23449241 0.27743909 0.2383998 0.1877621 0.3366674
##    Anno2019  Anno2020  Anno2021
## 1 0.1518938 0.3241506 0.3188167
## 2 0.2736504 0.7588655 1.1418025
## 3 0.3570880 0.3637034 0.3182485
```

```
## [1] "misura di non omogeneità del primo gruppo"
```

```
## [1] 14.02809
```

```
## [1] "misura di non omogeneità del secondo gruppo"
```

```
## [1] 31.50342
```

```
## [1] "misura di non omogeneità del terzo gruppo"
```

```
## [1] 26.47352
```

```
## [1] "misura di non omogeneità interna"
```

```
## [1] 72.00503
```

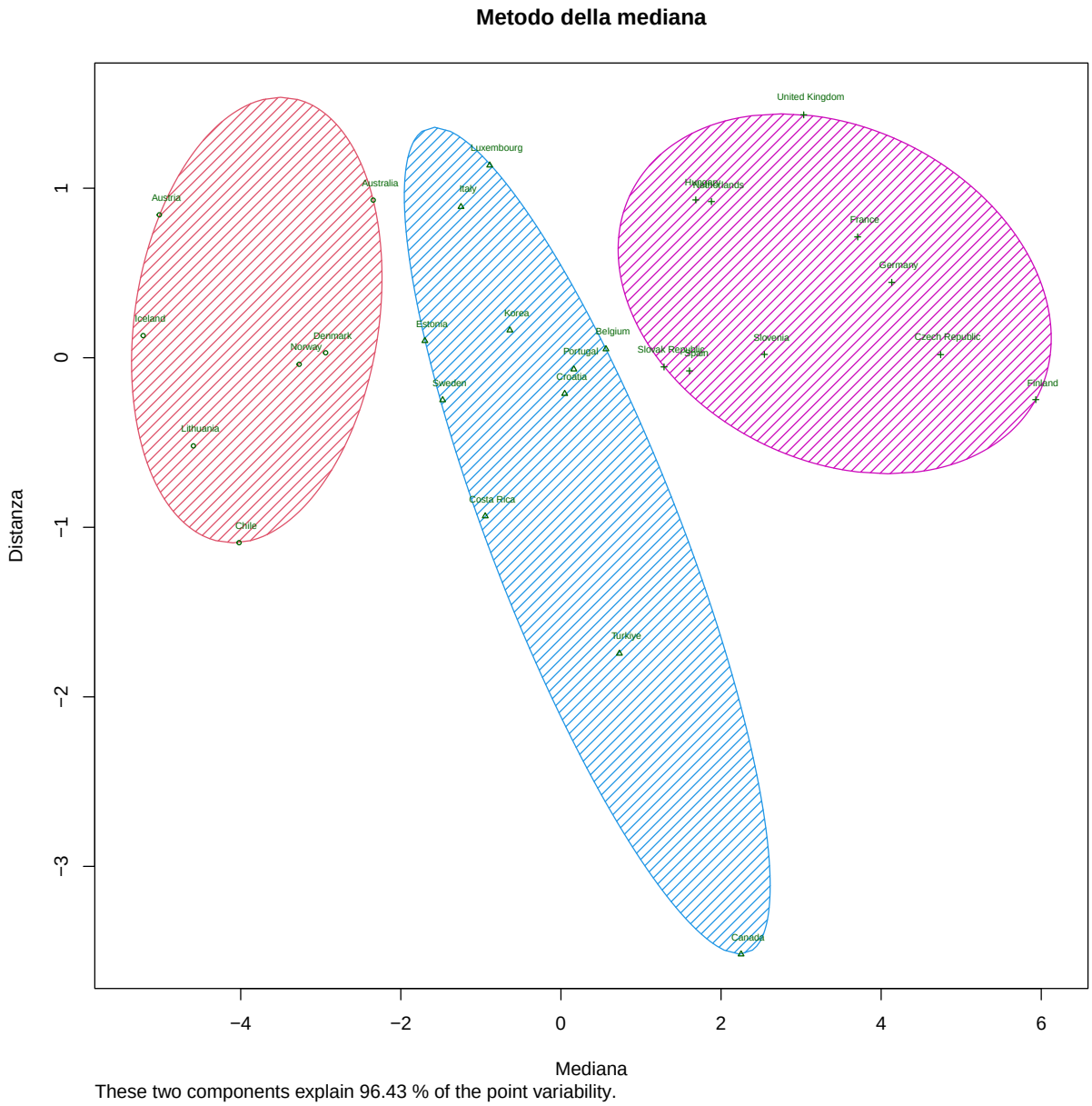
```
## [1] "misura di non omogeneità tra i cluster"
```

