Percorsi didattici, interdisciplinari ed innovativi per la Statistica

Maurizio Vichi

Sapienza Università di Roma

Presidente Federazione Europea delle Società Nazionali di Statistica

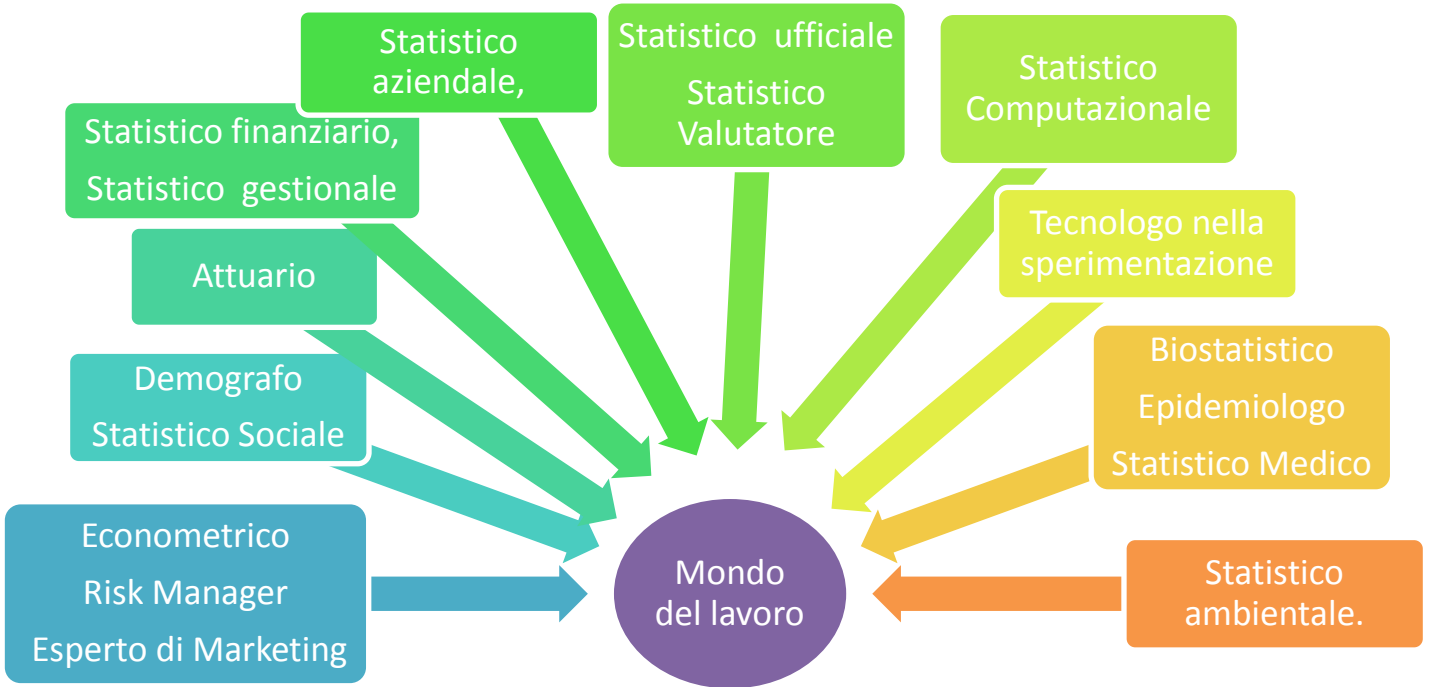
Scuola Estiva di Matematica per i Docenti delle Scuole Secondarie di 2°Grado
Montegrotto Terme, 22-25 Luglio 2014

La Statistica è multidisciplinare

1. Si utilizza per valutare una nazione moderna (statistica ufficiale);
2. E' utile nelle scienze sperimentali (fisica, chimica, biologia …) perché si basa sui principi galileiani di osservazione e raccolta dati e definizione di un modello, verifica del modello. Ad esempio si usa per monitorare il nostro habitat, misura i cambiamenti di clima, e per le previsioni meteo;
3. E' essenziale nelle aziende per misurare l'andamento del business, la qualità della produzione e il marketing (valutazione di efficacia efficienza, soddisfazione)
4. E' utile nella medicina per valutare l'efficacia dei trattamenti e prevenire malattie, per lo studio del genoma
5. …

Rappresenta una nuova conoscenza di base nelle società moderne

Conoscenza: saper leggere e interpretare e sintetizzare la realtà che ci circonda.



Come insegnare la statistica insieme ad altre materie

- La Statistica è fortemente multidisciplinare

Il prof di matematica che insegna la statistica può essere al centro di un progetto didattico multidisciplinare

E' di ausilio a molte materie scolastiche

- Statistica e Matematica;

  L'uso dell'algebra delle matrici per insegnare statistica

- Statistica e Informatica;

  La realizzazione di una indagine statistica

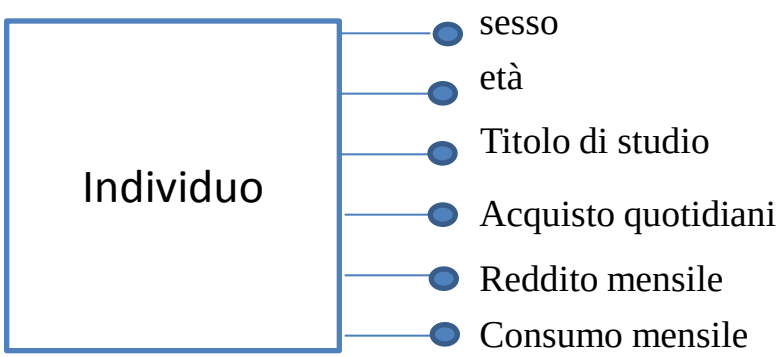
- Statistica e Fisica (o altre scienze dure);

  L'osservazione e la misurazione dei dati di un esperimento

Indagine statistica: Schema ER → Questionario → Base-dati (file) →Matrice Dati

Uso di ACCESS per la costruzione delle BASE DATI dell'Indagine Statistica

1 Schema ER del fenomeno



2 Questionario individuale

Sesso	**M=0; F=1 Età**
Titolo di studio	**Elem.=1, Media inf.=2, Media sup.=3, Laurea=4, Dottorato=5**
Acquisto quotidiano	**SI=1; NO=0**

Reddito mensile ……EUR Consumo mensile …….. EUR

3 Base dati mediante ACCESS e maschera di input Dati

Unità	*sesso*	*Età*	*Titolo di studio*	*Reddito mensile*	*Consumo mensile*	*Acquisto. quotidiano*
1	0	20	1	1050	800	1
2	0	19	3	1000	900	1
3	1	21	3	2000	1300	1
4	1	75	2	1200	1000	0
5	1	45	4	2200	800	1
6	1	35	5	2500	1500	0
7	1	21	2	1250	1000	0
8	0	60	2	1800	1100	1
9	0	18	2	1640	1350	1
10	1	12	1	1400	1000	1

4 Matrice dei dati

$$X=\begin{bmatrix} 0 & 20 & 1 & 1050 & 800 & 1 \\ 0 & 19 & 3 & 1000 & 900 & 1 \\ 1 & 21 & 3 & 2000 & 1300 & 1 \\ 1 & 75 & 2 & 1200 & 1000 & 0 \\ 1 & 45 & 4 & 2200 & 800 & 1 \\ 1 & 35 & 5 & 2500 & 1500 & 0 \\ 1 & 21 & 2 & 1250 & 1000 & 0 \\ 0 & 60 & 2 & 1800 & 1100 & 1 \\ 0 & 18 & 2 & 1640 & 1350 & 1 \\ 1 & 12 & 1 & 1400 & 1000 & 1 \end{bmatrix}$$

L’algebra delle Matrici

- Per calcolare i principali indici statistici
- Distribuzioni di frequenze semplici e doppie
- Media aritmetica e vettore delle medie
- Varianza
- Covarianza
- Matrice di varianze e covarianze
- Matrice di correlazione
- Retta di regressione

1. Matrici, Vettori

Definizione

Matrice: Una *matrice* **X** è una tabella rettangolare formata da ($N \times J$) elementi (oggetti, numeri, matrici). Il numero di elementi ($N \times J$) rappresenta la *dimensione* della matrice. La matrice **X** può essere vista come un insieme di *N righe* (righe orizzontali) o un insieme di *J colonne* (righe verticali).

Le matrici più diffuse in statistica sono costituite da numeri appartenenti al *campo dei reali*. Una matrice **X** è descritta da una lettera maiscola in grassetto, e si scrive equivalentemente nei seguenti modi

$$\mathbf{X} = [x_{ij} : i=1,\ldots,N;\ j=1,\ldots,J] ,$$

$$\mathbf{X} = [x_{ij}] \ \ (N \times J),$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots x_{1j} \ldots & x_{1J} \\ x_{21} & x_{22} & \ldots x_{2j} \ldots & x_{2J} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \ldots x_{ij} \ldots & x_{iJ} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{I2} & x_{Ij} & x_{NJ} \end{bmatrix}.$$

Vettore: Un vettore **x** è una particolare matrice avente una sola colonna, ossia, di dimensione ($N \times 1$). Un vettore è descritto da una lettera minuscola in grassetto ed è scritto come segue

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}.$$

Talvolta il vettore viene rappresentato come una riga, ossia è trasposto e si scrive

$$\mathbf{x}' = [x_1, \ldots, x_N].$$

1. Operazioni su matrici e vettori, proprietà e principali risultati

Addizione e sottrazione di due matrici $\mathbf{X} = [x_{ij}]$ $(N \times J)$ e $\mathbf{Y} = [y_{ij}]$ $(N \times J)$,

$$\mathbf{Z} = \mathbf{X} \pm \mathbf{Y},$$

dove

$$\mathbf{Z} = [z_{ij} = x_{ij} \pm y_{ij}] \ (N \times J).$$

Proprietà commutativa della somma $\mathbf{X} \pm \mathbf{Y} = \mathbf{Y} \pm \mathbf{X}$.
Proprietà associativa della somma $(\mathbf{X} \pm \mathbf{Y}) \pm \mathbf{Z} = \mathbf{X} \pm (\mathbf{Y} \pm \mathbf{Z})$.

