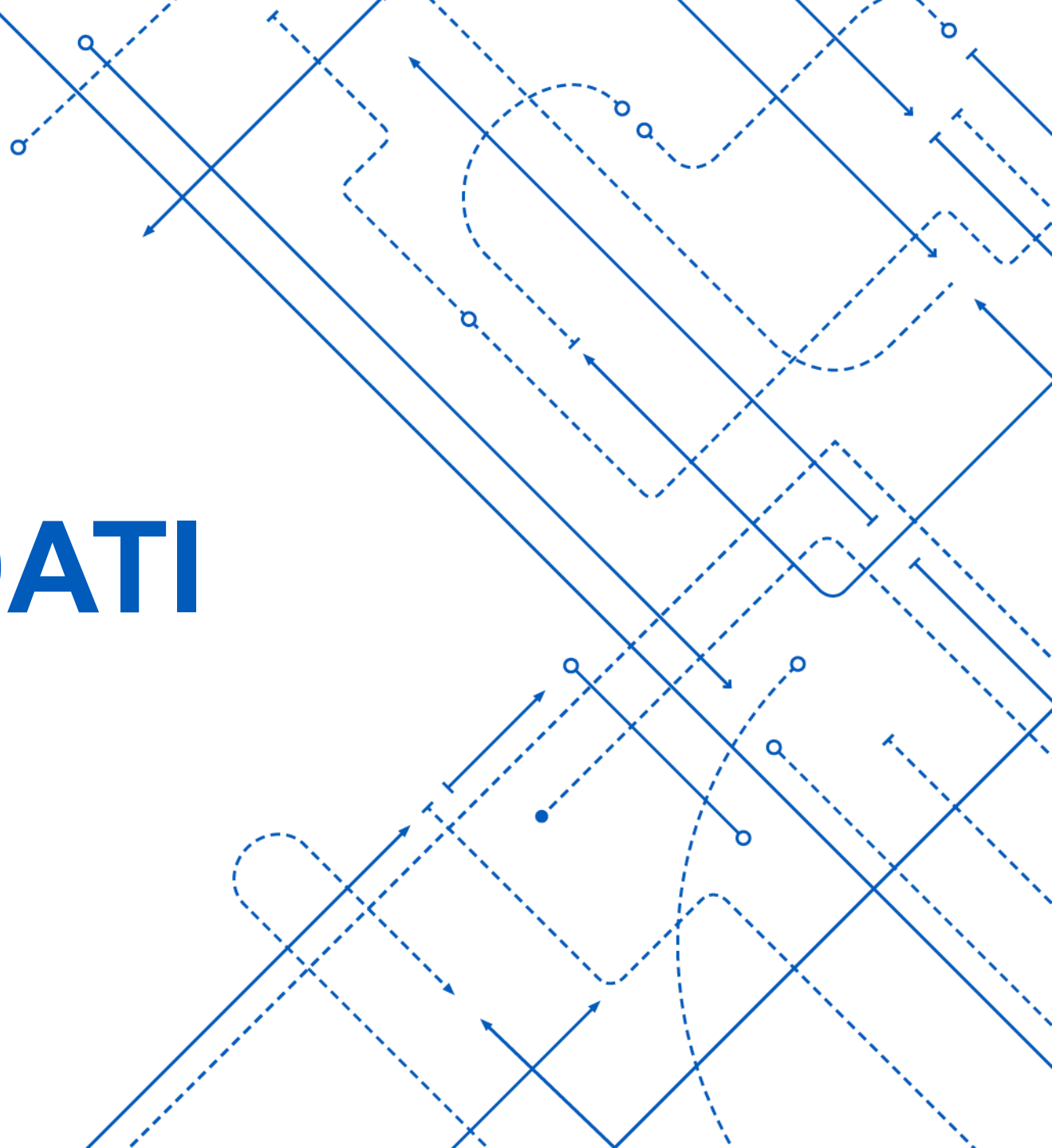
STATISTICA E ANALISI DEI DATI

Capitolo 6 – Funzioni di similarità

Dott. Stefano Cirillo
Dott. Luigi Di Biasi

a.a. 2023-2024



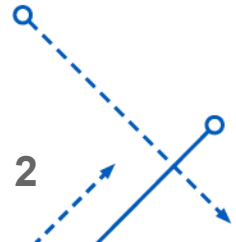
Misure di similarità

Per le tecniche di clustering occorre inizialmente calcolare la **matrice D delle distanze oppure** una **matrice S delle similarità**.

Dati due individui $\mathbf{I_i}$ e $\mathbf{I_j}$, una ***misura di similarità*** fornisce un valore numerico compreso tra 0 e 1, in cui:

- 0 indica l'assoluta assenza di similarità
- 1 la massima similarità (o somiglianza).

E' un indicatore quantitativo. E' definita come una funzione a valori reali.



