LEZIONI DI STATISTICA E

CALCOLO DELLE PROBABILITA'

UMBERTO MAGAGNOLI

Materiale per il Corso di lezioni di

"STATISTICA"

Laurea magistrale in "Matematica"
Facoltà di Scienze Matematiche, Fisiche e Naturali
Università di Ferrara
Anno accademico 2010-11

PARTE PRIMA
"Statistica Descrittiva"

0. Premessa

L'insegnamento di un corso di "Statistica" rivolto agli allievi di una Laurea magistrale in Matematica richiede specifiche attenzioni, che possono essere differenti, per alcuni aspetti peculiari, rispetto a quanto avviene per insegnamenti dedicati alla medesima disciplina ma in ambiti economici, finanziari o politico-sociali.

Infatti, la preparazione acquisita dagli studenti si avvicina di più, per gli aspetti metodologici, a quella di coloro che hanno seguito studi statistici o ingegneristici.

Inoltre, trattandosi di un unico insegnamento dedicato ai fondamenti della disciplina statistica si è ritenuto opportuno far precedere alla parte rivolta all'indagine dei fenomeni sperimentali con rilievo matematico probabilistico le linee di base dell'impiego razionale della statistica.

Tale argomento è proprio dell'ambito descrittivo ed è utilizzato in via preliminare in tutte le tipologie applicative; è richiesto anche come insegnamento negli istituti medi superiori di secondo grado, secondo le indicazioni ministeriali.

Il materiale predisposto, da cui sono tratti gli elementi illustrati nel corso delle lezioni, costituisce la prima parte dell'intero corso e, al momento, richiede ancora un controllo e un'accurata revisione, che s'intende fare anche sulla base dell'esposizione e dei suggerimenti che verranno dal confronto in aula.

Pertanto gli elementi qui proposti, non intendono essere sostitutivi della diretta partecipazione alle lezioni, che certamente costituisce la naturale modalità dell'apprendimento ed è vivamente consigliata, né può considerarsi sostitutiva della lettura dei manuali o di testi presenti in letteratura di cui si fornisce anche un succinto riferimento nella bibliografia. Tali letture, inoltre, possono consentire di

integrare i concetti e approfondire esemplificazioni e applicazioni, favorendo anche l'interazione con il docente.

L'intento è quello di facilitare lo studente nel seguire le lezioni e perciò questi appunti hanno una finalità didattica.

Il materiale qui proposto consiste in un'introduzione, relativa al significato della disciplina "Statistica", con particolare sottolineatura del ruolo sia metodologico sia operativo che essa svolge nel campo della ricerca sperimentale e osservazionale, in presenza di fenomeni aleatori, come ausilio per la presa di decisioni in condizioni d'incertezza.

La parte successiva s'incentra sulla descrizione dell'analisi univariata di grandezze quantitative ed è dedicata ai problemi della loro rappresentazione sintetica, in termini di distribuzione di frequenza e di indici di locazione e di variabilità.

La parte conclusiva è dedicata ad alcuni cenni riguardanti lo studio descrittivo dei fenomeni quantitativi bivariati e multivariati, con riferimento ai problemi di regressione di tipo polinomiale e multilineare.

U.M.

Febbraio 2011.

1. Ricerca di una definizione della disciplina Statistica

Il termine "Statistica" nel linguaggio comune è inteso, e confuso, con le "statistiche", cioè dati, tabelle, grafici, medie, indici, ecc., piuttosto che essere riferito a una disciplina scientifica.

E' utile cercare una definizione che abbia un carattere più vicino al concreto utilizzo dei metodi statistici e a un'interpretazione metodologica.

In primo luogo si ha una "concezione ordinaria" della Statistica, che riguarda l'impiego delle metodologie statistiche e concerne il trattamento e l'esposizione razionalmente ordinata dei dati relativi a un fenomeno e la loro analisi quali i seguenti.

- Raccolta di masse di "dati"
- Presentazione dei dati mediante: tabelle e grafici
- Calcolo di grandezze "globali":
    - medie,
    - indici di dispersione,
    - indici di correlazione,
    - funzioni di regressione, ecc.

A un ulteriore livello si pone la concezione scientifica della Statistica come disciplina avente un metodo proprio e che è in grado di proporre leggi e procedure operative, con un continuo sviluppo innovativo.

Sarà prevalente, in questa esposizione, il punto di vista della metodologia scientifica della Statistica, come disciplina che indaga le modalità di conduzione delle rilevazioni e la pianificazione della raccolta dei dati mediante il campionamento e la conduzione di relativi piani sperimentali, indicandone anche la validità e l'ottimalità.

La Statistica costituisce come una "interfaccia" per ogni ricerca applicata, indipendentemente dal settore scientifico, fisico-naturalistico o socio-economico, in cui si svolge.

Il ruolo di maggiore importanza metodologica della Statistica è dato dalla sua "concezione scientifica", alla quale verrà dedicato principalmente il contenuto delle presenti lezioni, che implicherà una formalizzazione matematica e logica dei problemi affrontati.

Alla concezione scientifica fanno riferimento i metodi e le teorie relative.

- Costruzioni di "modelli"
- Indagini campionarie
- Programmazione degli esperimenti
- Inferenza sulle leggi di distribuzione
- Stime parametriche e non parametriche
- Verifica d'ipotesi e decisioni, ecc.

Si può pertanto pervenire a una definizione sintetica, quale quella indicata:

"*STATISTICA: teoria e metodo per la raccolta, l'interpretazione dei dati e la scelta decisionale*"

A completamento di quanto fin qui esposto, si può aggiungere che la Statistica fornisce strumenti per la presa di "decisioni" in condizioni d'incertezza.

Qualora l'indagine comporti la raccolta di una numerosa massa di informazioni sul fenomeno allo studio, così da potersi ritenere che si disponga di tutto quanto è necessario per prendere decisioni, si può limitare l'impiego agli strumenti proposti dalla "concezione ordinaria" della disciplina che vengono ad assumere la denominazione di "Statistica Descrittiva". Quando ci si avvale di "rilevazioni parziali", spesso di numerosità limitata, è necessario ricorrere al metodo induttivo in cui: dal particolare si traggono conoscenze generalizzabili, al fine di ricavare conoscenze riguardanti l'interezza del fenomeno ed esprimere informazioni sulle possibili manifestazioni future. Questo modo di procedere si denomina "Statistica inferenziale" e a essa è

associato il concetto di "rischio di decisione errata", data l'incompletezza delle informazioni.

Il carattere scientifico della disciplina Statistica sta appunto nella consapevolezza del rischio insito in ogni decisione che richiede una "misura del grado d'incertezza" di ogni evento o decisione presa. A tale scopo ci si avvale del concetto di "probabilità", a cui è affidato il compito di misurare attraverso un numero compreso tra 0 e 1 il rischio di errori decisionali e, quindi, del verificarsi dell'evento corrispondente.

La limitatezza delle osservazioni, presenti in ogni indagine, è un motivo dell'incertezza dovuta alla casualità dei singoli risultati. Inoltre, data la complessità dei fenomeni, si evidenzia anche una causa di incertezza dovuta all'ignoranza del "modello" ipotizzato rispetto allo "stato del sistema" con cui si configura la realtà.

Si comprende, quindi, la necessità di ricorrere a un modello, che pur differendosi dal fenomeno, consente una sua rappresentazione nelle due componenti fondamentali: "strutturale" e "aleatoria".

La "componente strutturale" mette in luce i legami, le leggi o le regolarità che legano le diverse grandezze, avvalendosi di relazioni matematiche, che esprimono le relazione di causa-effetto, mentre, mediante la "componente aleatoria", viene espressa la diversità tra le osservazioni, pur svolte in condizioni di costanza ambientale, dovuta sia dell'incertezza della misurazione sia alla presenza di altri fattori detti "latenti".

Il modello, nella sua formulazione matematica, risponde alle esigenze di conoscenza razionale della realtà fenomenica, ne favorisce la comprensione e consente di individuare le scelte operative più congrue; inteso poi come ricerca di un'interpretazione della realtà, trova impiego in tutte le scienze applicate dove ha un ruolo l'osservazione.

La presenza della "variabilità" costituisce l'elemento aggiuntivo dei modelli statistici rispetto a quelli deterministici. La "variabilità accidentale" si verifica nei fenomeni ripetitivi in cui il risultato è diverso, pur in condizioni di stabilità dei fattori essenziali del fenomeno in oggetto.

L'importanza del modello interpretativo è evidenziata dalla possibilità di messa in discussione dei risultati, dalla valutazione dell'attendibilità, dalla ricerca della natura e dell'entità degli errori, consentendo di confutare il modello stesso, ciò permette di incentivare ulteriori ricerche.

- *Capacità interpretativa della realtà*
- *Valutazione dell'attendibilità dei risultati*
- *Natura e misura degli errori*
- *Ricerca di procedure ottimali*

Ogni ricerca richiede una sempre maggiore analiticità sia per l'osservazione dei dati sia per la predisposizione di una sperimentazione opportuna e per la costruzione di un modello.

Queste esigenze si trovano in contrasto con altri aspetti di molte ricerche, riferendosi principalmente all'onerosità dei costi, alle difficoltà di acquisizione dei dati (si pensi alla privacy), alla complessità dell'individuazione del modello e ai tempi di raccolta delle informazioni che possono non essere compatibili con la stabilità del fenomeno, che è spesso in continua trasformazione.

Tutto questo comporta l'accettazione di un certo grado d'incertezza delle decisioni, dovuto alla variabilità accidentale evidenziando ancora il ruolo della probabilità nell'indagine statistica.

Nella ricerca scientifica, pertanto, si deve ricorrere a una sorta di "compromesso" tra la "attendibilità" nell'indagine su quanto vi è di strutturale nel fenomeno e la presenza di un'accidentalità e il "costo"

che quest'indagine richiede. L'equilibrio che viene raggiunto corrisponde a quanto espresso sinteticamente col "Principio della parsimonia scientifica", che implica l'accettazione di un certo grado d'incertezza e la scelta di modelli il più possibile semplici per quanto riguarda la formalizzazione e il numero dei parametri.

E' possibile sintetizzare quanto è stato detto nella'affermazione:

"La STATISTICA permette di scoprire quanto di strutturale è presente nel fenomeno ripetitivo allo studio, accettando la presenza di variazioni inspiegabili, corrispondenti alla accidentale variabilità"

Il riferimento a fenomeni ripetitivi è relativo alla modalità di presentazione con risultanze differenti e di volta in volta imprevedibili, pur in condizioni di costanza di aspetti ritenuti essenziali.

Come disciplina scientifica la Statistica presenta come scopo quello di intervenire sulle analisi sperimentali al fine di "meglio" ottenere i risultati e/o "meglio" interpretarli. In questo intervento si presenta con le seguenti caratteristiche.

- *Autonomia* con il contenuto di altre discipline
- Si avvale di propri principi *Logico Matematici*

La definizione a cui si farà ricorso per la disciplina argomento di questo Corso di lezioni può essere espressa nel modo seguente.

"STATISTICA: settore delle Scienze Matematiche che è di ausilio alle discipline che ricorrono all'indagine sperimentale"

La conduzione di una ricerca quantitativa, che coinvolge l'impiego della disciplina statistica può essere schematizzata in 5 passi, posti in un percorso ciclico, in cui si evidenziano i momenti di "confutazione" e di "conferma" della teoria e del modello proposto.

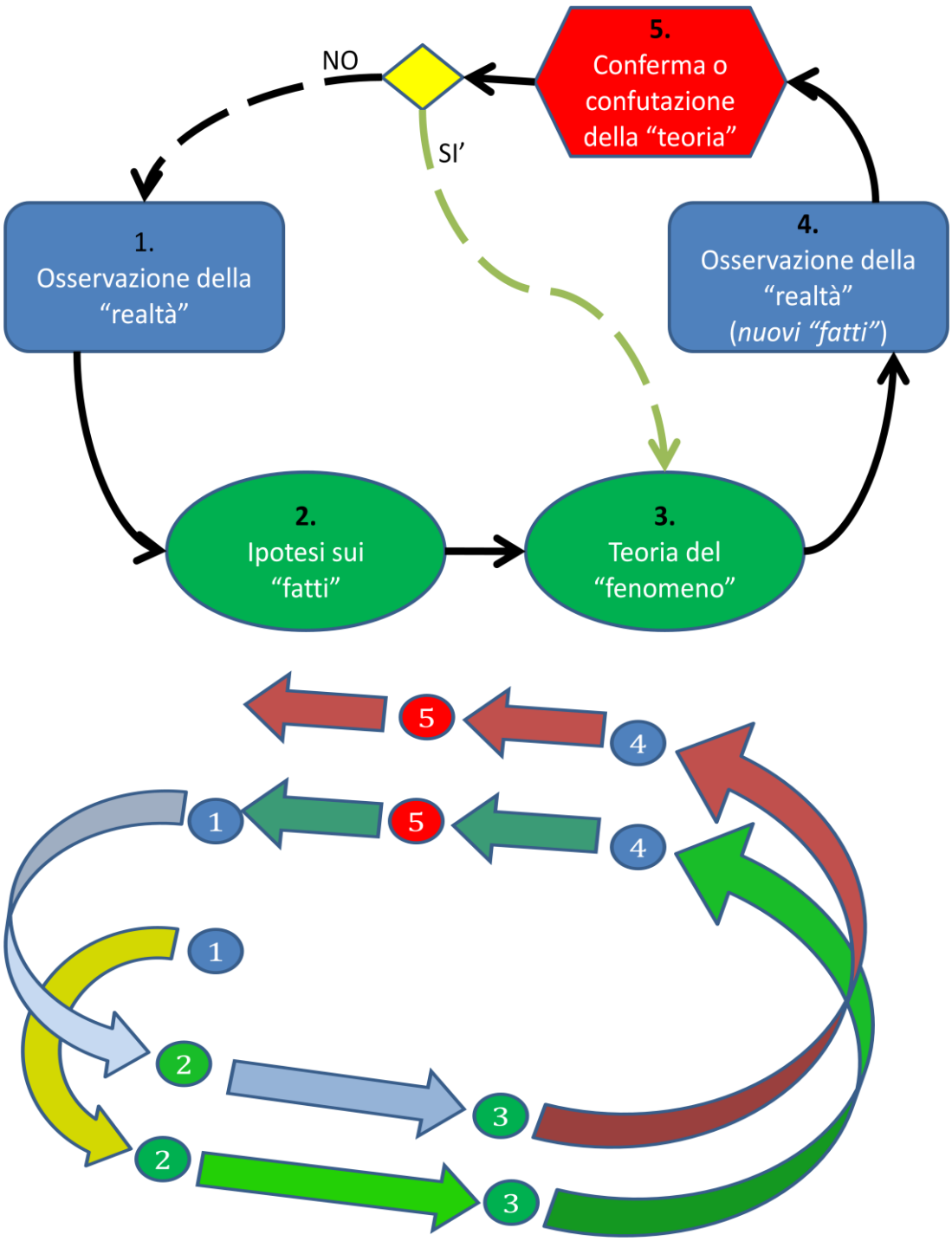
L’elemento di partenza di ogni ricerca sono le conoscenze pregresse del fenomeno che lo studioso possiede, le acquisizioni della letteratura e l’esperienza relativamente a fenomeni analoghi, ma decisive sono le proposte innovative e capacità di intuire e delineare una serie di ipotesi alternative e, quindi, di costruire una teoria.

Sulla base di una tale teoria, molto spesso abbozzata, vengono eseguite le osservazioni e/o le sperimentazioni, che dopo un’analisi accurata, nel rispetto e della logica e della razionalità delle decisioni, consentiranno di “confermare” o di “confutare” la teoria inizialmente formulata. Nel primo caso la teoria diventerà anche un punto di riferimento per altre ricerche o per applicazione di generale utilità. Nel secondo caso occorrerà disporre di ulteriori informazioni che porteranno a replicare i passi precedentemente condotti.

Al termine di ogni ciclo qualcosa è certamente cambiato: le conoscenze del fenomeno sono aumentate e si ha la possibilità di proporre ipotesi e teorie più “ricche” delle precedenti. L’andamento più che “circolare” è effettivamente “a spirale” o “elicoidale”, come si vede nel seguente grafico, e comporta un accrescimento e un miglioramento, almeno tendenziale, delle conoscenze.

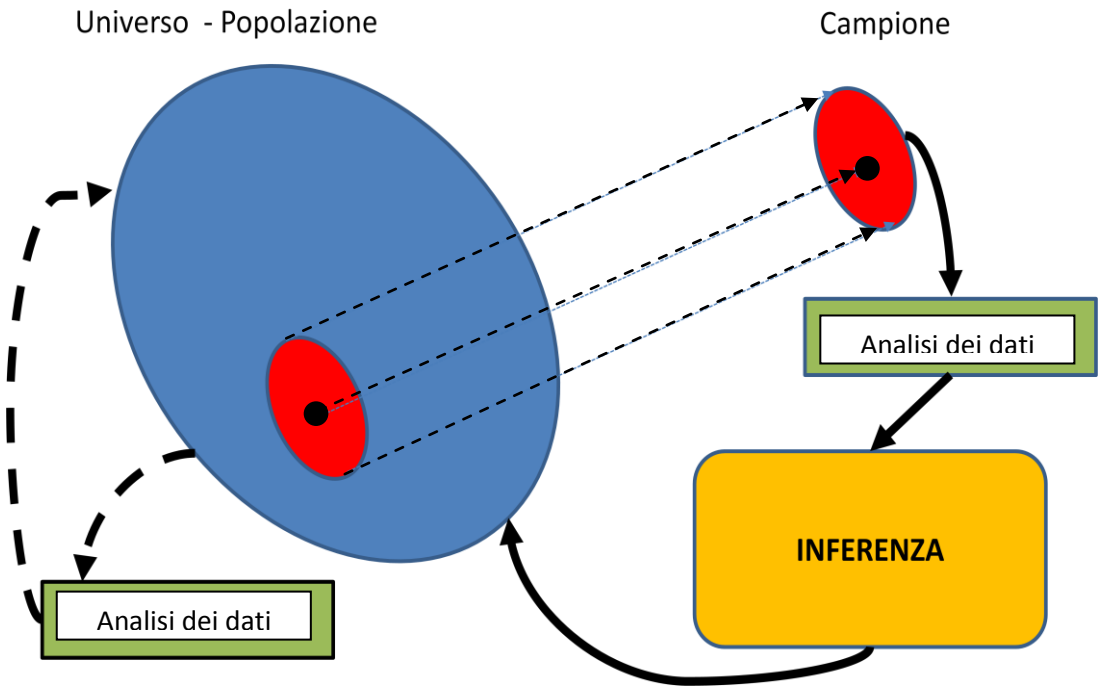
Dal punto di vista della raccolta dei dati e del loro conseguente trattamento, è possibile evidenziare due tipologie metodologiche.

Nel primo caso, qualora le informazioni riguardanti il fenomeno siano estese a tutti i dati dell’intera popolazione/universo allo studio, l’analisi statistica, utilizzando gli strumenti predisposti nell’ ambito della “Statistica descrittiva”, permette di ottenere una sintesi relativa alle caratteristiche dell’intera popolazione e con tale analisi si completa lo studio dal punto di vista quantitativo.



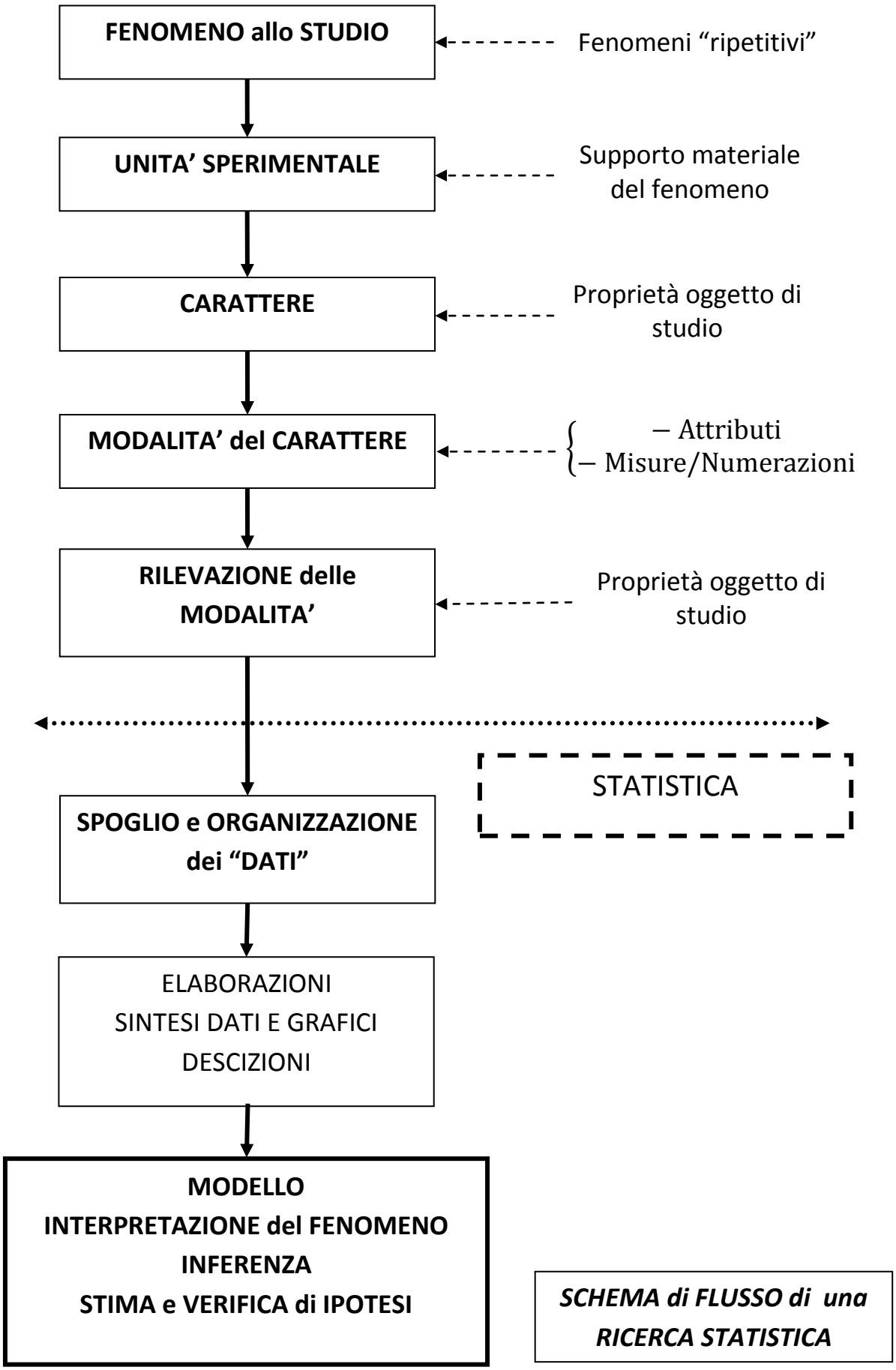
Nel secondo caso, qualora le informazioni riguardanti il fenomeno siano delle "rilevazioni parziali" relative a un "campione", estratto dalla popolazione complessiva, occorre un intervento "induttivo", dato dalla "Inferenza Statistica" che permetta di stimare o verificare ipotesi riguardanti l'intera popolazione, assegnando un grado di precisione e di attendibilità ai risultati numerici ottenuti. In questo caso, il risultato, dipendendo dal campione, varia, giustificando l'impiego del "Calcolo

delle Probabilità" con il proprio metodo "deduttivo", che ha in comune con le discipline matematiche.