```
## [1] 197.995
```

```
## [1] "rapporto tra between e total"
```

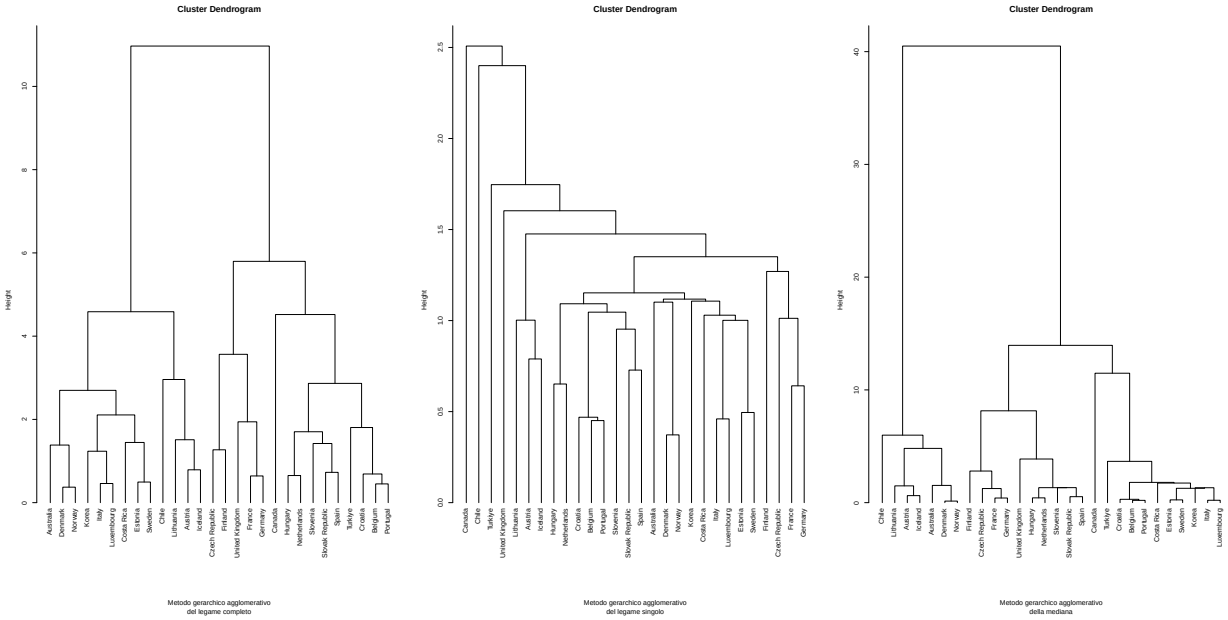
```
## [1] 0.7333147
```

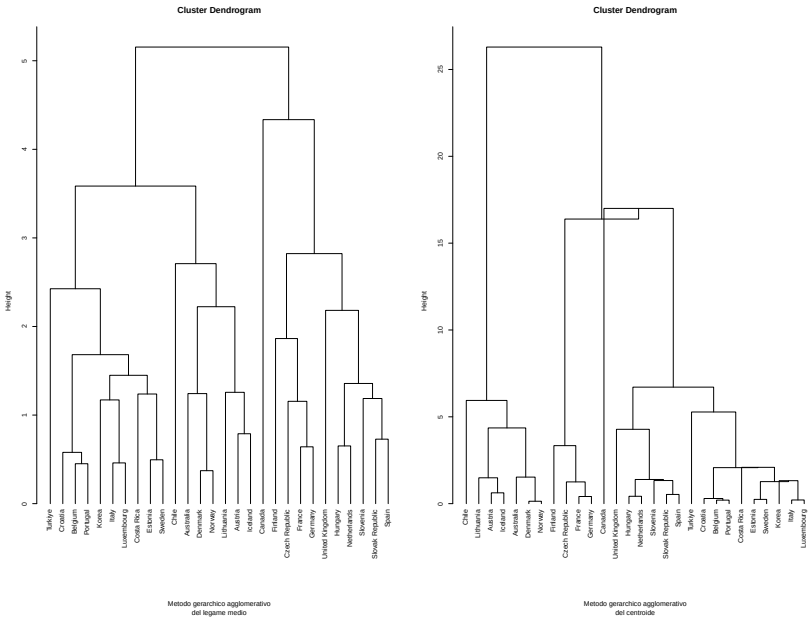
I risultati suggeriscono che il rapporto tra between e total risulta maggiore del 70% (73%) e quindi i risultati di tale metodo sono **soddisfacenti** per la suddivisione in tre cluster. Possiamo visualizzare un diagramma di Ven rappresentante i tre cluster e i punti appartenenti agli insiemi.



These two components explain 96.43 % of the point variability.

Voglio ora visualizzare tutti i dendrogrammi per riuscire a confrontarli





Conclusioni

	Metodo	Risultato between/total
Non gerarchico		74.5%
Gerarchico	del legame completo	72.4%
Gerarchico	del legame singolo	14.4%
Gerarchico	del legame medio	63.1%
Gerarchico	del centroide	70.7%
Gerarchico	della mediana	73.3%

Seconda parte

Scelta della variabile aleatoria

Nella seconda parte si vuole usare la *statistica inferenziale* che fa uso non solo dei dati ma anche della probabilità. Nelle analisi statistiche si analizza un fenomeno che si manifesta su una popolazione che può essere finita o infinita. La conoscenza delle caratteristiche di una popolazione finita può essere ottenuta osservando la totalità delle entità della popolazione oppure un sottoinsieme di essa che è chiamato campione estratto. L'**inferenza statistica** ha lo scopo di estendere le misure ricavate dall'esame di un campione alla popolazione da cui il campione è stato estratto. *Con l'inferenza statistica si desidera studiare una popolazione che risulta essere descritta da una variabile aleatoria.* Tale variabile deve essere osservabile, ovvero devo essere in grado di osservare i valori di tale variabile. La legge di probabilità deve contenere dei parametri che non sono noti e l'obiettivo è avere informazioni su questi parametri. Tali informazioni sono ricavabili estraendo dalla popolazione un campione ed effettuare le misure su tale campione per poi trasferirle all'intera popolazione. Affinchè le conclusioni dell'inferenza statistica siano valide, il campione deve essere scelto in modo tale da essere **rappresentativo** della popolazione. L'inferenza statistica si basa su due metodi fondamentali di indagine: la **stima dei parametri** e la **verifica delle ipotesi**.

Problema

Sono analizzati due server in base al numero di email che ricevono per un fissato numero di giorni. Si registrano le email arrivate al server1 per 30 giorni distinti e al server2 per 60 giorni distinti. Entrambi i campioni sono descritti da una **variabile aleatoria di Poisson**.

Distribuzione di Poisson

Una **variabile aleatoria discreta** X assume un numero finito o al più numerabile di valori $x_1, x_2, ...$ con rispettive probabilità $p_X(x_1), p_X(x_2), ...$ essendo $p_X(x_i) = P(X = x_i)$. Tramite R è possibile calcolare la funzione di probabilità, la funzione di distribuzione, la funzione per calcolare i quantili e la funzione che simula la variabile aleatoria mediante la generazione di sequenze di numeri pseudocasuali. Analizziamo prima una variabile aleatoria di Poisson, per poi passare al nostro problema. La distribuzione di Poisson interviene spesso nella descrizione di alcuni fenomeni coinvolgenti qualche tipo di conteggio, nel nostro caso il numero di email in arrivo nei server.

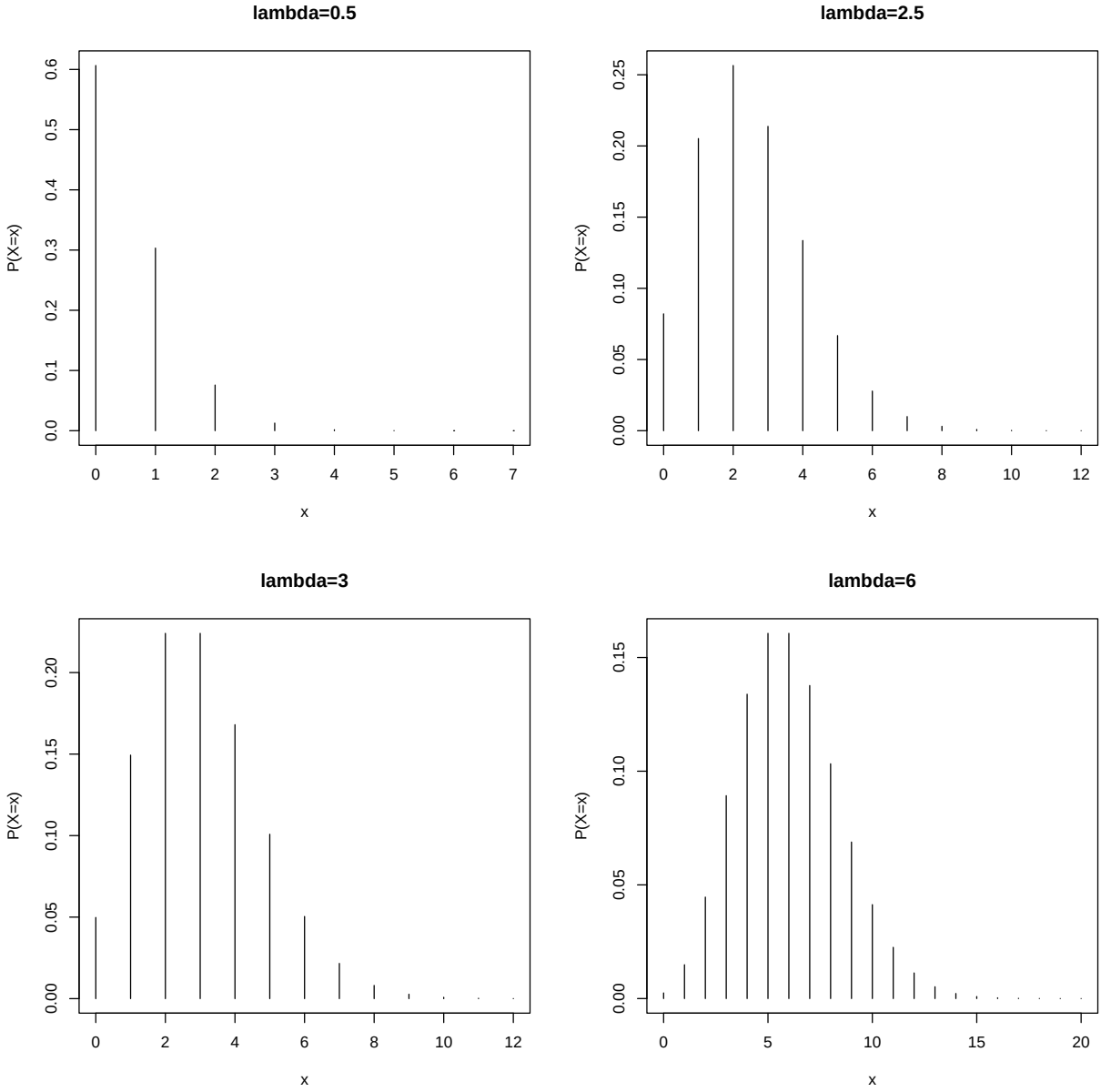
Una variabile aleatoria X avente funzione di probabilità $p_X(x) = P(X = x) = \begin{cases} \left(\frac{\lambda^x}{x!}\right) e^{-\lambda}, x = 0, 1, ...(\lambda > 0), \\ 0, altrimenti \end{cases}$

è detta di distribuzione di Poisson di parametro λ. Per una variabile aleatoria di Poisson $X \sim P(\lambda)$ si ha $E(X) = \lambda$ e $Var(X) = \lambda$.

Calcoliamo le probabilità di Poisson con la funzione *dpois(x, lambda)* dove x è il valore assunto dalla variabile aleatoria di Poisson considerata (nel nostro caso x = 0,1,...,7) e lambda è il vettore dei valori medi (nel nostro caso $\lambda = 2$).