Misure di similarità/2

Una ***funzione a valori reali:*** $s_{ij} = s(X_i, X_j)$ è detta **misura di similarità** se e soltanto se soddisfa le seguenti condizioni:

Chiamato anche **coefficiente di similarità**

$(i)\ s(X_i, X_i) = 1;$ **Implica che la misura di similarità vale uno se i due punti sono identici.**

$(ii)\ 0 \leq s(X_i, X_j) \leq 1;$ **Richiede che la misura di similarità sia compresa tra 0 e 1.**

$(iii)\ s(X_i, X_j) = s(X_j, X_i)\ per\ ogni\ X_i\ e\ X_j.$ **Impone la simmetria.**

La misura di similarità tra Xi e Xj deve assumere lo stesso valore della misura stessa, se valutata tra Xj e Xi

Coefficiente di Similarità

I **coefficienti di similarità** vanno a comporre **la matrice di similarità S**. L'elemento **S_{ij}** risulta essere l'elemento nella riga i-esima e colonna j-esima della matrice di similarità S.

$$S = \begin{pmatrix} 1 & s_{12} & \dots & s_{1n} \\ s_{21} & 1 & \dots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \dots & 1 \end{pmatrix}.$$

ESEMPIO

Il **coefficiente di similarità di Jaccard** è una misura di similarità per vettori binari.

$$s(X_i, X_j) = \frac{n_{11}}{n_{11} + n_{01} + n_{10}}.$$

Nota: E' il complemento della distanza di Jaccard!

(i) $s(X_i, X_i) = 1;$ se **X_i** = **X_j** risulta che $n_{01} = n_{10} = 0$ e quindi la prima condizione è valida.

(ii) $0 \leq s(X_i, X_j) \leq 1;$ le condizioni (ii) e (iii) sono soddisfatte

(iii) $s(X_i, X_j) = s(X_j, X_i)$ *per ogni* X_i *e* X_j.

Recap e Osservazioni importanti

- La misura di similarità assume valori tra 0 e 1;
- La misura di distanza assume valori non negativi;
- E' **sempre possibile trasformare una misura di distanza in una misura di similarità** con la formula:

$$s_{ij} = \frac{1}{1 + d_{ij}} \qquad (i, j = 1, 2, \ldots, n)$$

- **Non è sempre possibile** passare da una misura di similarità ad una misura di distanza, poiché la **funzione distanza deve soddisfare anche la disuguaglianza triangolare**.

Misure di non omogeneità tra cluster

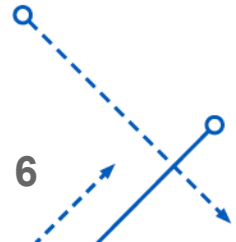
Definiamo ora delle misure di **non omogeneità:**

- **all'interno dei cluster;**
- **tra cluster distinti**.

Perchè?

Vogliamo che, al termine del procedimento di clustering/classificazione, gli individui appartenenti allo stesso cluster **siano:**

- **il più possibile omogenei tra di loro;**
- il più possibile **differenti** da quelli appartenenti **agli altri cluster** individuati.



Misure di non omogeneità totale/1

- Consideriamo un insieme :

$$I = \{I_1, I_2, \ldots, I_n\}$$ ← n individui

- E un insieme di **caratteristiche**:

$$C = \{C_1, C_2, \ldots, C_p\}$$ ← Osservabili e possedute da ciascun individuo

- Otteniamo la matrice delle misure (aka matrice dei descrittori o delle caratteristiche):

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

Misure di non omogeneità totale/2

Alla matrice **X** associamo una matrice **Wx** di cardinalità p×p, detta **matrice delle varianze e covarianze**:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \Rightarrow W_X = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1p} \\ w_{21} & w_{22} & \dots & w_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ w_{p1} & w_{p2} & \dots & w_{pp} \end{pmatrix},$$

Il generico elemento $w_{r\ell}$ della matrice W_x può essere definito come segue:

$$w_{r\ell} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ir} - \overline{x}_r)(x_{i\ell} - \overline{x}_\ell) \qquad (r, \ell = 1, 2, \dots, p)$$

Proprietà degli elementi della matrice $\mathbf{W}_x$

$$w_{r\ell} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ir} - \overline{x}_r)\,(x_{i\ell} - \overline{x}_\ell) \qquad (r, \ell = 1, 2, \ldots, p)$$

Se $r = \ell$:

- w_{rl} = Varianza campionaria relativa alla caratteristica r-esima effettuata su tutti gli n individui

Se $r \neq \ell$:

- w_{rl} = Covarianza campionaria relativa alla caratteristica r-esima e la caratteristica l-esima effettuata su tutti gli n individui

Calcolo degli elementi $w_{r\ell}$ di W_x

Abbiamo già visto che è possibile calcolare alcune statistiche tramite R.

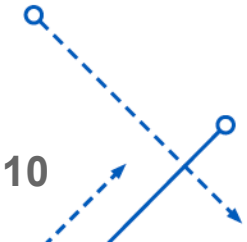
`apply(X, 2, mean)`	Calcola la media campionaria
`apply(X, 2, var)`	Calcola la varianza campionaria
`apply(X, 2, sd)`	Calcola la dev. standard campionaria

Ricordate apply su che oggetto lavora?

Vengono calcolati sulle colonne di X

`cov(X)`

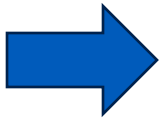
Calcola la matrice W_x delle varianze e covarianze campionarie tra le caratteristiche.