La struttura del Corso d'insegnamento della disciplina Statistica, sulla base di quanto è stato esposto, è organizzato in tre aree, strettamente collegate: 1) dedicata agli strumenti principali propri della "Statistica descrittiva"; 2) in cui vengono presentate le basi teoriche del "Calcolo delle Probabilità", con riferimento alle grandezze qualitative aleatorie – "variabili casuali"; 3) in cui verranno forniti i metodi, i teoremi e le procedure proprie della "Inferenza Statistica", relativamente al campionamento, ai problemi di stima parametrica e di verifica d'ipotesi.

Si ricorda che una ricerca statistica può schematizzarsi nei seguenti passi indicati nel diagramma di flusso



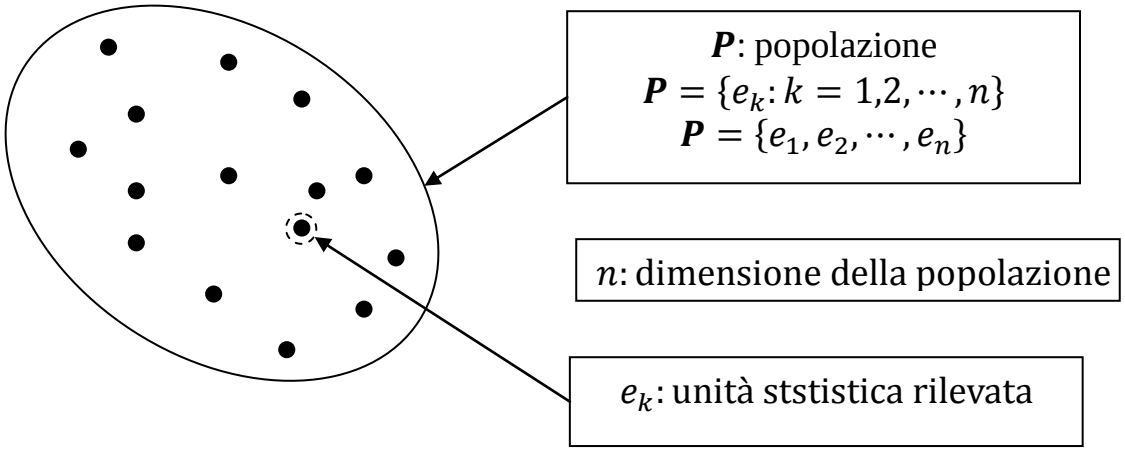
SCHEMA di FLUSSO di una RICERCA STATISTICA

2. La “Statistica Descrittiva”

Gli strumenti della Statistica Descrittiva intervengono con modalità identiche sia sulle osservazioni che provengono da un “universo” che da un “campione” estratto da esso. Conviene parlare di “popolazione” al posto di universo o campione.

Il trattamento svolto dalla Statistica Descrittiva sulle rilevazioni è chiamato spesso anche “Analisi dei Dati”.

La “popolazione” è costituita da un insieme di numerosità finita n di osservazioni, che sono dette “unità statistiche”.



Per ogni unità statistica vengono rilevate q grandezze $q = 1, 2, \cdots$ che sono dette anche “caratteri”. I caratteri sono ottenuti mediante una “astrazione”, rispetto al patrimonio informativo posseduto da ciascuna unità.

I singoli caratteri d’interesse vengono distinti con $X_1, X_2, \cdots, X_q$ e la generica unità statistica e_k possiede il vettore di caratteri:

$$e_k \equiv \{x_{k1}, x_{k2}, \cdots, x_{kq}\}$$

dove x_{k1} è il valore assunto dal carattere X_1 in concomitanza con la k-ma unità statistica e, analogamente, x_{k2}, per il carattere X_2, ecc..

Tutte le informazioni disponibili dalla rilevazione possono essere raccolte in una matrice $(n \times q)$, detta "matrice dei dati".

Matrice dei dati rilevati oggetto dell'indagine

n° unità	X_1	X_2	X_3	$\cdots$	X_q
1	x_{11}	x_{12}	x_{13}	$\cdots$	x_{1q}
2	x_{21}	x_{22}	x_{23}	$\cdots$	x_{2q}
…	…	…	…	…	…
k	x_{k1}	x_{k2}	x_{k3}		x_{kq}
…	…	…	…	…	…
…	…	…	…	…	…
n	x_{n1}	x_{n2}	x_{n3}	$\cdots$	x_{nq}

Caratteri dell'unità "k"

Valori rilevati del carattere X_2

La matrice o tabella dei dati permette un'analisi di lettura per "riga" o per "colonna":

- per riga permette di analizzare, a livello di ogni unità statistica, le modalità dei singoli caratteri che si sono manifestati;
- per colonna, con riferimento a un singolo carattere del fenomeno evidenzia le diversità che si sono verificate nella popolazione oggetto di studio. Tale analisi è quella che ha particolare rilievo in campo statistico.

Ogni carattere si presenta con tipi di "modalità" diverse che possono avere rilevanza dal punto di vista dell'analisi statistica.

Le principali tipologie di "modalità del carattere" possono classificarsi come segue.

- Qualitativo $\begin{cases} - \ non \ gerarchico \\ - \qquad gerarchico \end{cases}$

- Quantitativo $\begin{cases} - \ numerabile \\ - \ \ misurabile \end{cases}$

In relazione alla natura delle operazioni logico-matematiche eseguibili su tali tipi di modalità dei caratteri si possono distinguere in:

- Modalità qualitative "sconnesse" che sono misurate su "scala nominale".

- Modalità qualitative "ordinate" che sono misurate su "scala ordinale".

- Modalità quantitative misurate su "scala di intervalli". Il valore "zero" è convenzionale, es.: nel caso di valori di temperature in gradi centigradi. Per tali grandezze non ha senso valutare incrementi in forma percentuale.

- Modalità quantitative misurate su "scala di rapporti". Il valore "zero" è oggettivo ed esprime la mancanza di entità, es.: è il caso di valori di lunghezze, pesi, velocità, ecc.. Le modalità sono definite tutte positive o tutte negative. Per tali grandezze ha senso valutare incrementi in forma percentuale.

Per quanto riguarda la "cardinalità" potenziale, i caratteri quantitativi si distinguono in:

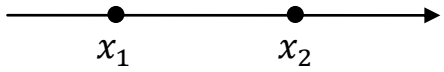
- "Discreti", costituiti da valori distinti numerabili finiti o da una infinità numerabile.

- "Continui", costituiti da valori appartenenti a una classe con potenza del continuo.

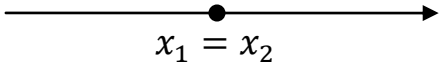
Ove è possibile esprimere o misurare una grandezza si preferisce la modalità "quantitativa" in quanto su di essa si possono svolgere operazioni di:

- "Ordinamento". Se x_1 e x_1 sono due modalità di un carattere, allora, può verificarsi che:

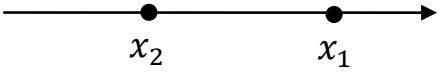
$x_1 < x_2$



$x_1 = x_2$

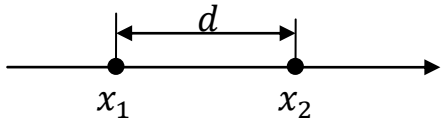


$x_1 > x_2$



- "Distanza". Se x_1 e x_1 sono due modalità di un carattere, allora, può verificarsi che:

$d = |x_1 - x_2| = |x_2 - x_1|$



$$\begin{cases} d = 0 \rightarrow x_1 = x_2 \\ d > 0 \rightarrow x_1 \neq x_2 \end{cases}$$

Se $x_1 < x_2 < x_3 \Longrightarrow |x_1 - x_3| = |x_1 - x_2| + |x_2 - x_3|$.

- Per le modalità quantitative è possibile inoltre svolgere le operazioni algebriche, ottenendo sintesi numeriche di facile determinazione e semplice comprensione o significato.

Esempio 1. Matrice dei dati

k	X_1	X_2	X_3
1	E	1	12
2	E	2	10
3	C	3	14
4	L	4	17
5	C	2	26
6	C	4	15
7	E	1	16
8	L	3	5
9	L	5	28
10	E	2	23
11	C	2	16
12	C	4	20
13	L	3	18
14	L	6	34
15	C	2	19
16	L	4	25
17	C	1	7
18	C	3	18
19	L	4	22
20	E	2	8

Fenomeno allo studio: informazioni riguardanti un complesso di appartamenti lungo la via di una città. *Numerosità*: $n = 20$

Caratteri: numero $q = 3$.

- X_1: tipo di appartamento. C≡Civile; E≡Economico; L≡Lusso;
- X_2: numero locali dell'appartamento;
- X_3: consumo energetico di metano nel trimestre scorso, in m^3.

Osservazioni

La "matrice dei dati" è spesso costituita da colonne più numerose, rispetto a quelle dell'esempio 1, in quanto i caratteri da tenere in considerazione e comunque rilevati comprendono aspetti di cui si vuol verificare l'influenza su quelli scelti specificatamente per l'indagine oggetto di interesse, questo avviene in particolare in inchieste e studi demoscopici. L'analisi dei dati si svolge, in un primo tempo, studiando i dati relativi a ogni singolo carattere (per "colonna") e, in secondo luogo, esaminando le relazioni tra due caratteri per volta e poi estendendo lo studio a più caratteri considerati congiuntamente.

Nella presentazione degli argomenti dedicati alla statistica descrittiva si seguirà una sequenza, presentando l'analisi dei caratteri unidimensionali, indi l'analisi bidimensionale e terminando con alcuni cenni allo studio multivariato.

3. Analisi descrittiva di un carattere unidimensionale

Si indichi con X il carattere preso in considerazione e con $\{x_k; k = 1,2, \cdots, n\}$ i valori rilevati per tale carattere nelle unità della popolazione P oggetto di studio, successione che viene spesso indicata come "serie di dati" relativi al carattere X, denominato sovente "variabile statistica" o più precisamente: a) "mutabile": se presenta modalità qualitative; b) "variabile": se presenta modalità quantitative.

In molte situazioni, per una lettura più valida dei dati, al posto della successione originaria, si può considerare la "serie ordinata", particolarmente nel caso di modalità quantitative, in ordine crescente.

$$\{x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}\} = \{x_{(k)}; k = 1,2, \cdots, n\}$$

Esempio 2

Riprendendo i dati dell'esempio 1, relativi, al carattere X_2: numero locali dell'appartamento, si ha

$$\{x_k\} = \{1,2,3,4,2,4,1,3,5,2,2,4,3,6,2,4,1,3,4,2\}$$
$$\Downarrow$$
$$\{1,1,1,2,2,2,2,2,2,3,3,3,3,4,4,4,4,4,5,6\}$$

e per i dati relativi al carattere X_3: consumo energetico di metano nel trimestre scorso, in m^3, si ha

$$\{x_k\} = \{12,10,14,17,26,15,16,5,28,23,16,20,18,34,19,25,7,18,22,8\}$$
$$\Downarrow$$
$$\{5,7,8,10,12,14,15,16,16,17,18,18,19,20,22,23,25,26,28,34\}$$

E' possibile svolgere l'ordinamento della serie dei dati anche per caratteri qualitativi di tipo gerarchico oppure convenzionalmente ordinando per modalità di carattere, ad es. in ordine alfabetico, come per i dati relativi al carattere X_1: tipo di appartamento, dell'esempio 1.

$$\{x_k\} = \{E, E, C, L, C, C, E, L, L, E, C, C, L, L, C, L, C, C, L, E\}$$
$$\Downarrow$$
$$\{C, C, C, C, C, C, C, C, E, E, E, E, E, L, L, L, L, L, L, L\}$$

oppure

$$\{E, E, E, E, E, C, C, C, C, C, C, C, C, L, L, L, L, L, L, L\}$$

L'ordinamento dei dati rilevati può aiutare la lettura del carattere allo studio ma la numerosità n che risulta spesso elevata rende necessaria un'organizzazione dei dati in forma tabellare mediante un intervento di"spoglio" che consiste nel contare le unità statistiche n_i aventi una specifica modalità distinta x_i, per $i = 1,2,\cdots,p \leq n$ del carattere X, essendo p il numero complessivo di tali modalità:

$$x_i \to n_i = \#(X = x_i)$$

dove $\#(\cdot)$ è l'operatore di conteggio delle unità della popolazione P oggetto di studio che rispettano la condizione posta in argomento.

Le numerosità n_i sono dette "frequenze semplici assolute" e sono numeri interi non negativi tali che:

$$\sum_{i=1}^{p} n_i = n$$

e la variabile statistica può rappresentarsi sinteticamente mediante le coppie, in alternativa alla rappresentazione mediante "serie" e viene detta, qualora il carattere sia di tipo quantitativo, "seriazione".

$$X \to \{x_i, n_i; i = 1,2, \cdots, p\}$$

Oltre alle frequenze semplici assolute si impiegano spesso le "frequenze semplici relative" per confrontate lo stesso carattere in popolazioni di numerosità complessiva diversa, che sono date da:

$$f_i = \frac{n_i}{n} \text{ per } i = 1,2, \cdots, p$$

con $0 \leq f_i \leq 1$ e $\sum_{i=1}^{p} f_i = 1$.

Esempio 3

Riprendendo i dati dell'esempio 1, relativi, al carattere X_1: tipo di appartamento, in cui le modalità distinte sono solo tre abbiamo la tabella

$$\{E, E, E, E, E, C, C, C, C, C, C, C, C, L, L, L, L, L, L, L\}$$

x_i	n_i	f_i
Economico	5	0,25
Civile	8	0,40
Lusso	7	0,35
Σ	20	1,00

Esempio 4

Per i dati dell'esempio 1, relativi al carattere X_2: numero locali dell'appartamento con modalità quantitative di tipo discreto, si ottiene una tabella analoga alla precedente ma dato l'ordinamento naturale evidenzia il modo di distribuirsi dei dati ed è detta "tabella di seriazione o di distribuzione". In situazioni analoghe è utile introdurre anche le "frequenze cumulate assolute" N_i e quelle relative F_i, definite come:

$$N_i = \#(X \leq x_i) = \sum_{j=1}^{i} n_j \leq n = N_{i-1} + i \quad \text{per } i = 1,2,\cdots,p$$

con $N_0 = 0$ e $N_p = n$;

$$F_i = \frac{N_i}{n} = \sum_{j=1}^{i} f_j \leq 1 = F_{i-1} + i \quad \text{per } i = 1,2,\cdots,p$$

con $F_0 = 0$ e $F_p = 1$.

{1,1,1,2,2,2,2,2,2,3,3,3,3,4,4,4,4,4,5,6}

x_i	n_i	f_i	N_i	F_i
1	3	0,15	3	0,15
2	6	0,30	9	0,45
3	4	0,20	13	0,65
4	5	0,25	18	0,90
5	1	0,05	19	0,95
6	1	0,05	20	1,00
Σ	20	1,00		

Se il carattere X preso in considerazione è di tipo quantitativo "continuo" e quindi le modalità distinte sono teoricamente infinite, come avviene per grandezze misurabili, conviene sintetizzare la raccolta dei dati stabilendo una successione di p classi di intervallo in $\Re$ opportune, sia come numerosità p che come estremi.

- Successioni di "intervalli"
$$I_i = (v_{i-1}, v_i] \equiv v_{i-1} \dashv v_i = \{x \in \Re: v_{i-1} < x \leq v_i\}$$

- Estremi degli intervalli
$$v_0 < v_1 < \cdots v_{i-1} < v_i < \cdots v_p$$

- Ampiezza degli intervalli
$$a_i = |v_{i-1} - v_i| = v_i - v_{i-1} > 0$$

- Scelta di v_0 e di v_p
$$v_0 \leq \min_k\{x_k\} = x'$$

$$v_p \leq \max_k\{x_k\} = x''$$

- Scelta ampiezza intervalli

  Se è possibile conviene considerare gli intervalli di ampiezza uguale
$$a_i = a = cost. \ \rightarrow \ a = \left(v_p - v_0\right)/p$$

- Spoglio dei dati

  Per ciascun intervallo I_i si individua il numero di unità statistiche n_i contenute in esso, "frequenze semplici assolute"
$$I_i \rightarrow n_i = \#(X \in I_i) = \#(v_{i-1} < X \leq v_i)$$

  con $0 \leq n_i \leq n$ e $\sum_{i=1}^{p} n_i = n.$

- Densità dei dati nell'intervallo

  Ogni intervallo I_i può presentare una ampiezza propria a_i; è opportuno misurare l'addensamento o concentrazione dei dati osservati nell'intervallo mediante una misura di "densità assoluta"

$$d_i^* = \frac{n_i}{a_i} = \frac{n_i}{v_i - v_{i-1}} \text{ per } i = 1,2,\cdots,p$$

  con $d_i^* \geq 0$ e $\sum_{i=1}^{p} d_i^* a_i = n.$

- Oltre alle frequenze assolute semplici n_i è possibile definire anche
  - "Frequenze semplici relative":

$$f_i = n_i/n \text{ per } i = 1,2,\cdots,p;$$

    con $0 \leq f_i \leq 1$ e $\sum_{i=1}^{p} f_i = 1.$

  - "Densità relative":

$$d_i = \frac{d_i^*}{n} = \frac{f_i}{a_i} \quad \text{per } i = 1,2,\cdots,p$$

    con $d_i \geq 0$ e $\sum_{i=1}^{p} d_i a_i = 1.$

  - "Frequenze cumulate assolute"

    Analogamente a quanto visto per i caratteri quantitativi con modalità di tipo discreto è possibile definire:

$$N_i = \#(X \leq v_i) = \sum_{j=1}^{i} n_j = N_{i-1} + n_i \text{ per } i = 1,2,\cdots,p$$

    con $N_0 = 0$ e $N_p = n$. Si osservi che N_i indica la numerosità di osservazioni con valori inferiori o uguali all'estremo superiore dell'intervallo v_i.

  - "Frequenze cumulate relative"

$$F_i = \frac{N_i}{n} = \sum_{j=1}^{i} f_j = F_{i-1} + f_i \quad \text{per } i = 1,2,\cdots,p$$

con $F_0 = 0$ e $F_p = 1$.

- Valore centrale della classe dell'intervallo I_i
  Al fine di adottare un valore rappresentativo dei diversi valori compresi nell'intervallo I_i, si ricorre all'impiego del valore centrale x_i dell'intervallo stesso, interpretandolo come elemento della classe di equivalenza dei valori contenuti in I_i

$$I_i \to x_i = \frac{v_{i-1} + v_i}{2} = v_{i-1} + \frac{a_i}{2} \quad \text{per} \quad i = 1,2,\cdots,p$$

Esempio 5

Come esempio si può considerare il caso del carattere X_3: consumo energetico di metano nel trimestre scorso, in m^3, presentato nell'esempio 1. Scelti i valori di $p = 4, v_0 = 0, v_4 = 40$ e $a = cost. = 10$, si ha:

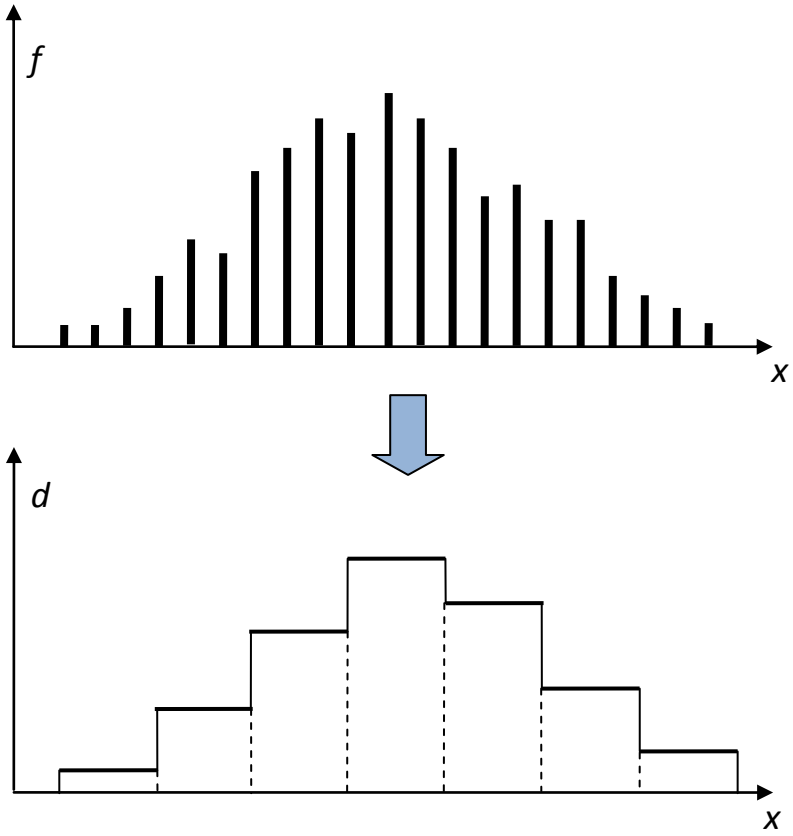
{5,7,8,10,12,14,15,16,16,17,18,18,19,20,22,23,25,26,28,34}

I_i	n_i	f_i	a_i	d_i^*	d_i	N_i	F_i	x_i
0 ⊣ 10	4	0,20	10	0,4	0,020	4	0,20	5
10 ⊣ 20	10	0,50	10	1,0	0,050	14	0,70	15
20 ⊣ 30	5	0,25	10	0,5	0,015	19	0,95	25
30 ⊣ 40	1	0,05	10	0,1	0,005	20	1,00	35
Σ	20	1,00						

La formazione di tabelle di frequenza può risultare pesante se svolta manualmente ma, attualmente, con semplici algoritmi digitali, è di facile ottenimento.

Osservazioni

Per una variabile statistica X, con modalità di tipo "discreto", può convenire rappresentare la distribuzione dei dati in forma di seriazione per classi di intervallo $\{I_i, f_i\}$ invece che in termini delle modalità discrete originarie. Si ricorre a ciò quando il numero delle modalità p originarie è molto grande. Si sceglie un numero nuovo di intervalli $p' \ll p$, e si scelgono gli estremi degli intervalli $I_i = (v_{i-1}, v_i]$ come per i caratteri di tipo continuo:



La rappresentazione per classi di intervallo comporta delle "approssimazioni", introdotte dall'operatore statistico, sia sulla distribuzione che sui suoi "indicatori sintetici e il grado di tale approssimazione dipende dalla scelta degli intervalli (sia in numero che negli estremi).