```
## [1] 0.135335283 0.270670566 0.270670566 0.180447044 0.090223522 0.036089409
## [7] 0.012029803 0.003437087
```

Possiamo poi confrontare la funzione di probabilità di Poisson per alcune scelte di λ .

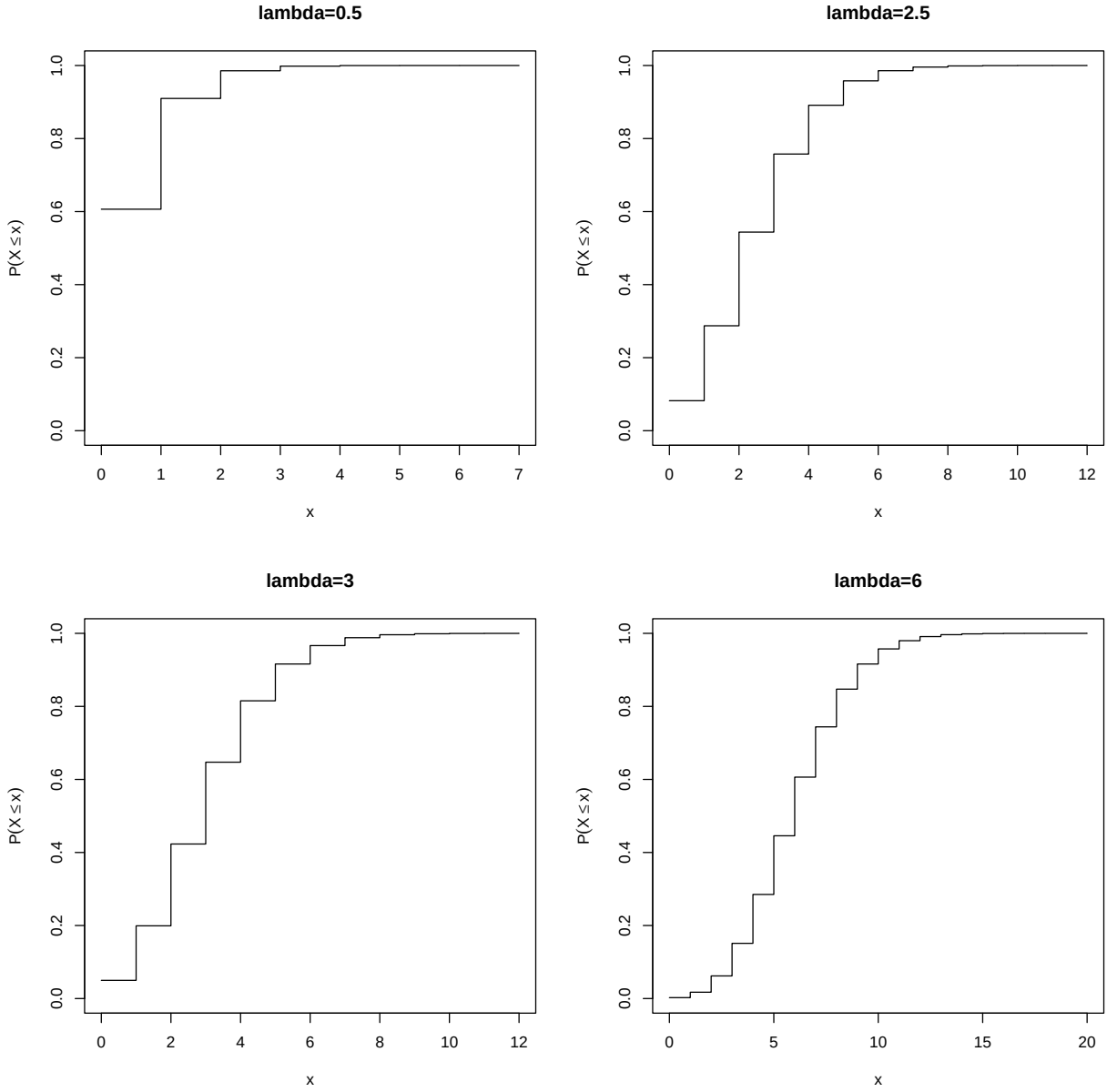


Nel primo caso ($\lambda = 0.5$), $p_X(x)$ è strettamente decrescente; nel secondo caso con $\lambda = 2.5$ notiamo che è presente un unico massimo in x=2; negli ultimi due casi avremo due massimi.

Per il calcolo della **funzione di distribuzione** si utilizza la funzione *ppois(x, lambda, lower.tail=TRUE)* dove se lower.tail è TRUE calcola $P(X \leq x)$, altrimenti calcola $P(X > x)$.