Calcolo in **R** della della matrice $\mathbf{W_x}$

Consideriamo la seguente matrice dei dati che si riferisce a due caratteristiche C1 e C2 osservate per cinque differenti individui I1, I2, I3, I4, I5, espresse nelle stesse unità di misura:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 6 & 3 \\ 8 & 2 \\ 8 & 0 \end{pmatrix} \end{array}$$



```
> X<-data.frame(c1=c(1,1,6,8,8),c2=c(1,2,3,2,0))
> row.names(X)<-c("I1","I2","I3","I4","I5")
> X # visualizza il data frame X
   c1 c2
I1  1  1
I2  1  2
I3  6  3
I4  8  2
I5  8  0
>
> apply(X,2,mean)
 c1  c2
4.8 1.6
> apply(X,2,var)
  c1   c2
12.7  1.3
>
> WI<-cov(X)
> WI # visualizza la matrice di covarianza
      c1    c2
c1 12.70 -0.35
c2 -0.35  1.30
```

Notate Cov(C1, C2) = −0.35. Cosa significa?

Le due caratteristiche C1 e C2 sono correlate negativamente!

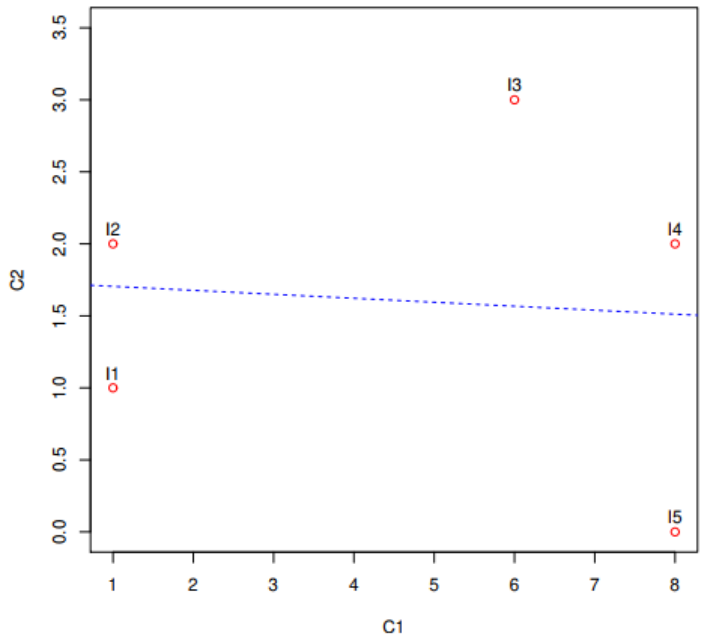
Misure di non omogeneità totale/6

ESEMPIO

Possiamo rappresentare i cinque punti relativi agli individui I1, I2, I3, I4, I5 in uno scatterplot.

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 6 & 3 \\ 8 & 2 \\ 8 & 0 \end{pmatrix} \end{array}$$

```
> plot(X$c1,X$c2,col="red",xlab="C1",
+ ylab="C2",ylim=c(0,3.5))
> text(X$c1,X$c2+0.1,c("I1","I2","I3","I4","I5"))
>
> abline(lm(X$c2~X$c1),lty=2,col="blue")
```



Matrice statistica di non omogeneità

Possiamo definire la **matrice statistica di non omogeneità** (statistical scatter matrix) per l'insieme I di individui, di cardinalit`a p × p come segue.

$$H_I = (n-1)W_I = \begin{pmatrix} h_{11} & h_{12} & \dots & h_{1p} \\ h_{21} & h_{22} & \dots & h_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ h_{p1} & h_{p2} & \dots & h_{pp} \end{pmatrix} \qquad h_{r\ell} = \sum_{i=1}^{n} (x_{ir} - \overline{x}_r)\,(x_{i\ell} - \overline{x}_\ell) = (n-1)\,w_{r\ell} \qquad (r, \ell = 1, 2 \dots, p)$$

Notate che se r = ℓ, h_{rr} corrisponde a n−1 volte la varianza campionaria della caratteristica r-esima effettuata su tutti gli n individui, ossia

$$h_{rr} = (n-1)\,\mathrm{Var}(C_r) = (n-1)\,s_r^2 \qquad (r = 1, 2 \dots, p).$$

Misura di non omogeneità statistica

Si definisce misura di non omogeneità statistica (statistical scatter) dell'insieme I di individui **la traccia** della matrice **H_I**:

è la somma degli elementi sulla diagonale principale di una matrice quadrata.

$$\mathrm{tr}H_I = \sum_{r=1}^{p} h_{rr} = (n-1)\sum_{r=1}^{p} s_r^2.$$

E possibile applicare tale definizione quando n > 1;
Quando invece n = 1, si suppone che la misura di non omogeneità statistica sia nulla.

DA RICORDARE:
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.

È intuitivo pensare che **più un insieme di dati è addensato** e **più piccola è la traccia della matrice** di non omogeneità.

