Capitolo 8 Rappresentazione analitica di variabili. Interpolazione

Sommario **1.** Introduzione. - **2.** Interpolazione, estrapolazione e perequazione. - **3.** Scelta della funzione teorica. - **4.** Metodo dei minimi quadrati. - **5.** Verifica del grado di accostamento.

1. Introduzione

Alla base di un fenomeno statistico è possibile evincere l'esistenza di una legge che ne regola l'andamento; scopo della **rappresentazione analitica** di una variabile X è quello di specificare la forma funzionale della legge sottostante il fenomeno investigato.

Una rilevazione statistica consente, sovente, di evincere una corrispondenza empirica tra le modalità di un carattere quantitativo, indicato genericamente con X, e le rispettive frequenze o intensità, indicate genericamente con Y.
La corrispondenza tra X e Y rappresenta una **funzione statistica**, e precisamente siano $x_1, x_2, \ldots, x_n$ le modalità del carattere X e $y_1, y_2, \ldots, y_n$ le corrispondenti frequenze, una funzione statistica è definita dalle n coppie di valori:

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

Tale corrispondenza può evincersi anche tra le modalità di tempo di una serie storica e le intensità del fenomeno investigato.
Per evidenziare il tipo di legame tra le variabili è di notevole ausilio il **diagramma a dispersione** (o **scatter plot**), ossia il diagramma empirico costituito dalle n coppie di osservazioni $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ sulle variabili X e Y rappresentate da una nuvola di punti.

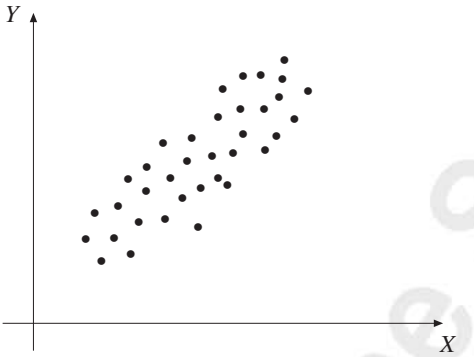

Fig. 1 - Diagramma a dispersione

Generalmente, una funzione statistica è rappresentata graficamente da una spezzata, in cui si assumono come **variabili indipendenti** le modalità del carattere e come **variabili dipendenti** le corrispondenti frequenze.

2. Interpolazione, estrapolazione e perequazione

L'interpolazione, l'estrapolazione e la perequazione, sono procedure di correzione dei dati rilevati per ovviare a tre frequenti inconvenienti:

a) i dati rilevati possono presentare lacune;
b) i dati disponibili possono provenire da rilevazioni statistiche distanti nel tempo;
c) alcuni dati rilevati possono essere affetti da errori.

A) Interpolazione

Non sempre, a seguito di una rilevazione statistica, si deducono, in corrispondenza delle modalità di un carattere, le rispettive frequenze o intensità.
L'**interpolazione** è il processo di determinazione di una successione di valori (tutti o in parte teorici) di frequenze o intensità, ottenuti in corrispondenza di valori osservati di modalità di un carattere quantitativo in una distribuzione di frequenza, o modalità di tempo in una serie storica.
Il procedimento si attua sia *analiticamente* sia *graficamente*.

La **rappresentazione analitica** consiste nel trovare una funzione matematica che rappresenti nel miglior modo possibile la distribuzione osservata del fenomeno.
La **rappresentazione grafica** consiste nel sostituire al diagramma rappresentativo dei dati osservati una curva teorica associata ad una funzione matematica.

