RELAZIONI STATISTICHE

Premessa

Molto spesso, nello studio di un fenomeno statistico, occorre individuare due variabili, questo con lo scopo di individuare le eventuali relazioni che possono esistere fra esse. Questo tipo di indagine può essere condotta utilizzando la tecnica dell'*interpolazione*.

Interpolazione

E' importante specificare sin d'ora che lo scopo dell'interpolazione può essere duplice:

1. In presenza di una distribuzione di dati che si presenta lacunosa, cioè mancante di qualche dato che non può più essere rilavato, questa tecnica permette di eliminare tale lacuna attribuendo un valore a ciascuno dei dati mancanti;
2. In presenza di una distribuzione di dati, alcuni dei quali si ritengano scaturiti da errori accidentali, l'interpolazione permette di eliminare l'influenza di tali errori correggendo opportunamente i dati interessati sostituendone ad essi degli altri.

In entrambi i casi il problema può essere risolto come segue:

- Si suppone che i dati considerati, rappresentati in un sistema di assi cartesiani, si dispongano secondo una ben determinata funzione, detta ***funzione interpolante***, che, di volta in volta, può essere una funzione di primo grado o di secondo grado, o anche di altro tipo;
- Si scrive, quindi, l'espressione analitica di tale funzione e in base a quest'ultima si calcolano i valori mancanti o quelli che devono essere rettificati.

In pratica, si tratta di vedere come occorre procedere per scrivere l'espressione analitica della funzione interpolante e a tale scopo si può porre una delle seguenti condizioni:

Prima condizione: dovendo inserire uno o più risultati fra alcuni dati noti, si vuole che la funzione interpolante *riproduca esattamente i dati noti*.

Seconda condizione: dovendo correggere alcuni dati che si ritengano influenzati da errori accidentali, si vuole sostituire la curva che rappresenta la distribuzione dei dati disponibili con un'altra curva, detta *curva interpolante*, che rappresenta la sequenza originaria in modo *approssimativo*, ma più regolare.

Se viene posta la prima condizione si parla di ***interpolazione per punti*** *(interpolazione matematica)*; invece, se viene posta la seconda condizione si parla di ***interpolazione fra punti (interpolazione statistica)***, che è il caso specifico del mio software).

Interpolazione per punti o interpolazione matematica

Con riferimento ad un certo fenomeno supponiamo di disporre dei seguenti dati:

X	Y
X_1	Y_1
X_2	Y_2

Il problema che si presenta è quello di dover inserire tra di essi altri dati: a tale scopo, *supponendo che il fenomeno abbia andamento lineare,* usiamo come funzione interpolante una retta.
Per risolvere il problema considerato si procede come segue:

1. Si considera l'equazione di una generica funzione lineare (cioè di primo grado):

   $$\mathbf{y = a+bx}$$ (che talvolta si può trovare scritta come y = mx+q) (1)

   che è rappresentata da una retta.
   A questo punto può essere necessario ricordare che *a* è l'*ordinata all'origine*, cioè l'ordinata del punto in cui la retta interseca l'asse delle ordinate; *b* è il *coefficiente angolare* che determina l'inclinazione della retta.

2. Quindi si considerano i punti $\mathbf{P_1(x_1;y_1)}$, $\mathbf{P_2(x_2;y_2)}$ e si determinano *a* e *b* in modo che la retta passi per questi punti.
   A tale scopo, sostituendo nella (1) il valore di x_1 posto di x e quello di y_1 al posto di y si ottiene l'equazione $\mathbf{a+bx_1 = y_1}$.
   Facendo altrettanto con x_2 e y_2 si ottiene una seconda equazione: $\mathbf{a+bx_2 = y_2}$.
   Entrambe le equazioni così ottenute presentano le incognite *a* e *b*. Per determinare il valore di queste ultime basta allora risolvere il sistema costituito dalle due equazioni, cioè il sistema:

   $$\begin{cases} \mathbf{a+bx_1 = y_1} \\ \mathbf{a+bx_2 = y_2} \end{cases}$$

3. Una volta trovati i valori di *a* e *b*, sostituendoli nella (1) si trova subito l'equazione della funzione interpolante tramite la quale si possono calcolare tutti i valori da inserire.