Misura di non omogeneità statistica /2

La **misura di non omogeneità statistica** $\mathrm{tr}H_I$ è anche esprimibile in termine della somma dei **quadrati delle distanze euclidee tra ogni vettore $X_1, X_2, \ldots, X_n$** e il vettore **X** delle medie campionarie.

d_2 è la Distanza euclidea

$\overline{X} = (\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_p)$ È un vettore di cardinalità **p** il cui generic elemento $\overline{x}_j$ rappresenta la **media campionaria relativa** alla j-esima caratteristica effettuata su **n** individi

$$\mathrm{tr}H_I = \sum_{i=1}^{n} d_2^2(X_i, \overline{X}) = \sum_{r=1}^{p} h_{rr} = (n-1) \sum_{r=1}^{p} s_r^2.$$

DIMOSTRAZIONE ALLA LAVAGNA

Misura di non omogeneità statistica /3

ESEMPIO

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 6 & 3 \\ 8 & 2 \\ 8 & 0 \end{pmatrix} \end{array}$$

```
> # PRIMO METODO
> n<-nrow(X)
>  if(n>1)
> trHI<-(n-1)*sum(apply(X,2,var))
> else
>  trHI<-0
>
> trHI # visualizza la misura di non omogeneita' statistica
[1] 56
```

TRACCIA

```
> # SECONDO METODO (n>1)
> n<-nrow(X) # numero di righe del data frame
> WI<-cov(X)
> HI<-(n-1)*WI
> HI # visualizza la matrice di non omogeneita' statistica
     c1   c2
c1 50.8 -1.4
c2 -1.4  5.2
>
> trHI<-sum(diag(HI))
> trHI # visualizza la misura di non omogeneita' statistica
[1] 56
```

$$H_I = \begin{pmatrix} 50.8 & -1.4 \\ -1.4 & 5.2 \end{pmatrix},$$

Misura di non omogeneità statistica /4

ESEMPIO

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 6 & 3 \\ 8 & 2 \\ 8 & 0 \end{pmatrix} \end{array}$$

```
> # TERZO METODO (n>1)
> d<-dist(X,method="euclidean",diag=FALSE,upper=FALSE)
> d # visualizza la matrice delle distanze
          I1        I2        I3        I4
I2 1.000000
I3 5.385165 5.099020
I4 7.071068 7.000000 2.236068
I5 7.071068 7.280110 3.605551 2.000000
>
> tr<-sum(d^2)/n
> tr # visualizza la misura di non omogeneita' statistica
[1] 56
```

Misure di non omogeneità tra cluster

Sono state considerate delle misure di non omogeneità relative **all'insieme totale** di individui della popolazione.

Occorre definire delle misure di non omogeneità **all'interno dei cluster** e delle misure di non omogeneità (o disparità) **tra cluster distinti**.

PERCHE'?

Al termine del procedimento di classificazione, gli individi:

- **appartenenti allo stesso cluster dovrebbero essere il più possibile omogenei;**
- **appartenenti allo stesso cluster dovrebbero essere il più possibile differenti** da quelli appartenenti agli altri cluster individuati.

Misure di non omogeneità tra cluster

Al termine del procedimento di classificazione, gli individi:

- **appartenenti allo stesso cluster dovrebbero essere il più possibile omogenei;**
- **appartenenti allo stesso cluster dovrebbero essere il più possibile differenti** da quelli appartenenti agli altri cluster individuati.

Come possiamo assicurare di soddifare queste due proprietà?

- PAUSA

Misure di non omogeneità tra cluster/2

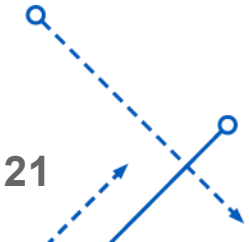
Consideriamo prima il caso base **tra due cluster**. Definiamo:

- **$I = \{I_1, I_2, \ldots, I_{n_1}\}$** e **$J = \{J_1, J_2, \ldots, J_{n_2}\}$ due cluster distinti** di individui di una popolazione.
- **$C = \{C_1, C_2, \ldots, C_p\}$** un insieme di caratteristiche.
- Grazie all'insieme delle caratteristiche possiamo definire:

$$\mathbf{X} = \{X_1, X_2, \ldots, X_{n_1}\} \qquad X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_1 1} & x_{n_1 2} & \cdots & x_{n_1 p} \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_{n_1} \end{pmatrix},$$

$$\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_{n_2}\} \qquad Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n_2 1} & y_{n_2 2} & \cdots & y_{n_2 p} \end{pmatrix} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{n_2} \end{pmatrix},$$

ATTENZIONE:
Da non confondere con le matrici di distanza o similarità.