Si parla di **interpolazione per punti** o **matematica** se le variabili X e Y non sono affette da errori, e la distribuzione è costituita da n coppie di valori (x_i, y_i) cui corrispon-

dono *n* punti-immagine nel piano cartesiano. La rappresentazione analitica si attua mediante una funzione matematica cui corrisponde una curva che passi **per** tutti i punti disponibili (**Fig. 2 (a)**), i quali si devono trovare sulla curva teorica e le loro coordinate devono soddisfarne l'equazione.
Fondamentale, appare quindi, ricercare una funzione matematica che determini esattamente l'insieme di punti (x, y) effettivamente osservati.
Si parla di **interpolazione fra punti** o **statistica** se una delle due variabili *X* e *Y* o entrambe sono affette da errori; in tal caso l'interpolazione consiste nel determinare valori teorici delle variabili non affetti da errori. La rappresentazione analitica si attua determinando una funzione matematica associata ad una curva che passi **tra** i punti-immagine del diagramma a dispersione della distribuzione osservata (**Fig. 2 (b)**).
La funzione di interpolazione ha in questo caso, una finalità sostitutiva dei dati che si ritengono affetti da errori, con dati che si ritengono più attendibili attraverso l'elaborazione di una funzione il cui grafico si approssimi il più possibile al diagramma a dispersione originario; alla distribuzione dei dati di partenza, viene quindi sostituita una nuova distribuzione approssimata ma non affetta da errori.

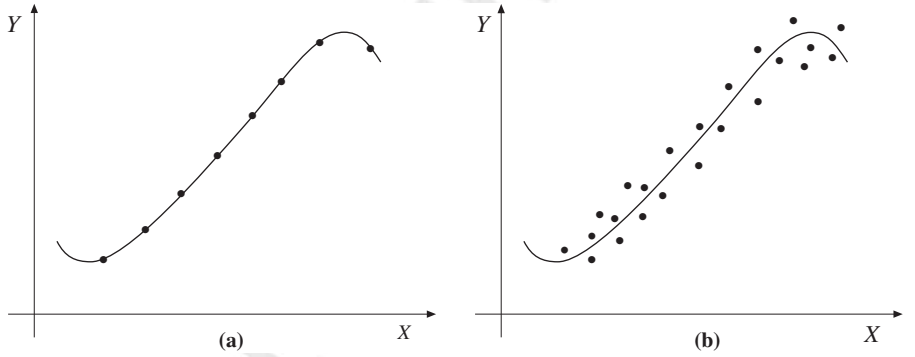

Fig. 2 - Interpolazione per punti (a); interpolazione fra punti (b)

Rappresentazione analitica in un procedimento di interpolazione

Per realizzare una corretta rappresentazione analitica lo statistico deve:

1. specificare la funzione matematica che riproduca meglio la funzione statistica individuata, in altri termini deve mutuare dalla matematica una **funzione teorica** in grado di rappresentare con una legge matematica la distribuzione empirica;
2. determinare numericamente i **parametri** che compaiono nella funzione matematica;
3. verificare il **grado di accostamento** tra valori empirici (o osservati) delle frequenze o intensità e valori teorici ottenuti per mezzo della funzione matematica. È ovvio che questo momento riguarda solo l'interpolazione fra punti.

Nei successivi paragrafi ci occuperemo di questi tre momenti della rappresentazione analitica servendoci anche di esempi, prima, però, diamo alcuni cenni sui concetti di estrapolazione e di perequazione.

B) Estrapolazione

Talvolta si rende necessario disporre di frequenze o intensità esterne al campo di osservazione del carattere investigato.
L'**estrapolazione** è il processo di determinazione di una successione di valori teorici di frequenze o intensità, ottenuti in corrispondenza di modalità di un carattere quantitativo in una distribuzione di frequenza, o modalità di tempo in una serie storica, esterne all'intervallo di osservazione.
Il procedimento si attua sia *analiticamente* sia *graficamente*.
Esso presuppone che il fenomeno si sia svolto in passato con una certa regolarità. In quanto basato solo sulla regolarità, in passato, del fenomeno che rappresenta, l'estrapolazione è poco attendibile non tenendo conto di cause perturbatrici che potrebbero verificarsi in futuro.