Di seguito si trova un esempio che serve a chiarire quanto appena esposto:
Abbiamo la seguente distribuzione di dati:

X	*Y*
3	*15*
7	*39*

Ci proponiamo di inserire i valori di y che corrispondano a x=4 x=5 e x=6.
Partendo dall'equazione y = a+bx sostituiamo 3 al posto di x e 15 al posto di y. Così facendo scriviamo la prima equazione 15 = a+3b. Quindi, sostituendo 7 al posto di x e 39 al posto di y scriviamo la seconda equazione 39 = a+7b.
A questo punto consideriamo il sistema

$$\begin{cases} a+3b=15 \\ a+7b=39 \end{cases}$$

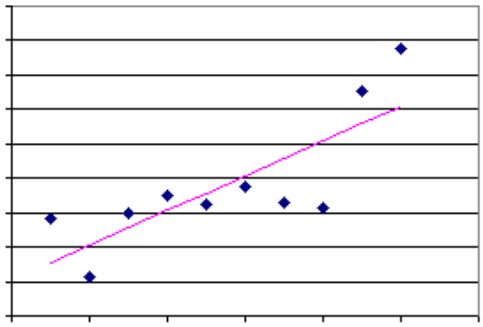
Per risolvere il sistema, sottraendo membro a membro si ottiene 4b=24, cioè b=6.
Sostituendo in una delle equazioni, per esempio nella prima, si ha: $3 \cdot 6 + a = 15$, cioè a = -3.
Ne segue che l'equazione della funzione interpolante è la seguente: y = -3+6x cioè y = 6x-3.
Possiamo ora calcolare i valori desiderati. Precisamente:

per x=4 si ha $y=6 \cdot 4-3=21$;
per x=5 si ha $y=6 \cdot 5-3=27$;
per x=6 si ha $y=6 \cdot 6-3=33$;

La sequenza dei dati di partenza può allora essere completata come segue:

X	*Y*
3	*15*
4	*21*
5	*27*
6	*33*
7	*39*

Osservazione: Graficamente, la sequenza di valori ottenuti può essere rappresentata mediante punti che giacciono sulla retta che congiunge P_1 e P_2.
Naturalmente ciò non può accadere, anzi di solito non accade, se i valori inseriti vengono ricavati da osservazioni concrete. In questo si può costruire un *diagramma a dispersione* come quello della figura che segue.



Sulla retta si troverebbero i punti corrispondenti ai valori teorici che si trovano con la funzione interpolante, mentre i valori osservati in concreto si discostano dalla retta.
Ovviamente, i valori ottenuti interpolando sono tanto più plausibili quanto più si ha modo di ritenere che i dati effettivi (non conosciuti) si dispongono linearmente.

Interpolazione fra punti o interpolazione statistica

Facendo riferimento ad un certo fenomeno supponiamo di disporre dei seguenti dati:

X	Y
x_1	y_1
x_2	y_2
x_3	y_3
x_4	y_4
…	…
x_n	y_n

Il problema che si presenta è quello di sostituire ai dati considerati, ritenuti influenzati da errori accidentali, altri dati approssimati, ma più regolari, *supponendo sempre che il fenomeno abbia andamento lineare.*
In questo caso (*interpolazione fra punti*), al contrario di quanto accade con l'interpolazione per punti, la risoluzione del problema si presenta piuttosto complicata, infatti nell'interpolazione per punti si dispone di due soli punti e, poiché per due punti passa una sola retta (la retta interpolante) di equazione $y = a + bx$, questa resta subito definita in modo univoco. Nel caso che ora stiamo esaminando, invece, si dispone di *n* punti, per cui equazioni di rette interpolanti se ne possono scrivere tante quante sono le possibili coppie di punti, quindi, si tratta di vedere quale fra tutte le possibili rette interpolanti può essere ritenuta più idonea per risolvere il problema considerato.

In pratica si assume come tale la retta di equazione

y = a + bx *dove* ***b=S/S*** ***(1)*** *e* ***$a = M_y - b \cdot M_x$*** ***(2)***

essendo precisamente ***"M_x" la media dei valori di x, "M_y" la media dei valori di y, "S" la somma dei prodotti degli scarti di x e di y dalle rispettive medie M_x ed M_y e "S*" la somma dei quadrati degli scarti di x dalla rispettiva media M_x .***

Tale retta è ritenuta la più idonea in quanto soddisfa la condizione seguente:
la somma dei quadrati degli scostamenti dei dati teorici (dati che si ottengono interpolando) dai rispettivi dati empirici (dati di cui si dispone) è minima rispetto alla somma dei quadrati degli scarti che si otterrebbero con qualsiasi altra retta interpolante. E' questo il cosiddetto metodo dei minimi quadrati.