4. Rappresentazioni grafiche

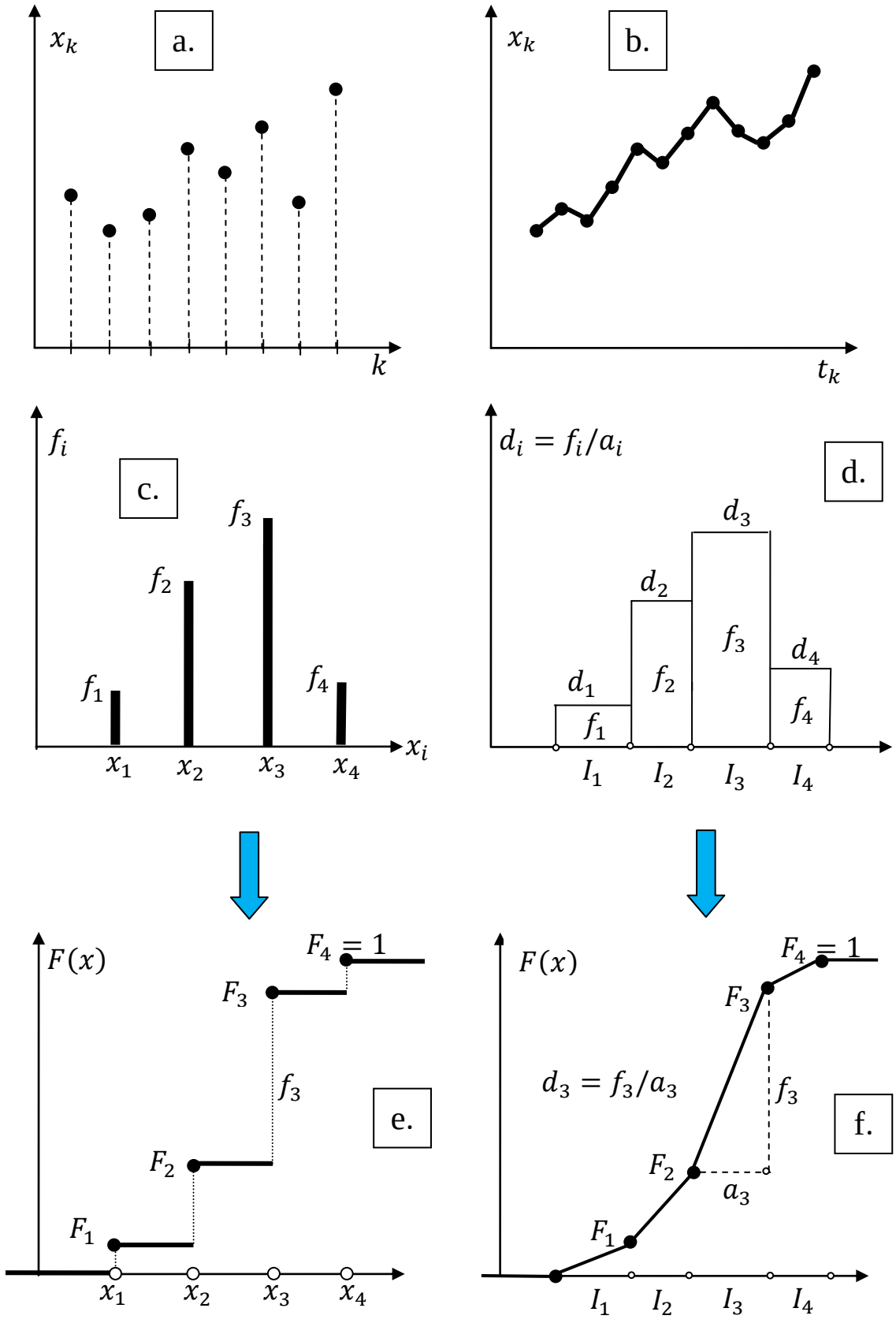
E' un modo alternativo e complementare di rappresentazione dei dati rispetto a quello tabellare, che permette di avere una visione d'insieme del fenomeno allo studio. In figura sono presentati esempi di:

a. serie di un carattere quantitativo $\{x_k; k = 1,2, \cdots n\}$;
b. serie temporale di un carattere quantitativo $\{(t_k, x_k); k = 1,2, \cdots n\}$;
c. seriazione nel caso di grandezza quantitativa discreta$\{x_{i,}; f_{i,} i = 1,2, \cdots p\}$;
d. seriazione nel caso di grandezza quantitativa per classe di intervalli $\{I_{i,}; d_{i,} i = 1,2, \cdots p\}$;
e. andamento delle frequenze cumulate nel caso di seriazione discreta e corrispondente funzione di distribuzione $F(x)\ \forall x \in \Re$;
f. andamento delle frequenze cumulate nel caso di seriazione per classe di intervalli e corrispondente funzione di distribuzione $F(x)\ \forall x \in \Re$.

Per "funzione di distribuzione" si intende la frequenza, in termini relativi, di valori del carattere X inferiori o uguali al generico valore x:

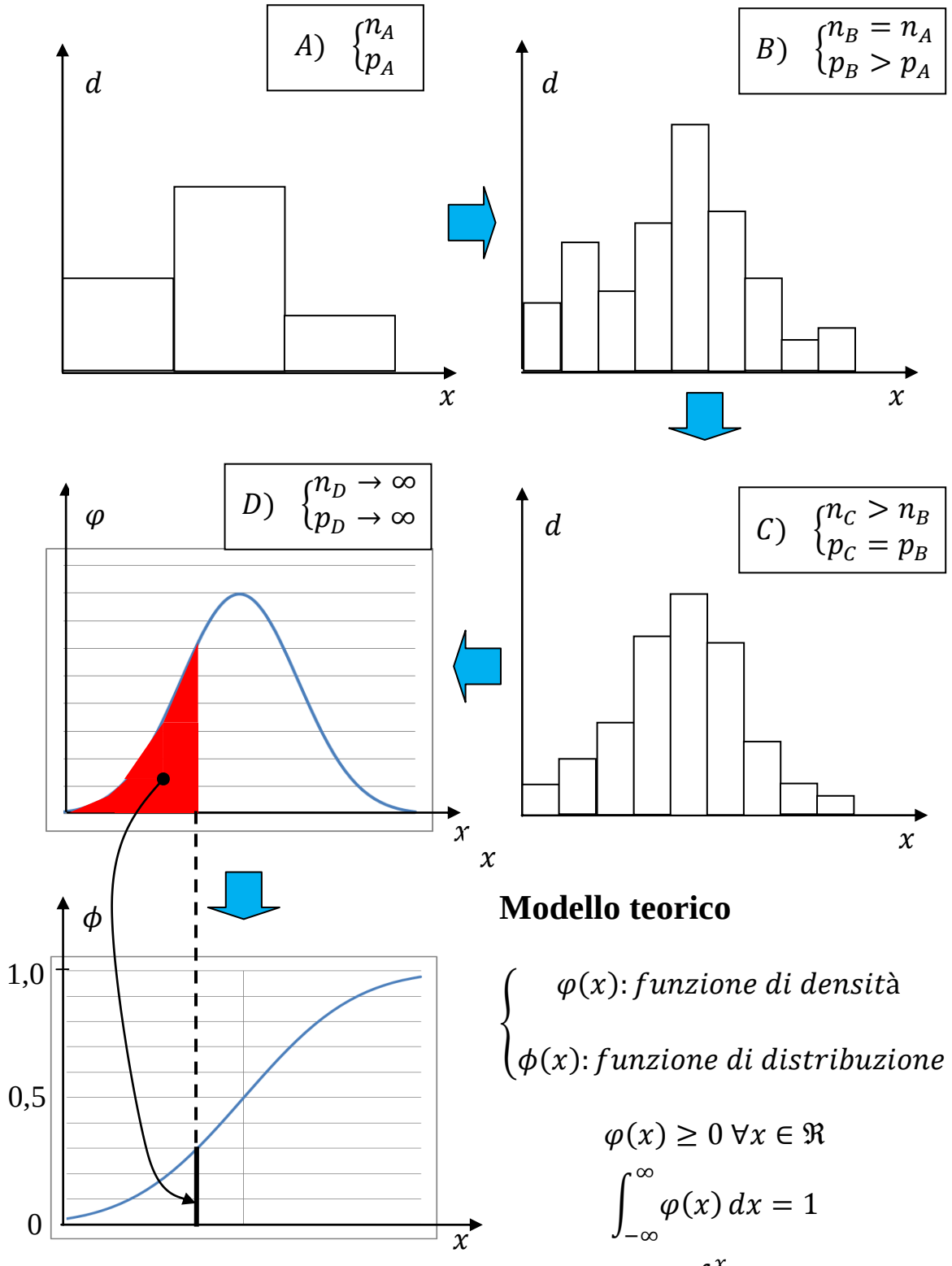
$$F(x) = \frac{1}{n}\#(X \leq x) \text{ per } x \in \Re$$

Si osservi che tale funzione ha un andamento monotono non decrescente, che passa in corrispondenza dei punti definiti dalle frequenze cumulate (x_i, F_i), nel caso di seriazione discreta, oppure (v_i, F_i), nel caso di seriazione per classe di intervalli. Nel primo caso la funzione $F(x)$ presenta salti di continuità e ha un andamento a gradini, nel secondo caso si presenta con continuità a tratti lineari in relazione al grafico della densità considerato costante per ogni classe di intervallo.



In molte situazioni le rilevazioni disponibili sono molto numerose, quindi può risultare opportuno approssimare l'andamento del grafico

della seriazione per classi di intervallo a una funzione continua, come è illustrato in figura.



Modello teorico

$$\begin{cases} \varphi(x): funzione\ di\ densità \\ \phi(x): funzione\ di\ distribuzione \end{cases}$$

$$\varphi(x) \geq 0 \ \forall x \in \Re$$

$$\int_{-\infty}^{\infty} \varphi(x)\, dx = 1$$

$$\phi(x) = \int_{-\infty}^{x} \varphi(u)\, du$$

5. Rappresentazioni alternative di una variabile quantitativa X

Le grandezze quantitative, originate da conteggio o da misure, costituiscono la più frequente modalità di manifestazione di un fenomeno; come è stato già evidenziato, le rilevazioni si possono rappresentare in modo differente.

"Serie": successione dei valori osservati

$$X \to \{x_k; k = 1,2, \cdots, n\};$$

"Serie ordinata": successione dei valori osservati posti in ordine crescente (non decrescente), con $x_{(k-1)} \leq x_{(k)}$

$$X \to \{x_{(k)}; k = 1,2, \cdots, n\};$$

"Seriazione" per modalità discrete, con $x_{i-1} < x_i$

$X \to \{x_i, n_i; i = 1,2, \cdots, p\}$, con frequenze semplici assolute

$X \to \{x_i, f_i; i = 1,2, \cdots, p\}$, con frequenze semplici relative

$X \to \{x_i, F_i; i = 1,2, \cdots, p\}$, con frequenze cumulate relative

$X \to F(x) = \#(X \leq x)/n$, funzione di distribuzione per $x \in \Re$;

"Seriazione" per modalità continue (o classi di intervallo $I_i = v_{i-1} \dashv v_i$)

$X \to \{I_i, n_i; i = 1,2, \cdots, p\}$, con frequenze semplici assolute

$X \to \{I_i, f_i; i = 1,2, \cdots, p\}$, con frequenze semplici relative

$X \to \{I_i, F_i; i = 1,2, \cdots, p\}$, con frequenze cumulate relative

oppure, caratterizzando l'intervallo I_i con il valore centrale dello stesso $x_i = (v_{i-1} + v_i)/2$

$$X \to \{x_i, n_i; i = 1,2, \cdots, p\} \qquad X \to \{x_i, f_i; i = 1,2, \cdots, p\}$$

$$X \to \{v_i, F_i; i = 1,2, \cdots, p\}$$

$X \to F(x) = \#(X \leq x)/n$, funzione di distribuzione per $x \in \Re$, con $F(x) = F_i$ per $x = v_i; i = 1,2, \cdots, p$

Si definisce anche una "funzione di densità":

$$d(x) \geq 0 \text{ per } x \in \Re,$$

con $d(x) = d_i = f_i/(v_i - v_{i-1})$ per $x \in I_i$; oppure $d(x) = 0$, in qualunque altro caso.

Tutte queste formulazioni risultano equivalenti nella loro rappresentazione dei dati osservati e verranno impiegate in seguito in modo alternativo o in quello più opportuno per lo specifico scopo.

6. Rappresentazione sintetica di una variabile quantitativa X

Le rappresentazioni in forma di successione dei dati o in tabelle di frequenza pur facilitando i confronti e i paragoni tra fenomeni analoghi o riferiti a situazioni spaziali o temporali diverse, spesso non permettono di dare risposte immediate e univoche. Si ricorre allora a delle sintesi dei dati stessi che evidenziano mediante un unico valore (o almeno con pochi valori) la proprietà/e del carattere allo studio.

In particolare ci si soffermerà su due classi di tali indicatori sintetici:

a) indici di "locazione" o "posizione";
b) indici di "dispersione" o di "variabilità".

La presentazione di tali classi di indicatori sarà completata con una famiglia di indicatori, detti "momenti" dei dati osservati che comprendono sia indicatori di posizione sia indicatori di variabilità, e altri che misurano aspetti del carattere quantitativo unidimensionale oggetto di interesse.

7. Sintesi di una variabile quantitativa unidimensionale

Per effettuare confronti tra diverse grandezze quantitative raccolte in "serie" o in "seriazione" un primo strumento è quello di sintetizzare i dati mediante un indice di "posizione" o "locazione" che possa rappresentarli nel loro complesso.

Considerata una variabile statistica X, definita mediante le osservazioni raccolte in: $\{x_k; k = 1,2,\cdots,n\}$ o $\{x_i, n_i; i = 1,2,\cdots,p\}$ ecc., indicato con $\theta = \theta_X = \theta(X)$, un generico indice di posizione è una funzione dei dati osservati di X

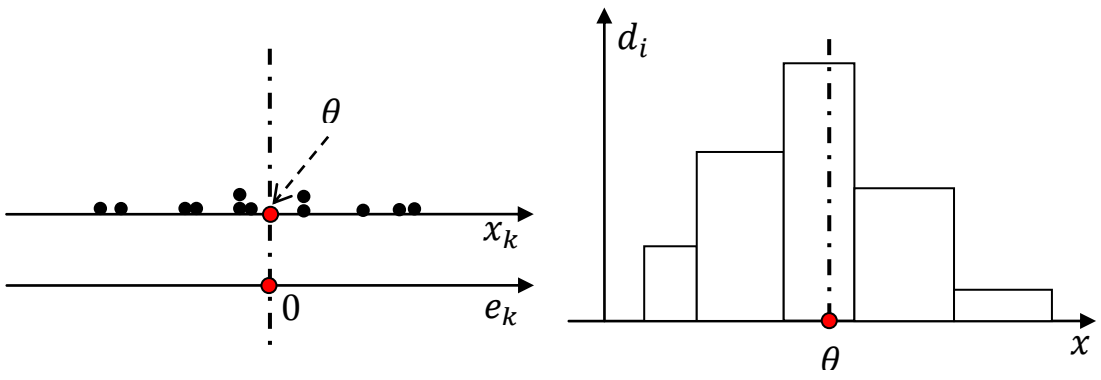
$$\theta(X) = \theta(x_k; k = 1,2,\cdots,n) \text{ ecc.}$$

E' possibile pensare la variabile X come somma di due componenti: una "strutturale" individuata dall'indice di posizione θ; l'altra dalla componente "aleatoria" E:

$$X = \theta + E \;\rightarrow\; x_k = \theta + e_k; k = 1,2,\cdots,n$$

La componente aleatoria E, detta anche "variabile accidentale", "errore", "scarto o scostamento", evidenzia la variabilità presente nei dati osservati e quindi ha le caratteristiche proprie di una variabile statistica e può rappresentarsi in forma di serie o seriazione.

$$E = X - \theta$$
$$\{e_k = x_k - \theta\,; k = 1,2,\cdots,n\}$$
$$\{e_i = x_i - \theta_i, f_i; i = 1,2,\cdots,p\} \text{ (frequenze relative)}$$



8. Proprietà degli indici di posizione

L'indice di posizione θ di una variabile statistica X, dovendo rappresentare i valori osservati, deve essere un numero compreso tra il valore "minimo" e quello "massimo", estremi inclusi:

$$\theta \in [x', x''] \rightarrow x' \leq \theta \leq x''$$

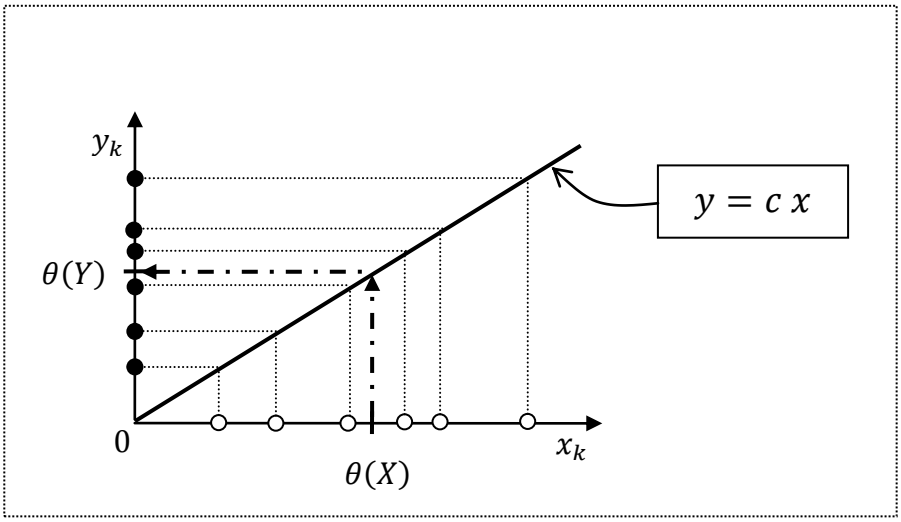
dove $x' = \min\{x_k; k = 1,2,\cdots,n\} = x_{(1)}$ e $x'' = \max\{x_k; k = 1,2,\cdots,n\} = x_{(n)}$ nel caso di serie di dati, $x' = x_1$ e $x'' = x_p$ nel caso di seriazione discreta, $x' = v_0$ e $x'' = v_p$ nel caso di seriazione per classi di intervallo. Questa proprietà che tutti gli indicatori di posizione devono avere è detta "proprietà di Cauchy".

Altre proprietà che gli indici di posizione possono presentare e che permettono di caratterizzare e differenziare i diversi indici proposti sono le seguenti.

1) *Proprietà "moltiplicativa"*: qualora una variabile statistica X presenti un cambiamento "di scala" anche l'indice di posizione $\theta(X)$ comporta un uguale cambiamento.

Se tale proprietà è valida, indicata con $Y = cX$, dove $c = cost. \neq 0$, allora:

$$\theta(Y) = \theta(cX) = c\,\theta(X)$$



2) *Proprietà di "monotonicità"*: se una variabile statistica Y presenta valori corrispondenti "maggiori o uguali" a quelli di una variabile statistica X allora l'indice di posizione $\theta(Y)$ "non è inferiore" a $\theta(X)$.

Se tale proprietà è valida allora per $X \equiv \{x_k\}$ e $Y \equiv \{y_k\}$ con $\forall\, y_k \geq x_k$ si ha

$$\theta(Y) \geq \theta(X)$$

3) *Proprietà di "linearità"*: se una variabile statistica Z è legata ad altre variabili da una relazione "lineare", ad esempio, $Z = a + bX + cY \rightarrow \{z_k = a + bx_k + cy_k; k = 1,2,\cdots,n\}$, con $a, b, c \in \Re$ valori costanti, allora l'indice di posizione θ gode della proprietà di "linearità" se:

$$\theta(Z) = a + b\theta(X) + c\theta(Y)$$

In questo caso l'operatore $\theta(\cdot)$ è detto "lineare" e la proprietà 1) ne costituisce un caso particolare per $a = c = 0$.

L'indice di posizione θ può intendersi come una applicazione dall'insieme dei dati $\{x_k; k = 1,2,\cdots,n\} \in \Re^n$ in $\Re$, nel rispetto della proprietà di Cauchy, in particolare si tratta di una funzione $\theta(X): \Re^n \rightarrow \Re$ in cui l'argomento $X \equiv \{x_k; k = 1,2,\cdots,n\}$ è costituito da n componenti "scambiabili" cioè tali che hanno rilevanza solo i valori osservati non l'ordine con cui si manifestano, in quanto la permutazione degli stessi origina un identico valore per θ.

9. La media aritmetica

L'indice di posizione più frequentemente impiegato, sia per la sua semplicità euristica che per il ruolo svolto nella teoria probabilistica e nella statistica inferenziale, è quello della "media aritmetica". Disponendo i dati nella forma di "serie" $\{x_k; k = 1,2,\cdots,n\}$, la media

aritmetica è data dalla somma delle osservazioni divisa per il loro numero

$$m = M(X) = \frac{\sum_{k=1}^{n} x_k}{n}$$

espressione che diviene nel caso di seriazioni pari a:

$$m = M(X) = \frac{\sum_{i=1}^{p} x_i n_i}{n} = \sum_{i=1}^{p} x_i f_i$$

dove x_i indica le modalità distinte nella situazione di dati per valori discreti o i valori centrali nella situazione mediante classi di intervallo.

Osservazioni

- Spesso al posto del simbolo m vengono utilizzati: il simbolo μ, se l'analisi è estesa all'intero universo del fenomeno allo studio, il simbolo $\bar{x}$, se l'analisi riguarda dati campionari.
- Se tutte le osservazioni sono identiche come valore, allora la variabile X oggetto di interesse è detta "degenere", ne consegue che tutti gli indici di posizione compresa la media aritmetica coincidono con l'unico valore in comune x

  $$X \equiv \{x_k = x = cost.\ \forall k\} \rightarrow \theta(X) \equiv M(X) \equiv x \equiv x' \equiv x''$$

- Nella situazione in cui nel calcolo di un indice di posizione, in particolare del calcolo della media aritmetica, si utilizzano i valori centrali delle classi di intervallo si ottiene un valore approssimato rispetto a quello direttamente ottenibile dalla successione dei valori $\{x_k\}$ o $\{x_{(k)}\}$.
- In molti fenomeni fisici ed economici (es.: quantità di sostanze inquinanti, reddito personale, costi di materiali, ecc.) la grandezza complessiva del fenomeno, data dalla domma dei valori osservati, ha un suo significato ed è detta "intensità totale" Q:

$$Q = \sum_{k=1}^{n} x_k = \sum_{i=1}^{n} x_i n_i = n\, m$$

da cui $m = Q/n$.