Moltiplicazione di una matrice $\mathbf{X} = [x_{ij}]$ $(N \times J)$ *per uno scalare* a

$$a\mathbf{X} = \mathbf{X}a = [ax_{ij}].$$

Proprietà associativa del prodotto di una matrice per uno scalare $(ab)\mathbf{X} = (a)b\mathbf{X}$.
Proprietà distributiva della somma $a(\mathbf{X} \pm \mathbf{Y}) = (a\mathbf{X} \pm a\mathbf{Y})$.

Moltiplicazione di due vettori (Prodotto scalare) $\mathbf{x} = [x_i]$ $(N \times 1)$ ed $\mathbf{y} = [y_{ij}]$ $(N \times 1)$,

$$\underset{(1\times 1)}{\mathbf{x}'\mathbf{y}} = [x_1, x_2, \ldots, x_N] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = x_1y_1 + x_2y_2 + \ldots + x_Ny_N$$

Proprietà

$\mathbf{x} = [x_i]$ $(N \times 1)$, $\mathbf{y} = [y_{ij}]$ $(N \times 1)$,

1. $\mathbf{x}'\mathbf{y} = \mathbf{y}'\mathbf{x}$;
3. $(a\mathbf{x})'(b\mathbf{x}) = ab\,\mathbf{x}'\mathbf{y}$;
4. $(\mathbf{x} + \mathbf{y})'\mathbf{z} = \mathbf{x}'\mathbf{z} + \mathbf{y}'\mathbf{z}$;
5. $(\mathbf{x} + \mathbf{y})'(\mathbf{w} + \mathbf{z}) = \mathbf{x}'(\mathbf{w} + \mathbf{z}) + \mathbf{y}'(\mathbf{w} + \mathbf{z})$.

Norma di un vettore $\mathbf{x} = [x_i]$ $(N \times 1)$:

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}} = \sqrt{\sum_{i=1}^{N} x_i^2} \geq 0$$ è il quadrato della distanza Euclidea dall'origine

Moltiplicazione di due vettori (Prodotto matriciale) $\mathbf{x} = [x_i]$ $(J \times 1)$ e $\mathbf{y} = [y_{ij}]$ $(N \times 1)$:

$$\underset{(N\times J)}{\mathbf{y}\mathbf{x}'} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} [x_1, x_2, \ldots, x_J] = \begin{bmatrix} y_1x_1 & y_1x_2 & \ldots & y_1x_J \\ y_2x_1 & y_2x_2 & \ldots & y_2x_J \\ \vdots & \vdots & \vdots & \vdots \\ y_Nx_1 & y_Nx_2 & \ldots & y_Nx_J \end{bmatrix}.$$

Moltiplicazione di due matrici (riga per colonna) $\mathbf{X} = [x_{ij}]$ $(N \times J)$ ed $\mathbf{Y} = [y_{ij}]$ $(J \times H)$:

$$\mathbf{XY} = \left[\sum_{j=1}^{J} x_{ij}y_{ij}\right] = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_N' \end{bmatrix} [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_J] = \begin{bmatrix} \mathbf{x}_1'\mathbf{y}_1 & \mathbf{x}_1'\mathbf{y}_2 & \ldots & \mathbf{x}_1'\mathbf{y}_J \\ \mathbf{x}_2'\mathbf{y}_1 & \mathbf{x}_2'\mathbf{y}_2 & \ldots & \mathbf{x}_2'\mathbf{y}_J \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{x}_N'\mathbf{y}_1 & \mathbf{x}_N'\mathbf{y}_2 & \ldots & \mathbf{x}_N'\mathbf{y}_J \end{bmatrix}.$$

Trasposta di $\mathbf{X} = [x_{ij}]$ $(N \times J)$:

$$\mathbf{X}' = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1J} \\ x_{21} & x_{22} & \cdots & x_{2J} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NJ} \end{bmatrix}' = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{N1} \\ x_{12} & x_{22} & \vdots & x_{N2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1N} & x_{2N} & \cdots & x_{JN} \end{bmatrix}$$

La matrice trasposta di **X** si denota con un apice o con una *T* ad esponente (**X**' o $\mathbf{X}^T$) ed è la matrice in cui le colonne diventano righe e le righe diventano colonne.

Proprietà

X = $[x_{ij}]$ $(N \times J)$, **Y** =$[y_{ij}]$ $(N \times J)$, *a* scalare

1. $(\mathbf{X} + \mathbf{Y})' = \mathbf{X}' + \mathbf{Y}'$;
2. $(a\mathbf{X})' = a\mathbf{X}'$;
3. $(\mathbf{X}'\mathbf{Y})' = \mathbf{Y}'\mathbf{X}$;
4. $(\mathbf{X}')' = \mathbf{X}$;
5. $(\mathbf{X}')^+ = (\mathbf{X}^+)'$
6. $rk(\mathbf{X}') = rk(\mathbf{X})$;
7. $\mathbf{X}'\mathbf{X}$ e $\mathbf{X}\mathbf{X}'$ sono simmetriche;
8. $[\mathbf{X}\ \mathbf{Y}]' = \begin{bmatrix} \mathbf{X}' \\ \mathbf{Y}' \end{bmatrix}$;

X = $[x_{ij}]$ $(N \times J)$, **Y** =$[y_{ij}]$ $(J \times H)$

9. $(\mathbf{X}\mathbf{Y})' = \mathbf{Y}'\mathbf{X}'$;

X = $[x_{ij}]$ $(N \times J)$, **Y** =$[y_{ij}]$ $(N \times H)$, **Z** = $[z_{ij}]$ $(J \times P)$, **W** =$[w_{ij}]$ $(J \times H)$

12. $\begin{bmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{Z} & \mathbf{W} \end{bmatrix}' = \begin{bmatrix} \mathbf{X}' & \mathbf{Z}' \\ \mathbf{Y}' & \mathbf{W}' \end{bmatrix}$.

Rango di una matrice $\mathbf{X} = [x_{ij}]$ $(N \times J)$,

si denota con $rk(\mathbf{X})$ è il massimo numero di righe o colonne di $\mathbf{X}$ linearmente indipendenti.

Traccia di una matrice $\mathbf{X} = [x_{ij}]$ $(N \times N)$:

$$tr(\mathbf{X}) = \sum_{i=1}^{N} x_{ii}$$

Proprietà

$\mathbf{X} = [x_{ij}]$ $(N \times J)$, $\mathbf{Y} = [y_{ij}]$ $(J \times N)$,

1. $tr(\mathbf{XY}) = \text{tr}(\mathbf{YX})$;
2. $tr(\mathbf{XX}^{+}) = rk(\mathbf{X})$;

$\mathbf{X} = [x_{ij}]$ $(N \times N)$, $\mathbf{Y} = [y_{ij}]$ $(N \times N)$, a scalare

3. $tr(\text{X} \pm \text{Y}) = tr(\text{X}) \pm tr(\text{Y})$;
4. $tr(a\text{X}) = a\ tr(\text{X})$;
5. $tr(\mathbf{X}') = tr(\mathbf{X})$.

Norma di una matrice $\mathbf{X} = [x_{ij}]$ $(N \times J)$:

1. *Norma Euclidea* : $\|\text{X}\| = \sqrt{tr\left(\mathbf{X}'\mathbf{X}\right)} = \left[\sum_{i}^{N}\sum_{j=1}^{J} x_{ij}^{2}\right]^{1/2}$

2. Norma di Mahalanobis: $\|\text{X}\|_{\mathbf{W}} = \sqrt{tr\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)}$, dove $\mathbf{W}$ è definita positiva. Spesso $\mathbf{W} = \mathbf{\Sigma}_{\mathbf{X}}^{-1}$, l'inversa della matrice di varianze e covarianze di $\mathbf{X}$.

Proprietà

$\mathbf{X} = [x_{ij}]$ $(N \times J)$, $\mathbf{Y} = [y_{ij}]$ $(N \times J)$, a scalare

1. Se $\mathbf{X} \neq \mathbf{0}$, $\|\mathbf{X}\| > 0$;
2. $\|a\mathbf{X}\| = |a|\ \|\mathbf{X}\|$;
3. $\|\mathbf{X} + \mathbf{Y}\| = \leq \|\mathbf{X}\| + \|\mathbf{Y}\|$.

Determinante di $\mathbf{X} = [x_{ij}]$ $(N \times N)$,

$$det(\mathbf{X}) = |\mathbf{X}| = \sum_{\sigma \in P_N} sng(\sigma) \prod_{i=1}^{N} x_{i\sigma(i)}$$

dove la somma si riferisce a tutti i possibili prodotti ciascuno dei quali e formato da un elemento per ciascuna riga di $\mathbf{X}$. I possibili prodotti sono pari alle permutazioni P_N ed il prodotto è preso con il segno negativo o positivo a seconda che la permutazione σ sia dispari o pari.

determinante di una matrice 2×2

$$det(\mathbf{X}) = \begin{vmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{vmatrix} = x_{11}x_{22} - x_{12}x_{21};$$

determinate di una matrice 3×3

$$det(\mathbf{X}) = \begin{vmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{vmatrix} = \begin{vmatrix} x_{11} & x_{12} & x_{13} & x_{11} & x_{12} \\ x_{21} & x_{22} & x_{23} & x_{21} & x_{22} \\ x_{31} & x_{32} & x_{33} & x_{31} & x_{32} \end{vmatrix} = x_{11}x_{22}x_{33}+x_{12}x_{23}x_{31}+x_{13}x_{21}x_{32} - x_{31}x_{22}x_{13}-x_{32}x_{23}x_{11}-x_{33}x_{21}x_{12}.$$

Proprietà

$\mathbf{X} = [x_{ij}]$ $(N \times N)$, a scalare

1. $det(a\mathbf{X}) = a^N \, det(a\mathbf{X})$;
2. $det(\mathbf{X}') = det(\mathbf{X})$;
3. Se $\mathbf{X}$ è non singolare $det(\mathbf{X}^{-1}) = det(\mathbf{X})^{-1}$;
4. $rk(\mathbf{X}) < N$, allora $det(\mathbf{X}) = 0$;
5. $rk(\mathbf{X}) = N$, allora $det(\mathbf{X}) \neq 0$;

$\mathbf{X} = [x_{ij}]$ $(N \times N)$, $\mathbf{Y} = [y_{ij}]$ $(N \times N)$,

6. $det(\mathbf{XY}) = det(\mathbf{X}) \, det(\mathbf{Y})$;

Matrice identità: $\mathbf{I}_N$, $(N \times N)$,

$$\mathbf{I}_N = \begin{bmatrix} 1 & \ddots & 0 \\ 0 & & 1 \end{bmatrix}.$$

Rappresenta l'elemento neutro del prodotto di matrici. Ovvero per $\mathbf{X} = [x_{ij}]$ $(N \times N)$, $\mathbf{X}\mathbf{I}_N = \mathbf{I}_N\mathbf{X} = \mathbf{X}$.