```
## [1] 0.6065307 0.9097960 0.9856123 0.9982484 0.9998279 0.9999858 0.9999990
## [8] 0.9999999
```

Possiamo visualizzare le diverse funzioni di distribuzione.



In R è possibile calcolare anche i quantili (percentili) della distribuzione di Poisson mediante la funzione *qpois(z, lambda)* dove z è il valore assunto dalla probabilità relativa al percentile z*100-esimo. Il risultato di tale funzione è il percentile z*100-esimo, ossia il più piccolo numero intero k assunto dalla variabile aleatoria di Poisson X tale che $P(X \leq k) \geq z$.

```
## [1]   0   1   2   3 Inf
```

In questo caso per $\lambda = 2.5$ il primo quartile (25-esimo percentile) è $Q_1 = 1$, il secondo quartile o mediana (50-esimo percentile) è $Q_2 = 2$ e il terzo quartile (75-esimo percentile) è $Q_3 = 3$. Il minimo è $Q_0 = 0$ e il massimo è $Q_4 = \infty$.

E' possibile inoltre simulare la variabile aleatoria di Poisson generando una sequenza di numeri pseudocasuali mediante la funzione *rpois(N,lambda)* dove N è la lunghezza della sequenza da generare e calcolarne le frequenze relative. Vogliamo generare una sequenza di 60 numeri pseudocasuali simulando una variabile aleatoria di Poisson di valore medio $\lambda = 2.5$