Misure di non omogeneità tra cluster/3

Possiamo ancora definire:

$$\overline{x}_j = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{ij} \qquad (i = 1, 2, \ldots, p),$$

La **media campionaria** relativa alla caratteristica j-esima effettuata su tutti gli **n_1** individui del primo cluster

$$\overline{y}_j = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{ij} \qquad (i = 1, 2, \ldots, p).$$

La **media campionaria** elative alla caratteristica j-esima effettuata su tutti gli **n_2** individui del secondo cluster

Statistica e analisi dei dati

Notate che:

- gli elementi della matrice H_I si ottengono come prodotto di $\mathbf{n_1} - 1$ per gli elementi della matrice della varianze e covarianze tra le caratteristiche della matrice X

- gli elementi della matrice HJ si ottengono come prodotto di $\mathbf{n_2} - 1$ per gli elementi della matrice della varianze e covarianze tra le caratteristiche della matrice Y

$$H_I = \begin{pmatrix} h_{11}^{(1)} & h_{12}^{(1)} & \dots & h_{1p}^{(1)} \\ h_{21}^{(1)} & h_{22}^{(1)} & \dots & h_{2p}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ h_{p1}^{(1)} & h_{p2}^{(1)} & \dots & h_{pp}^{(1)} \end{pmatrix}, \qquad h_{r\ell}^{(1)} = \sum_{i=1}^{n_1} (x_{ir} - \overline{x}_r)\,(x_{i\ell} - \overline{x}_\ell) \qquad (r, \ell = 1, 2 \dots, p),$$

$$H_J = \begin{pmatrix} h_{11}^{(2)} & h_{12}^{(2)} & \dots & h_{1p}^{(2)} \\ h_{21}^{(2)} & h_{22}^{(2)} & \dots & h_{2p}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ h_{p1}^{(2)} & h_{k2}^{(2)} & \dots & h_{pp}^{(2)} \end{pmatrix} \qquad h_{r\ell}^{(2)} = \sum_{i=1}^{n_2} (y_{ir} - \overline{y}_r)\,(y_{i\ell} - \overline{y}_\ell) \qquad (r, \ell = 1, 2 \dots, p).$$

Misure di non omogeneità tra cluster/5

Definiamo:

$H_{I \cap J}$ Le **matrici di non omogeneità tra i due** cluster

$H_{I \cup J}$ **Le matrici di non omogeneità dell'unione** di due cluster

Inoltre, vogliamo **definire le relative misure di non omogeneità statistiche**.

RICORDATE: E' la traccia!

Misure di non omogeneità tra cluster/6

La matrice $H_{I \cup J}$ si può esrpimere come:

$$H_{I \cup J} = H_I + H_J + H_{I \cap J}.$$

La somma di tre matrici di cardinalità p × p:

- La matrice di non omogeneità statistica **H_I** relativa al cluster I;
- La matrice di non omogeneità statistica **H_J** relativa al cluster J;
- La matrice di non omogeneità statistica $H_{I \cap J}$

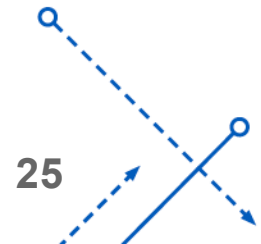
Possiamo inoltre ottenere la **misura di non omogeneità statistica relativa all'unione** dei cluster I e J, ossia:

$$\operatorname{tr} H_{I \cup J} = \operatorname{tr} H_I + \operatorname{tr} H_J + \operatorname{tr} H_{I \cap J}.$$

RICORDATE: E' la traccia!

La misura di **non omogeneità statistica tra i cluster** può essere calcolata come

$$\operatorname{tr} H_{I \cap J} = \operatorname{tr} H_{I \cup J} - \operatorname{tr} H_I - \operatorname{tr} H_J.$$



Misure di non omogeneità tra cluster/7

ESEMPIO: Consideriamo due cluster G_1 e G_2 dell'insieme di individui I

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 6 & 3 \\ 8 & 2 \\ 8 & 0 \end{pmatrix} \end{array}$$

$G_1 = \{I_1, I_2\}$ e

$G_2 = \{I_3, I_4, I_5\}.$

Misure di non omogeneità tra cluster/8

ESEMPIO: Per G_1

```
> X1<-data.frame(c1=c(1,1),c2=c(1,2))
> row.names(X1)<-c("I1","I2")
> X1 # visualizza il primo data frame
   c1 c2
I1  1  1
I2  1  2
>
> apply(X1,2,mean)
 c1  c2
1.0 1.5
>
> apply(X1,2,var)
 c1  c2
0.0 0.5
>
> W1<-cov(X1)
> W1 # visualizza la matrice delle varianze covarianze
    c1  c2
c1   0 0.0
c2   0 0.5
> # SECONDO METODO (n>1)
> n1<-nrow(X1)
> H1<-(n1-1)*W1
> H1 # visualizza la matrice di non omogeneita' statistica
    c1  c2
c1   0 0.0
c2   0 0.5
>
> tr1<-sum(diag(H1))
> tr1 # visualizza la misura di non omogeneita' statistica
[1] 0.5
```