Matrice diagonale di $\mathbf{X} = [x_{ij}]$ $(N \times N)$:

$$dg(\mathbf{X}) = \begin{bmatrix} x_{11} & \ddots & 0 \\ 0 & & x_{NN} \end{bmatrix} = \mathbf{x} = [x_i] \ (N \times 1),$$

$$diag(\mathbf{x}) = diag\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}\right) = \begin{bmatrix} x_{11} & \ddots & 0 \\ 0 & & x_{NN} \end{bmatrix}.$$

Matrice Indicatrice $\mathbf{X} = [x_{ij}]$ $(N \times K)$, se è ad elementi binari e tale che $\mathbf{X}\,\mathbf{1}_K = \mathbf{1}_N$,
ossia ogni riga di $\mathbf{X}$ ha un solo elemento diverso da zero.

Proprietà

1. $\mathbf{X}$ indicatrice $\Leftrightarrow \mathbf{X}'\mathbf{X} = \text{diag}(\mathbf{x})$, dove $\mathbf{x} = \mathbf{X}'\mathbf{1}_N$;
2. $\mathbf{X}$ indicatrice normalizzata : $\mathbf{X}_n = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2}$ è ortogonale, ovvero $\mathbf{X}_n'\,\mathbf{X}_n = \mathbf{I}_K$

Matrice Inversa di $\mathbf{X} = [x_{ij}]$ $(N \times N)$,
con determinante $det(\mathbf{X}) \neq 0$ (non singolare o invertibile)

$\mathbf{X}^{-1}$ $(N \times N)$: $\mathbf{X}^{-1}\mathbf{X} = \mathbf{I}_N$

Proprietà

$\mathbf{X} = [x_{ij}]$ $(N \times N)$, $\mathbf{Y} = [y_{ij}]$ $(N \times N)$, a scalare

1. $(\mathbf{XY})^{-1} = \mathbf{Y}^{-1}\mathbf{X}^{-1}$;
2. $(a\mathbf{X})^{-1} = a^{-1}\,\mathbf{X}^{-1}$;
3. $|\mathbf{X}^{-1}| = |\mathbf{X}|^{-1}$;
4. $(diag(x_1, \ldots, x_N))^{-1} = \text{diag}(x_1^{-1}, \ldots, x_N^{-1})$.

Matrice Inversa di Moore-Penrose di $\mathbf{X} = [x_{ij}]$ $(N \times J)$:

$\mathbf{X}^+$ $(J \times N)$: $\mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}$; $\mathbf{X}^+\mathbf{X}\mathbf{X}^+ = \mathbf{X}^+$; $(\mathbf{X}\mathbf{X}^+)'$

$= \mathbf{X}\mathbf{X}^+$; $(\mathbf{X}^+\mathbf{X})' = \mathbf{X}^+\mathbf{X}$;

Proprietà

1. $\mathbf{X}^+$ esiste ed è unica;
2. $(\mathbf{X}^+)^+ = \mathbf{X}$;
3. $(\mathbf{X}')^+ = (\mathbf{X}^+)'$;
4. $rk(\mathbf{X}) = J \Rightarrow \mathbf{X}^+ = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$;
5. $rk(\mathbf{X}) = N \Rightarrow \mathbf{X}^+ = \mathbf{X}'\,(\mathbf{X}\mathbf{X}')^{-1}$;
6. $\mathbf{X}'\mathbf{X}\mathbf{X}^+ = \mathbf{X}'$;
7. $\mathbf{X}^+\mathbf{X}\mathbf{X}' = \mathbf{X}'$;
8. $\mathbf{X}'(\mathbf{X}^+)'\mathbf{X}^+ = \mathbf{X}^+$;
9. $\mathbf{X}^+(\mathbf{X}^+)'\mathbf{X}' = \mathbf{X}^+$;
10. $(\mathbf{X}'\mathbf{X})^+ = \mathbf{X}^+(\mathbf{X}^+)'$;
11. $(\mathbf{X}\mathbf{X}')^+ = (\mathbf{X}^+)'\mathbf{X}^+$;
12. $\mathbf{X}\mathbf{X}^+$ e $\mathbf{X}^+\mathbf{X}$ sono idempotenti.

Matrice $\mathbf{X} = [x_{ij}]$ $(N \times N)$ *idempotente*

$\mathbf{X}\mathbf{X} = \mathbf{X}$.

Matrice ortogonale:

$\mathbf{X} = [x_{ij}]$ $(N \times N)$, con determinante $det(\mathbf{X}) \neq 0$ (non singolare) è ortogonale se

$\mathbf{X}' = \mathbf{X}^{-1}$.

Proprietà

1. $\mathbf{X}$ ortogonale $\Leftrightarrow \mathbf{X}'\mathbf{X} = \mathbf{X}\mathbf{X}' = \mathbf{I}_N$;
2. $\mathbf{X}$ ortogonale $\Leftrightarrow \mathbf{X}'$ ortogonale;
3. $\mathbf{X}$ ortogonale $\Leftrightarrow \mathbf{X}^{-1}$ ortogonale;
4. Una matrice di permutazione è ortogonale;
6. La matrice indicatrice è ortogonale.

1. Le strutture statistiche dei dati

Alla base di una completa conoscenza statistica (quantitativa) di un fenomeno in tutte le sue manifestazioni, vi è, una *popolazione* o *universo*, che supponiamo costituita da un numero finito di N *unità statistiche*.
Si supponga di poter osservare o misurare su tale una popolazione J *variabili statistiche*.

1.1 La matrice dei dati

Le *modalità* delle J *variabili* sono disposte in una *matrice dei dati* di dimensione $(N \times J)$,

$$\underset{(N \times J)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1J} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2J} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{iJ} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{N1} & x_{I2} & & x_{Ij} & & x_{NJ} \end{bmatrix} = [\, x_{ij} : i=1,\dots,N;\ j=1,\dots,J],$$

dove l'elemento x_{ij} rappresenta la modalità della variabile j-esima osservata o misurata sull'unità i-esima.
In pratica, i dati osservati, o misurati, sono organizzati in forma tabellare secondo due *modi*: le unità statistiche e le variabili, che rappresentano rispettivamente le righe e le colonne di $\mathbf{X}$.
La matrice dei dati $\mathbf{X}$ può essere vista come un insieme di J colonne di dati, ovvero, composta da J vettori di dimensione $(N \times 1)$, che rappresentano le J variabili,

$$\mathbf{X} = [\mathbf{x}_{.1}, \mathbf{x}_{.2}, \dots, \mathbf{x}_{.j}, \dots, \mathbf{x}_{.J}],$$

dove in particole con $\mathbf{x}_{.j}$ si indica il vettore associato alla variabile j-esima, che si scrive

$$\underset{(N \times 1)}{\mathbf{x}_{.j}} = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{Nj} \end{bmatrix} = [x_{ij} : i=1,..,N].$$

Più frequentemente la matrice dei dati è vista come un insieme di N righe, ovvero, N vettori ($J \times 1$), che rappresentano le N unità statistiche. Queste rappresentano un insieme di *N osservazioni multivariate*

$$\mathbf{X} = [\,\mathbf{x}_{1.}, \mathbf{x}_{2.}, \ldots, \mathbf{x}_{i.}, \ldots, \mathbf{x}_{N.}\,]',$$

dove in particolare con $\mathbf{x}_{i.}$ si indica il vettore associato all'unità statistica i-esima, che si scrive

$$\underset{(J \times 1)}{\mathbf{x}_{i.}} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iJ} \end{bmatrix} = [x_{ij} : j=1,..,J].$$

1.1.2 Variabili qualitative

Nel caso la *variabile* X_j sia *qualitativa* (o *categoriale*) ed assuma un numero finito N_j di diverse modalità (*categorie*), queste ultime possono essere arbitrariamente codificate con i primi N_j numeri interi da 1 a N_j che sono riportati nel vettore $\underline{\mathbf{x}}_j$. Ciò non corrisponde ad una quantificazione del carattere qualitativo che rimane tale, ma semplicemente una convenzione usata con il solo scopo di semplificare la scrittura dei dati; infatti, le modalità qualitative sono sostituite con dei numeri scelti arbitrariamente. Ciò generalmente aiuta a minimizzare gli errori di registrazione dei dati sui supporti informatici.