```
##  [1] 2 1 3 2 3 3 1 2 4 1 2 3 0 1 2 2 2 2 0 1 4 4 1 2 2 2 2 2 4 5 3 6 4 2 2 4 6 1
## [39] 0 1 6 8 5 2 3 3 2 0 5 6 1 3 3 8 6 4 4 2 1 6
```

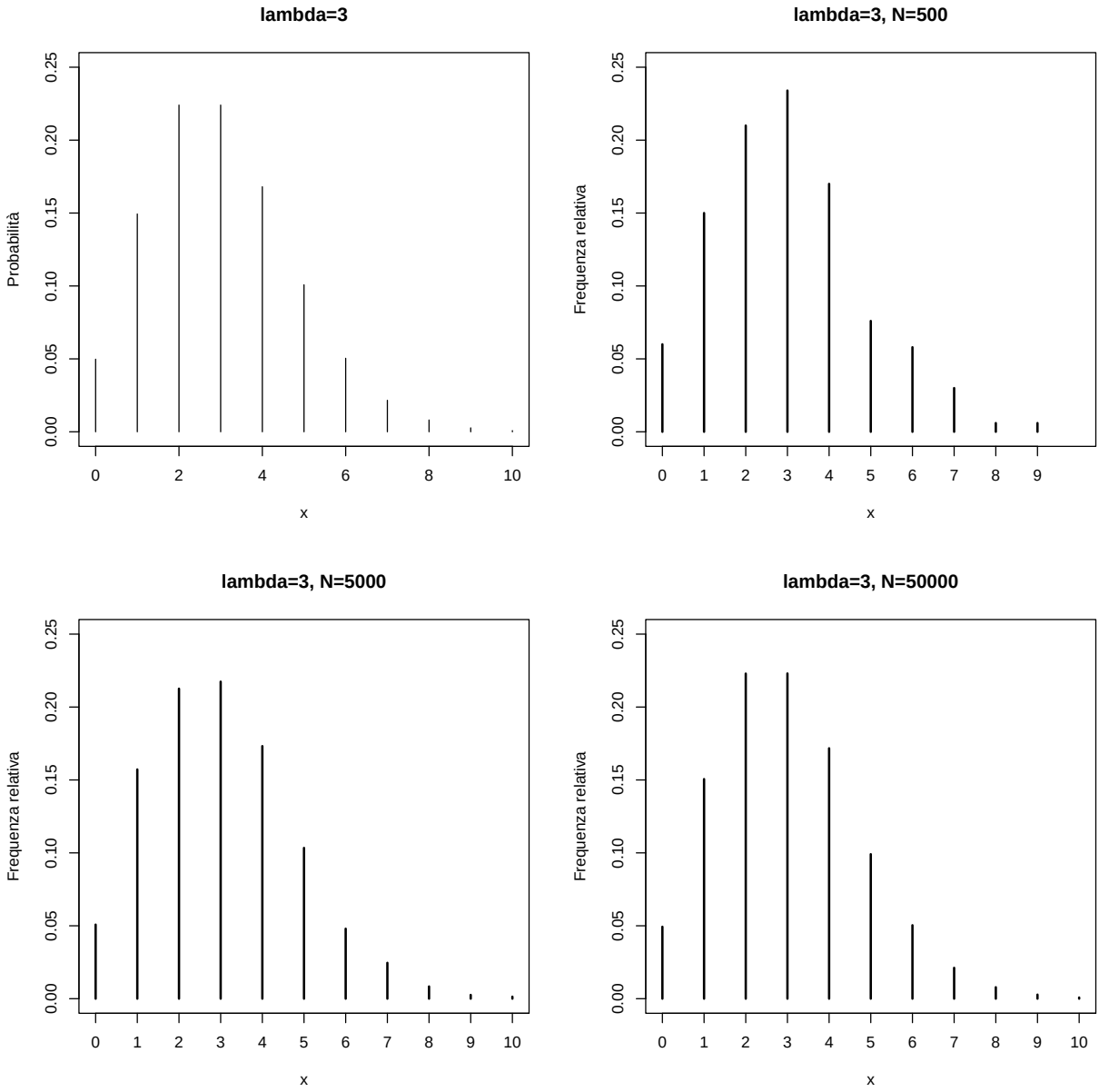
```
## [1] "frequenze assolute"
```

```
## sim
##  0  1  2  3  4  5  6  8
##  4 10 18  9  8  3  6  2
```

```
## [1] "frequenze relative"
```

```
## sim
##          0          1          2          3          4          5          6
## 0.06666667 0.16666667 0.30000000 0.15000000 0.13333333 0.05000000 0.10000000
##          8
## 0.03333333
```

E' possibile poi confrontare la funzione di probabilità di Poisson teorica con quella simulata all'aumentare di N.



Possiamo notare che all'aumentare della lunghezza della sequenza generata, il grafico delle frequenze relative si avvicina sempre di più al grafico della funzione di probabilità di Poisson.

Stima puntuale

Si vuole studiare una popolazione descritta da una variabile aleatoria osservabile X la cui funzione di distribuzione ha una forma nota ma contiene un parametro non noto. Il termine osservabile significa che si possono osservare i valori assunti dalla variabile aleatoria X e quindi il parametro non noto è presente soltanto nella legge di probabilità. Nei metodi di indagine dell'inferenza statistica si considera un campione casuale $X_1, X_2, ..., X_n$ di ampiezza n estratto dalla popolazione e si vogliono ottenere informazioni sui parametri non noti facendo uso di alcune variabili aleatorie dette statistiche e stimatori. Una **statistica** è una funzione misurabile e osservabile del campione casuale. Essendo la statistica osservabile, i valori da essa assunti dipendono soltanto dal campione osservato estratto dalla popolazione e i parametri non noti sono presenti soltanto nella funzione di distribuzione della statistica. Uno **stimatore** è una statistica i cui valori possono

essere usati per stimare un parametro non noto della popolazione. I valori assunti da tale stimatore sono detti stime del parametro non noto. Gli stimatori tipicamente utilizzati sono la media campionaria e la varianza campionaria.

Metodi per la ricerca di stimatori I principali metodi per trovare uno stimatore sono il **metodo dei momenti** e il **metodo della massima verosimiglianza**. L'idea alla base del *metodo dei momenti* è di porre il momento campionario assoluto uguale al corrispondente momento campionario. Occorre definire i momenti campionari, ovvero la media aritmetica delle potenze r-esime delle n osservazioni effettuate sulla popolazione. Affinchè il metodo dei momenti sia utilizzabile occorre che $E(X^r) = M_r(x_1, x_2, ..., x_n)$ $(r = 1, 2, ..., k)$ ammetta un'unica soluzione. In particolare, si desidera determinare lo stimatore del valore medio λ di una popolazione di Poisson descritta dalla variabile aleatoria $X \sim P(\lambda)$. Occorre stimare il parametro λ. Le stime dei parametri ottenute con tale metodo dipendono dal campione osservato e quindi al variare dei possibili campioni osservati si ottengono gli stimatori dei parametri non noti della popolazione, detti *stimatori del metodo dei momenti*. Poichè in una popolazione di Poisson $E(X) = \lambda$, si ha che lo stimatore è proprio la media campionaria $\bar{X}$: $\hat{\lambda} = \frac{(x_1+x_2+...+x_n)}{n} = \bar{x}$