C) Perequazione

Alcuni dati possono essere affetti da errori accidentali a danno o a vantaggio di quelli che immediatamente li seguono o li precedono nella successione. La tecnica statistica che consente di eliminare tali errori dai dati e, in particolare, irregolarità di andamento nelle serie storiche, è definita **perequazione**.
Il metodo di perequazione più semplice è la **media mobile** che consiste nel sostituire a ciascun termine il valore medio aritmetico di un gruppo di tre, cinque, sette, ... termini, di cui il termine dubbio è quello centrale.
La funzione che consente di adattare ai dati osservati dati teorici è detta **funzione perequatrice**.
La tecnica della perequazione, rappresenta infatti, una metodologia di livellazione dei dati di una serie non regolare; i dati eccessivamente irregolari e quindi presumibilmente affetti da errori di varia natura, sono sostituiti da dati teorici ricavati attraverso un'apposita funzione matematica.

3. Scelta della funzione teorica

La prima fase di una rappresentazione analitica consiste nella scelta di una funzione matematica che rappresenti le *n* coppie di punti:

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

A) Interpolazione per punti

Nell'interpolazione per punti, la scelta della funzione rappresentatrice dei dati osservati verte su una funzione che rappresenti l'esatta espressione della funzione incogni-

ta che lega le modalità x_i , $i = 1, 2, \ldots, n$, alle rispettive frequenze y_i; infatti, essa deve passare **per** tutte le n coppie di punti - immagine. A tal fine si opta per una funzione che contenga tanti parametri quante sono le coppie di punti. Generalmente si fa riferimento ad un **polinomio completo di ordine $n - 1$**, del tipo:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_{n-1} X^{n-1}$$

i cui parametri o coefficienti $\beta_0, \beta_1, \ldots, \beta_{n-1}$ sono n.

Per il *principio di identità dei polinomi*, fissati gli n punti in cui è nota la funzione, il polinomio di grado $n - 1$ scelto è *unico*.

Pertanto, dati due punti nel piano, l'equazione prescelta è **lineare** nei due parametri da stimare ed è del tipo:

$$Y = \beta_0 + \beta_1 X$$

e la sua rappresentazione è riportata in **figura 3(a)**.

Se, invece, sono dati tre punti non allineati nel piano, l'equazione di **secondo grado** prescelta ha tre parametri da stimare ed è del tipo:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$

e la sua rappresentazione è riportata in **figura 3(b)**.

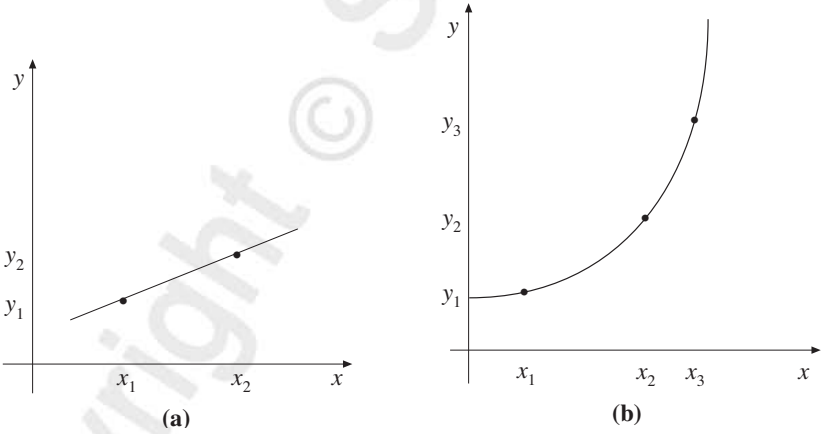

Fig. 3 - Retta (a); parabola (b)

L'interpolazione per punti viene utilizzata anche per estrapolare dati da quelli esistenti, ossia per estendere l'intervallo di validità della funzione al di là dei valori osservati sulla base dell'ipotesi che il fenomeno abbia ad es. lo stesso andamento riscontrato nel passato.