$G_1 = \{I_1, I_2\}$

Si nota che per il primo gruppo si ha:

$$s_1^2 = 0,\ s_2^2 = 0.5,\ \mathrm{Cov}(C_1, C_2) = 0$$

$$\mathrm{tr}\, H_I = \quad 0.5.$$

Misure di non omogeneità tra cluster/9

ESEMPIO: Per G_2

```
> X2<-data.frame(c1=c(6,8,8),c2=c(3,2,0))
> row.names(X2)<-c("I3","I4","I5")
> X2 # visualizza il secondo data frame
   c1 c2
I3  6  3
I4  8  2
I5  8  0
>
> apply(X2,2,mean)
       c1        c2
7.333333  1.666667
>
> apply(X2,2,var)
       c1        c2
1.333333  2.333333
>
> W2<-cov(X2)
> W2 # visualizza la matrice delle varianze covarianze
          c1        c2
c1  1.333333 -1.333333
c2 -1.333333  2.333333
> # SECONDO METODO (n>1)
> n2<-nrow(X2)
> H2<-(n2-1)*W2
> H2 # visualizza la matrice di non omogeneita' statistica
          c1        c2
c1  2.666667 -2.666667
c2 -2.666667  4.666667
>
> tr2<-sum(diag(H2))
> tr2 # visualizza la misura di non omogeneita' statistica
[1] 7.333333
```

$$G_2 = \{I_3, I_4, I_5\}.$$

Si nota che per il secondo gruppo si ha:

$$s_1^2 = 1.333333,\ s_2^2 = 2.333333,\ \mathrm{Cov}(C_1, C_2) = -1.333333$$
$$\mathrm{tr}\, H_J = 7.333333.$$

Misure di non omogeneità tra cluster/9

ESEMPIO: Calcoliamo ora:

1. Le **misure di non omogeneità statistiche** dei cluster;
2. la misura di non omogeneità **interna** ai cluster (within);
3. la misura di non omogeneità **tra i cluster** (between)

```
> X<-data.frame(c1=c(1,1,6,8,8),c2=c(1,2,3,2,0))
>  row.names(X)<-c("I1","I2","I3","I4","I5")
> X # visualizza il data frame X
   c1 c2
I1  1  1
I2  1  2
I3  6  3
I4  8  2
I5  8  0
> # PRIMO METODO (n>1)
> n<-nrow(X)
>  trHI<-(n-1)*sum(apply(X,2,var))
> trHI # visualizza la misura di non omogeneita' statistica
[1] 56
>
> X1<-data.frame(c1=c(1,1),c2=c(1,2))
> row.names(X1)<-c("I1","I2")
> X1 # visualizza il primo data frame
   c1 c2
I1  1  1
I2  1  2
```

```
>  # PRIMO METODO (n1>1)
> n1<-nrow(X1)
> tr1<-(n1-1)*sum(apply(X1,2,var))
> tr1 # visualizza la misura di non omogeneita' di G1
[1] 0.5
>
> X2<-data.frame(c1=c(6,8,8),c2=c(3,2,0))
> row.names(X2)<-c("I3","I4","I5")
> X2 # visualizza il secondo data frame
   c1 c2
I3  6  3
I4  8  2
I5  8  0
> # PRIMO METODO (n2>1)
> n2<-nrow(X2)
> tr2<-(n2-1)*sum(apply(X2,2,var))
> tr2 # visualizza la misura di non omogeneita' di G2
[1] 7.333333
>
> trWithin<- tr1+tr2
> trWithin # visualizza la misura di non omogeneita' interna
[1] 7.833333
>
> trBetween <- trHI-trWithin
> trBetween # visualizza la misura di non omogeneita' tra i cluster
[1] 48.16667
```

Misure di non omogeneità tra cluster/9

In conclusione:

- le misure di non omogeneità statistica per il primo e per il secondo cluster sono $\operatorname{tr} H_I = 0.5$ e $\operatorname{tr} H_J = 7.333333$,
- la misura di non omogeneità statistica tra i cluster è $\operatorname{tr} H_{I \cap J} = 48.16667$
- la misura di non omogeneità relativa all'unione dei due cluster è $\operatorname{tr} H_{I \cup J} = 56$,

Si nota che:

- la misura di non omogeneità di ciascuno dei cluster è minore della misura di non omogeneità ottenuta unendo i due cluster $\operatorname{tr} H_{I \cup J} = 56.$
- la misura di non omogeneità all'interno (within) $\operatorname{tr} H_I + \operatorname{tr} H_J = 0.5 + 7.333333 = 7.833333$

e inferiore della misura di non omogeneità tra i cluster (between) $\operatorname{tr} H_{I \cap J} = 48.16667.$

Misure di non omogeneità: caso generale

E' possible partizionare un insieme I = {**I**1, **I**2, . . . , **I**n} di **n** individui in m particolari cluster estendendo la formula $H_{I \cup J} = H_I + H_J + H_{I \cap J}.$ nel modo seguente:

$$H_{G_1 \cup \ldots \cup G_m} = \sum_{\ell=1}^{m} H_{G_\ell} + \sum_{i<j} H_{G_i \cap G_j} + \sum_{i<j<k} H_{G_i \cap G_j \cap G_k} + \ldots + H_{G_1 \cap \ldots \cap G_m}$$

Denotando con:

- $T = H_{G_1 \cup \ldots \cup G_m} = H_I$ la matrice di non omogeneità statistica relativa all'insieme totale I = {I1, I2, . . . , In} degli n individui
- $S = H_{G_1} + H_{G_2} + \ldots + H_{G_m}$ la somma delle matrici di non omogeneità statistica relative ai singoli m cluster (within)
- $B = \sum_{i<j} H_{G_i \cap G_j} + \sum_{i<j<k} H_{G_i \cap G_j \cap G_k} + \ldots + H_{G_1 \cap \ldots \cap G_m}$ la matrice di non omogeneità statistica tra i vari cluster considerati (between)

Misure di non omogeneità: caso generale

L'equazione $$H_{G_1 \cup \ldots \cup G_m} = \sum_{\ell=1}^{m} H_{G_\ell} + \sum_{i<j} H_{G_i \cap G_j} + \sum_{i<j<k} H_{G_i \cap G_j \cap G_k} + \ldots + H_{G_1 \cap \ldots \cap G_m}$$

Può essere riscritta come:

$$T = S + B$$

Le matrici T , S e B hanno cardinalit`a p × p.

Per ogni fissata matrice X dei dati (di cardinalità n × p corrispondente all'insieme I = {I1, I2, . . . , In} degli n individui) si ha che la matrice T è fissata.

Invece, le matrici S e B dipendono strettamente dalla partizione in cluster dell'insieme I di individui considerata. Per ogni partizione dell'insieme I degli n individui in m fissati cluster, otteniamo un'equazione matriciale del tipo $T = S + B$

Misure di non omogeneità: caso generale

Abbiamo quindi che:

$$\text{tr}\, T = \text{tr}\, S + \text{tr}\, B, \quad \Rightarrow \quad 1 = \frac{\text{tr}\, S}{\text{tr}\, T} + \frac{\text{tr}\, B}{\text{tr}\, T}.$$

Poichè tr T è univocamente determinata per ogni matrice X di cardinalità
n× p, fissato il numero m di suddivisioni, **i cluster dovrebbero essere individuate** in modo da:

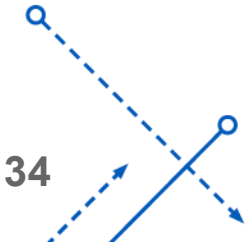
- **minimizzare** la misura di non omogeneità statistica **all'interno dei cluster (within);**
- **massimizzare** la misura di non omogeneità **statistica tra I gruppi** (between).

Scelta del partizionamento ottimale

Le misure di non omogeneità statistiche sono utilizzate per valutare, fissato il numero di cluster, **la bontà della suddivisione** in cluster

Una volta scelta la misura di distanza (o di similarità) si pone il problema di procedere alla scelta di un idoneo algoritmo di raggruppamento delle unità osservate.

I metodi di raggruppamento si distinguono in tre tipi: enumerazione completa, gerarchici, non gerarchici.



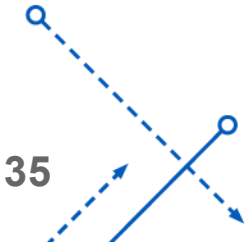
Scelta del partizionamento ottimale

A questo punto sappiamo:

1. Calcolare una “misura” di “non omogeneità” tra cluster generic;
2. Di voler cluster “quanto più omogenei possibili” tra di loro (internamente) e quanto più “non omogenei possibili” tra di loro “esternamente”.
    1. Gli elementi del cluster A devono essere quanto più omogenei tra di loro e quanto meno omogenei con gli elementi degli altri cluster;

Rimane una questione molto importante aperta!

Come determiniamo i cluster **ovvero gli elementi che devono essere inseriti in ogni cluster** e il **numero ottimale di cluster da creare**?



Come scegliamo il clustering migliore?

Un modo semplice che risolve il problema di clustering consiste nel determinare un partizionamento che soddisfa alcuni **criteri di ottimalità**.

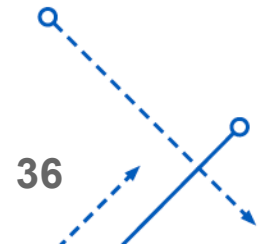
Questo criterio può essere dato in termini di una relazione funzionale (funzione obiettivo) che rifletta il livello di desiderabilità dei vari raggruppamenti.

Assegnata una funzione obiettivo: (Al momento conosciamo la misura di non omogeneità statistica interna ai cluster)

La **soluzione ottima** (ovvero il miglior clustering di n elementi in m cluster) può essere ottenuta:

- **valutando la funzione obiettivo per ogni possibile partizione dell'insieme degli n individui in cluster;**
- **scegliendo quella partizione che fornisce il valore ottimo della funzione obiettivo!**

Il minimo valore per la misura di non omogeneità statistica interna ai cluster.



Metodo dell'enumerazione completa

Quasi Brute Force

Ci facciamo aiutare dal calcolo combinatorio.

Supponiamo di considerare un insieme $\{I_1, I_2, \ldots, I_n\}$ di **n** individui e sia **m** il numero di cluster che abbiamo scelto di creare.

Il calcolo combinatorio ci fornisce il seguente risultato:

$$R(n,m) = \sum_{k=0}^{m} \binom{m}{k} (-1)^k (m-k)^n,$$

Il numero di modi in cui è possibile sistemare **n** biglie distinte in **m** urne distinte tali che **nessuna delle m urne sia vuota**

ATTENZIONE: L'ordine delle biglie deve essere irrilevante per valere questo risultato!