```
## [1] "stima del parametro lambda"
```

```
## [1] 2.866667
```

La stima del parametro λ per il nostro campione con il metodo dei momenti è $\hat{\lambda} = 2.46$. Per il *metodo della massima verosimiglianza* occorre definire la funzione di verosimiglianza che è la *funzione di probabilità congiunta* oppure la funzione densità di probabilità congiunta del campione casuale. Tale metodo consiste nel massimizzare la funzione di verosimiglianza rispetto ai parametri $\theta_1, \theta_2, ..., \theta_k$, quindi cerca di determinare da quale funzione di probabilità congiunta oppure di densità di probabilità congiunta è più verosimile che provenga il campione osservato. I valori di $\theta_1, \theta_2, ..., \theta_k$ che massimizzano la funzione di verosimiglianza sono indicati con $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$ e sono le *stime di massima verosimiglianza* dei parametri non noti $\theta_1, \theta_2, ..., \theta_k$ della popolazione. Tali stime dipendono dal campione osservato e quindi al variare dei possibili campioni osservati si ottengono gli stimatori di massima verosimiglianza dei parametri non noti della popolazione. Anche in questo caso la stima di verosimiglianza per parametro λ per la popolazione di Poisson risulta essere la media campionaria $\bar{E}$.

Intervalli di fiducia approssimati

Alla stima puntuale di un parametro non noto di una popolazione che è costituita da un singolo valore reale, spesso si preferisce sostituire un intervallo di valori detto **intervallo di confidenza**, ossia si cerca di determinare in base ai dati del campione, un limite superiore e uno inferiore entro i quali sia compreso il parametro non noto con un certo **coefficiente di confidenza**. Se la dimensione del campione è elevata (n>30) è possibile utilizzare il *teorema centrale di convergenza* per determinare un intervallo di confidenza di grado $1 - \alpha$ per il parametro non noto di una popolazione. Se il valore medio $E(X) = \mu$ e $Var(X) = \sigma^2$ della popolazione dipendono da un parametro non noto θ della popolazione, si nota la variabile aleatoria Z_n può essere interpretata come variabile aleatoria di pivot poichè:

-dipende dal campione casuale;

-dipende dal parametro non noto θ;

-per grandi campioni la sua funzione di distribuzione è approssimativamente normale standard e quindi non contiene il parametro θ da stimare.

Possiamo dunque applicare il *metodo pivotale in forma approssimata*. Nel nostro caso, consideriamo una popolazione di Poisson descritta da una variabile aleatoria $X \sim P(X)$ e il valore medio di una variabile aleatoria è $E(X) = \lambda$ e la varianza è $Var(X) = \lambda$ ed entrambi dipendono dal parametro non noto λ. Ricaviamo che $E(\bar{X}_n) = \lambda, Var(\bar{X}_n) = \frac{\lambda}{n}$. Applicando il teorema centrale di convergenza si ha che la variabile aleatoria converge in distribuzione ad una variabile aleatoria normale standard. Si supponga che il numero N(t) di email che arrivano ad un server nell'intervallo (0,t) sia distribuito secondo Poisson. Se in 100 osservazioni effettuate in intervalli di tempo di t = 30 giorni si riscontra che in media sono state ricevute 6 email, si vuole determinare una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per il parametro α.

```
## [1] 1.959964
```

```
## [1] 0.1846244+0i 0.2166561-0i
```

L'intervallo di confidenza approssimato è quindi (0.1846, 0.2166). La stima puntuale, ovvero $6/30 = 0.2$, è compresa nell'intervallo. Possiamo inoltre calcolare la stima approssimata dell'intervallo di confidenza.

```
## [1] 0.183997
```

```
## [1] 0.216003
```

Confronto tra due popolazioni di Poisson

Spesso si è interessati a stimare la differenza tra le medie di due distinte popolazioni. Consideriamo una prima popolazione di Poisson descritta da una variabile $X \sim P(\lambda_1)$ ed una seconda popolazione di Poisson descritta da una variabile $Y \sim P(\lambda_2)$ e siano $X_1, X_2, ..., X_{n1}$ e $Y_1, Y_2, ..., Y_{n2}$ due campioni casuali indipendenti di ampiezza n_1 e n_2 estratti da due popolazioni di Poisson. Vogliamo determinare un intervallo di confidenza di grado $1 - \alpha$ per la differenza $\lambda_1 - \lambda_2$ tra i parametri delle due popolazioni per grandi valori di n_1 e n_2. Sono analizzati due server in base al numero di email che ricevono per un fissato numero di giorni. Si registrano le email arrivate al server1 per 30 giorni distinti e al server2 per 60 giorni distinti. Entrambi i campioni sono descritti da una **variabile aleatoria di Poisson**. Si desidera determinare l'intervallo di confidenza di grado $1 - \alpha = 0.97$.

Inizio generando due campioni casuali.