Verifica delle proprietà

Proprietà di Cauchy – Essendo

$$x' = \min\{x_k\} \leq x_k \leq \max\{x_k\} = x'' \ \forall k$$

sommando membro a membro per tutti i valori di k, si ha

$$\sum_k x' \leq \sum_k x_k \leq \sum_k x'' \rightarrow x'n \leq \sum_k x_k \leq x''n$$

dividendo tutti i membri per n, si ha

$$x' \leq \frac{1}{n}\sum_k x_k \leq x'' \rightarrow x' \leq m \leq x''$$

quindi la "media aritmetica soddisfa la proprietà di Cauchy". Si può precisare che, a esclusione del caso in cui *X* è "degenere", si ha

$$x' < m < x''$$

Proprietà "moltiplicativa" – Se si considera la variabile $Y = cX$, con $c \neq 0$, allora

$$y_k = cx_k, \forall k \rightarrow \sum_k y_k = \sum_k cx_k = c\sum_k x_k$$

e quindi

$$M(Y) = \frac{c}{n}\sum_k x_k = M(cX) = cM(X)$$

Si è verificato che la "media aritmetica" soddisfa la proprietà "moltiplicativa".

Proprietà di "linearità" – Sia Z legata ad altre variabili dalla relazione "lineare", $Z = a + bX + cY$, con $a, b, c \in \Re$, allora $\{z_k = a + bx_k + cy_k; k = 1,2,\cdots,n\}$, la media aritmetica di Z risulta

$$M(Z) = M(a + bX + cY)$$
$$= \frac{1}{n}\sum\nolimits_k (a + bx_k + cy_k) = \frac{1}{n}\left[an + b\sum_k x_k + c\sum_k y_k\right]$$
$$= a + bM(X) + cM(Y)$$

Quindi, la "media aritmetica soddisfa la proprietà di linearità". L'operatore $M(\cdot)$ è un "operatore lineare" e gode delle proprietà di tali operatori e conviene impiegarlo al posto delle relazioni espresse mediante le sommatorie che, a seconda del tipo di rappresentazione dei dati, possono essere formalmente diverse esso, inoltre, presentano analogie con "sommatoria" e "derivata".

Proprietà di "monotonicità" – Se due variabili statistiche X e Y sono tali che $x_k \leq y_k \; \forall k$, in tal caso sinteticamente si indicherà $X \leq Y$, allora:

$$\sum\nolimits_k x_k \leq \sum\nolimits_k y_k$$

Si ha, dividendo per n entrambi i membri:

$$\frac{1}{n}\sum\nolimits_k x_k \leq \frac{1}{n}\sum\nolimits_k y_k \;\rightarrow\; M(X) \leq M(Y)$$

Quindi, la "media aritmetica soddisfa la proprietà di monotonicità". Inoltre se nelle n osservazioni ve ne sia una, ad es. $k = 1$, tale che $x_1 < y_1$, mentre per le rimanenti $n - 1$ valga la condizione di uguaglianza $\{x_k = y_k; k = 2,\cdots,n\}$, si ha la proprietà di "monotonicità stretta":

$$M(X) < M(Y)$$

Si può indicare, dalle verifiche sulle proprietà dell'indice "media aritmetica", che esso soddisfa tutte le proprietà precedentemente elencate, giustificandone l'impiego diffuso in aggiunta alle sua facilità di calcolo.

Proprietà specifiche della media aritmetica

La media aritmetica presenta alcune proprietà riguardanti gli "scarti" o "scostamenti" $E = X - m$, ossia la componente aleatoria della variabile oggetto di studio.

1. *La media (o la somma) degli scarti dalla media aritmetica di X è nulla.*

   Infatti, applicando a E l'operatore lineare $M(\cdot)$ si ha

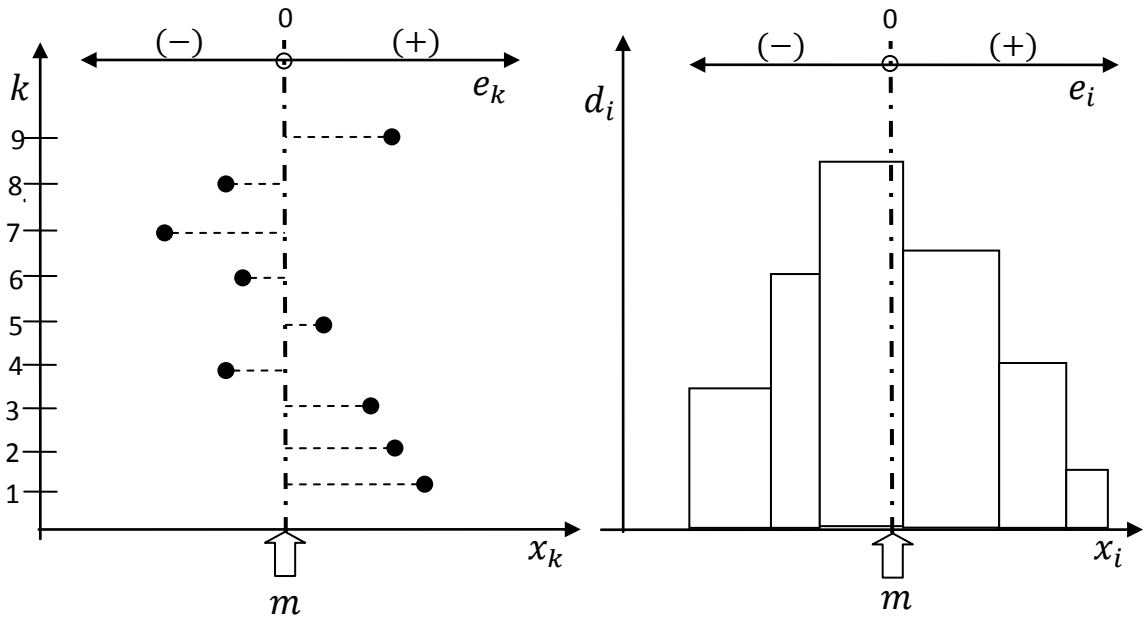
$$M(E) = M(X - m) = M(X) - m = m - m = 0$$

   da cui:

$$\sum_{k=1}^{n} (x_k - m) = 0$$

$$\frac{1}{n}\sum_{i=1}^{p} (x_i - m)n_i = 0 \text{ e } \sum_{i=1}^{p} (x_i - m)f_i = 0$$

   Questa proprietà permette di interpretare la media aritmetica come il valore "baricentrico" dei dati osservati.



2. La media aritmetica minimizza la media (o la somma) dei quadrati degli scarti da un generico indice θ.

Sia $\psi(\theta) = M\{(X-\theta)^2\}$ la media dei quadrati degli scarti da un generico indice θ, la funzione $\psi(\theta)$ può essere minimizzata uguagliando a zero la derivata prima: $\psi'(\theta) = 0$. Trattandosi di operatori lineari gli operatori $M(\cdot)$ e $d/d\theta$ possono essere scambiati:

$$\psi'(\theta) = \frac{d}{d\theta} M\{(X-\theta)^2\} = M\left\{\frac{d}{d\theta}[(X-\theta)^2]\right\} = M\{-2(X-\theta)\}$$
$$= -2M(X-\theta) = -2(M(X)-\theta) = 0$$

da cui si ottiene

$$(M(X)-\theta) = 0 \rightarrow \theta = M(X) \rightarrow \theta = m$$

ed essendo la derivata seconda $\psi''(\theta) = \frac{d}{d\theta}[-2(M(X)-\theta)] = 2 > 0$, il punto $\theta = m$ è di minimo assoluto per $\psi(\theta)$

$$m = \underset{\theta}{\operatorname{argmin}}\, M\{(X-\theta)^2\}$$

dove il valore di minimo di $\psi(\theta)$ è dato da $M\{(X-m)^2\} = M\{(E)^2\} = \sigma^2 \geq 0$ in cui σ^2, come si vedrà nel seguito, è un indicatore di dispersione di X, denominato "varianza".

Esempi di calcolo della media aritmetica

Esempio 6

Riprendendo i dati dell'Esempio 2 relativi alla serie di osservazioni di consumo di gas in $n = 20$ appartamenti

$$\{x_k\} = \{12,10,14,17,26,15,16,5,28,23,16,20,18,34,19,25,7,18,22,8\}$$

La media aritmetica $m = M(X) = \sum_{k=1}^{n} x_k/n$ risulta pari a m =353/20= 17,65 m^3, come è indicato nella tabella seguente in cui vengono evidenziati anche i valori degli scarti dalla media aritmetica $\{e_k = x_k - m\}$, la cui somma è nulla. Ordinando in ordine crescente i valori in tabella $\{x_k\}$ vengono evidenziati il valor minimo $x' = 5$ e il valor massimo $x'' = 34$ potendosi verificare che $x' < m < x'' \rightarrow 5 < 17,17 < 34$.

k	x_k	e_k	$x_{(k)}$	$e_{(k)}$
1	12	-5,65	5	-12,65
2	10	-7,65	7	-10,65
3	14	-3,65	8	-9,65
4	17	-0,65	10	-7,65
5	26	8,35	12	-5,65
6	15	-2,65	14	-3,65
7	16	-1,65	15	-2,65
8	5	-12,65	16	-1,65
9	28	10,35	16	-1,65
10	23	5,35	17	-0,65
11	16	-1,65	18	0,35
12	20	2,35	18	0,35
13	18	0,35	19	1,35
14	34	16,35	20	2,35
15	19	1,35	22	4,35

16	25	7,35	23	5,35
17	7	-10,65	25	7,35
18	18	0,35	26	8,35
19	22	4,35	28	10,35
20	8	-9,65	34	16,35
Σ	353	0	353	0

Esempio 7

Per lo stesso fenomeno, considerato in precedenza, si esegua il calcolo della media aritmetica sulla base dei dati raccolti in seriazione, come è riportato nell'esempio 5.

Considerando le frequenze assolute n_i e i valori centrali delle classi x_i, si ha come media aritmetica $m = \sum_{i=1}^{p} x_i n_i / n$ pari a $m = \frac{330}{20} = 16{,}5$ oppure, impiegando le frequenze relative f_i, si ottiene lo stesso risultato $m = \sum_{i=1}^{p} x_i f_i = 16{,}5$, valore che differisce, per motivi di approssimazione, da quello ottenuto nell'esempio 6.

I_i	x_i	n_i	$x_i n_i$	f_i	$x_i f_i$	e_i	$e_i f_i$
0 ⊣ 10	5	4	20	0,20	1	-11,5	-2,3
10 ⊣ 20	15	10	150	0,50	7,5	-1,5	-0,75
20 ⊣ 30	25	5	125	0,25	6,25	8,5	2,125
30 ⊣ 40	35	1	35	0,05	1,75	18,5	0,925
Σ		20	330	1,00	16,5		0

Esempio 8

Si consideri la tabella di seriazione riguardante il fenomeno, a caratteri discreti, presentato nell'esempio 4.

x_i	n_i	f_i	$x_i n_i$	$x_i f_i$
1	3	0,15	3	0,15
2	6	0,30	12	0,60
3	4	0,20	12	0,60
4	5	0,25	20	1,00
5	1	0,05	5	0,25
6	1	0,05	6	0,30
Σ	20	1,00	58	2,90

Il numero medio di locali per appartamento risulta pari $m = \sum_{i=1}^{p} \frac{x_i n_i}{n} = \frac{58}{20} = \sum_{i=1}^{p} x_i f_i = 2{,}9$.

10. Altri tipi di indici di posizione

Oltre alla media aritmetica vengono impiegati anche altri indicatori di posizione che si distinguono in:

- indici di posizione "analitici", ottenuti mediante operazioni algebriche sui dati come avviene per la media aritmetica;
- indici di posizione "non analitici" ottenuti mediante operazioni di "ordinamento" dei dati o l'individuazione dell'intensità che ha la massima frequenza semplice.

La media quadratica

Se la variabile statistica assume valori "non negativi" $X \geq 0 \rightarrow \{x_k \geq 0; k = 1,2,\cdots,n\}$ si definisce come "media quadratica" dei dati la funzione $m_{[2]}$

$$m_{[2]} = +\sqrt{M(X^2)} = \sqrt{\frac{\sum_{k=1}^{n} x_k^2}{n}} = \sqrt{\frac{\sum_{i=1}^{p} x_i^2 n_i}{n}} = \left[\sum_{i=1}^{p} x_i^2 f_i\right]^{1/2}$$

La media quadratica gode della proprietà di "*Cauchy*", ossia:

$$x' < m_{[2]} < x''$$

Se la variabile statistica non è degenere.

La media quadratica gode, inoltre, delle proprietà *"moltiplicativa"* e di*"monotonicità"*, ma non gode di quella di *"linearità"*, come è possibile dimostrare (tali dimostrazioni sono lasciate ai lettori data l'analogia con le proprietà della media aritmetica).

La media geometrica

Qualora la variabile statistica assuma valori solo "positivi" $X > 0 \rightarrow \{x_k > 0; k = 1,2,\cdots,n\}$ si definisce come "media geometrica" dei dati la funzione $m_{[0]}$

$$m_{[0]} = \left[\prod_{k=1}^{n} x_k\right]^{1/n} = \left[\prod_{i=1}^{p} {x_i}^{n_i}\right]^{1/n} = \prod_{i=1}^{p} {x_i}^{f_i}$$

Il logaritmo di $m_{[0]}$ risulta definito come media aritmetica della variabile $X > 0$ e quindi dei suoi valori:

$$\log\left(m_{[0]}\right) = M\left(\log(X)\right) = \frac{1}{n}\sum_{k=1}^{n} \log(x_k)$$

La media geometrica gode delle stesse proprietà della media quadratica, quindi tutte quelle della media aritmetica a esclusione di quella di essere un operatore lineare.

Per una variabile $X > 0$ e non degenere, le tre medie $m_{[0]}$, $m_{[1]} = m$, $m_{[2]}$ si presentano in ordine crescente:

$$x' < m_{[0]} < m_{[1]} = m < m_{[2]} < x''$$

A titolo di verifica si consideri il seguente esempio.

Esempio 9

Si riprendano i dati dell'esempio 8 e si determinino la media quadratica e geometrica oltre alla già mota media aritmetica $m = \sum_{i=1}^{p} x_i f_i = 2{,}9$.

x_i	f_i	$x_i f_i$	$\ln(x_i)$	$f_i \ln(x_i)$	x_i^2	$x_i^2 f_i$
1	0,15	0,15	0,0000	0,0000	1	0,15
2	0,30	0,60	0,6931	0,2079	4	1,20
3	0,20	0,60	1,0986	0,2197	9	1,80
4	0,25	1,00	1,3863	0,3466	16	4,00
5	0,05	0,25	1,6094	0,0805	25	1,25
6	0,05	0,30	1,7918	0,0896	36	1,80
Σ	1	2,90		0,9443		10,20

Per la media geometrica, impiegando i logaritmi in base *e*, si ha $\sum_{i=1}^{p} f_i \ln(x_i) = 0{,}9443$ da cui $m_{[0]} = e^{0{,}9443} = 2{,}57$; per la media quadratica essendo $\sum_{i=1}^{p} x_i^2 \, f_i = 10{,}20$ da cui $m_{[2]} = \sqrt{10{,}20} = 3{,}19$. Si verifica la proprietà di ordinamento

$$x' = 1 < m_{[0]} = 2{,}57 < m = 2{,}9 < m_{[2]} = 3{,}19 < x'' = 6$$

e si può dimostrare con semplicità la proprietà di ordinamento crescente tra le medie analitiche considerate nel caso semplice di $n = 2$. Siano $x_1 < x_2$ i valori osservati di una variabile statistica $X > 0$, risultando così:

$$x' = x_1;\ x'' = x_2;$$

$$m_{[0]} = (x_1 x_2)^{1/2};\ \ m_{[1]} = m = \frac{x_1 + x_2}{2};\ m_{[2]} = \left(\frac{x_1^2 + x_2^2}{2}\right)^{1/2}$$

da cui, elevando al quadrato, si ha

$$m_{[0]}^2 = (x_1 x_2);\ \ m^2 = \left(\frac{x_1 + x_2}{2}\right)^2 ;\ m_{[2]}^2 = \left(\frac{x_1^2 + x_2^2}{2}\right)$$

$$m^2 = \left(\frac{x_1 + x_2}{2}\right)^2 = \frac{x_1^2 + x_2^2 + 2x_1 x_2}{4} = \frac{1}{2}\left[\left(\left(\frac{x_1^2 + x_2^2}{2}\right) + (x_1 x_2)\right)\right]$$

Quindi m^2 è la media aritmetica di $m_{[2]}^2$ e $m_{[0]}^2$ ed è compresa tra i due valori

$$m^2 = \frac{m_{[0]}^2 + m_{[2]}^2}{2} \ \rightarrow\ \min\{m_{[0]}^2, m_{[2]}^2\} < m^2 < \max\{m_{[0]}^2, m_{[2]}^2\}$$

Essendo inoltre:

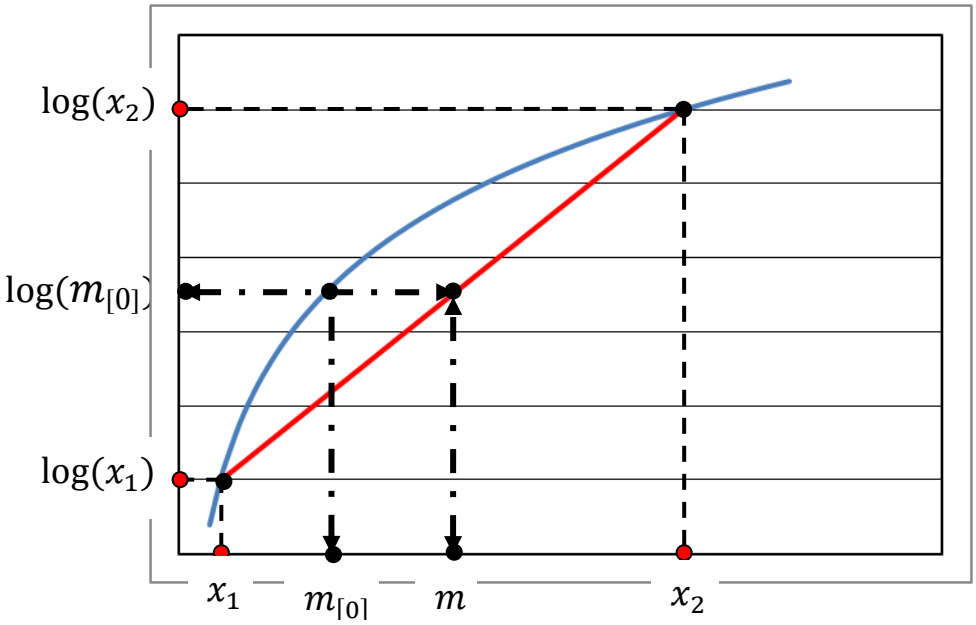
$$\log(m_{[0]}) = \frac{\log(x_1) + \log(x_2)}{2}$$

poiché la funzione logaritmo è monotona crescente con concavità verso il basso, come è evidenziato dalla figura, si ha

$$m_{[0]} < m$$

quindi si dimostra che:

$$x' = m_{[0]} < m < m_{[2]} = x''$$



Osservazione

Per variabili statistiche $X > 0$ viene costruita una classe di indici di posizione analitici detti "medie potenziate", ad esse appartengono le medie analitiche considerate finora, definite nel modo seguente.

Media potenziata di ordine "*r*"

$$m_{[r]} = +\left(M(X^r)\right)^{1/r} = +\left(\frac{\sum_{k=1}^{n} x_k^r}{n}\right)^{1/r} = +\left(\sum_{i=1}^{p} x_i^r f_i\right)^{1/r}$$

per $r \in \Re$.

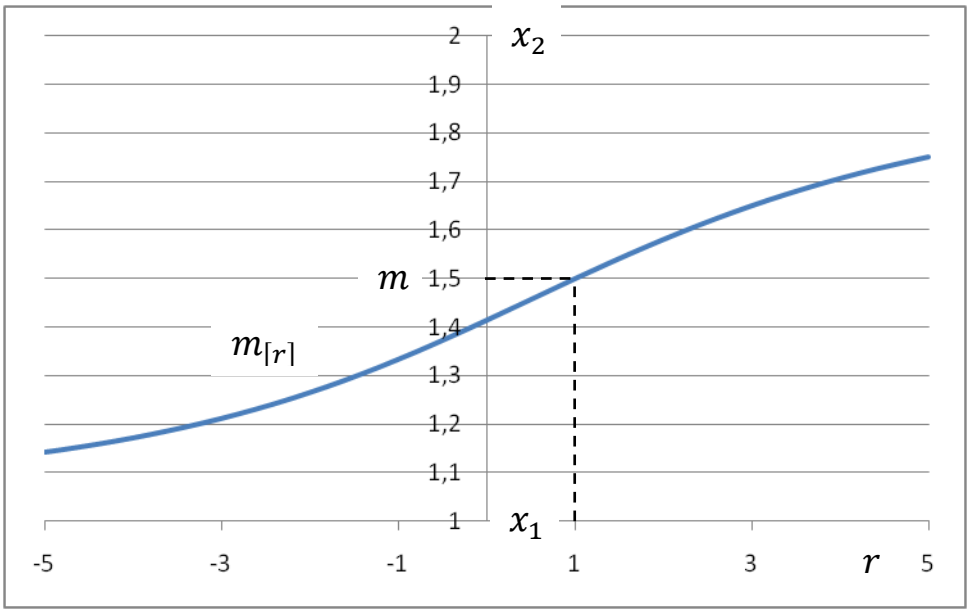
Si dimostra che:

$$\lim_{r\to-\infty} m_{[r]} = x' \quad \text{e} \quad \lim_{r\to\infty} m_{[r]} = x''$$

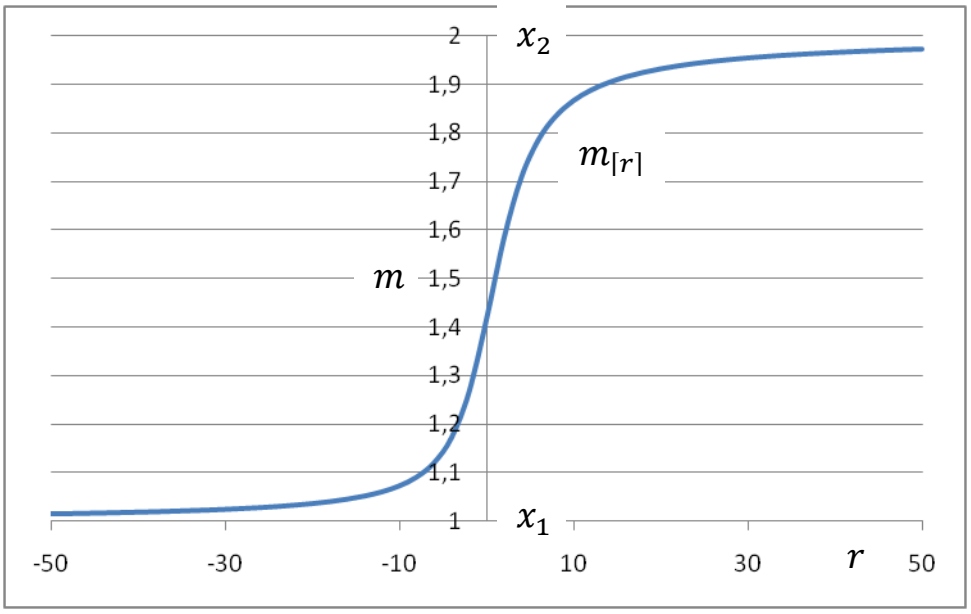
Inoltre per $r = 1$ abbiamo la media aritmetica, per $r = 2$ la media quadratica e per $\lim_{r\to 0} m_{[r]} = m_{[0]}$ la media geometrica.