Le variabili qualitative con N_j modalità (*variabili categoriali*) ($N_j > 1$) possono essere codificate in forma binaria associando ad ogni variabile X_j una matrice indicatrice $\mathbf{B}_j$ questa volta di dimensioni ($N \times N_j$). Le colonne di $\mathbf{B}_j$ corrispondono alle N_j modalità che la variabile X_j può assumere. Ogni unità avrà N_j valori di cui (N_j -1) pari a zero e il rimanente uguale a 1. Se l'unità assume la modalità *l*-esima allora l'1 è posizionato nella colonna *l*-esima. Anche in questo caso si ha che la somma di ogni riga è paria ad 1, ossia

$$\mathbf{B}_j \, \mathbf{1}_{N_j} = \mathbf{1}_N.$$

La matrice $\mathbf{X} = [\mathbf{x}_{.1}, \mathbf{x}_{.2}, \ldots, \mathbf{x}_{.j}, \ldots, \mathbf{x}_{.J}]$ relativa a J variabili qualitative X_j (j=1,…,J) avrà associata la matrice indicatrice a blocchi

$$\mathbf{B} = [\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_J],$$

di dimensioni ($N \times L$) dove $L = \sum_{j=1}^{J} N_j$. La matrice $\mathbf{B}$ ha la proprietà di avere somma per riga costante e pari a J. Infatti, si ha

$$\mathbf{B} \, \mathbf{1}_L = J\mathbf{1}_N,$$

ovvero la somma si ogni riga di B è paria a J.
In generale si ha che

$$\mathbf{B}'_j \, \mathbf{1}_N = \mathbf{B}'_j \mathbf{B}_j \, \mathbf{1}_{N_j},$$

ovvero la somma delle colonne di $\mathbf{B}_j$ sono pari alle frequenze delle modalità della variabile X_j. Infatti, $\mathbf{B}'_j \mathbf{B}_j = diag(\mathbf{B}'_j \, \mathbf{1}_N)$, è una matrice diagonale i cui elementi non nulli sono pari alle frequenze delle modalità della variabile X_j.
Si noti che dalla variabile qualitativa codificata in forma binaria $\mathbf{B}_j$ si ritorna al vettore $\mathbf{x}_j$ mediante la seguente espressione

$$\mathbf{x}_j = \mathbf{B}_j \underline{\mathbf{x}}_{j.}$$

dove gli elementi di $\underline{\mathbf{x}}_j$ sono le codifiche numeriche delle N_j modalità qualitative.

1.2 Le distribuzioni di frequenze semplici

Data la variabile qualitativa X_j che presenta N_j diverse modalità, riportate nel vettore $\underline{\mathbf{x}}_j=[\, x_{1_j}\,,\ x_{2_j}\,,\ \ldots,\, x_{N_j}\,]'$ la distribuzione di frequenze della variabile è

Modalità $\underline{\mathbf{x}}_j$	Frequenze
x_1	n_1
x_2	n_2
…	…
x_k	n_k
…	…
x_{N_j}	n_{N_j}
Totale	N

Per scriverla in forma compatta si consideri la matrice binaria $\mathbf{B}_j$ associata al vettore $\mathbf{x}_j$, la distribuzione semplice di frequenze è rappresentata dalla coppia di vettori

$$(\ \underline{\mathbf{x}}_j,\ \mathbf{c}_j\ ),$$

con

$$\underset{(N_j \times 1)}{\mathbf{c}_j} = \mathbf{B}'_j\mathbf{B}_j\ \mathbf{1}_{N_j}\ ,$$

dove $\mathbf{c}_j$ =[n_{k_j} : k=1,…,N_j] è il *vettore delle frequenze assolute* di dimensione (N_j × 1) i cui elementi n_{k_j} sono le *frequenze assolute* delle unità che presentano la modalità k-esima della variabile j-esima.

Dalle frequenze assolute si passa alle *frequenze relative* mediante la seguente espressione che rappresenta il *vettore delle frequenze relative*

$$\underset{(N_j \times 1)}{{}_r\mathbf{c}_j} = (\ \mathbf{c}'_j\mathbf{1}_{N_j}\ )^{-1}\ \mathbf{c}_j = (\ \mathbf{1}'_{N_j}\mathbf{B}'_j\mathbf{B}_j\mathbf{1}_{N_j}\ )^{-1}\mathbf{B}'_j\mathbf{B}_j\ \mathbf{1}_{N_j}\ ,$$

dove ${}_r\mathbf{c}_j$ =[f_{k_j} : k=1,…,N_j] è un vettore di dimensione (N_j × 1) i cui elementi f_{k_j} sono le *frequenze relative* delle unità che presentano la modalità k-esima della variabile j-esima.

Date due variabili qualitative X_j e X_m con N_j ed N_m diverse modalità riportate nei vettori $\underline{\mathbf{x}}_j$ e $\underline{\mathbf{x}}_m$, la *tabella di contingenza* relativa alle due variabili si scrive

Modalità $\underline{\mathbf{x}}_j$	Modalità $\underline{\mathbf{x}}_m$					
		x_{1_m}	x_{2_m}	…	x_{N_m}	Totale riga
	x_{1_j}	$n_{1_j 1_m}$	$n_{1_j 2_m}$	…	$n_{1_j N_m}$	$n_{1_j \cdot}$
	x_{2_j}	$n_{2_j 1_m}$	$n_{2_j 2_m}$	…	$n_{2_j N_m}$	$n_{2_j \cdot}$
	…	…	…	…	…	…
	x_{N_j}	$n_{N_j 1_m}$	$n_{N_j 2_m}$	…	$n_{N_j N_m}$	$n_{N_j N_m}$
	Totale colonna	$n_{\cdot_j 1_m}$	$n_{\cdot_j 2_m}$	…	$n_{\cdot_j N_m}$	N

La distribuzione doppia espressa in forma matriciale, considerando le matrici binarie $\mathbf{B}_j$ e $\mathbf{B}_m$, associate a X_j e X_m, si scrive

$$\begin{bmatrix} X_j \backslash X_m & \underline{\mathbf{x}}'_m & \\ \underline{\mathbf{x}}_j & \mathbf{C}_{jm} & \mathbf{c}_j \\ & \mathbf{c}'_m & N \end{bmatrix}$$

dove $\mathbf{c}_j$ e $\mathbf{c}'_m$ sono i vettori delle frequenze marginali della tabella doppia; in $\underline{\mathbf{x}}_j$ e $\underline{\mathbf{x}}'_m$ sono riportate le modalità delle variabili X_j e X_m.

La tabella doppia è sinteticamente rappresentata dalla distribuzione delle frequenze, ovvero dalla *matrice delle frequenze assolute*

$\mathbf{C}_{jm}$,

con

$$\underset{(N_j \times N_m)}{\mathbf{C}_{jm}} = \mathbf{B}'_j \mathbf{B}_m ,$$

dove $\mathbf{C}_{jm} = [\, n_{k_j l_m} : k=1,\ldots,N_j;\ l=1,\ldots,N_m]$ è una matrice di dimensioni $(N_j \times N_m)$ i cui elementi $n_{k_j l_m}$ sono le *frequenze assolute* delle unità che presentano la modalità k-esima ed l-esima rispettivamente delle variabili X_j ed X_m .

Dalle frequenze assolute si passa alle *frequenze relative di riga*, e quindi alla *tabella di contingenza normalizzata per riga* mediante la seguente *matrice delle frequenze relative di riga*

$$ {}_{r}\mathbf{C}_{jm} = (\mathbf{B}'_j\mathbf{B}_j)^{-1}\,\mathbf{C}_{jm} = (\mathbf{B}'_j\mathbf{B}_j)^{-1}\,\mathbf{B}'_j\mathbf{B}_m = \mathbf{B}^{+}_j\,\mathbf{B}_m , $$
$$ (N_j \times N_m) $$

dove ${}_{r}\mathbf{C}_{jm} = [\ {}_{r}f_{k_j l_m} : k=1,\ldots,N_j;\ l=1,\ldots,N_m]$ è una matrice di dimensioni $(N_j \times N_m)$ i cui elementi ${}_{r}f_{k_j l_m}$ $(l=1,\ldots,N_m)$ sono le frequenze relative (di riga) delle unità al variare della modalità l-esima della variabile m-esima, e della modalità k-esima delle variabili j-esima.

La matrice $\mathbf{B}^{+}_j$ è la matrice inversa di Moore-Penrose di $\mathbf{B}_j$ supposta di rango massimo (generalmente pari a N_j).

Analogamente la *tabella di contingenza normalizzata per colonna*, ove cioè le colonne sono frequenze relative è rappresentata dalla *matrice delle frequenze relative di colonna*

$$ {}_{c}\mathbf{C}_{jm} = \mathbf{C}_{jm}\,(\mathbf{B}'_m\mathbf{B}_m)^{-1} = \mathbf{B}'_j\,\mathbf{B}_m\,(\mathbf{B}'_m\mathbf{B}_m)^{-1} = \mathbf{B}'_j\,\mathbf{B}^{+\prime}_m = (\mathbf{B}^{+}_m\,\mathbf{B}_j)', $$
$$ (N_j \times N_m) $$

dove ${}_{c}\mathbf{C}_{jm} = [\ {}_{c}f_{k_j l_m} : k=1,\ldots,N_j;\ l=1,\ldots,N_m]$ è una matrice di dimensioni $(N_j \times N_m)$ i cui elementi ${}_{c}f_{k_j l_m}$ $(k=1,\ldots,N_j)$ sono le frequenze relative (di colonna) delle unità al variare della modalità k-esima della variabile j-esima e della modalità l-esima delle variabili m-esima. La matrice $\mathbf{B}^{+}_m$ è la matrice inversa di Moore-Penrose della matrice $\mathbf{B}_m$ supposta di rango massimo.