```
## [1] "campione 1"
```

```
##  [1]  8  6  2  7  4  8  7  8  4  5  3  5  5  7  6  4  7  6  5  8  8  7  9  7 11
## [26]  9  4  1  9  7
```

```
## [1] 30
```

```
## [1] "campione 2"
```

```
##  [1]  5 11  4  4  6  6  7  4  5  4  3  8  4  4  9  8  7  4  4  1  9  3  3  8  6
## [26]  6  9  7  4  3  3  6  7  6  2  6  4  5  4  4  5  8  5  6  9  6  5  4  5  6
## [51]  3  3  7  7  5  5  7 10  5 11
```

```
## [1] 60
```

Le frequenze assolute dei due server sono:

```
## [1] "frequenze assolute campione 1"
```

```
## campione1
##  1  2  3  4  5  6  7  8  9 11
##  1  1  1  4  4  3  7  5  3  1
```

```
## [1] "media campione 1"
```

```
## [1] 6.233333
```

```
## [1] "frequenze assolute campione 2"
```

```
## campione2
##  1  2  3  4  5  6  7  8  9 10 11
##  1  1  7 13 10 10  7  4  4  1  2
```

```
## [1] "media campione 2"
```

```
## [1] 5.583333
```

Determiniamo adesso l'intervallo di confidenza per $\lambda_1 - \lambda_2$ di grado $1 - \alpha = 0.97$.

```
## [1] 2.17009
```

```
## [1] -0.5402572
```

```
## [1] 1.840257
```

Dunque una stima dell'intervallo di confidenza è (-1.6803, 0.7469). Poichè questo intervallo include la possibilità che $\lambda_1 = \lambda_2$, non si può concludere che il numero medio di email arrivate ai due server siano differenti con un grado di fiducia del 97%.

Verifica delle ipotesi

Le aree più importanti dell'inferenza statistica sono la *stima dei parametri* e la *verifica delle ipotesi*. La verifica delle ipotesi interviene spesso nelle ricerche di mercato, nelle indagini sperimentali e industriali, nei sondaggi di opinione. Gli elementi che costituiscono il punto di partenza del procedimento di verifica delle ipotesi sono una popolazione descritta da una variabile aleatoria X caratterizzata da una funzione di probabilità o densità di probabilità.

Un'**ipotesi statistica** è un'affermazione o una congettura sul parametro non noto θ. Se l'ipotesi statistica specifica completamente $f(x;\theta)$ è detta *ipotesi semplice*, altrimenti è chiamata *ipotesi composita*. L'ipotesi soggetta a verifica viene in genere denotata con H_0 e viene chiamata *ipotesi nulla*. Si chiama *test di ipotesi* il procedimento o regola con cui si decide, sulla base dei dati del campione, se accettare o rifiutare H_0. La costruzione del test richiede la formulazione, in contrapposizione all'ipotesi nulla, di una proposizione alternativa, chiamata *ipotesi alternativa* e indicata con H_1. Il problema della verifica delle ipotesi consiste nel determinare un **test** che permetta di suddividere, mediante opportuni criteri, l'insieme dei possibili campioni in due sottoinsiemi: una regione di accettazione A dell'ipotesi nulla ed una regione di rifiuto R dell'ipotesi nulla. Spesso si usa dire che l'ipotesi nulla H_0 deve essere verificata in alternativa all'ipotesi H_1. Si può però incorrere in due tipi di errori:

- rifiutare l'ipotesi nulla H_0 nel caso in cui tale ipotesi sia vera (errore di tipo I);

- accettare l'ipotesi nulla H_0 nel caso in cui tale ipotesi sia falsa (errore di tipo II).

E' importante definire il concetto di *misura della regione critica* di un test che fornisce la probabilità massima di commettere un errore di tipo I al variare di θ, ossia la probabilità massima di rifiutare l'ipotesi nulla quando essa è vera. Solitamente la probabilità di commettere un errore di tipo I si sceglie uguale a 0.05 (statisticamente significativo), 0.01 (statisticamente molto significativo), 0.001 (statisticamente estremamente significativo); *quanto è minore il valore di α tanto maggiore è la credibilità di un eventuale rifiuto dell'ipotesi nulla.* I test statistici possono essere di due tipi:

- **test bilaterali**;

- **test unilaterali**.

Nel nostro caso, consideriamo una popolazione di Poisson descritta dalla variabile aleartoria $X \sim P(\lambda)$. Vogliamo costruire dei test unilaterali e bilaterali per il valore medio $E(X) = \lambda$. Il *test bilaterale* può essere così formulato: $H_0 : \lambda = \lambda_0$ $H_1 : \lambda = /\lambda_0$ mentre il *test unilaterale sinistro e destro* sono rispettivamente i seguenti $H_0 : \lambda \leq \lambda_0$ $H_0 : \lambda \geq \lambda_0$ $H_0 : \lambda > \lambda_0$ $H_0 : \lambda < \lambda_0$ avendo fissato a priori un livello di significatività α. Supponiamo che il numero N(t) di email che arrivano ad un server nell'intervallo (0,t) sia distribuito secondo Poisson con valore medio $E[N(t)] = \lambda t$. In 100 osservazioni effettuate in intervalli di tempo di t = 30 giorni si riscontra che in media sono state ricevute 6 email. Precedentemente si è mostrato che una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per il parametro λ è (0.1846, 0.2166). Vogliamo adesso verificare l'ipotesi $H_0 : 30\lambda \leq 0.2$ in alternativa a $H_1 : 30\lambda > 0.2$ con un livello di significatività $\alpha = 0.05$. Occorre considerare un **test unilaterale sinistro.**