B) Interpolazione fra punti

L'interpolazione fra punti ha lo scopo di individuare una funzione matematica il cui grafico passi **fra** le *n* coppie di punti - immagine.

Pertanto, non essendovi esatta riproduzione dei dati osservati, si opta per un **polinomio di grado *s* < *n***, del tipo:

$$Y^* = \beta_0 + \beta_1 X + \beta_2 X^2 + \text{É} + \beta_s X^s$$

in cui alla *Y* si è apposto un asterisco a indicare che il valore della frequenza o intensità che si ottiene sostituendo alla *X* la modalità osservata del carattere (o della serie storica) è **teorico**, giacché ottenuto approssimando la legge statistica con una legge matematica di tipo noto. Lo statistico, nella fattispecie, sarà aiutato, nella scelta della funzione da adattare ai dati, dal diagramma empirico, ed una volta individuato la funzione teorica che appare più adatta, dovrà fissare una condizione di accostamento (par.5) e verificarla analiticamente allo scopo di selezionare la funzione teorica migliore tra tutte quelle possibili.

4. Metodo dei minimi quadrati

Esistono diversi metodi per determinare i parametri di una funzione matematica in un procedimento di **interpolazione**; in questo paragrafo esamineremo il **metodo dei minimi quadrati *OLS*** (dall'inglese *Ordinary Least Squares*) da Karl Friederich Gauss nel 1795 e da Adrien Marie Legendre nel 1805.

Il metodo dei minimi quadrati, si fonda sull'ipotesi che i valori della variabile dipendente siano affetti da errore, mentre i valori assunti dalla variabile indipendente siano corretti. Da un punto di vista analitico, esso consiste nel determinare i valori della curva teorica che rendono minima la somma dei quadrati degli scarti tra valori teorici e valori osservati, poiché si ipotizza che i valori assunti dalla variabile Y siano affetti da errori accidentali che tendono a compensarsi all'aumentare del numero dei casi osservati; tale metodo è utilizzato sia per funzioni lineari che paraboliche. La scelta del tipo di funzione non può, ovviamente, essere frutto di una valutazione aprioristica; sarà, infatti, necessario osservare la nuvola di punti (**diagramma a dispersione**) per valutare l'andamento del fenomeno. Si formuleranno delle ipotesi di lavoro esaminando la relazione esistente fra le variabili X ed Y; infatti, se la relazione fra le due variabili è di proporzionalità diretta, oppure se gli incrementi dei valori della variabile Y sono, per incrementi dei valori di X, costanti, si opterà per una funzione lineare (**retta**). Se, invece, la relazione che lega le variabili X ed Y è di proporzionalità inversa (al crescere di X diminuisce Y o viceversa), si opterà per una funzione parabolica (**iperbole**). Infine, se la relazione tra le variabili X ed Y è di tipo esponenziale, e quindi ad es. i valori di X crescono secondo una

legge di proporzionalità aritmetica, mentre i valori di Y secondo una legge di proporzionalità geometrica, si opterà per una **curva esponenziale**. Tali funzioni esponenziali sono ad es. molto utilizzate per spiegare i meccanismi di riproduzione dei microrganismi in biologia, oppure in economia per illustrare le prospettive di crescita esponenziale del capitale negli investimenti.

Se la funzione interpolatrice è:

$$Y^* = f(X;\beta_0,\beta_1,\beta_2,\dots)$$

il metodo dei minimi quadrati è tale per cui:

$$G(\beta_0,\beta_1,\beta_2,\dots) = \sum_{i=1}^{n}\left(y_i^* - y_i\right)^2 = \sum_{i=1}^{n}\left[f(x_i;\beta_0,\beta_1,\beta_2,\dots) - y_i\right]^2 = \min$$

dove $y_i^*, i = 1,2,\dots,n$, sono i valori teorici mentre $y_i, i = 1,2,\dots,n$, sono i valori osservati della variabile *Y*.