Per quanto riguarda il calcolo di *a* e *b* occorre:

- ✓ calcolare anzitutto le medie M_x ed M_y
- ✓ calcolare quindi gli scarti di x da M_x e di y da M_y , cioè $x - M_x = x'$ e $y - M_y = y'$
- ✓ calcolare il prodotto degli scarti: $(x - M_x)(y - M_y) = x' \cdot y'$
- ✓ calcolare quindi la somma del prodotto degli scarti $x' \cdot y'$, cioè S
- ✓ calcolare i quadrati degli scarti di x da M_x , cioè $(x - M_x)^2 = x'^2$
- ✓ calcolare ancora la somma dei quadrati degli scarti x'^2, cioè S*

Soltanto a questo punto, applicando la (1) e la (2), si possono ottenere il valore di *a* e quello di *b* che permettono di scrivere l'equazione della retta interpolante $y = a + bx$.

Proprio questo carattere algoritmico dell'interpolazione statistica mi ha permesso di realizzare il software che si trova in questa sezione del mio sito.

Esempio di interpolazione statistica

Consideriamo i dati della tabella che segue: produzione di frumento (in milioni di q) per gli anni dal 1974 al 1982. Come si vede i valori di x sono gli anni, che per semplicità vengono contati ponendo il 1974 come anno zero (il 1975 come anno 1 e così via…), e i valori di y sono i valori della produzione. In tale tabella, oltre ai valori di x e di y, sono indicati i valori x', y', x'·y', x'^2, i valori interpolati (teorici), gli scarti fra valori teorici e valori empirici. Si ha:

M_x =36/9=4 *M_y = 825,60/9=91,73* *S=44,4* *S*=60*

$b = 44,40/60 = 0,74$ $a = 91,73 – 0,74 \cdot 4 = 88,77$

quindi la retta interpolante è $y = 0,74x + 88,7$.
Da questa retta si ottengono i valori interpolati dando ad x i valori 0, 1, 2, … , 8.

Anni	x	y	x'	y'	x'y'	x'^2	Valori teorici
1974	0	85,6	-4	-6,13	24,52	16	88,7
1975	1	91,8	-3	0,07	-0,21	9	89,44
1976	2	94,6	-2	2,87	-5,74	4	90,18
1977	3	88,5	-1	-3,23	3,23	1	90,92
1978	4	86,4	0	-5,33	0	0	91,66
1979	5	98,2	1	6,47	6,47	1	92,4
1980	6	90,5	2	-1,23	-2,46	4	93,14
1981	7	99,3	3	7,57	22,71	9	93,88
1982	8	90,7	4	-1,03	-4,12	16	94,26
	36	825,6			44,4	60	

Perequazione

Un problema connesso all'interpolazione è quello della *perequazione*. In sostanza, perequare un fenomeno statistico vuol dire eliminare tutte quelle variazioni di carattere accidentale che ne turbano la regolarità e non permettono di cogliere l'effettiva tendenza di esso.
Nella pratica l'eliminazione di tutte queste variazioni si ottiene si ottiene sostituendo ai dati empirici i dati teorici che si ottengono mediante l'interpolazione statistica.

Extrapolazione

Si parla di *extrapolazione* quando, noti certi dati relativi ad un fenomeno statistico, se ne vogliono ottenere altri, che, però, contrariamente a quanto accade nell'interpolazione, non sono interni all'intervallo di osservazione, bensì esterni. La cosa può essere interessante quando si vogliono fare previsioni sull'andamento futuro di un certo fenomeno. In questo caso, una volta costruita la retta interpolante, si tratta di estenderne la validità anche ai valori di x esterni all'intervallo di osservazione.
Naturalmente è chiaro che nel formulare previsioni occorre usare sempre molta prudenza: è assolutamente necessario poter supporre validamente che all'esterno del campo di osservazione il fenomeno considerato si sviluppi secondo la retta interpolante costruita.