Le medie potenziate di ordine "*r*" godono delle stesse delle altre medie presentate a esclusione della "linearità", proprietà quest'ultima che rimane propria della media aritmetica.

Al variare di r, la funzione $m_{[r]}$ è monotona crescente tendendo asintoticamente a x' per $r \to -\infty$ e a x'' per $r \to +\infty$, come è evidenziato dal grafico sottostante.



Andamento delle medie potenziate per $X \equiv \{x_1 = 1, x_2 = 2\}$



Andamento delle medie potenziate per $X \equiv \{x_1 = 1, x_2 = 2\}$

Si ricorda che la media potenziata di ordine $r = -1$ è detta "media armonica".

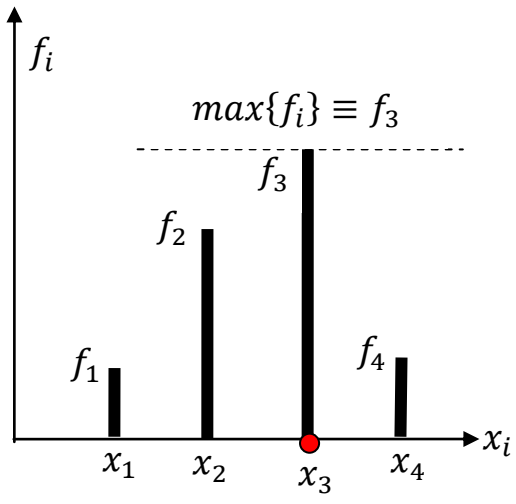
11. Moda o valore modale

E' un indice di posizione non analitico ed è definito 'intensità della variabile statistica che presenta la "massima" frequenza o densità di frequenza. Per individuare tale valore occorre, coerentemente alla definizione, disporre i dati in seriazione discreta o per classi di intervallo.

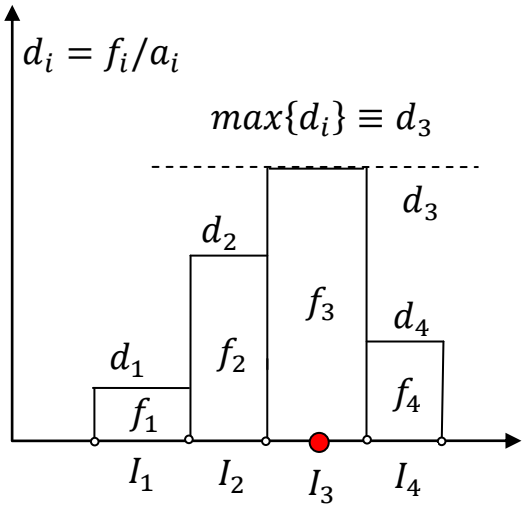
Indicata la moda con $\tilde{x} = Mo(X)$ e con $X \equiv \{x_i, ; f_i; i = 0,1, \cdots, p\}$ nel caso di valori discreti e $X \equiv \{I_i, ; d_i = f_i/a_i; i = 0,1, \cdots, p\}$ nel caso di classi di intervallo, si ha

$$\tilde{x} = x_l \quad \Rightarrow \; f_l = \max_i \{f_i; i = 0,1, \cdots, p\}$$
$$\tilde{x} \in I_l \quad \Rightarrow \; d_l = \max_i \{d_i; i = 0,1, \cdots, p\}$$

dove I_l è la "classe o l'intervallo modale" e in tal caso la moda si può scegliere coincidente con il valore centrale $\tilde{x} = x_l = (v_{l-1} + v_l)/2$.



$\tilde{x} \equiv x_3$



$\tilde{x} \in I_3 \rightarrow \tilde{x} \equiv x_3$

Osservazione

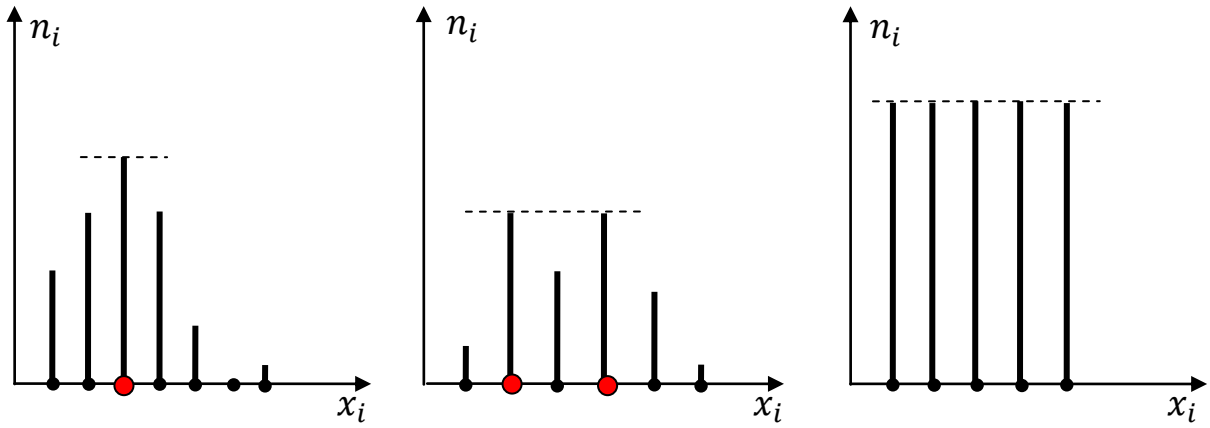
Il valore modale può non essere unico, si hanno infatti variabili statistiche: bimodali, trimodali, ecc. o amodali. Si vedano gli esempi sotto riportati riguardanti variabili discrete, con frequenze assolute.

Esempio 10

x_i	n_i
1	6
2	9
3	12
4	9
5	3
6	0
7	1
	40

x_i	n_i
5	2
10	9
15	6
20	9
25	5
30	1
	32

x_i	n_i
10	15
20	15
30	15
40	15
50	15
	75



La "moda" è certamente un indice di posizione in quanto soddisfa la proprietà di "Cauchy", infatti essendo una modalità del carattere o il valore centrale di una classe d'intervallo è sempre compreso tra il valore minimo e il valore massimo delle osservazioni

$$x' \leq \tilde{x} \leq x''$$

Per lo stesso motivo la "moda" gode della proprietà "moltiplicativa" e di quella "lineare", limitata al caso di trasformazione semplice $X \to Y = a + bX$:

$$\tilde{x} = Mo(X)$$

$$\tilde{y} = Mo(Y) = Mo(a + bX) = a + bMo(X) = a + b\tilde{x}$$

La proprietà di "monotonicità" non è sempre verificata, come si evidenzia nell'esempio riportato.

Esempio 11

$$X \Longrightarrow \{1,1,2,2,2,2,3,3,3,3,3,4\} \Longrightarrow \tilde{x} = 3$$

$$Y \Longrightarrow \{1,1,2,2,2,2,3,3,3,4,4,4\} \Longrightarrow \tilde{y} = 2$$

x_i	n_i
1	2
2	4
3	5
4	1
	12

y_i	n_i
1	2
2	4
3	3
4	3
	12

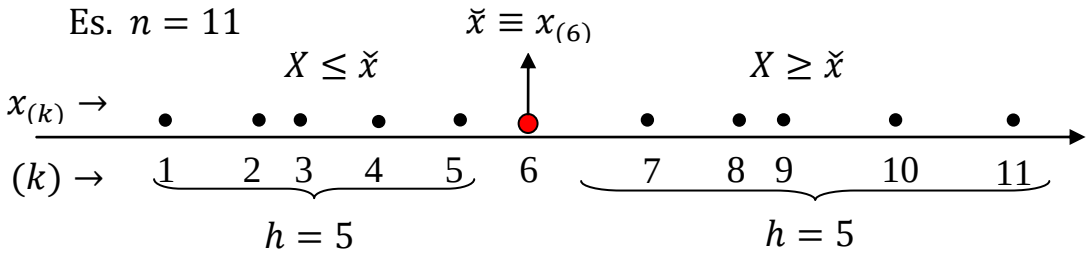
Pur essendo $\{y_k \geq x_k \,; k = 1,2,\cdots,n\}$ abbiamo $\tilde{y} < \tilde{x}$.

12. Mediana o valore mediano

E' un indice di posizione non analitico ed è definito 'intensità della variabile statistica che si colloca nel "posto centrale" nella sequenza ordinata dei dati. Per individuare tale valore occorre disporre i dati di una "serie" in forma ordinata, mentre per quelli in "seriazione, discreta o per classi di intervallo, l'ordine è individuato dalle frequenze cumulate.

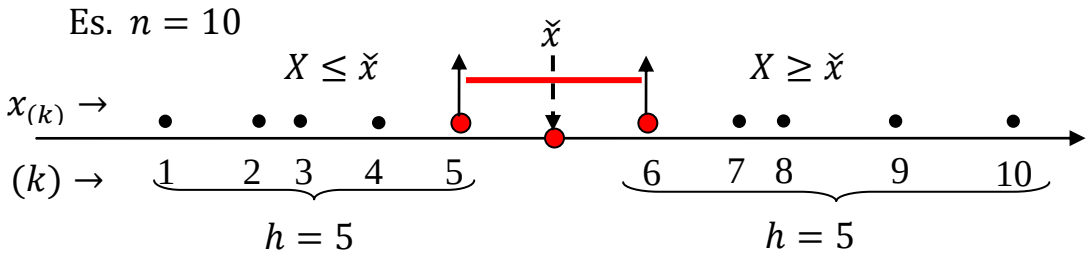
Indicata la mediana con $\breve{x} = Me(X)$ e con $X \equiv \{x_{(k)}; k = 0,1,\cdots,n\}$ la serie ordinata in ordine non decrescente, si ha

- se $n = 2h + 1$, (dispari), allora $\breve{x} = Me(X) = x_{(h+1)} = x_{\left(\frac{n+1}{2}\right)}$



- se $n = 2h$, (pari), allora esistono due unità "centrali", con valori differenti o coincidenti $x_{(h)} = x_{\left(\frac{n}{2}\right)}$ e $x_{(h+1)} = x_{\left(\frac{n}{2}+1\right)}$ e come mediana può considerarsi

$$\breve{x} = Me(X) \in \left[x_{\left(\frac{n}{2}\right)}, x_{\left(\frac{n}{2}+1\right)}\right] \rightarrow \breve{x} = Me(X) = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$



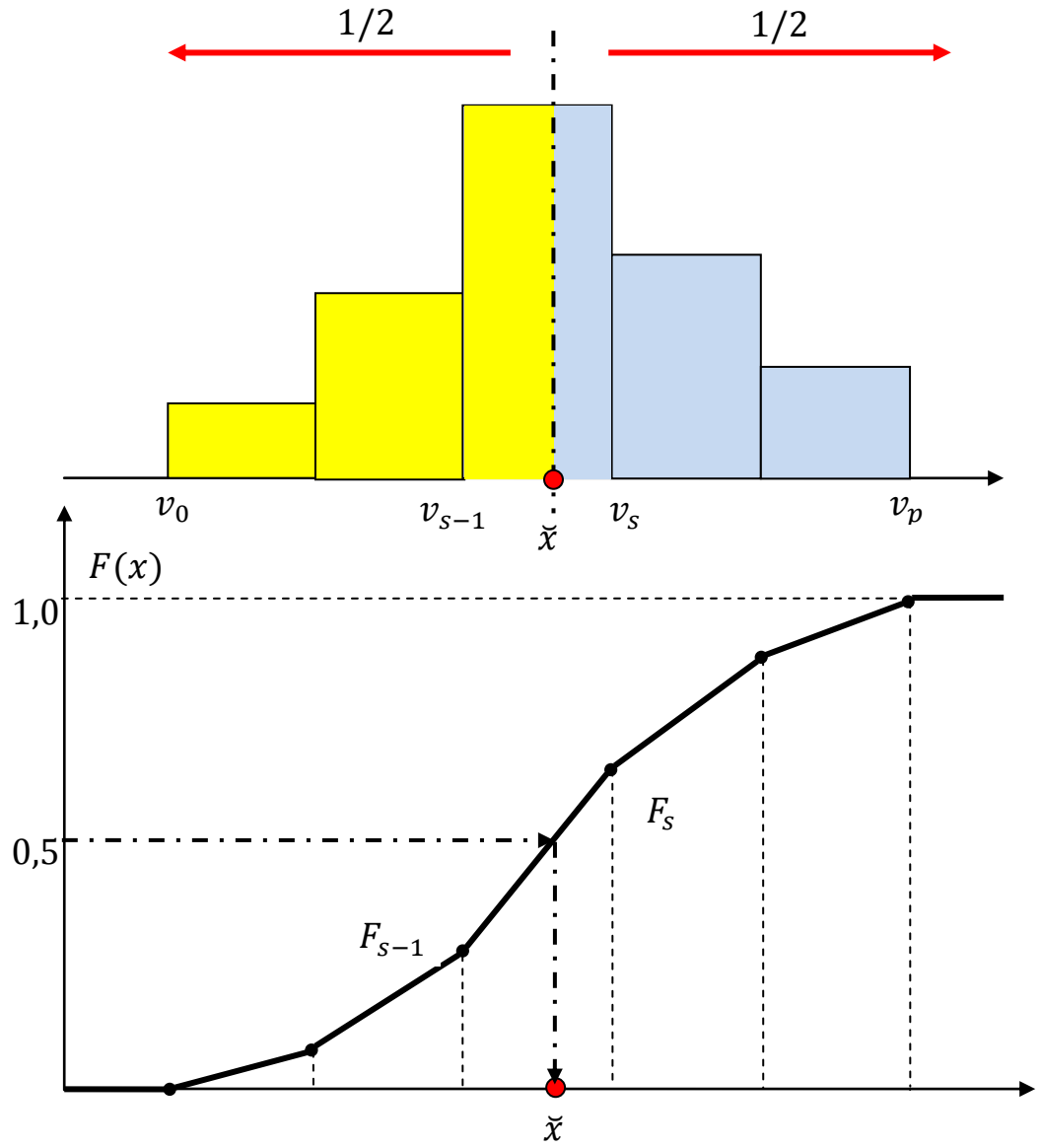
Come spesso accade nei casi concreti nell'analisi dei dati statistici la numerosità n è elevata e le osservazioni di X sono raccolte in una tabella per classi di intervallo, pertanto è opportuno determinare la mediana come il valore che separa i dati in due gruppi successivi di frequenza relativa pari a 0,5 (50%), determinando $\breve{x}$ dal grafico delle frequenze cumulate:

$$F(x) = \frac{1}{n}\#(X \leq x) \text{ per } x \in \Re$$

mediante la condizione:

$$\breve{x} \rightarrow F(\breve{x}) = \frac{1}{2}$$

Per tale motivo la mediana è detta anche valore 50% e indicata con $x_{0,5}$ o $x_{50\%}$.



Disponendo i dati in seriazione per classi di intervallo conviene prescindere da n e operare mediante le frequenze relative (semplici e cumulate) $\{I_i, f_i\} \equiv \{I_i, F_i\}$; il valore mediano si ottiene mediante una approssimazione lineare della funzione $F(x)$ individuando in un primo tempo l'intervallo mediano I_s :

$$I_s = (v_{s-1}, v_s] \Leftrightarrow F_{s-1} \leq 0{,}5 \leq F_s$$

poi la mediana:

$$\breve{x} = v_{s-1} + \frac{0{,}5 - F_{s-1}}{f_s}(v_s - v_{s-1})$$

La mediana, come può si può verificare, gode delle proprietà principali richieste agli indici di posizione: di "Cauchy",

"moltiplicativa", di "monotonicità" e similmente a quanto avviene all'operatore "moda" non gode della proprietà di "linearità generale" che invece è tipica della "media aritmetica".

La mediana essendo un valore centrale è poco sensibile a variazioni dei valori "estremi" (sia piccoli sia grandi) ed è stabile rispetto a errori di rilevazione di dati estremi (fondo scala nelle misurazioni analogiche).

La mediana presenta una proprietà riguardante gli "scarti" o "scostamenti" $E = X - \breve{x}$, ossia la componente aleatoria della variabile oggetto di studio.

- *La mediana $\breve{x}$ minimizza la media (o la somma) dei valori assoluti degli scarti da un generico indice $\theta \in \Re$.*

Sia

$$\psi(\theta) = M(|X - \theta|) = \frac{1}{n}\sum_{k=1}^{n} |x_k - \theta|$$

si ha

$$\min_{\theta} \psi(\theta) \Longrightarrow \theta = \breve{x}$$

Per dimostrare la proprietà si consideri inizialmente $n = 2$:

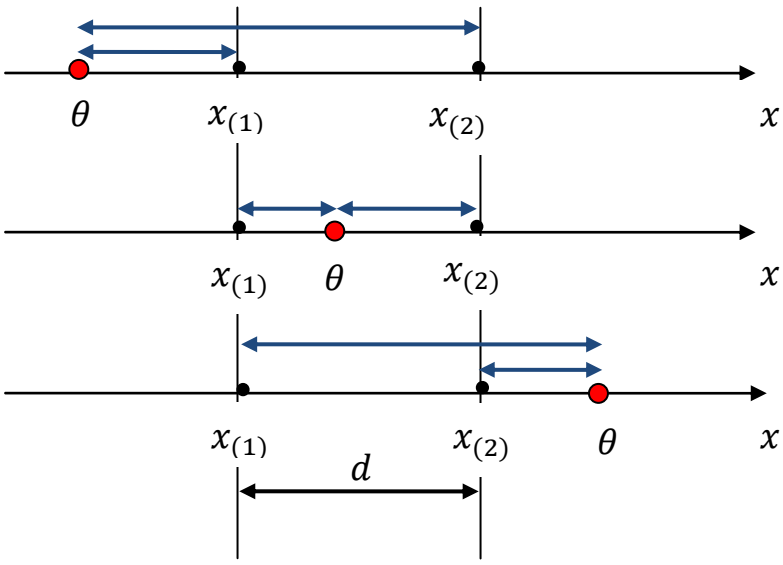
$$\{x_{(k)}\} = \{x_{(1)} < x_{(2)}\}$$

Si ha

$$2\psi(\theta) = \sum_{k=1}^{2} |x_k - \theta| = \begin{cases} 2|x_{(1)} - \theta| + d & \text{se } \theta < x_{(1)} \\ d & \text{se } x_{(1)} \leq \theta \leq x_{(2)} \\ 2|x_{(2)} - \theta| + d & \text{se } \theta > x_{(2)} \end{cases}$$

essendo $d = |x_{(1)} - x_{(2)}| = x_{(2)} - x_{(1)} > 0$ la distanza tra i due valori osservati.

La condizione di minimo di $\sum_{k=1}^{2}|x_k - \theta|$ si verifica per ogni valore θ compreso tra $x_{(1)}$ e $x_{(2)}$, estremi inclusi: $\min_\theta \psi(\theta) \Longrightarrow x_{(1)} \leq \theta \leq x_{(2)}$ e il valore di minimo risulta pari a $\min_\theta \psi(\theta) = d = x_{(2)} - x_{(1)}$.



Se $n = 2h$, si può generalizzare il risultato precedente riordinando gli scarti dell'espressione:

$$n\psi(\theta) = \sum_{k=1}^{n}|x_k - \theta| = \sum_{k=1}^{h}\left[\left|x_{(k)} - \theta\right| + \left|x_{(n-k+1)} - \theta\right|\right]$$

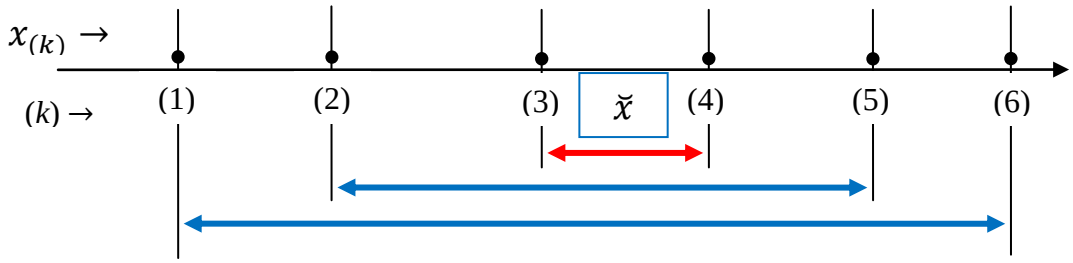
si individua, poi, una sequenza di h intervalli $J_k = \left[x_{(k)}, x_{(n-k+1)}\right]$, per $k = 1,2,\cdots,h = n/2$, contenuti uno nell'altro:

$$J_1 \supset J_2 \supset \cdots \supset J_h = \left[x_{(h)}, x_{(h+1)}\right]$$

per minimizzare l'espressione $n\psi(\theta)$ è sufficiente scegliere il valore di θ in:

$$\theta \in \bigcap_{k=1}^{h} J_k = J_h = \left[x_{(n/2)}, x_{(n/2+1)}\right]$$

che equivale a scegliere la mediana $\breve{x}$, come è illustrato in figura nel caso di $n = 6$, con valori distinti per semplicità.



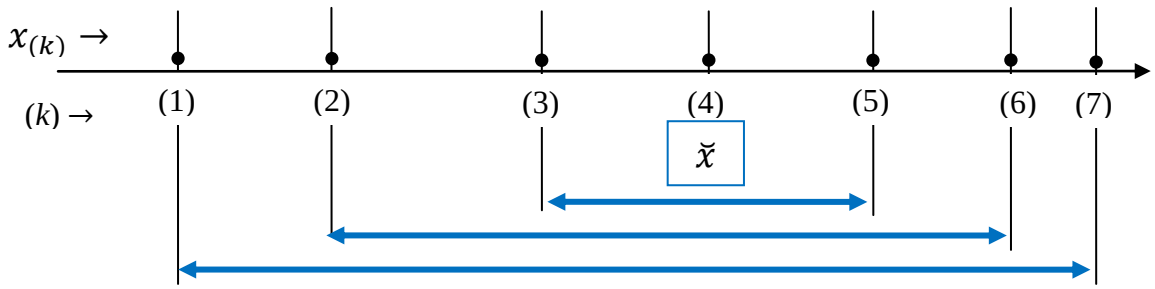
Se $n = 2h + 1$ il riordino degli scarti porta alla espressione:

$$\sum_{k=1}^{n} |x_k - \theta| = \sum_{k=1}^{h} \left[\left| x_{(k)} - \theta \right| + \left| x_{(n-k+1)} - \theta \right| \right] + \left| x_{(h+1)} - \theta \right|$$

per minimizzare la sommatoria presente al secondo membro è sufficiente scegliere θ contenuto nell'intervallo $J_h = \left[x_{(h)}, x_{(h+2)} \right]$, con $h = (n-1)/2$, essendo inoltre $x_{(h+1)} \in J_l$, per minimizzare la somma complessiva basta porre $\theta = x_{(h+1)} \rightarrow \left| x_{(h+1)} - \theta \right|$, risultando, quindi, θ pari alla mediana:

$$\min_{\theta} \psi(\theta) \Longrightarrow \theta = x_{(h+1)} = x_{\left(\frac{n+1}{2}\right)} = \breve{x}$$

In figura viene presentata la situazione per $n = 7$, con valori, per semplicità distinti.