Funzione lineare

Date due variabili *X* e *Y* se la funzione teorica è lineare nei parametri, cioè è del tipo:

$$Y^* = \beta_0 + \beta_1 X \qquad \textbf{(4.1)}$$

i parametri da determinare sono β_0 e β_1. Il metodo dei minimi quadrati è tale per cui:

$$G(\beta_0,\beta_1) = \sum_{i=1}^{n}(y_i^* - y_i)^2 = \min \qquad \textbf{(4.2)}$$

ossia:

$$G(\beta_0,\beta_1) = \sum_{i=1}^{n}(\beta_0 + \beta_1 x_i - y_i)^2 = \min \qquad \textbf{(4.3)}$$

da cui, derivando rispetto ai due parametri e eguagliando a zero si ottengono le cosiddette **equazioni normali**:

$$\frac{\partial G}{\partial \beta_0} = 2\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i - y_i)\cdot 1 = 0$$

$$\frac{\partial G}{\partial \beta_1} = 2\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i - y_i)x_i = 0$$

per determinare i valori dei due parametri si ottiene il **sistema normale**:

$$\begin{cases} n\beta_0 + \beta_1\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \beta_0\sum_{i=1}^{n} x_i + \beta_1\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \qquad \textbf{(4.4)}$$

da cui:

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad \textbf{(4.5)}$$

$$\beta_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad \textbf{(4.6)}$$

La retta dei minimi quadrati (4.1), che passa per il baricentro della distribuzione avente coordinate (μ_x, μ_y), può essere scritta nel modo seguente:

$$Y^* - \mu_y = \beta_1 (X - \mu_x) \qquad \textbf{(4.7)}$$

Il coefficiente β_1, in termini di scarti della media, è il seguente:

$$\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)}{\sum_{i=1}^{n} (x_i - \mu_x)^2}$$

per cui la (4.7) diviene:

$$Y^* - \mu_y = \left[\frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)}{\sum_{i=1}^{n} (x_i - \mu_x)^2} \right] (X - \mu_x) \qquad \textbf{(4.8)}$$

Funzione del secondo ordine

Per interpolare una funzione intera del secondo ordine:

$$Y^* = \beta_0 + \beta_1 X + \beta_2 X^2 \qquad \textbf{(4.9)}$$

con il metodo dei minimi quadrati, il sistema normale è il seguente:

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i + \beta_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 + \beta_2 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i \\ \beta_0 \sum_{i=1}^{n} x_i^2 + \beta_1 \sum_{i=1}^{n} x_i^3 + \beta_2 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i \end{cases} \qquad \textbf{(4.10)}$$

Trasformazione di variabili

Per illustrare gli esercizi sui minimi quadrati, quando le x_i rappresentano gli anni di una serie storica (1999, 2000, 2001, …), per evitare di lavorare su cifre enormi, utilizziamo una convenzione: a partire dal primo anno della serie, indichiamo gli anni, rispettivamente, con i numeri da 0, 1, 2, …, *n*. Utilizziamo cioè una trasformazione di variabili attraverso delle ascisse di comodo x_i^t tali che:

$$x_i^t = \text{anno}(i) - \text{primo anno della serie}$$

Ovviamente, quando andremo a utilizzare la funzione ottenuta dovremo continuare a usare la convenzione, ossia sostituiremo agli anni tali valori.

