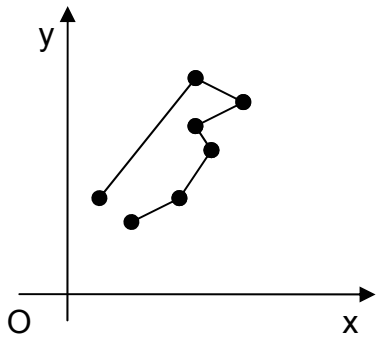
Interpolazione statistica

Premessa sulle relazioni statistiche

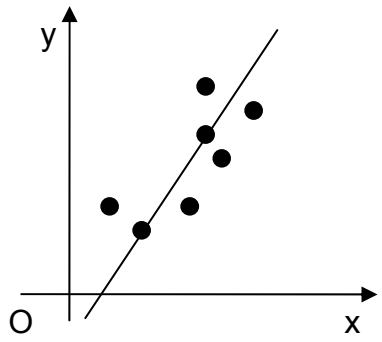
Nell'esame di fenomeni di qualsiasi natura si cerca di esprimere, sia mediante relazioni matematiche, sia mediante specifici indici, i legami rilevati, o ipotizzati, fra le grandezze che interagiscono nei fenomeni stessi.
Nello studio dei legami fra due variabili statistiche, partendo da un insieme di coppie (x_i, y_i) di dati rilevati, si determina, se possibile, una funzione $y = f(x)$ che rappresenti il fenomeno.
Vari sono gli scopi della ricerca di tale funzione; fra essi ricordiamo:
- descrivere sinteticamente la relazione fra due variabili osservate;
- determinare la legge di distribuzione dei dati statistici;
- ricavare eventuali dati intermedi mancanti;
- correggere valori affetti da errori accidentali o perturbati da cause secondarie.

Per trovare una funzione che rappresenti il fenomeno si può procedere in due modi:
- determinare una funzione che assuma esattamente i valori (x_i, y_i) rilevati; questo procedimento viene detto interpolazione per punti noti, o **interpolazione matematica**;
- determinare una funzione il cui grafico "si accosti" il più possibile ai punti del diagramma a dispersione; questo procedimento viene detto interpolazione (o perequazione) fra punti noti, o **interpolazione statistica**.

In Statistica, al contrario di quanto accade in Matematica, scegliendo una funzione polinomiale che sia soddisfatta da tutte le coppie assegnate di valori, essendo il numero delle coppie di valori piuttosto elevato, risulterebbe tanto complessa da determinare, quanto di scarsa utilità.
Per tale motivo nelle applicazioni statistiche si preferisce cercare una funzione il cui grafico "si avvicini" al grafico rappresentativo delle coppie di valori rilevati.
Il procedimento che trattiamo utilizza il metodo dei minimi quadrati che illustriamo.



Interpolazione matematica



Interpolazione statistica

Nell'interpolazione matematica la funzione interpolante cercata passa per i valori (x_i, y_i) mentre nell'interpolazione statistica la funzione interpolante cercata passa tra i valori (x_i, y_i).

Metodo dei minimi quadrati

Si considerino due variabili X ed Y sulle quali sono effettuate n rilevazioni espresse dalle coppie:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

Si presentano due problemi:
- scegliere il tipo di funzione che si ritiene esprima meglio la relazione tra X ed Y;
- determinare i parametri della funzione scelta.

Tale funzione è detta funzione **interpolante**.

Per quanto riguarda la scelta della funzione interpolante, non esistono criteri generali validi per ogni caso e si possono solo dare delle indicazioni.
La scelta della funzione dipende da un'eventuale relazione tra le variabili riscontrata dall'osservazione dei valori assunti dalle stesse. Ad esempio, se gli incrementi dei valori di Y, per incrementi costanti di X, sono quasi costanti, la curva che meglio rappresenta il fenomeno è la retta. Se invece il confronto tra i valori osservati presenta caratteristiche diverse, allora la curva che meglio rappresenta il fenomeno deve in generale avere le stesse caratteristiche.
La scelta della funzione dipende anche dallo scopo per cui si fa la ricerca. Ad esempio, se lo scopo è puramente descrittivo del fenomeno, allora si cercherà una funzione semplice. Se invece lo scopo è investigativo, ossia se si vuole ricavare la legge, o un modello matematico del fenomeno, allora la funzione sarà più complessa.
Indichiamo con $\hat{y}_i$ i valori teorici sulla curva corrispondenti ai valori x_i rilevati. Sostituendo ai valori y_i rilevati i valori $\hat{y}_i$ teorici, si commettono errori dati dalla differenza:

$$d_i = y_i - \hat{y}_i$$

che possono essere positivi, negativi o nulli.
Occorre minimizzare questi errori, ma non è corretto minimizzare la somma delle d_i, in quanto gli errori positivi potrebbero compensare quelli negativi. Il criterio corretto per ottenere un buon accostamento è quello di minimizzare la somma dei quadrati delle d_i, precisamente:

la condizione di accostamento data dal **metodo dei minimi quadrati** è: determinare la funzione interpolante in modo che sia **minima** la somma dei quadrati delle differenze fra i valori osservati y_i ed i valori teorici $\hat{y}_i$, cioè

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 .$$

Ricavata la funzione che si ritiene più rappresentativa della distribuzione, bisogna verificare che i valori teorici approssimino i valori empirici, ossia che il grado di accostamento sia accettabile.
A questo scopo si calcolano, per prima cosa, le differenze

$$d_i = y_i - \hat{y}_i$$

che dovranno essere, il più possibile, di segni alternati; quindi si calcolano gli **indici di accostamento**.
Gli indici di accostamento più usati sono l'**indice lineare relativo I_1** e l'**indice quadratico relativo I_2** aventi le seguenti espressioni:

$$I_1 = \frac{\sum |y_i - \hat{y}_i|}{\sum \hat{y}_i}, \qquad I_2 = \frac{\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}}{\frac{\sum \hat{y}_i}{n}}.$$

Fra i due indici è preferibile il secondo poiché il metodo dei minimi quadrati opera sui quadrati delle differenze.
I valori ottenuti vanno considerati in relazione al fenomeno; comunque, in linea di massima, per avere un buon accostamento non devono superare il valore 0,1 (in certi casi non devono superare 0,01); ovviamente, tanto più piccoli sono i valori di I_1 e di I_2, tanto migliore è l'accostamento.
Se lo scopo della ricerca della funzione è quello di avere un modello matematico del fenomeno, attualmente è stato introdotto un indice detto **coefficiente di determinazione** che tiene conto dello scarto quadratico medio dei valori y_i .e
Indicata con

$$\bar{y} = \frac{\sum y_i}{n}$$

la media aritmetica dei valori y_i, il coefficiente di determinazione ha la seguente espressione:

$$\delta = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}.$$

Quanto più δ è "vicino" ad 1, tanto più il modello rappresenta bene il fenomeno.

Funzione interpolante: retta

Applichiamo il metodo dei minimi quadrati nel caso in cui come funzione interpolante venga scelta la retta.
La funzione scelta è una funzione di primo grado in x ed y, avente quindi un'espressione del tipo:

$$y = a + bx,$$

con a e b parametri reali.
Imponendo che sia minima la funzione

$$\varphi(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 ,$$

si ottengono le espressioni

$$\begin{cases} a = \dfrac{\sum y_i \cdot \sum x_i^2 - \sum x_i y_i \cdot \sum x_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \\ b = \dfrac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \end{cases}.$$

È importante notare che, indicate con

$$\bar{x} = \frac{\sum x_i}{n} \qquad \text{e} \qquad \bar{y} = \frac{\sum y_i}{n},$$

le medie aritmetiche rispettivamente di x_i e di y_i, vale la relazione

$$a = \bar{y} - b\bar{x}$$

e l'equazione della retta interpolante è suscettibile di essere scritta nella forma

$$y - \bar{y} = b\left(x - \bar{x}\right).$$

La retta interpolante passa per il punto di coordinate $\left(\bar{x}, \bar{y}\right)$, detto baricentro della distribuzione.

SCHEMA DI ESERCIZIO

N. valori	x	y	xy	x^2	$\hat{y}$	$d = y - \hat{y}$	$\lvert y - \hat{y}\rvert$	$\left(y - \hat{y}\right)^2$	$y - \bar{y}$	$\left(y - \bar{y}\right)^2$
1										
2										
…										
…										
…										
…										
…										
n										
Σ (TOT.)										