Questa proprietà, propria della mediana, ha un ruolo analogo a quella della media aritmetica che è stata indicata come dei "minimi quadrati".

13. Valori quantili

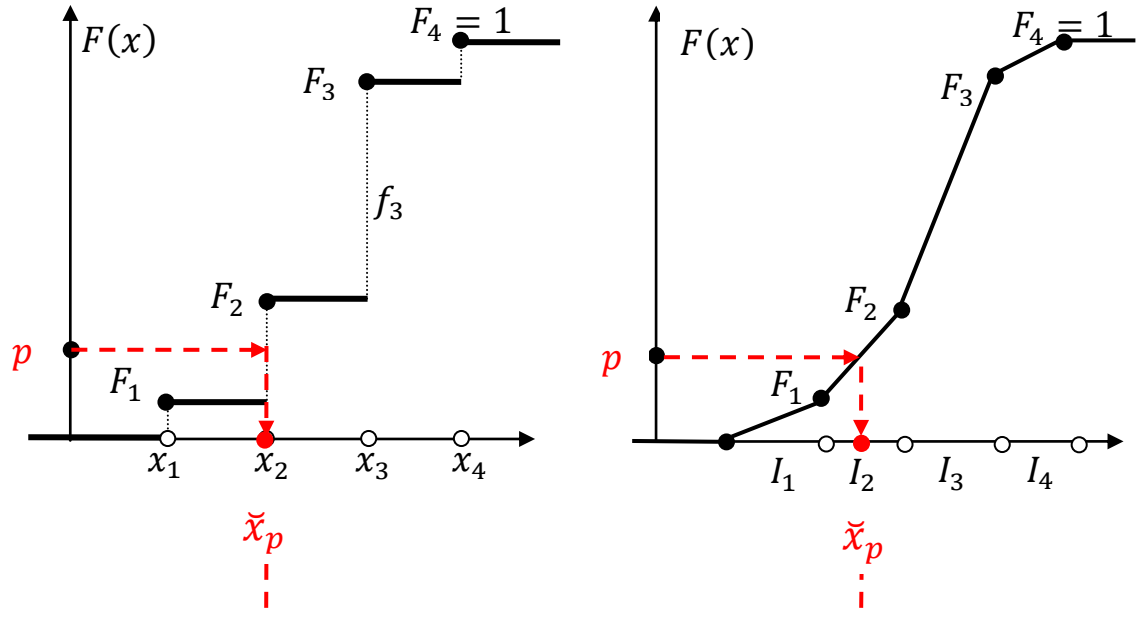
Se la mediana, come valore centrale sintetizza bene una variabile statistica osservata, per certi fenomeni può essere d'interesse costruire indici di posizione che rappresentino il valore che non è superato da una frazione di unità statistiche pari a p, con $0 < p < 1$ o, in termini percentuali, "punti percentili".

Tale valore è detto valore o punto "p-quantile" e indicato con $\breve{x}_p$. Il valore p-quantile, analogamente a quanto avviene per la mediana, che corrisponde al quantile per $p = 0{,}5$, si determina mediante le frequenze cumulate F_i e la funzione di ripartizione $F(x)$.

$$\breve{x}_p \rightarrow \; F\left(x = \breve{x}_p\right) = p$$

Al fine di eseguire confronti tra distribuzioni diverse, spesso si assumono valori percentili pari a: 5%, 10%, 20%, 50%, 80%, 90% e 95% (es.: carico di rottura di un materiale pari al 90%, livello di reddito di sussistenza di una popolazione al 5%).

In certe analisi si considerano i valori "quartili": 1° quartile che corrisponde a $p = 25\%$; 2° quartile che corrisponde a $p = 50\%$ (mediana); 3° quartile che corrisponde a $p = 75\%$.



Esempio 12

Si considerino i seguenti dati relativi agli stipendi mensili di 220 dipendenti di una azienda (in €) raccolti per classi di intervallo.

i	I_i	n_i	N_i	f_i	F_i
1	0 ⊣ 750	50	50	0,2273	0,2273
2	750 ⊣ 1000	75	125	0,3409	0,5682
3	1000 ⊣ 1300	60	185	0,2727	0,8409
4	1300 ⊣ 1500	20	205	0,0909	0,9318
5	1500 ⊣ 3000	15	220	0,0682	1,0000
		220		1,0000	

Si richiede di determinare la mediana e il punto 90° percentile, cioè il reddito che è superato dal 10% dei dipendenti.

Intervallo mediano

$$I_2 = (750, 1000] \quad \text{essendo} \quad 0{,}2273 < 1/2 \leq 0{,}5682$$

Valore mediano

$$\breve{x} = 750 + \frac{(0{,}5 - 0{,}2273)}{0{,}3409}(1000 - 750) = 950$$

Intervallo 90° percentile

$$I_4 = (1300, 1500] \quad \text{essendo} \quad 0{,}8409 < 0{,}9 \leq 0{,}9318$$

90° percentile

$$\breve{x}_{0,9} = 1300 + \frac{(0{,}9 - 0{,}8408)}{0{,}0909}(1500 - 1300) = 1430{,}25$$

14. La scelta degli indici di posizione

Disponendo di numerosi indici di posizione: media aritmetica, geometrica, quadratica, medie potenziate e ancora moda, mediana, quantili, ecc., spesso è ci si chiede quale sia opportuno applicare. Occorre tener presente, inizialmente, la presenza di elementi, quali:

- modalità con cui sono disponibili i dati;
- proprietà generali e specifiche dell'indice di posizione;
- grado di complessità delle elaborazioni richieste;

- capacità di “robustezza” o “sensibilità” alle variazioni dei dati;
- ruolo che l’indice ha nell’ambito dei fenomeni a cui i dati si riferiscono.

Sono stati proposti diversi criteri di scelta, fra questi si considerano i seguenti.

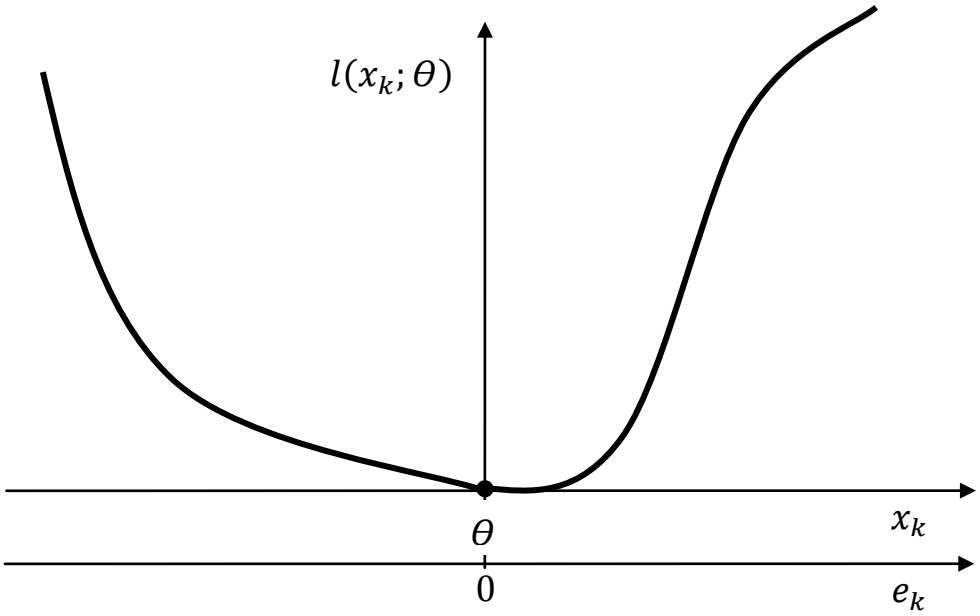
A. Minimizzazione della perdita di informazione “globale/media” che l’impiego di un solo valore in luogo di tutti i dati comporta. La scelta è condotta secondo un criterio di “minimo danno”.
B. Mantenimento di una condizione di “invarianza” nei confronti di una funzione complessiva dei dati. La scelta comporta l’individuazione di una”media obiettivo” (secondo Chisini).

15. Minimizzazione della funzione di perdita

Sia X una variabile statistica individuata da $\{x_k; k = 1,2,\cdots,n\}$ e sia θ un generico indice di posizione, indichiamo con $l(x_k;\theta) \geq 0$ la funzione che esprime l’entità della “perdita” di informazione qualora si sostituisca il dato reale x_k con il valore sintetico considerato θ, spesso misurato in termini economici e quindi di natura additiva.

$$l(x_k;\theta) = l(x_k - \theta) = l(e_k) \begin{matrix} \nearrow = 0 \text{ se } x_k = \theta \rightarrow e_k = 0 \\ \searrow > 0 \text{ se } x_k \neq \theta \rightarrow e_k \neq 0 \end{matrix}$$

per $k = 1,2,\cdots,n$ e dove $e_k = x_k - \theta$ è lo scarto o scostamento di x_k da θ.



Definitala funzione di "perdita complessiva" come:

$$L(X;\theta) = \sum_{k=1}^{n} l(x_k - \theta) = \sum_{i=1}^{p} l(x_i - \theta) n_i$$

e il valor medio:

$$\bar{L}(X;\theta) = \frac{L(X;\theta)}{n} = \sum_{i=1}^{p} l(x_i - \theta) f_i$$

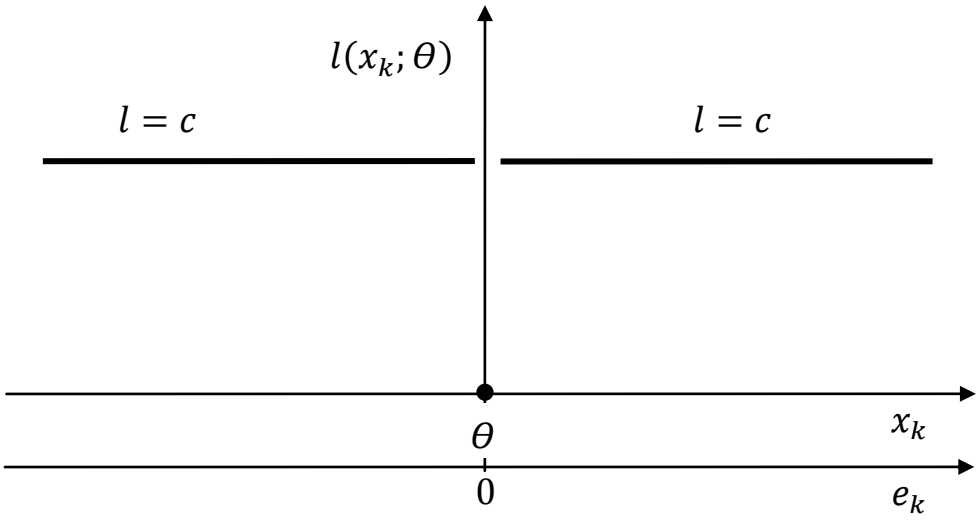
quale indice di posizione si sceglie θ in modo da minimizzare $L(X;\theta)$ o equivalentemente $\bar{L}(X;\theta)$:

$$\bar{\theta} = \arg\min_{\theta} L(X;\theta) \Leftrightarrow \bar{\theta} = \arg\min_{\theta} \bar{L}(X;\theta) \quad \text{per} \quad x' \leq \theta \leq x''$$

Si considerino le seguenti tre funzioni di perdita di largo impiego.

1. <u>Funzione costante</u>

Sia $l(x_k - \theta) = \begin{cases} 0 & \text{se } x_k = \theta \\ c = cost > 0 & \text{se } x_k \neq \theta \end{cases}$



Considerando $\bar{L}(X;\theta)=\sum_{i=1}^{p} l(x_i-\theta)f_i$ si ha

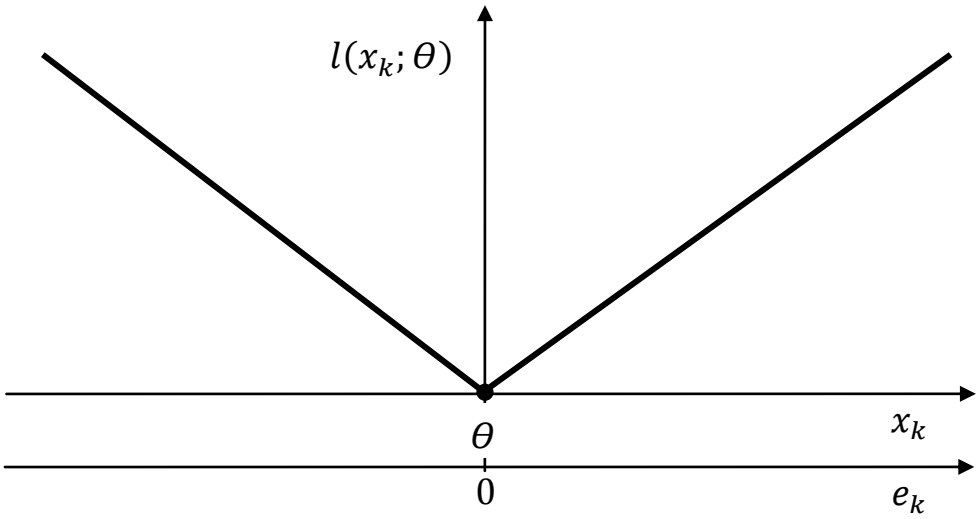
- se $\theta \neq x_i \ \forall\, i$, $\bar{L}(X;\theta)=\sum_{i=1}^{p} cf_i = c$ valore costante che non dipende da θ;
- se $\theta = x_s$ (con $x_i \neq \theta$ per $i=1,2,\cdots,s-1,s+1,\cdots,p$) $\bar{L}(X;\theta)=\sum_{i=1}^{s-1} cf_i + 0 + \sum_{i=s+1}^{p} cf_i = c\sum_{\substack{i=1\\ i\neq s}}^{p} f_i = c(1-f_s)$ valore dipendente da x_s.

Il valore θ che minimizza $\bar{L}(X;\theta)$ è quello che rende massima la frequenza f_s cioè il "valore modale" $\tilde{x}$, infatti:

$$\min_{\theta} \bar{L}(X;\theta) = c \min_{s}(1-f_s) \Longrightarrow \max_{s} f_s \Leftrightarrow x_s \equiv \tilde{x} \Longrightarrow \bar{\theta}=\tilde{x}$$

2. <u>Funzione lineare</u>

Sia $l(x_k-\theta)=c|x_k-\theta|$ con $c>0$

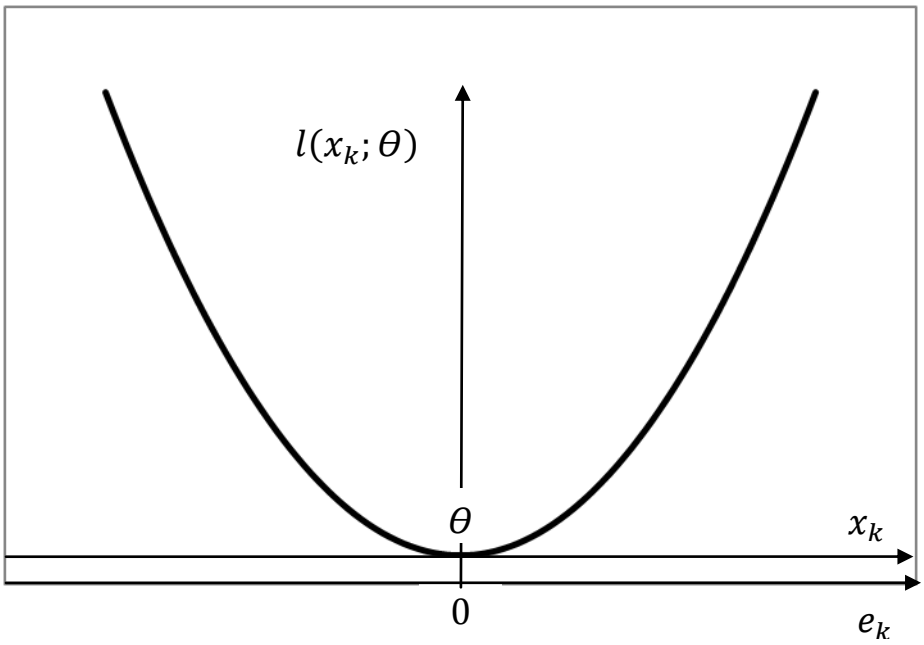


Per minimizzare $L(X;\theta) = c\sum_{k=1}^{n}|x_k - \theta|$ occorre minimizzare "la somma o la media dei valori assoluti degli scarti" e questa è una tipica proprietà della "mediana" $\tilde{x}$ della variabile *X*.

$$\min_{\theta} \bar{L}(X;\theta) = c \min_{\theta} \sum_{k=1}^{n} |x_k - \theta| \Longrightarrow \bar{\theta} = \tilde{x}$$

3. <u>Funzione quadratica</u>

Sia $l(x_k - \theta) = c(x_k - \theta)^2$ con $c > 0$



Occorre minimizzare $\bar{L}(X;\theta) = c\sum_{i=1}^{p}(x_i - \theta)^2 f_i$ ovvero minimizzare "la media o la somma dei quadrati degli scarti" e questa

è una proprietà caratteristica della "media aritmetica" m della variabile X.

$$\min_{\theta} \bar{L}(X;\theta) = c \min_{\theta} \sum_{i=1}^{p} (x_i - \theta)^2 f_i \Longrightarrow \bar{\theta} = m$$

16. Media "obiettivo" secondo Chisini

Nello studio di molti fenomeni, naturali, fisici ed economici, spesso esiste una funzione dei dati che ha una particolare rilevanza rispetto al tipo di indagine oggetto di interesse.

Sia data una variabile X con n intensità osservate $\{x_k; k = 1,2,\cdots,n\}$ e sia definibile una funzione "obiettivo" di interesse che congloba in sé il fenomeno allo studio $G = G(x_1, x_2, \cdots, x_k, \cdots, x_n)$, indichiamo con $\bar{\bar{\theta}}$ la "media obiettivo" cioè l'intensità che sostituita a ogni osservazione lascia inalterato – invariante – il valore globale di $G(\cdot)$

$$G = G(x_1, x_2, \cdots, x_k, \cdots, x_n) = G\left(\bar{\bar{\theta}}, \bar{\bar{\theta}}, \cdots, \bar{\bar{\theta}}, \cdots, \bar{\bar{\theta}}\right) = g(\bar{\bar{\theta}})$$

Se $g(\cdot)$ è una funzione invertibile si ottiene la "media obiettivo" $\bar{\bar{\theta}}$ come funzione dei valori $\{x_k; k = 1,2,\cdots,n\}$

$$\bar{\bar{\theta}} = g^{-1}[G(x_1, x_2, \cdots, x_k, \cdots, x_n)] = \bar{\bar{\theta}}(x_1, x_2, \cdots, x_k, \cdots, x_n) = \bar{\bar{\theta}}(X)$$

La funzione $\bar{\bar{\theta}}(\cdot)$ ha la struttura di un indice di posizione e deve rispettare la condizione propria di tali indici ossia la proprietà di Cauchy: $x' \leq \bar{\bar{\theta}} \leq x''$.