```
## [1] 1.644854
```

```
## [1] 22.13594
```

Si nota che $z_\alpha = 1.6448$ e $z_o s = 129.6919$ cade nella regione di rifiuto. Occorre quindi rifiutare l'ipotesi nulla con un livello di significatività del 5%.

Criterio del chi-quadrato

In molti problemi reali, si desidera verificare se il campione osservato può essere stato estratto da una popolazione descritta da una variabile aleatoria X con funzione di distribuzione $F_X(x)$. A questo scopo si usa il criterio di verifica delle ipotesi del **chi-quadrato**. Col criterio del chi-quadrato si desidera verificare l'ipotesi che una certa popolazione sia caratterizzata da una funzione di distribuzione $F_x(x)$ con k parametri non noti da stimare. Denotiamo con H_0 l'ipotesi soggetta a verifica e con H_1 l'ipotesi alternativa. Il test chi-quadrato con livello di significatività α mira a verificare l'ipotesi nulla H_0: X ha una funzione di distribuzione $F_X(x)$ in alternativa all'ipotesi H_1: X non ha una funzione di distribuzione $F_X(x)$; α è la probabilità massima di rifiutare l'ipotesi nulla quando essa è vera. Occorre determinare un test con livello di significatività α che permetta di determinare una regione di accettazione e di rifiuto dell'ipotesi nulla. Il test di verifica delle ipotesi considerato è bilaterale. Per un campione sufficientemente numeroso di ampiezza n, il test chi-quadrato bilaterale di misura α è il seguente:

-si accetti l'ipotesi H_0 se $\chi^2_{1-\alpha/2,r-k-1} < \chi^2 < \chi^2_{\alpha/2,r-k-1}$;

-si rifiuti l'ipotesi H_0 se $\chi^2 < \chi^2_{1-\alpha/2,r-k-1}$ oppure $\chi^2 > \chi^2_{\alpha/2,r-k-1}$.

Applicazione per una popolazione di Poisson Nel nostro caso andiamo a generare un campione casuale che registra le email ricevute da un server in 60 giorni e calcoliamo le frequenze assolute.

```
## [1] "campione casuale"
```

```
##  [1] 3 3 2 2 3 5 4 1 2 3 5 3 3 1 4 1 1 5 3 2 2 1 3 4 1 3 5 4 2 5 1 1 2 3 4 5 5 3
## [39] 1 2 5 5 4 5 2 1 1 3 1 5 1 2 4 4 3 1 2 1 2 4
```

```
## [1] "frequenze assolute"
```

```
## campione
##  1  2  3  4  5
## 15 12 13  9 11
```

Si nota che nei 60 giorni sono state ricevute: 1 email in 15 giorni, 2 email in 12 giorni, 3 email in 13 giorni, 4 email in 9 giorni, 5 email in 11 giorni. Si desidera verificare se il numero di email ricevute sia descrivibile con una variabile aleatoria X di Poisson di parametro λ, ossia: $p_X(x) = \frac{\lambda^x}{x!}e^{-\lambda}(x = 0, 1, ...)$. I dati del campione permettono di ottenere una stima del parametro λ. Infatti, ricordando che uno stimatore corretto con varianza uniformemente minima del parametro λ di una dsitribuzione di Poisson risulta essere la media campionaria, si ha:

```
## [1] "valore stima"
```

```
## [1] 2.816667
```

Supponiamo di considerare 5 categorie corrispondenti agli intervalli $I_1 = 1, I_2 = (1, 2], I_3 = (2, 3], I_4 = (3, +\infty)$. Le probabilità associate agli intervalli possono essere così calcolate:

```
## [1] "probabilità associate agli intervalli"
```

```
## [1] 0.1684506 0.2372347 0.2227370 0.3715777
```

```
## [1] "somma delle probabilità"
```

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

Si nota che $p_1 + p_2 + p_3 + p_4 = 1$.

```
## [1] 16.84506
```

Essendo $min(n * p[1], n * p[2], n * p[3], n * p[4]) > 5$ (16.84), la condizione per cui ogni classe contenga in media almeno 5 elementi è soddisfatta. Il numero di elementi del campione appartenente ai quattro intervalli è

```
## [1] "numero di elementi per ogni intervallo"
```

```
## [1] 15 12 13 20
```

```
## [1] "somma degli elementi"
```

```
## [1] 60
```

Dunque nell'intervallo I_1 ci sono 15 elementi, nell'intervallo I_2 ci sono 12 elementi, nell'intervallo I_3 ci sono 13 elementi e nell'intervallo I_4 ci sono 20 elementi. Inoltre, tutti gli elementi cadono in un intervallo come dimostra la somma ottenuta degli elementi di ogni intervallo che corrisponde alla lunghezza del campione considerato (60). Calcoliamo adesso il χ^2 definito come $\chi^2 = \sum_{i=1}^{r} (\frac{n_i - np_i}{\sqrt{np_i}})^2$:

```
## [1] 2.965501
```

che risulta essere $\chi^2 = 2.9655$. In questo caso il numero di categorie è r = 4 e occorre porre k = 1 poichè la probabilità di Poisson contiene un parametro non noto. Abbiamo quindi $r - k - 1 = 2$ e scegliendo $\alpha = 0.01$ occorre calcolare $\chi^2_{1-\alpha/2,2}$ e $\chi^2_{\alpha/2,2}$:

```
## [1] 0.05063562
```

```
## [1] 7.377759
```

risulta essere $\chi^2_{1-\alpha/2,r-k-1} = 0.0506$ e $\chi^2_{\alpha/2,r-k-1} = 7.377$. Essendo $0.0506 < \chi^2 < 7.3777$, l'ipotesi H_0 di una popolazione di Poisson può essere accettata.