Esempio

La tabella seguente riporta la distribuzione dei libri di una data collana venduti da una casa editrice negli anni dal 2003 *al* 2008*:*

Anni	Libri venduti
2003	800
2004	980
2005	1.040
2006	1.200
2007	1.240
2008	1.550
Totale	6.810

Tabella 1

a) *determinare l'equazione della retta dei minimi quadrati;*
b) *tracciare il diagramma della distribuzione e la retta dei minimi quadrati.*

a) I calcoli per giungere alla determinazione della retta dei minimi quadrati sono contenuti nel seguente schema in cui si è effettuata una trasformazione di variabili del tipo:

$$x_i^t = \text{anno}(i) - 2003$$

dove il 2003 è l'anno $i = 0$, e in cui le variabili x_i^t solo le ascisse di comodo.

x_i	y_i	x_i^t	$(x_i^t)^2$	$x_i^t y_i$
2003	800	0	0	0
2004	980	1	1	980
2005	1.040	2	4	2.080

2006	1.200	3	9	3.600
2007	1.240	4	16	4.960
2008	1.550	5	25	7.750
Totale	6.810	15	55	19.370

Schema 1

da cui si deduce il sistema normale:

$$\begin{cases} 6\beta_0 + 15\beta_1 = 6.810 \\ 15\beta_0 + 55\beta_1 = 19.370 \end{cases}$$

Pertanto:

$$\beta_0 = \frac{374.550 - 290.550}{330 - 225} = \frac{84.000}{105} = 800$$

$$\beta_1 = \frac{116.220 - 102.150}{330 - 225} = \frac{14.070}{105} = 134$$

da cui, la retta dei minimi quadrati in termini di ascisse di comodo ha equazione:

$$Y^* = 800 + 134X^t$$

b) In un sistema di riferimento cartesiano ortogonale, si pongono sull'asse delle ascisse gli anni e sull'asse delle ordinate le frequenze osservate in corrispondenza dei diversi anni. Si tracciano, quindi, i punti-immagine empirici e la retta dei minimi quadrati.

Il grafico è il seguente:

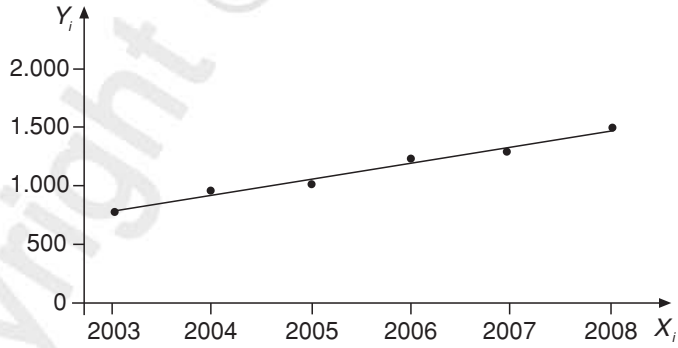

Fig. 4 - L'interpolazione è lineare

Dallo stesso si evince che l'accostamento è valido.

5. Verifica del grado di accostamento

In un procedimento di interpolazione **fra** punti, una volta scelta la funzione da adattare alla distribuzione empirica e i relativi parametri, è compito dello statistico verificare il **grado di accostamento tra funzione teorica e funzione statistica**; in altri termini, si rende inevitabile misurare la **dispersione** dei dati osservati intorno alla curva rappresentatrice della funzione prescelta.
Diversi indici sono stati elaborati per misurare il grado di approssimazione del procedimento di interpolazione, tutti si basano sugli n scarti tra valori teorici (y_i^*)e valori osservati (y_i).
In genere, in Statistica si ritiene accettabile il grado di accostamento fra i valori osservati ed i valori teorici ricavati attraverso interpolazione, se gli indici di accostamento assumono valore inferiore a 0,1; qualora il valore fosse superiore al valore limite prefissato, sarà opportuno optare per un differente modello interpolante.

A) Indici assoluti

✔ Un primo indice è dato dalla **media aritmetica dei valori assoluti degli scarti**:

$$\alpha_1 = \frac{\sum_{i=1}^{n} |y_i^* - y_i|}{n} \qquad \textbf{(5.1)}$$

È evidente la necessità di considerare i valori assoluti degli scarti nella sommatoria al numeratore della frazione, in quanto, in caso contrario, essa sarebbe nulla se il metodo di stima dei parametri applicato è quello dei minimi quadrati.