Infine, la *tabella di contingenza normalizzata per il numero totale* di unità si ottiene dividendo le frequenze assolute per il numero N di unità; in tal caso *la matrice delle frequenze relative al totale* delle unità è

$$ {}_{T}\mathbf{C}_{jm} = N^{-1}\,\mathbf{C}_{jm} = [\ {}_{T}f_{k_j l_m} : k=1,\ldots,N_j;\ l=1,\ldots N_l]. $$
$$ (N_j \times N_m) $$

2.1 La media

Il valore atteso $E(X_j)$ della variabile aleatoria X_j è la *media aritmetica* μ_j delle modalità, riportate nel vettore $\mathbf{x}_j$, osservate sulle N unità statistiche della popolazione sulle quali si manifesta X_j

$$E(X_j) = \mu_j = \left(\mathbf{1}'_N\mathbf{1}_N\right)^{-1}\mathbf{1}'_N\mathbf{x}_j = \mathbf{1}^{+}_N\mathbf{x}_j = \frac{1}{N}\mathbf{1}'_N\mathbf{x}_j = \frac{1}{N}\mathbf{x}'_j\mathbf{1}_N,$$

dove $\mathbf{1}_N$ è un vettore a N componenti unitarie, ossia il vettore N-dimensionale $\mathbf{1}_N = [1, 1, \ldots, 1]'$.

2.2 Il vettore delle medie

Il valore atteso $E(\mathbf{x})$ del vettore aleatorio $\mathbf{x}$ è il *vettore delle medie aritmetiche* delle variabili aleatorie $X_1, X_2, \ldots, X_J$, ossia il vettore le cui componenti sono i J valori attesi o medie μ_j delle modalità, riportate nella matrice $\mathbf{X}$, osservate sulle N unità statistiche della popolazione,

$$E(\mathbf{x}) = \underset{(J\times 1)}{\boldsymbol{\mu}_{\mathbf{X}}} = E[X_1, X_2, \ldots, X_J]' = [E(X_1), \mathrm{E}(X_2), \ldots, \mathrm{E}(X_J)]'$$

$$= [\mu_1, \mu_2, \ldots, \mu_j, \ldots, \mu_J]'$$

$$\underset{(J\times 1)}{\boldsymbol{\mu}_{\mathbf{X}}} = \left(\frac{1}{N}\mathbf{1}'_N\mathbf{X}\right)' = \left(\left(\mathbf{1}'_N\mathbf{1}_N\right)^{-1}\mathbf{1}'_N\mathbf{X}\right)' = \left(\mathbf{1}^{+}_N\mathbf{X}\right)' = \left(\frac{1}{N}\mathbf{X}'\mathbf{1}_N\right)$$

$$= \frac{1}{N}\left[\mathbf{x}_{.1}, \mathbf{x}_{.2}, \ldots, \mathbf{x}_{.J}\right]'\mathbf{1}_N = [\mu_1, \mu_2, \ldots, \mu_J]'.$$

Il vettore delle medie si può scrivere anche a partire dal vettori delle osservazioni multivariate $\mathbf{x}_{1.}, \mathbf{x}_{2.}, \ldots, \mathbf{x}_{i.}, \ldots, \mathbf{x}_{N.}$,

$$\underset{(J\times 1)}{\boldsymbol{\mu}_{\mathbf{X}}} = \left(\frac{1}{N}\mathbf{1}'_N\mathbf{X}\right)' = \frac{1}{N}\left[\mathbf{x}_{1.}, \ldots, \mathbf{x}_{N.}\right]\mathbf{1}_N = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{i.} = [\mu_1, \mu_2, \ldots, \mu_J]',$$

2.3 Il vettore delle medie delle distribuzioni parziali

Quando si osserva una variabile quantitativa X_j e una variabile qualitativa X_m con associata matrice indicatrice $\mathbf{B}_m$, per le distribuzioni parziali della variabile X_j condizionatamente alle modalità della variabile X_m si possono calcolare le medie che possono essere disposte nel *vettore delle medie condizionate*

$$\underset{(N_m \times 1)}{\boldsymbol{\mu}_{j|m}} = [\mu_{j|1}, \mu_{j|2}, \ldots \mu_{j|N_m}]' = (\mathbf{B}_m' \mathbf{B}_m)^{-1} \mathbf{B}_m' \mathbf{x}_j = \mathbf{B}_m^{+} \mathbf{x}_j .$$

Il *vettore delle medie campionarie delle distribuzione parziali* di X_j condizionatamente alle modalità di X_m per le n osservazione campionarie multivariate $\mathbf{x}_{1.}, \mathbf{x}_{2.}, \ldots, \mathbf{x}_{i.}, \ldots,$

2.4 La varianza

La *varianza* $Cov(X_j) = \sigma_j^2$ (che si scrive anche con la notazione σ_{jj}) della variabile aleatoria X_j è

$$Cov(X_j) = E(X_j - \mu_j)^2 = \sigma_j^2 .$$

La varianza della variabile j-esima si scrive a partire dal vettore degli scarti ${}_c\mathbf{x}_j$ associato alla variabile j-esima

$$\sigma_j^2 = \frac{1}{N} {}_c\mathbf{x}_j' \, {}_c\mathbf{x}_j = \frac{1}{N}(\mathbf{x}_j - \mathbf{1}_N \mu_j)'(\mathbf{x}_j - \mathbf{1}_N \mu_j) = \frac{1}{N}\left(\mathbf{x}_j - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N'\mathbf{x}_j\right)'\left(\mathbf{x}_j - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N'\mathbf{x}_j\right) =$$

$$= \frac{1}{N}\mathbf{x}_j'\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N'\right)'\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N'\right)\mathbf{x}_j = \frac{1}{N}\mathbf{x}_j'\mathbf{J}\mathbf{x}_j ,$$

dove ${}_c\mathbf{x}_j = \mathbf{x}_j - \mathbf{1}_N \mu_j$ è il vettore degli scarti dalla media aritmetica.
Dalla precedente espressione si ricava anche

$$\sigma_j^2 = \frac{1}{N}\mathbf{x}_j'\mathbf{x}_j - \left(\frac{1}{N}\mathbf{1}_N'\mathbf{x}_j\right)^2 .$$

Infine, la varianza σ_j^2 si può calcolare anche come norma euclidea del vettore $\mathbf{x}_j$, ossia

$$\sigma_j^2 = \frac{1}{N}\left\| {}_c\mathbf{x}_j \right\|^2 .$$

2.9 La covarianza

La *covarianza* $Cov(X_j, X_m) = \sigma_{jm}$ tra le variabili aleatorie quantitative X_j, X_m si scrive

$$\sigma_{jm} = \frac{1}{N} {}_c\mathbf{x}'_j \, {}_c\mathbf{x}_m$$

$$= \frac{1}{N}\left(\mathbf{x}_j - \mathbf{1}_N \mu_j\right)'\left(\mathbf{x}_m - \mathbf{1}_N \mu_m\right) = \frac{1}{N}\left(\mathbf{x}_j - \frac{1}{N}\mathbf{1}_N\mathbf{1}'_N\mathbf{x}_j\right)'\left(\mathbf{x}_m - \frac{1}{N}\mathbf{1}_N\mathbf{1}'_N\mathbf{x}_m\right) =$$

$$= \frac{1}{N}\mathbf{x}'_j\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}'_N\right)'\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}'_N\right)\mathbf{x}_m = \frac{1}{N}\mathbf{x}'_j\mathbf{J}\mathbf{x}_m .$$

Dalla precedente espressione si ricava anche

$$\sigma_{jm} = \frac{1}{N}\mathbf{x}'_j\mathbf{x}_m - \left(\frac{1}{N}\mathbf{1}'_N\mathbf{x}_j\right)\left(\frac{1}{N}\mathbf{1}'_N\mathbf{x}_m\right).$$

2.10 Il coefficiente di correlazione

$$r_{jm} = \frac{\sigma_{jm}}{\sigma_j\sigma_m} = \frac{\frac{1}{N}\mathbf{x}'_j\mathbf{J}\mathbf{x}_m}{\frac{1}{N}\sqrt{\mathbf{x}'_j\mathbf{J}\mathbf{x}_j\mathbf{x}'_m\mathbf{J}\mathbf{x}_m}} = \frac{\mathbf{x}'_j\mathbf{J}\mathbf{x}_m}{\sqrt{\mathbf{x}'_j\mathbf{J}\mathbf{x}_j\mathbf{x}'_m\mathbf{J}\mathbf{x}_m}} .$$

2.11 La matrice di varianze e covarianze

Le varianze e le covarianze tra le coppie di J variabili si possono disporre nella matrice **di varianze e covarianze** quadrata $\Sigma_{\mathbf{X}}$ di dimensione J, simmetrica.

$$\underset{(J\times J)}{\Sigma_{\mathbf{X}}} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1J} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2J} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{J1} & \sigma_{J2} & \dots & \sigma_J^2 \end{bmatrix}.$$

La matrice di varianze e covarianze si ricava a partire dalle osservazioni multivariate,

$$\Sigma_{\mathbf{X}} = Cov(\mathbf{x}) = E(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})'$$

$$= \frac{1}{N}\left[\mathbf{x}_{1.} - \boldsymbol{\mu}_{\mathbf{X}}, \mathbf{x}_{2.} - \boldsymbol{\mu}_{\mathbf{X}}, \dots, \mathbf{x}_{N.} - \boldsymbol{\mu}_{\mathbf{X}}\right]\left[\mathbf{x}_{1.} - \boldsymbol{\mu}_{\mathbf{X}}, \mathbf{x}_{2.} - \boldsymbol{\mu}_{\mathbf{X}}, \dots, \mathbf{x}_{N.} - \boldsymbol{\mu}_{\mathbf{X}}\right]'$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_{i.} - \boldsymbol{\mu}_{\mathbf{X}}\right)\left(\mathbf{x}_{i.} - \boldsymbol{\mu}_{\mathbf{X}}\right)' = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{i.}\mathbf{x}_{i.}' - \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\mu}'_{\mathbf{X}},$$

essendo $\mathbf{X}'\mathbf{X} = [\mathbf{x}_{1.}, \mathbf{x}_{2.}, \dots, \mathbf{x}_{N.}][\mathbf{x}_{1.}, \mathbf{x}_{2.}, \dots, \mathbf{x}_{N.}]' = \sum_{i=1}^{N}\mathbf{x}_{i.}\mathbf{x}_{i.}'$.