17. Principali tipi di "medie obiettivo"

Si distinguono due tipi di strutture di funzioni obiettivo, che rispettano la scambiabilità tra i dati: a) di natura "additiva"; b) di natura "moltiplicativa".

a) Struttura “additiva”

Sia

$$G = G(x_1, x_2, \cdots, x_k, \cdots, x_n) = \sum_{k=1}^{n} \gamma(x_k)$$

allora

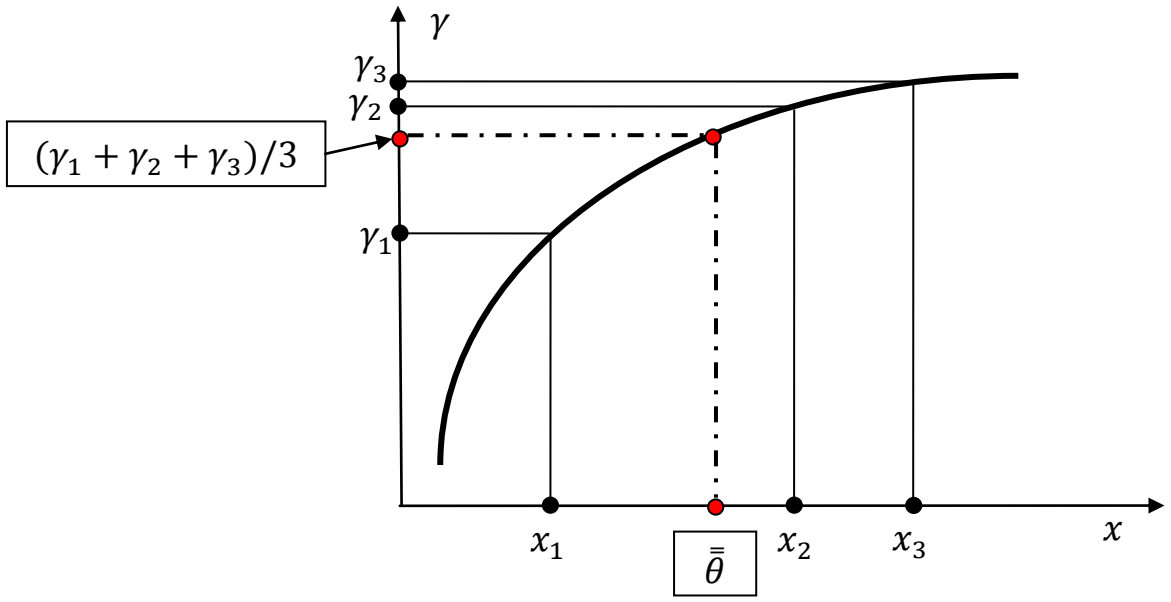
$$G(\bar{\bar{\theta}}, \bar{\bar{\theta}}, \cdots, \bar{\bar{\theta}}, \cdots, \bar{\bar{\theta}}) = \sum_{k=1}^{n} \gamma(\bar{\bar{\theta}}) = n\gamma(\bar{\bar{\theta}}) = g(\bar{\bar{\theta}})$$

dall’uguaglianza dei primi membri delle due equazioni precedenti si ha

$$\sum_{k=1}^{n} \gamma(x_k) = n\gamma(\bar{\bar{\theta}}) \Rightarrow \gamma(\bar{\bar{\theta}}) = \frac{1}{n}\sum_{k=1}^{n} \gamma(x_k)$$

$\gamma(\bar{\bar{\theta}})$ risulta pari alla media aritmetica dei valori $\gamma(x_k)$ e se $\gamma(\cdot)$ è una funzione invertibile la media obiettivo è pari a:

$$\bar{\bar{\theta}} = \gamma^{-1}[M\{\gamma(X)\}]$$



Se, ad esempio, $\gamma(x) = a\,x^r$, con $x > 0$ e $-\infty < r < \infty$, allora

$$a\,\bar{\bar{\theta}}^r = \frac{1}{n}\sum_{k=1}^{n} a\;x^r \Rightarrow\; \bar{\bar{\theta}}^r = \frac{1}{n}\sum_{k=1}^{n} x^r \Rightarrow \bar{\bar{\theta}} = m_{[r]}$$

La media "obiettivo" $\bar{\bar{\theta}}$ coincide con la media potenziata di ordine *r* $m_{[r]}$ e quindi, in particolare, si ha per le seguenti funzioni globali:

$$\begin{cases} G = x_1 + x_2 + \cdots + x_n \;\Rightarrow\; \bar{\bar{\theta}} = m_1 = m \\ G = x_1^2 + x_2^2 + \cdots + x_n^2 \;\Rightarrow\; \bar{\bar{\theta}} = m_2 \\ G = \dfrac{1}{x_1} + \dfrac{1}{x_2} + \cdots + \dfrac{1}{x_n} \;\Rightarrow\; \bar{\bar{\theta}} = m_{-1} \end{cases}$$

b) <u>Struttura "moltiplicativa"</u>

Sia

$$G = G(x_1, x_2, \cdots, x_k, \cdots, x_n) = \prod_{k=1}^{n} \gamma(x_k)$$

allora

$$G\left(\bar{\bar{\theta}}, \bar{\bar{\theta}}, \cdots, \bar{\bar{\theta}}, \cdots, \bar{\bar{\theta}}\right) = \prod_{k=1}^{n} \gamma\left(\bar{\bar{\theta}}\right) = \left[\gamma\left(\bar{\bar{\theta}}\right)\right]^n = g\left(\bar{\bar{\theta}}\right)$$

dall'uguaglianza dei primi membri delle due equazioni precedenti si ha

$$\prod_{k=1}^{n} \gamma(x_k) = \left[\gamma\left(\bar{\bar{\theta}}\right)\right]^n \Rightarrow \gamma\left(\bar{\bar{\theta}}\right) = \left[\prod_{k=1}^{n} \gamma(x_k)\right]^{1/n}$$

$\gamma\left(\bar{\bar{\theta}}\right)$ risulta pari alla media geometrica dei valori $\gamma(x_k)$ e se $\gamma(\cdot)$ è una funzione invertibile la media obiettivo è pari a:

$$\bar{\bar{\theta}} = \gamma^{-1}[M_0\{\gamma(X)\}]$$

Se $\gamma(x) \equiv x$ allora $G(x_1, x_2, \cdots, x_n) = x_1 \cdot x_2 \cdot \cdots \cdot x_n = \prod_{k=1}^{n} x_k$ ne consegue che

$$\bar{\bar{\theta}}^n = \prod_{k=1}^{n} x_k \Rightarrow \bar{\bar{\theta}} = \left[\prod_{k=1}^{n} x_k\right]^{1/n} = m_0$$

quindi la media "obiettivo" è la media geometrica dei valori di *X*.

Osservazioni

- La scelta della media più opportuna comporta il disporre o lo scegliere una particolare funzione globale "obiettivo";
- Per definire la funzione obiettivo occorre conoscere in modo non superficiale il fenomeno allo studio e lo scopo specifico della ricerca, potendosi solo così stabilire la "caratteristica invariante" da considerare.
- Si comprende, pertanto, che in molte situazioni, non disponendo di informazioni adeguate, si ricorra frequentemente all'impiego della "media aritmetica" e della "mediana", date le importanti proprietà di tali indici di posizione.

Esempio 13

Siano $X \equiv \{v_1, v_2, \cdots, v_n\}$ le osservazioni riguardanti la velocità di un mobile (Km/h) di cui si voglia conoscere la "velocità media". E' possibile considerare le due seguenti situazioni.

a) Le velocità sono state assunte da uno stesso mobile nel percorrere in successione uno stesso spazio (es.: giro di pista) *s*.

   La funzione obiettivo è "il tempo complessivo impiegato dal mobile" quindi la velocità media $\bar{\bar{v}}$ è quella che mantiene inalterato tale tempo complessivo

$$T = T(v_1, v_2, \cdots, v_n) = \frac{s}{v_1} + \frac{s}{v_2} + \cdots + \frac{s}{v_n} = s\sum_{k=1}^{n}\frac{1}{v_k}$$

$$T = T(\bar{\bar{v}}, \bar{\bar{v}}, \cdots \bar{\bar{v}}) = \frac{s}{\bar{\bar{v}}} + \frac{s}{\bar{\bar{v}}} + \cdots + \frac{s}{\bar{\bar{v}}} = s\frac{n}{\bar{\bar{v}}}$$

da cui si ottiene

$$s\sum_{k=1}^{n}\frac{1}{v_k}=s\frac{n}{\bar{\bar{v}}}\ \Rightarrow\ \bar{\bar{v}}=\frac{n}{\sum_{k=1}^{n}\frac{1}{v_k}}\Rightarrow\ \bar{\bar{v}}=m_{[-1]}$$

La velocità media è pari alla media "armonica" delle osservazioni.

b) Le velocità sono state mantenute dal mobile in tratti di percorso successivi per una durata temporale costante *t*.

La funzione obiettivo è "il percorso (spazio) complessivo effettuato dal mobile" quindi la velocità media $\bar{\bar{v}}$ è quella che mantiene inalterato tale spazio complessivo

$$S=S(v_1,v_2,\cdots,v_n)=v_1t+v_2t+\cdots+v_nt=t\sum_{k=1}^{n}v_k$$

$$S=S(\bar{\bar{v}},\bar{\bar{v}},\cdots\bar{\bar{v}})=\bar{\bar{v}}t+\bar{\bar{v}}t+\cdots+\bar{\bar{v}}t=tn\bar{\bar{v}}$$

da cui si ottiene

$$t\sum_{k=1}^{n}v_k=tn\bar{\bar{v}}\ \Rightarrow\ \bar{\bar{v}}=\frac{\sum_{k=1}^{n}v_k}{n}\Rightarrow\ \bar{\bar{v}}=m_{[1]}=m$$

La velocità media è pari alla media "aritmetica" delle osservazioni.

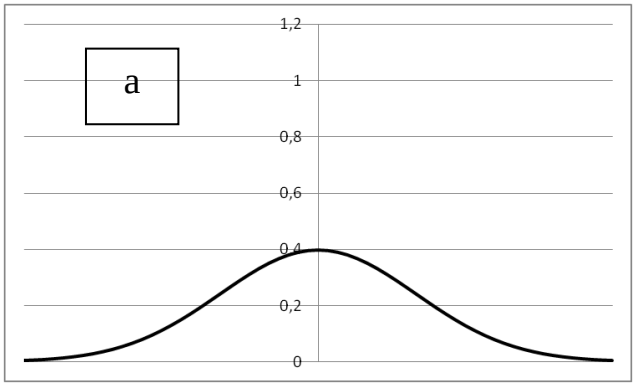
18. Concetto e misure di variabilità

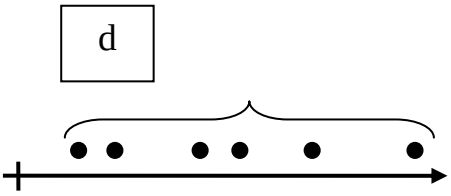
Oltre alla necessità di avere un'indicazione del livello medio dei valori presentati da un grandezza unidimensionale X è utile disporre di uno strumento sintetico che evidenzi l'entità della variabilità ossia della diversità di valore tra le osservazioni.

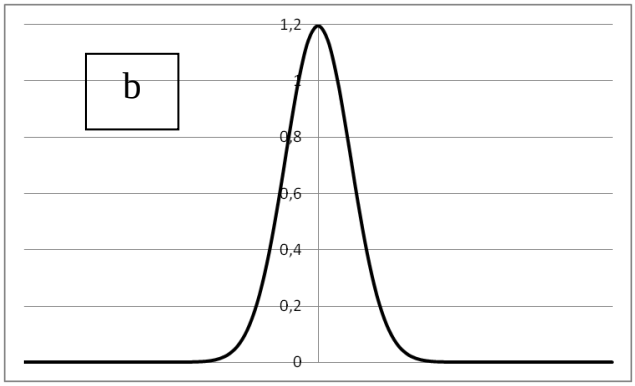
Senza entrare nel merito della domanda relativa a chi o a che cosa siano imputabili le differenze tra le osservazioni si assegna un ruolo generale di "variabilità accidentale" ai risultati ottenuti mediante le indagini "statistiche".

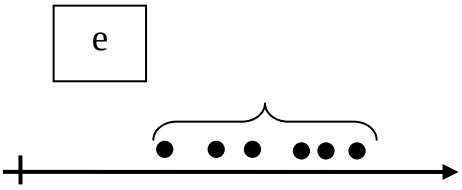
Come situazione di confronto generale si dispone di quella di "a-variabilità" corrispondente a una variabile statistica X avente tutte le osservazioni uguali di valore: $X \to \{x_k = \text{cost.}\ \forall\, k\}$ che sarà detta "variabile degenere".

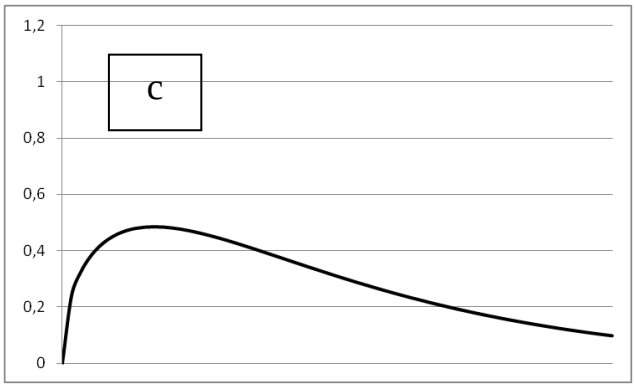
Sorge l'esigenza di misurare mediante opportuni indici la variabilità per confrontare differenti distribuzioni di variabili aventi o non aventi pari indice di posizione.

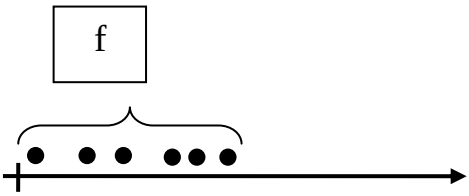












Come misura della variabilità di X si ricorre a "indici di dispersione" o a "indici di concentrazione", genericamente indicati con $\delta(X) = \delta(x_k; k = 1,2, \cdots, n) \geq 0$ che sintetizzano i dati mediante un valore non negativo. Qualora la X sia una variabile "degenere" $\delta(X)$ è identicamente nullo: $\delta(X) = 0$.

19. Tipologie di indici di dispersione

Come elementi base per misurare la dispersione, essendo X una grandezza quantitativa, si ricorre alle "distanze" in termini assoluti tra:

- *Ciascun valore x_k e un valore centrale θ, indice di posizione*

$$d_k = d(x_k;\, \theta) = |x_k - \theta| \geq 0 \quad \text{per} \quad k = 1,2,\cdots,n$$

disponendo di *n* valori. Come indice di posizione θ si impiega o la media aritmetica *m* o la mediana $\breve{x}$.

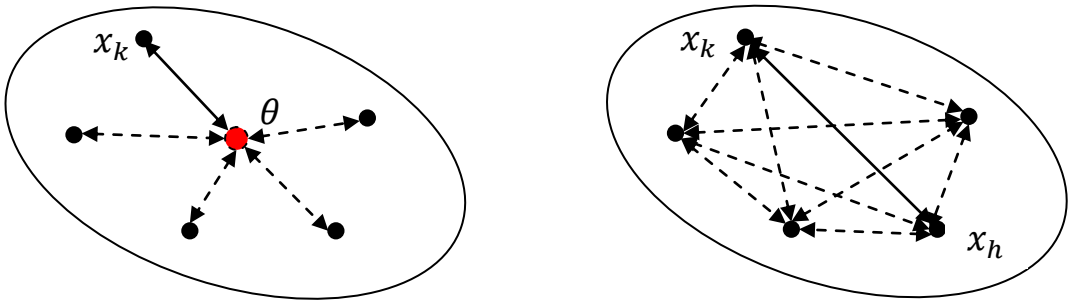
- *Ciascuna coppia di valori x_k e x_h*

$$d_{kh} = d(x_k, x_h) = |x_k - x_h| \geq 0 \quad \text{per} \quad k,h = 1,2,\cdots,n \ \text{con} \ k \neq h$$

disponendo di $n(n-1)$ valori.

Impiegando gli operatori sintetici è possibile ottenere misure di dispersione rispettivamente indicati come:

- *Indici di dispersione riferiti a un centro*;
- *Indici di dispersione globali.*



20. Principali indici di dispersione rispetto a un centro

I principali indici di dispersione rispetto a un centro si ottengono impiegando gli operatori "medie potenziate" di ordine $\boldsymbol{r = 1}$ **o** $\boldsymbol{r = 2}$ rispettivamente "media aritmetica" o "media quadratica" delle distanze $\boldsymbol{d_k}$ e come indice di posizione $\boldsymbol{\theta}$ si considera, rispettivamente, la mediana e la media aritmetica delle osservazioni.

$$r = 1, \theta = \breve{x} \Rightarrow {}_{\breve{x}}s_{[1]} = M\{|X - \breve{x}|\} = \frac{1}{n}\sum_{k=1}^{n}|x_k - \breve{x}|$$

$$r = 2, \theta = m \Rightarrow {}_{m}s_{[2]} = s = \sqrt{M\{(X-m)^2\}} = \left[\frac{1}{n}\sum_{k=1}^{n}(x_k - m)^2\right]^{1/2}$$

Tali indici sono detti "scostamenti medi assoluti" e in particolare s che è quello di più largo impiego essendo in concomitanza con la media aritmetica è detto "scarto quadratico medio (s.q.m)" o "standard deviation" o "écart type". Spesso al posto del simbolo s viene utilizzato il simbolo σ, se l'analisi è estesa all'intero universo del fenomeno allo studio.

Se la variabile X è degenere sia ${}_{\breve{x}}s_{[1]}$ sia s assumono il loro valor minimo pari a zero. Dal punto di vista dimensionale sia ${}_{\breve{x}}s_{[1]}$ sia s si esprimono con le stesse unità di misura delle osservazioni di X.

Frequentemente a fianco dello s.q.m. viene impiegato come indice di dispersione il suo quadrato s^2, detto "varianza di X ", che è la media aritmetica dei quadrati degli scarti dalla media aritmetica di X.

$$s^2 = Var\{X\} = M\{(X-m)^2\} = \frac{1}{n}\sum_{k=1}^{n}(x_k - m)^2 = \sum_{i=1}^{p}(x_i - m)^2 f_i$$

La somma dei quadrati degli scarti $\frac{1}{n}\sum_{k=1}^{n}(x_k - m)^2 = \frac{1}{n}\sum_{i=1}^{p}(x_i - m)^2 n_i$ è detta "devianza" e indicata con $Dev\{X\}$.

Spesso si utilizza l'operatore varianza $Var\{X\}$, le cui proprietà derivano dall'operatore media aritmetica $M\{X\}$.

Per il calcolo di s^2 e di s conviene impiegare la seguente relazione che non comporta il calcolo dei singoli scarti $x_k - m$:

$$s^2 = M\{(X-m)^2\} = M\{X^2 - 2mX + m^2\}$$
$$= M\{X^2\} - 2m\,M\{X\} + m^2 = M\{X^2\} - m^2$$

da cui la varianza di *X* può definirsi come media aritmetica dei quadrati di *X* meno il quadrato della media aritmetica di *X*.

Esempio 14

Riprendendo i dati riportati nell'esempio 10 si ottengono la varianza e lo s.q.m.

x_i	n_i	N_i	$x_i n_i$	$x_i - m$	$(x_i - m)^2$	$(x_i - m)^2 n_i$	x_i^2	$x_i^2 n_i$
1	6	6	6	-1,95	3,8025	22,815	1	6
2	9	15	18	-0,95	0,9025	8,1225	4	36
3	12	27	36	0,05	0,0025	0,0300	9	108
4	9	36	36	1,05	1,1025	9,9225	16	144
5	3	39	15	2,05	4,2025	12,6075	25	75
7	1	40	7	4,05	16,4025	16,4025	49	49
	40		118			69,9000		418

$m = 118/40 = 2{,}95$

$s^2 = 69{,}90/40 = 1{,}7475$ oppure

$s^2 = \frac{418}{40} - 2{,}95^2 = 10{,}45 - 8{,}7025 = 1{,}7475;$ $s = \sqrt{1{,}7475} = 1{,}3219$

21. Alcuni indici di dispersione “globali”

Tra i diversi indici di dispersione “globali”, che per costruzione si basano solo sulle distanze tra le osservazioni e quindi non dipendono dall’indice di posizione scelto, ci si limita a illustrare i seguenti tre che sono di frequente impiego per la loro semplicità.

- Il “campo di variazione” detto anche “gamma” o “range”

$$\Gamma = \max\{X\} - \min\{X\} = x'' - x'$$

  Γ è in generale maggiore di zero; si ha $\Gamma = 0$ solo se la X è “degenere”. In Γ, per definizione, è contenuto il 100% dei dati osservati.

- La “differenza interquartile”

$$Dq = \breve{x}_{0,75} - \breve{x}_{0,25}$$

  dove $\breve{x}_{0,25}$ è il 1° quartile e $\breve{x}_{0,75}$ è il 3° quartile della variabile X, oggetto di studio. $Dq \geq 0$ in particolare è pari a zero se la X è “degenere”. In Dq, per definizione, è contenuto il 50% dei dati osservati più centrali.

- La “differenza media assoluta di ordine $r = 1$”

$$\Delta = \frac{\sum_{k=1}^{n}\sum_{h=1}^{n} d_{kh}}{n(n-1)} = \frac{\sum_{k=1}^{n}\sum_{h=1}^{n}|x_k - x_h|}{n(n-1)}$$

  Le somme al numeratore, delle espressioni precedenti, dovrebbero limitarsi ai valori con $k \neq h$, ma risultando $d_{kk} = |x_k - x_k| = 0$ non occorre una tale precisazione. Δ è la media aritmetica di tutte le $n(n-1)$ distanze tra le osservazioni. $\Delta > 0$ ad esclusione del caso di variabile X “degenere”.

  Se i dati sono raccolti in seriazione, si ha

$$\Delta = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} d_{ij} n_i n_j}{n(n-1)} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} |x_i - x_j| n_i n_j}{n(n-1)}$$

o ancora, se si dispone delle sole frequenze relative

$$\Delta \cong \sum_{i=1}^{p}\sum_{j=1}^{p} d_{ij} f_i f_j = \sum_{i=1}^{p}\sum_{j=1}^{p} |x_i - x_j| f_i f_j$$

approssimazione valida tanto più quanto più *n* è elevato.

Esempio 15

Si consideri la seguente serie di dati, per $n = 6$:

$$\{x_k\} = \{5; 7; 12; 8; 10; 7\}$$

Conviene determinare le diverse distanze disponendo i dati in ordine non decrescente

$$\{x_{(k)}\} = \{5; 7; 7; 8; 10; 12\}$$

Organizzando i valori per il calcolo delle distanze in una tabella

d_{kh}		$x_{(k)}$						Σd_{kh}
		5	7	7	8	10	12	
$x_{(h)}$	5	[redacted]	2	2	3	5	7	19
	7	2	[redacted]	0	1	3	5	11
	7	2	0	[redacted]	1	3	5	11
	8	3	1	1	[redacted]	2	4	11
	10	5	3	3	2	[redacted]	2	15
	12	7	5	5	4	2	[redacted]	23
							$\Sigma d_{kh} =$	90

si ottiene

$$\Delta = (2 \times 45)/(6 \times 5) = 3{,}00$$

22. Indici di dispersione “assoluti” e “relativi”

Gli indici di dispersione finora considerati: ${}_{\tilde{x}}s_{[1]}$, ${}_{m}s_{[2]} = s$, Γ, Dq e Δ si presentano tutti con “dimensione” omogenea con quella con cui si esprimono i valori della variabile X, per questo motivo sono detti “assoluti”. Un cambiamento di “scala” dei valori osservati si ripercuote parimenti sull’entità di tali indici di dispersione come pure su quelli di posizione. Spesso la variabile oggetto di interesse presenta modalità quantitative misurate su “scala di rapporti” in cui, quindi, le modalità sono definite tutte positive o negative. E’ opportuno eliminare l’effetto dimensionale esprimendo la dispersione dei dati in termini “relativi” o “percentuali” in forma adimensionale.

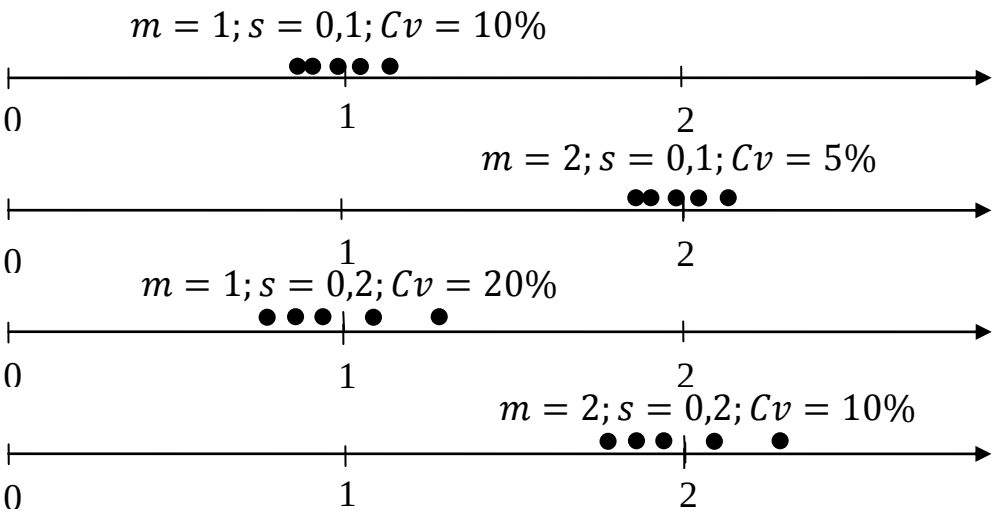
Una famiglia di indici di dispersione “relativi” si ottiene dividendo l’indice di dispersione assoluto per un indice di posizione.