✔ Un secondo indice è fornito dalla **media quadratica degli scarti**:

$$\alpha_2 = \sqrt{\frac{\sum_{i=1}^{n} (y_i^* - y_i)^2}{n}} \qquad \textbf{(5.2)}$$

B) Indici relativi

I seguenti indici sono ottenuti rapportando gli indici assoluti alla media aritmetica dei valori osservati, e per questo sono numeri puri.

✔ Un primo indice è, dunque:

$$\rho_1 = \frac{\sum_{i=1}^{n} |y_i^* - y_i| \Big/ n}{\sum_{i=1}^{n} y_i \Big/ n} = \frac{\sum_{i=1}^{n} |y_i^* - y_i|}{\sum_{i=1}^{n} y_i} \qquad \textbf{(5.3)}$$

✔ Un secondo indice è:

$$\rho_2 = \frac{\sqrt{\dfrac{\sum_{i=1}^{n}\left(y_i^* - y_i\right)^2}{n}}}{\sum_{i=1}^{n} y_i \Big/ n} = \frac{\sqrt{n\sum_{i=1}^{n}\left(y_i^* - y_i\right)^2}}{\sum_{i=1}^{n} y_i} \qquad \textbf{(5.4)}$$

C) Indici normalizzati

— Un primo indice si ottiene dall'espressione (5.3):

$$r_1 = \frac{100}{100 + \rho_1} \qquad \textbf{(5.5)}$$

— Un secondo indice si ottiene, invece, dall'espressione (5.4):

$$r_2 = \frac{100}{100 + \rho_2} \qquad \textbf{(5.6)}$$

Entrambi gli indici assumono:

✔ **valore minimo** 0 quando ρ_1, o ρ_2, sono infinitamente grandi, per cui la funzione teorica non è in grado assolutamente di rappresentare la distribuzione empirica;

✔ **valore massimo** 1 quando ρ_1, o ρ_2, sono prossimi allo zero, per cui la funzione teorica rappresenta in maniera efficace la distribuzione empirica.

Esempio

Verificare la bontà dell'accostamento della retta dei minimi quadrati ottenuta inerpolando la distribuzione riportata nella tabella 1*, utilizzando l'espressione* (5.5).

Per verificare la bontà dell'accostamento è necessario calcolare i valori teorici y_i^*. Essi si desumono dal seguente schema:

x_i	y_i	y_i^*	$y_i - y_i^*$	$\lvert y_i - y_i^* \rvert$
0	800	800	0	0
1	980	934	46	46
2	1.040	1.068	–28	28
3	1.200	1.202	–2	2
4	1.240	1.336	–96	96
5	1.550	1.470	80	80
Totale	6.810	6.810	0	252

Schema 2

L'indice di accostamento relativo espresso dalla (5.3) è:

$$\rho_1 = \frac{\sum_{i=1}^{n} |y_i - y_i^*|}{\sum_{i=1}^{n} y_i} = \frac{252}{6.810} = 0,037$$

Pertanto, il valore dell'indice r_1 espresso dalla (5.5) è:

$$r_1 = \frac{100}{100 + 0,037} = 0,9996$$

Il risultato evidenzia un opportuno accostamento.

Questionario

1. Che differenza c'è tra **interpolazione matematica** e **interpolazione statistica**? (**par. 2**)
2. Illustrare il **metodo interpolativo dei minimi quadrati** per una distribuzione di dati tratta dalla realtà italiana. (**par. 4**)
3. Derivare le **equazioni normali** del **metodo dei minimi quadrati** per la funzione:

   $$Y^* = \beta_0 + \beta_1 X + \beta_2 X^2$$

   (**par. 4**)
4. Come si ottengono, a partire dagli **indici assoluti**, gli **indici normalizzati** in grado di misurare il grado di approssimazione del procedimento di interpolazione? (**par. 5**)