La matrice di varianze e covarianze si ottiene a partire dalla matrice $\mathbf{X}$ centrata

$$\Sigma_{\mathbf{X}} = \frac{1}{N}\mathbf{X}_c'\mathbf{X}_c = \frac{1}{N}\mathbf{X}'\mathbf{J}\mathbf{X}.$$

La matrice di varianze e covarianze si può ricavare a partire dai vettori $\mathbf{x}_{.1}, \mathbf{x}_{.2}, \dots, \mathbf{x}_{.j}, \dots, \mathbf{x}_{.J}$ associati alle J variabili,

$$\Sigma_{\mathbf{X}} = \frac{1}{N}\mathbf{X}'\mathbf{J}\mathbf{X} = \frac{1}{N}\left[\mathbf{x}_{.1}, \mathbf{x}_{.2}, \dots, \mathbf{x}_{.J}\right]'\mathbf{J}\left[\mathbf{x}_{.1}, \mathbf{x}_{.2}, \dots, \mathbf{x}_{.J}\right]$$

$$= \frac{1}{N}\begin{bmatrix}\mathbf{x}_{.1}' \\ \mathbf{x}_{.2}' \\ \dots \\ \mathbf{x}_{.J}'\end{bmatrix}\mathbf{J}\left[\mathbf{x}_{.1}, \mathbf{x}_{.2}, \dots, \mathbf{x}_{.J}\right] = \begin{bmatrix} \frac{1}{N}\mathbf{x}_{.1}'\mathbf{J}\mathbf{x}_{.1} & \frac{1}{N}\mathbf{x}_{.1}'\mathbf{J}\mathbf{x}_{.2} & \dots & \frac{1}{N}\mathbf{x}_{.1}'\mathbf{J}\mathbf{x}_{.J} \\ \frac{1}{N}\mathbf{x}_{.2}'\mathbf{J}\mathbf{x}_{.1} & \frac{1}{N}\mathbf{x}_{.2}'\mathbf{J}\mathbf{x}_{.2} & \dots & \frac{1}{N}\mathbf{x}_{.2}'\mathbf{J}\mathbf{x}_{.J} \\ \dots & \dots & \dots & \dots \\ \frac{1}{N}\mathbf{x}_{.J}'\mathbf{J}\mathbf{x}_{.1} & \frac{1}{N}\mathbf{x}_{.J}'\mathbf{J}\mathbf{x}_{.2} & \dots & \frac{1}{N}\mathbf{x}_{.J}'\mathbf{J}\mathbf{x}_{.J} \end{bmatrix}.$$

La matrice di covarianza essendo $\Sigma_{\mathbf{X}} = \frac{1}{N}\mathbf{X}'\mathbf{J}\mathbf{X}$ si può anche scrivere essendo $\mathbf{J} = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}'_N$

$$\Sigma_{\mathbf{X}} = \frac{1}{N}\mathbf{X}'\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}'_N\right)\mathbf{X} = \frac{1}{N}\mathbf{X}'\mathbf{X} - \frac{1}{N}\mathbf{X}'\mathbf{1}_N\mathbf{1}'_N\mathbf{X}\frac{1}{N} = \frac{1}{N}\mathbf{X}'\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\mu}'_{\mathbf{X}}.$$

2.12 La matrice di correlazione

I coefficienti di correlazioni tra le coppie di J variabili si possono disporre nella matrice **di varianze e covarianze** quadrata **R** di dimensione J, simmetrica e semidefinta positiva.

$$\underset{(J\times J)}{\mathbf{R_X}} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1J} \\ r_{21} & 1 & \dots & \sigma_{2J} \\ \vdots & \vdots & \vdots & \vdots \\ r_{J1} & r_{J2} & \dots & 1 \end{bmatrix}.$$

La matrice di correlazione si ottiene

$$\underset{(J\times J)}{\mathbf{R_X}} = \mathbf{D}^{-1} \underset{(J\times J)}{\sum_{\mathbf{X}}} \mathbf{D}^{-1},$$

dove: $\mathbf{D} = dg(\sigma_1,\dots\sigma_J)$ è una matrice diagonale che ha gli elementi non nulli pari agli scostamenti quadratici medi delle J variabili. Dalla matrice di correlazione si ricava la matrice di varianze e covarianze

$$\underset{(J\times J)}{\sum_{\mathbf{X}}} = \mathbf{D}\,\mathbf{R_X}\,\mathbf{D}.$$

2.13 Il Chi-quadrato

La dipendenza fra due variabili qualitative X_j ed X_m come è noto si misura mediante il chi-quadrato verificando la dissimilarità tra la tabella di contingenze osservata C_{jm} e quella di indipendenza $N\ {}_r\mathbf{c}_j\ {}_r\mathbf{c}'_m$, dove ${}_r\mathbf{c}_h=(\mathbf{c}'_h\mathbf{1}_{N_h})^{-1}\mathbf{c}_h=(\mathbf{1}'_{N_h}\mathbf{B}'_h\mathbf{B}_h\mathbf{1}_{N_h})^{-1}\mathbf{B}'_h\mathbf{B}_h\ \mathbf{1}_{N_h}$, ($h$=$j$, m) sono le distribuzioni delle frequenze relative delle due variabili X_j e X_m.

Il chi-quadrato espresso come quadrato delle distanza euclidea tra la tabella osservata e la tabella di indipendenza si scrive

$$\chi^2_{jm} = \sum_{k=1}^{N_j}\sum_{l=1}^{N_m}\frac{\left(n_{kl}-\frac{n_{k.}n_{.l}}{N}\right)^2}{\frac{n_{k.}n_{.l}}{N}} = N\sum_{k=1}^{N_j}\sum_{l=1}^{N_m}\frac{({}_Tf_{kl}-{}_Tf_{k.\,T}f_{.l})^2}{{}_Tf_{k.\ T}f_{.l}}$$

$$= N^3\ tr[(\mathbf{B}'_j\mathbf{B}_j)^{-1}({}_T\mathbf{C}_{jm}-{}_r\mathbf{c}_j\ {}_r\mathbf{c}'_m)(\mathbf{B}'_m\mathbf{B}_m)^{-1}({}_T\mathbf{C}_{jm}-{}_r\mathbf{c}_j\ {}_r\mathbf{c}'_m)']$$

Il chi-quadrato χ^2_{jm} tra le variabile qualitative X_j e X_m si scrive anche

$$\chi^2_{jm} = N\left(\sum_{k=1}^{N_j}\sum_{l=1}^{N_m}\frac{n^2_{kl}}{n_{k.}n_{.l}}-1\right)$$

$$= N(tr({}_r\mathbf{C}'_{jm}\ {}_c\mathbf{C}_{jm})\ \text{-1}) = N(tr(\mathbf{B}'_m\ \mathbf{B}'^{+}_j\ \mathbf{B}'_j\ \mathbf{B}'^{+}_m)\text{-1})$$

$$= N(tr(\mathbf{P}_{\mathbf{B}_j}\ \mathbf{P}_{\mathbf{B}_m})\ \text{-1}),$$

dove $\mathbf{P}_{\mathbf{B}_j} = \mathbf{B}_j(\mathbf{B}'_j\ \mathbf{B}_j)^{-1}\mathbf{B}'_j = \mathbf{B}_j\mathbf{B}^{+}_j$ è una matrice di proiezione della matrice $\mathbf{B}_j$ e nella precedente espressione si è utilizzata la proprietà circolare della traccia del prodotto di tre matrici **A**, **B**, **C**: tr(**ABC**)=tr(**BCA**)=tr(**CAB**). Possiamo anche scrivere

$$\chi^2_{jm} = N\sum_{k=1}^{N_j}n_{k.}\left({}_r\mathbf{c}_{j|k}-{}_r\mathbf{c}_m\right)'(\mathbf{B}'_m\mathbf{B}_m)^{-1}\left({}_r\mathbf{c}_{j|k}-{}_r\mathbf{c}_m\right)$$

$$= N\sum_{l=1}^{N_m}n_{.l}\left({}_r\mathbf{c}_{m|l}-{}_r\mathbf{c}_j\right)'(\mathbf{B}'_j\mathbf{B}_j)^{-1}\left({}_r\mathbf{c}_{m|l}-{}_r\mathbf{c}_j\right).$$

Le ultime due espressioni del chi-quadrato individuano due modi alternativi per definire l'indipendenza ovvero che le distribuzioni parziali sono simili ovvero le distribuzioni parziali delle frequenze relative di riga sono tra loro uguali e quindi uguali a quelle della distribuzione marginale, formalmente ${}_r\mathbf{c}_{j|l}={}_r\mathbf{c}_m$ per l=1,…,N_j, ${}_r\mathbf{c}_{m|k}={}_r\mathbf{c}_j$ per k=1,…,N_m.