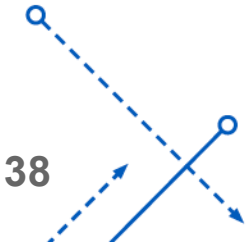
Numero di possibili cluster/1

$$R(n,m) = \sum_{k=0}^{m} \binom{m}{k} (-1)^k (m-k)^n,$$

Il numero di modi in cui è possibile sistemare **n** biglie distinte in **m** urne distinte tali che **nessuna delle m urne sia vuota**

Se pensiamo alle **urne come cluster** e alle **biglie come individui** di una popolazione, potrebbe essere possibile sfruttare la formulazione precedente per determinare in **quanti diversi cluster possono esistere dati n individi.**

ATTENZIONE**: le m urne sono supposte distinte**. In una partizione **di n individui in m cluster** di cui nessuno è vuoto, **l'ordine degli m cluster è esso stesso irrilevante**.



Numero di possibili cluster/2

$$R(n,m) = \sum_{k=0}^{m} \binom{m}{k} (-1)^k (m-k)^n,$$

Il numero di modi in cui è possibile sistemare **n** biglie distinte in **m** urne distinte tali che **nessuna delle m urne sia vuota**

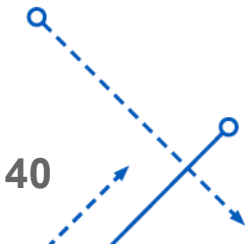
$$S(n,m) = \frac{1}{m!} \sum_{k=0}^{m} \binom{m}{k} (-1)^k (m-k)^n,$$

Il numero totale di modi di partizionare n individui in un fissato numero m di cluster.

Il risultato è detto Numero di Stirling del secondo tipo.

$$B_n = \sum_{m=1}^{n} S(n,m).$$

Se il numero m dei cluster non è fissato a priori, il numero totale di partizioni di n individui è dato dai numeri di Bell.



Numero di Stirling del secondo tipo

Come calcoliamo il numero di Stirling?

$$S(n,m) = \frac{1}{m!}\sum_{k=0}^{m}\binom{m}{k}(-1)^k(m-k)^n,$$

Metodo 1: Usiamo R e scriviamo una funzione per calcolarlo.

```
> stirling2<-function(n,m){
+       s<-0
+       if((m>=1)&(m<=n)){
+       for(k in seq(0,m)){
+               s<-s+(choose(m,k)*(-1)^k*(m-k)^n/factorial(m))}
+               return(c(s))
+               }}
```

Metodo 2: Usiamo una relazione di ricorrenza.

$$S(n,1) = 1$$
$$S(n,n) = 1$$
$$S(n,m) = S(n-1,m-1) + m\,S(n-1,m) \qquad (2 \le m \le n-1)$$

Esempi di calcolo del numero di Stirling

ESEMPIO

Se abbiamo 6 elementi, in quanti modi possiamo clusterizzarli se usiamo tre cluster (m = 3)?

$$S(n,m) = \frac{1}{m!}\sum_{k=0}^{m}\binom{m}{k}(-1)^k(m-k)^n,$$

$$S(n,1) = 1$$
$$S(n,n) = 1$$
$$S(n,m) = S(n-1,m-1) + m\,S(n-1,m) \qquad (2 \le m \le n-1)$$

$$S(6,3) = \frac{1}{3!}\sum_{k=0}^{3}\binom{3}{k}(-1)^k(3-k)^6 = \frac{1}{6}[3^6 - 3\,2^6 + 3\,1^6] = 90.$$

E se abbiamo n elementi, in quanti modi possiamo clusterizzarli se usiamo solo due cluster (m = 2)?

$$S(n,2) = \frac{1}{2}\sum_{k=0}^{2}\binom{2}{k}(-1)^k(2-k)^n = \frac{1}{2}(2^n - 2) = 2^{n-1} - 1,$$

Calcolo dei numeri di Bell

$$B_n = \sum_{m=1}^{n} S(n,m).$$

```
> sumstirling2<-function(n){
+       s<-0
+       for(k in seq(1,n))
+       s<-s+stirling2(n,k)
+       return(c(s))
+       }
>
> sumstirling2(6)
[1] 203
```

$$B_6 = \sum_{m=1}^{6} S(6,m)$$

$$B_6 = \sum_{m=1}^{6} S(6,m) = 1 + \frac{62}{2} + \frac{540}{6} + \frac{1560}{24} + \frac{1800}{120} + 1 = 203,$$

Problemi dell'enumerazione completa

E' impraticabile a meno che n (il numero degli individui) e m (il numero di cluster) non siano piccoli in quanto computazionalmente onerose. (Pensate a cosa accade se non conosciamo nemmeno m)

Prevedono il calcolo delle funzioni di non omogeneità **per ogni possibile partizione dell'insieme** totale di n individui in m cluster.

Nella pratica si adottano metodi di raggruppamento gerarchici e non gerarchici che operano su una sottoclasse delle partizioni degli n individui in cluster.