Statistica e Fisica

La statistica per un fisco sperimentale è uno dei principali strumenti di lavoro

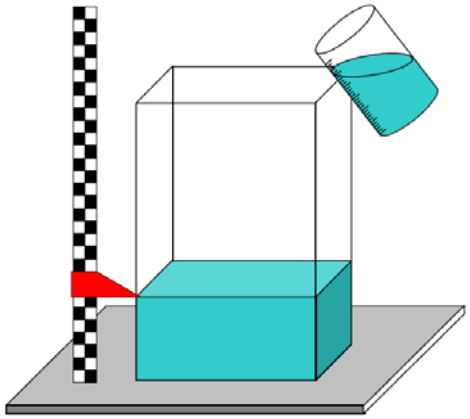
- rilevazione statistica – per la realizzazione di un esperimento di fisica, chimica, medicina
- Rilevazione dei dati
- Randomizzazione dell'esperimento
- Piano degli esperimenti (caso controllo, disegno ottimo) (Biostatistica)

  Replicazione, blocking (gruppi), Otogonalità (confronto)

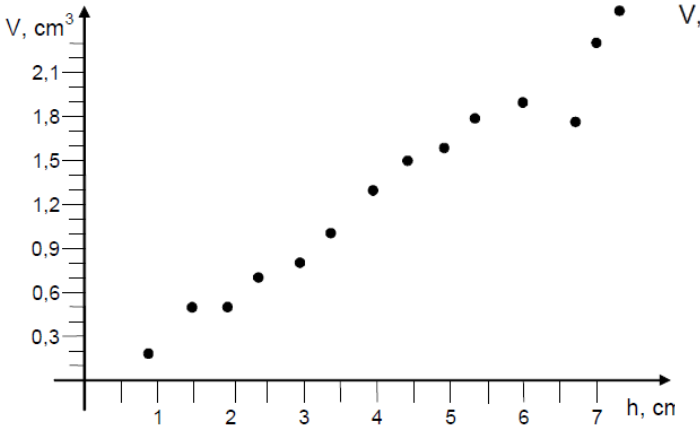
ESEMPIO: Rilevazione dati di un esperimento sulla relazione fra variabili

Esperimento volume dell'acqua in un beaker graduato e l'altezza della colonna misurata in cm con un righello graduato

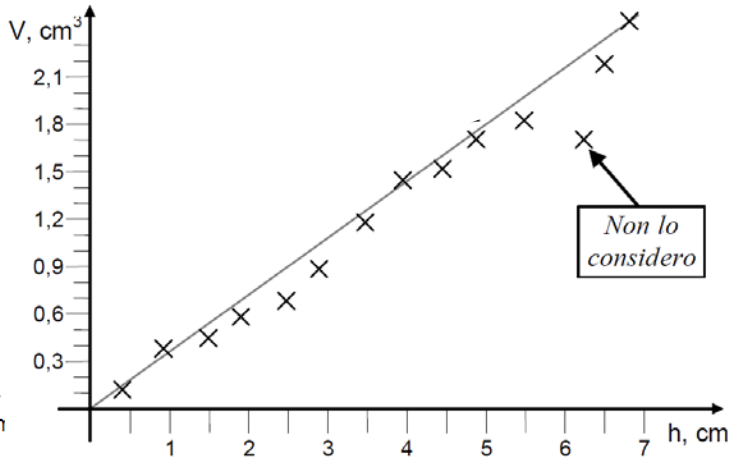
Misurazione



rappresentazione grafica dei dati misurati



interpolazione con una funzione lineare retta di regressione (uso excell)



- SOFTWARE
- Excel
- ACESS
- Matlab o Scilab
- R

Microsoft Excel

Unità	sesso	Età	Titolo di studio	Reddito mensile	Consumo mensile	Acquisto. quotidiano
1	0	20	1	1050	800	1
2	0	19	3	1000	900	1
3	1	21	3	2000	1300	1
4	1	75	2	1200	1000	0
5	1	45	4	2200	800	1
6	1	35	5	2500	1500	0
7	1	21	2	1250	1000	0
8	0	60	2	1800	1100	1
9	0	18	2	1640	1350	1
10	1	12	1	1400	1000	1

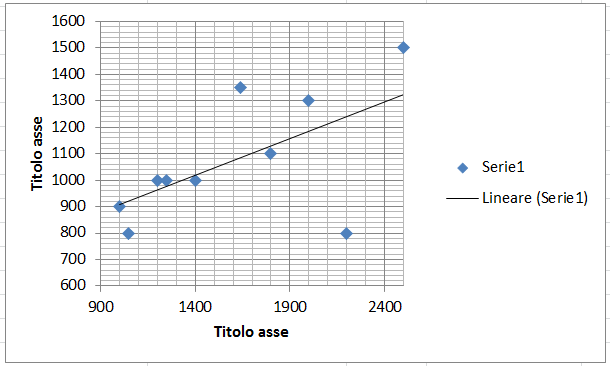
File Home Inserisci Layout di pagina Formule Dati Revisione Visualizza Acrobat

Taglia Copia Copia formato — Calibri 11 — Testo a capo — Generale — Unisci e allinea al centro — Normale Neutrale Valore non v... Valore valido — Formattazione condizionale — Formatta come tabella — Inserisci Elimina Formato — Somma automatica Riempimento Cancella — Ordina e filtra Trova e seleziona

Appunti Carattere Allineamento Numeri Stili Celle Modifica

E24

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	unità	sesso	età	Titolo di studio	Reddito	Sonsumi	Acquisto								
2	1	0	20	1	1050	800	1								
3	2	0	19	3	1000	900	1								
4	3	1	21	3	2000	1300	1			Conteggio di sesso	sesso				
5	4	1	75	2	1200	1000	0			Titolo di studio	0	1	Totale complessivo		
6	5	1	45	4	2200	800	1			1	1	1	2		
7	6	1	35	5	2500	1500	0			2	2	2	4		
8	7	1	21	2	1250	1000	0			3	1	1	2		
9	8	0	60	2	1800	1100	1			4		1	1		
10	9	0	18	2	1640	1350	1			5		1	1		
11	10	1	12	1	1400	1000	1			Totale complessivo		4	6	10	
12															
13															
14		Media di Reddito	sesso												
15		Titolo di studio	0	1	Totale complessivo										
16		1	1050	1400	1225										
17		2	1720	1225	1472,5										
18		3	1000	2000	1500										
19		4		2200	2200										
20		5		2500	2500										
21		Totale complessivo	1372,5	1758,333333	1604										
22															
23															
24															
25															
26															
27															
28															
29															
30															
31															
32															



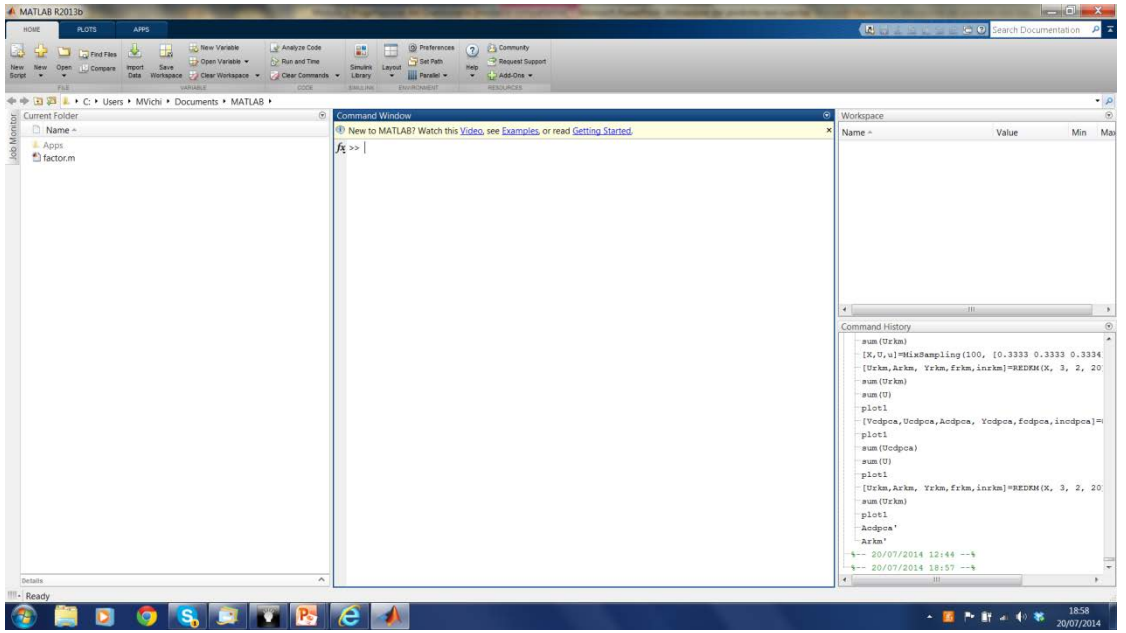
Foglio1 Foglio2 Foglio3

Pronto 100%

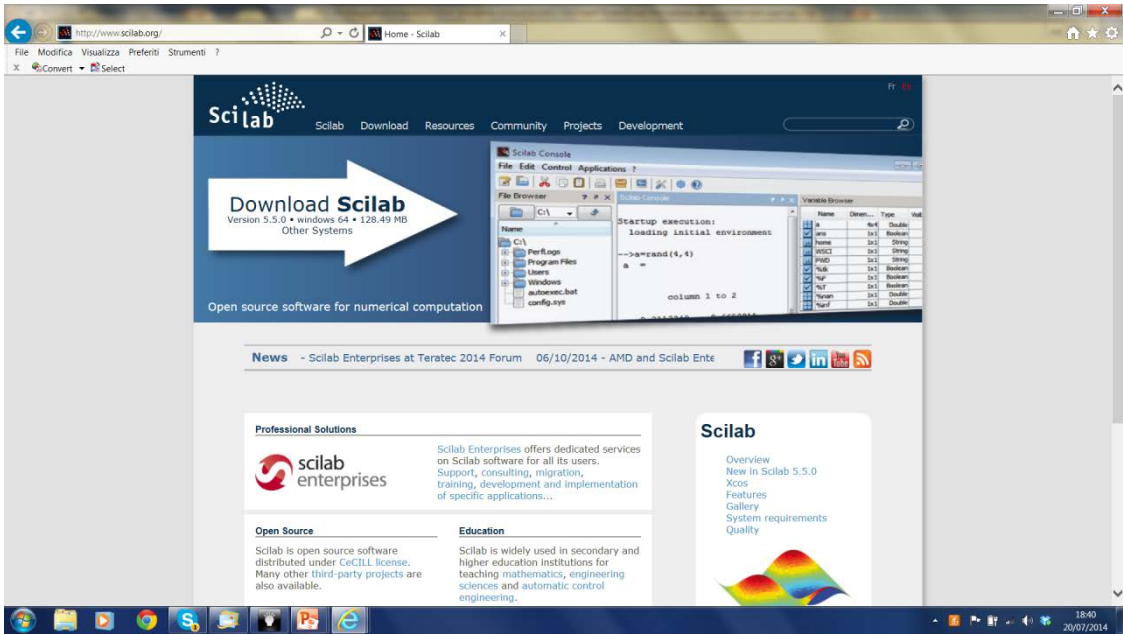
IT 19:31 04/04/2013

Il software Matlab o Scilab o GNU Octave

Software commerciale



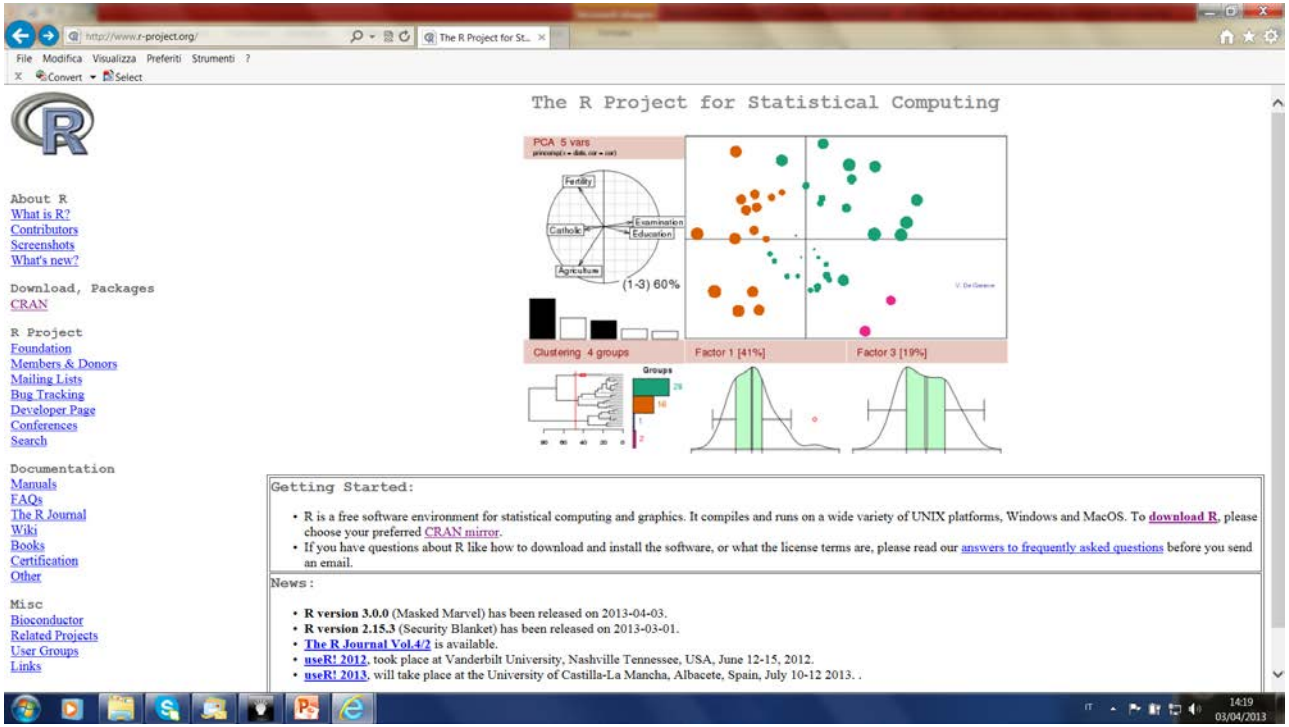
Software libero facilmente scaricabile



Il software R

http://www.r-project.org/

R è un ambiente «open source» di sviluppo specifico per l'analisi statistica dei dati che utilizza un linguaggio di programmazione derivato



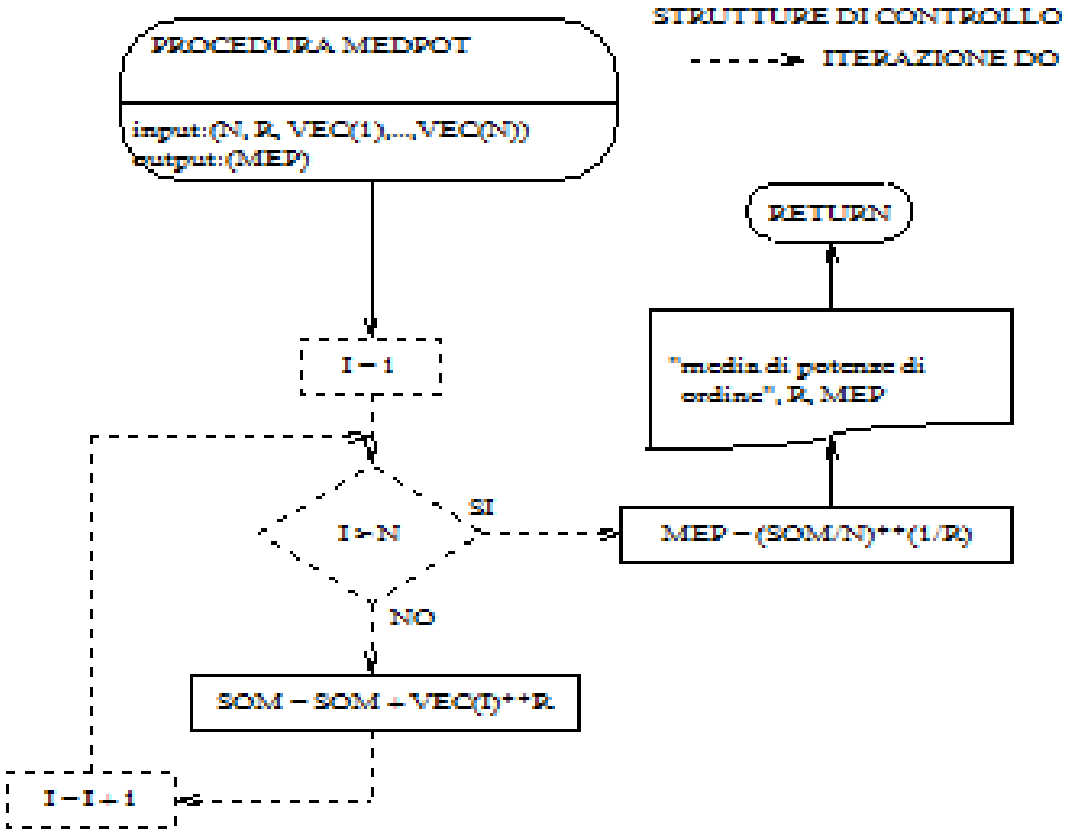
Si possono trattare matrici e vettori che sono gli strumenti di base della statistica multivariata.

- Esistono packages per ogni analisi statistica conosciuta anche molto avanzata.

INDICI STATISTICI come ALGORITMI di CALCOLO

- indicatori di sintesi dei dati: medie, indici di variabilità di correlazione, di regressione viste come algoritmi di calcolo

$$Mr = \sqrt[r]{\frac{1}{n}\sum_{i=1}^{n} x_i^r}$$




```
Program MediaVettore;
Uses Crt;
Const Max=30;
Var V:ARRAY [1..Max] OF integer;
    Dim,i,Tot:integer;
    Med:real;

Begin
 Clrscr;
 Writeln ('Programma per calcolare il valore
medio di un vettore numerico');
 REPEAT
  writeln;
  Write('Inserisci la dimensione del vettore: ');
  Readln (N);
  Writeln;
 UNTIL (N>=1) AND (N<=Max);
 FOR i:=1 TO N DO
  Begin
   Write ('Dammi il numero di posizione ',i,': ');
   Readln (VEC[i]);
  End;
 SOM:=VEC[1];
 Med:=0;
 FOR i:=2 TO N DO SOM:=SOM+VEC[i];
 Med:=(SOM/N);
 Writeln ('La media è: ',Med:5:2);
 Readln;
End.
```

Statistica e Fisica (o altre scienze dure)

La statistica per un fisco sperimentale è uno dei principali strumenti di lavoro

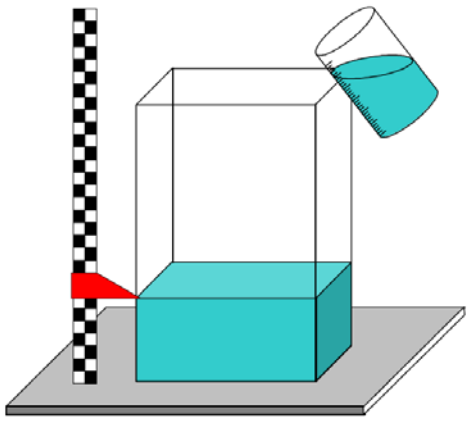
- rilevazione statistica – per la realizzazione di un esperimento di fisica, chimica, medicina
- Rilevazione dei dati
- Randomizzazione dell'esperimento
- Piano degli esperimenti (caso controllo, disegno ottimo) (Biostatistica)

  Replicazione, blocking (gruppi), Otogonalità (confronto)

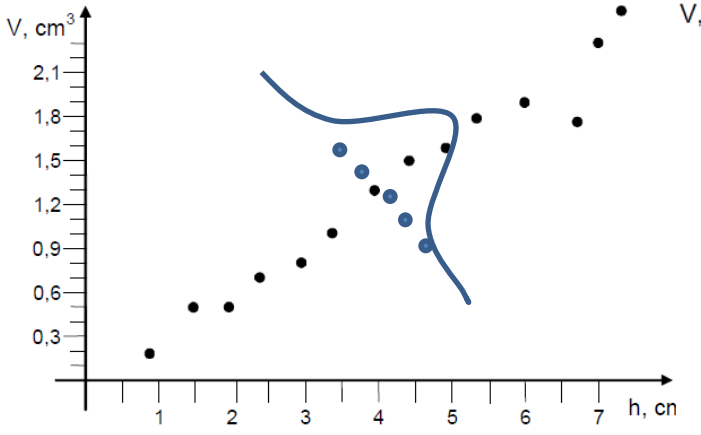
ESEMPIO: Rilevazione dati di un esperimento sulla relazione fra variabili

Esperimento volume dell'acqua in un beaker graduato e l'altezza della colonna misurata in cm con un righello graduato

Misurazione



rappresentazione grafica dei dati misurati



interpolazione con una funzione lineare retta di regressione (uso excell)