L’indice di dispersione relativo più impiegato è il “coefficiente di variazione” Cv, dato da:

$$Cv = Cv(X) = s/m$$

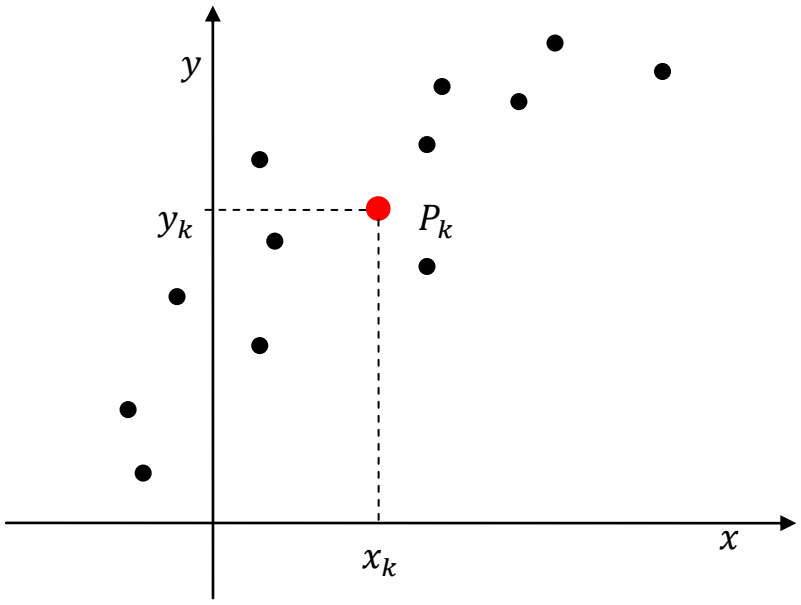
con $X > 0 \rightarrow \{x_k > 0 \; \forall k\}$, risultando $m > 0$ e $s \geq 0$ e conseguentemente:

$$Cv \geq 0 \quad \text{con} \quad Cv = 0 \text{ solo se } X \text{ è "degenere"}$$



23. Analisi descrittiva congiunta di due grandezze quantitative: la regressione polinomiale

Nei paragrafi precedenti si sono presentati i principali strumenti di studio descrittivo di una grandezza (variabile statistica) ma spesso si richiede di analizzare il comportamento congiunto di due grandezze, indicate con $\{X, Y\}$. In corrispondenza di ogni unità statistica e_k osservata, con $k = 1,2,\cdots,n$, si dispone di un punto $P_k \equiv (x_k, y_k)$. L'insieme dei punti P_k in un grafico cartesiano rappresenta l'intera popolazione che si concretizza come la "nube dei dati".



Tra le analisi descrittive di particolare interesse in questa sede ci si limita allo studio del legame funzionale (strutturale) tra la variabile *X* (esplicativa o regressore) e la variabile *Y* (dipendente o regressa) introducendo un modello $y = y(x)$ i cui parametri siano tali da accostare la funzione $y = y(x)$ ai punti dati, rispettando un appropriato criterio.

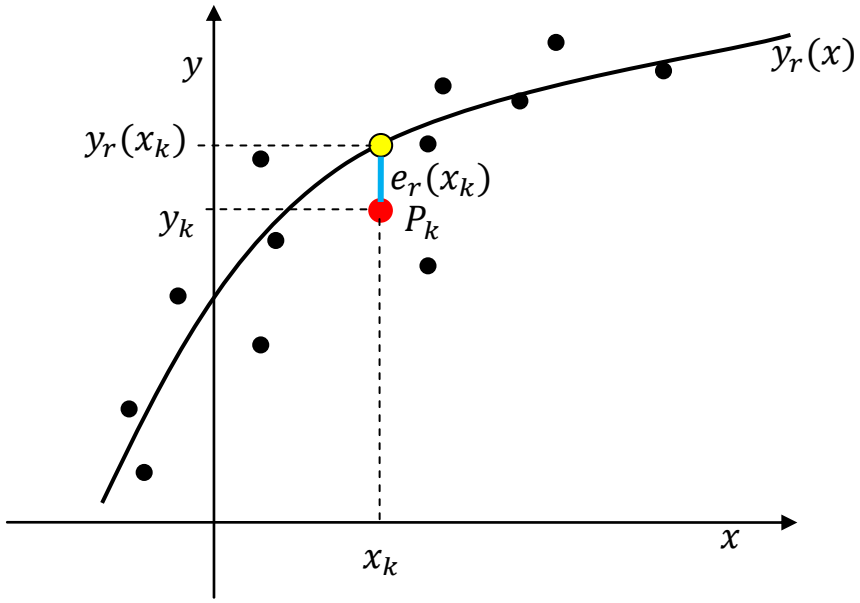
In relazione alle conoscenze "a priori" del fenomeno si sceglie la funzione $y = y(x)$; in particolare, spesso si adotta un polinomio in *x* di grado *r* :

$$y_r(x) = \sum_{s=0}^{r} a_{s,r} x^s \quad \text{con} \quad x \in \Re \quad \text{e} \quad r = 0, 1, 2, \cdots$$

Esprimendo la variabile regressa Y come $Y = Y_r + E_r = y_r(X) + E_r$ che evidenzia le due componenti: strutturale e casuale, si ha

$$E_r = Y - Y_r = Y - y_r(X)$$

La componente casuale dipende, oltre che dai dati $(X, Y) \rightarrow \{(x_k, y_k); k = 1,2, \cdots, n\}$, anche da r e dai parametri $\{a_{s,r}; s = 0,1, \cdots, r\}$.



La scelta dei valori da assumere per $\{a_{s,r}, s = 0,1, \cdots, r\}$ è demandata alla minimizzazione di una funzione di perdita media di tipo quadratico che corrisponde al criterio di accostamento dei "minimi quadrati":

$$\min_{a_{s,r}; s=0,1,\cdots,r} \Psi\left(a_{s,r}, s = 0,1, \cdots, r; (X, Y)\right) = \min_{a_{s,r}; s=0,1,\cdots,r} M(E_r^2)$$

$$\min_{a_{s,r}; s=0,1,\cdots,r} \Psi = \min_{a_{s,r}; s=0,1,\cdots,r} M(E_r^2) = \min_{a_{s,r}; s=0,1,\cdots,r} M([Y - y_r(X)]^2)$$

$$\min_{a_{s,r};s=0,1,\cdots,r} \Psi = \min_{a_{s,r};s=0,1,\cdots,r} M\left(\left[Y - \sum_{s=0}^{r} a_{s,r}X^s\right]^2\right)$$

Essendo $M(\cdot)$ un operatore lineare e la funzione Ψ continua e derivabile rispetto ai parametri, la condizione di minimo è soddisfatta dall'uguaglianza a zero delle derivate parziali di Ψ rispetto ai parametri $a_{h,r}$ per $h = 0,1,\cdots,r$

$$\frac{\partial \Psi}{\partial a_{h,r}} = (-2)M(E_r X^h) = 0 \Rightarrow M(E_r X^h) = 0 \text{ per } h = 0,1,\cdots,r$$

$$M\left(\left[Y - \sum_{s=0}^{r} a_{s,r}X^s\right]X^h\right) = 0 \text{ per } h = 0,1,\cdots,r$$

$$\sum_{s=0}^{r} a_{s,r}M\{X^{s+h}\} = M\{YX^h\} \text{ per } h = 0,1,\cdots,r$$

Si ottiene, così, un sistema lineare di $(r+1)$ equazioni in $(r+1)$ incognite, dei parametri $a_{h,r}$ per $h = 0,1,\cdots,r$, dove la "matrice dei coefficienti" è data da medie delle potenze di *X*, mentre il vettore dei "termini noti" è dato da medie di *Y* per potenze di *X*, che si calcolano dai dati osservati $\{(x_k, y_k); k = 1,2,\cdots,n\}$.

$$m_{h0} = M\{X^h\} = \frac{1}{n}\sum_{k=1}^{n} x_k^h \quad \text{per} \quad h = 0,1,\cdots 2r$$

$$m_{h1} = M\{X^h Y\} = \frac{1}{n}\sum_{k=1}^{n} x_k^h y_k \quad \text{per} \quad h = 0,1,\cdots r$$

Gli elementi della matrice dei coefficienti e del vettore dei termini noti fanno parte della classe dei "momenti" (dall'origine) della variabile bidimensionale $\{X, Y\}$, si veda per maggiori dettagli il Paragrafo 25.

$$m_{hl} = M\{X^h Y^l\} = \frac{1}{n}\sum_{k=1}^{n} x_k^h y_k^l \quad \text{per} \quad h, l = 0,1,2,\cdots$$

Per quanto riguarda la scelta del grado r del polinomio per motivi legati alla "parsimonia scientifica" sarà un valore possibilmente piccolo e certamente $r < n$.

Il sistema lineare di equazioni simultanee (equazioni normali) si presenta come:

$$\begin{cases} a_{0,r} & +m_{10}a_{1,r} & \cdots & +m_{r0}a_{r,r} & =m_{01} \\ m_{10}a_{0,r} & +m_{20}a_{1,r} & \cdots+m_{(r+1)0}a_{r,r} & & =m_{11} \\ \cdots & \cdots & \cdots & \cdots & \cdots \cdots \\ m_{r0}a_{0,r} & +m_{(r+1)0}a_{1,r} & \cdots+m_{(2r)\cdots 0}a_{r,r} & & =m_{r1} \end{cases}$$

Risolto il quale, si ottengono i valori dei parametri del modello polinomiale che rispettano il criterio di accostamento, specificatamente indicati: $\dot{a}_{0,r}, \dot{a}_{1,r}, \cdots, \dot{a}_{r,r}$. Disponendo di tali parametri è possibile definire il modello polinomiale

$$\dot{y}_r(x) = \sum_{s=0}^{r} \dot{a}_{s,r}x^s \ = \dot{a}_{0,r} + \dot{a}_{1,r}x + \cdots + \dot{a}_{r,r}x^r$$

in particolare, determinare i valori della variabile Y corrispondenti alle osservazioni di X

$$\{X, \dot{y}_r(X)\} \equiv \left\{x_k, \dot{y}_k = \sum_{s=0}^{r} \dot{a}_{s,r}x_k^s \ = \dot{a}_{0,r} + \dot{a}_{1,r}x_k + \cdots + \dot{a}_{r,r}x_k^r\right\}$$

e i valori della componente accidentale $\dot{E}_r \equiv Y - \dot{y}_r(X) \rightarrow \left\{\dot{e}_k = y_k - \sum_{s=0}^{r} \dot{a}_{s,r}x_k^s \ = y_k - (\dot{a}_{0,r} + \dot{a}_{1,r}x_k + \cdots + \dot{a}_{r,r}x_k^r)\right\}$.

La media aritmetica di $\dot{E}_r$, dalla prima equazione del sistema, è pari a zero: $M\left(\dot{E}_r\right) = 0$.

Come misura dell'accostamento si impiega la varianza dei "residui" $Var\left(\dot{E}_r\right) = M\left(\dot{E}_r^2\right) = \frac{1}{n}\sum_{k=1}^{n}(y_k - \dot{y}_k)^2 \leq Var\ (Y)$

da cui si ottiene un indice "standardizzato", che è detto "indice di determinazione"

$$R_r^2 = 1 - \frac{Var\left(\dot{E}_r\right)}{Var\ (Y)} \quad \text{con} \quad 0 \leq R_r^2 \leq 1$$

Oltre alla varianza dei residui si considera anche la varianza "spiegata" che misura la variabilità dei valori ottenuti dal modello $\dot{y}_k$

$$Var\{\dot{y}_r(X)\} = M\{[\dot{y}_r(X)]^2\} = \frac{1}{n}\sum_{k=1}^{n}(\dot{y}_k - M(Y))^2 \leq Var\ (Y)$$

Potendosi dimostrare che

$$Var\left\{\dot{Y}_r = \dot{y}_r(X)\right\} + Var\left(\dot{E}_r = Y - \dot{Y}_r\right) = Var\ (Y)$$

Tale identità è nota come "analisi o scomposizione della varianza" ed evidenzia come la varianza totale di Y sia pari alla somma della varianza spiegata dal modello più la corrispondente varianza residua, per ogni grado del modello polinomiale.

L'indice di determinazione è dato anche da:

$$R_r^2 = \frac{Var\left\{\dot{Y}_r\right\}}{Var\ (Y)} \quad \text{con} \quad 0 \leq R_r^2 \leq 1$$

Esempio 16

Si consideri, $r = 0, r = 1, r = 2$.

- Per $r = 0$

  Si ha: $y_0(X) = a_{0,0}$ (valore costante)

$$M(E_0) = 0 \quad \rightarrow \quad M\left(Y - a_{0,0}\right) = 0$$

$$\dot{a}_{0,0} = M(Y) = m_{h1}$$

$$Var\left(\dot{E}_0\right) = Var\ (Y) \rightarrow Var\left\{\dot{Y}_0\right\} = 0 \rightarrow R_0^2 = 0$$

- Per $r = 1$

Si ha: $y_1(X) = a_{0,1} + a_{1,1}X$ (funzione rettilinea)

$$\begin{cases} M(E_1) = 0 \\ M(E_1X) = 0 \end{cases} \rightarrow \begin{cases} M\big(Y - (a_{0,1} + a_{1,1}X)\big) = 0 \\ M\big(XY - (a_{0,1}X + a_{1,1}X^2)\big) = 0 \end{cases}$$

$$\begin{cases} M\big(a_{0,1} + a_{1,1}X\big) = M(Y) \\ M\big(a_{0,1}X + a_{1,1}X^2\big) = M(XY) \end{cases}$$

$$\begin{cases} a_{0,1} + a_{1,1}M(X) = M(Y) \\ a_{0,1}M(X) + a_{1,1}M(X^2) = M(XY) \end{cases} \rightarrow \begin{cases} a_{0,1} + m_{10}a_{1,1} = m_{01} \\ m_{10}a_{0,1} + m_{20}a_{1,1} = m_{11} \end{cases}$$

Se il rango della matrice dei coefficienti è pieno si determinano i parametri $\dot{a}_{0,1}$ e $\dot{a}_{1,1}$ come soluzioni del sistema.

- <u>Per $r = 2$</u>

  Si ha: $y_1(X) = a_{0,2} + a_{1,2}X + a_{1,2}X^2$ (funzione parabolica)

$$\begin{cases} M(E_2) = 0 \\ M(E_2X) = 0 \\ M(E_2X^2) = 0 \end{cases} \rightarrow \begin{cases} M\big(Y - (a_{0,2} + a_{1,2}X + a_{2,2}X^2)\big) = 0 \\ M\big(XY - (a_{0,2}X + a_{1,2}X^2 + a_{1,2}X^3)\big) = 0 \\ M\big(X^2Y - (a_{0,2}X^2 + a_{1,2}X^3 + a_{1,2}X^4)\big) = 0 \end{cases}$$

$$\begin{cases} M\big(a_{0,2} + a_{1,2}X + a_{2,2}X^2\big) = M(Y) \\ M\big(a_{0,2}X + a_{1,2}X^2 + a_{2,2}X^3\big) = M(XY) \\ M\big(a_{0,2}X^2 + a_{1,2}X^3 + a_{2,2}X^4\big) = M(X^2Y) \end{cases}$$

$$\begin{cases} a_{0,2} + a_{1,2}M(X) + a_{2,2}M(X^2) = M(Y) \\ a_{0,2}M(X) + a_{1,2}M(X^2) + a_{2,2}M(X^3) = M(XY) \\ a_{0,2}M(X^2) + a_{1,2}M(X^3) + a_{2,2}M(X^4) = M(X^2Y) \end{cases} \rightarrow$$

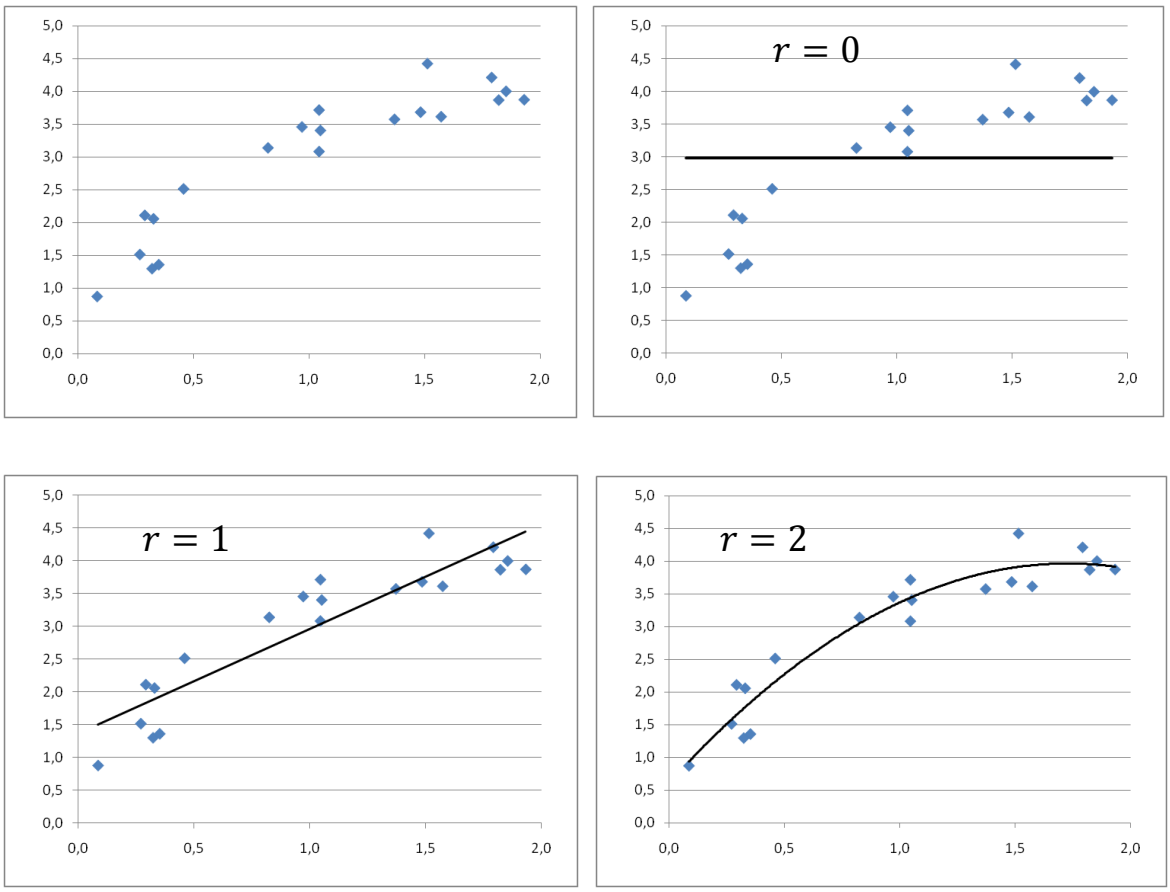
$$\begin{cases} a_{0,2} + m_{10}a_{1,2} + m_{20}a_{2,2} = m_{01} \\ m_{10}a_{0,2} + m_{20}a_{1,2} + m_{30}a_{2,2} = m_{11} \\ m_{20}a_{0,2} + m_{30}a_{1,2} + m_{40}a_{2,2} = m_{21} \end{cases}$$

Se il rango della matrice dei coefficienti è pieno si determinano i parametri $\dot{a}_{0,2}$, $\dot{a}_{1,2}$ e $\dot{a}_{2,2}$ come soluzioni del sistema.

Esempio 17

Si considerino le seguenti 20 rilevazioni riguardanti lo studio dell'intensità di capo magnetico (Y) al variare della corrente elettrica (X) in un solenoide, ottenute in un laboratorio. Si desideri determinare il legame funzionale tra le due grandezze considerando modelli polinomiali di grado $r = 0, 1, \text{e } 2$.

k	x_k	y_k	x_k^2	x_k^3	x_k^4	$x_k y_k$	$x_k^2 y_k$	$\dot{y}_1(x_k)$	$\dot{y}_2(x_k)$
1	1,93	3,87	3,734	7,216	13,944	7,476	14,446	4,437	3,915
2	0,46	2,51	0,211	0,097	0,044	1,152	0,529	2,094	2,158
3	1,79	4,21	3,210	5,751	10,304	7,538	13,505	4,213	3,957
4	1,05	3,08	1,092	1,142	1,193	3,219	3,365	3,026	3,440
5	1,05	3,40	1,104	1,161	1,220	3,574	3,756	3,036	3,449
6	1,48	3,68	2,203	3,269	4,852	5,460	8,103	3,724	3,895
7	1,82	3,86	3,321	6,051	11,026	7,039	12,828	4,262	3,952
8	1,51	4,42	2,292	3,469	5,252	6,688	10,125	3,772	3,911
9	0,32	1,30	0,104	0,034	0,011	0,418	0,135	1,878	1,749
10	1,05	3,71	1,093	1,142	1,194	3,880	4,055	3,027	3,440
11	0,09	0,87	0,007	0,001	0,000	0,074	0,006	1,500	0,939
12	0,97	3,45	0,943	0,916	0,889	3,354	3,257	2,909	3,320
13	0,29	2,11	0,085	0,025	0,007	0,614	0,179	1,828	1,650
14	0,33	2,06	0,108	0,036	0,012	0,676	0,222	1,888	1,769
15	0,27	1,51	0,073	0,020	0,005	0,408	0,110	1,794	1,580
16	1,85	4,00	3,437	6,372	11,814	7,410	13,739	4,312	3,944
17	0,35	1,36	0,124	0,043	0,015	0,477	0,168	1,923	1,839
18	1,37	3,57	1,881	2,580	3,539	4,896	6,715	3,545	3,820
19	0,82	3,14	0,679	0,560	0,461	2,585	2,130	2,675	3,047
n=20	1,57	3,61	2,475	3,894	6,126	5,681	8,938	3,866	3,935
$\boldsymbol{\Sigma}$	20,39	59,71	28,18	43,78	71,91	72,62	106,31		
$\boldsymbol{\Sigma}/n$	1,019	2,985	1,409	2,189	3,595	3,631	5,316		



- Per $r = 0$

Si ha: $\dot{a}_{0,0} = M(Y) = 2{,}985$; $Var(\dot{E}_0) = Var\ (Y) = 1{,}114$; $R_0^2 = 0$.

- Per $r = 1$

$$\begin{cases} a_{0,1} + 1{,}019a_{1,1} = 2{,}985 \\ 1{,}019a_{0,1} + 1{,}409a_{1,1} = 3{,}631 \end{cases}$$

I parametri del modello risultano: $\dot{a}_{0,1} = 1{,}365$; $\dot{a}_{1,1} = 1{,}590$; $Var(\dot{E}_1) = 0{,}180$; $Var(\dot{Y}_1) = 0{,}934$; $R_1^2 = 0{,}839$.

- Per $r = 2$

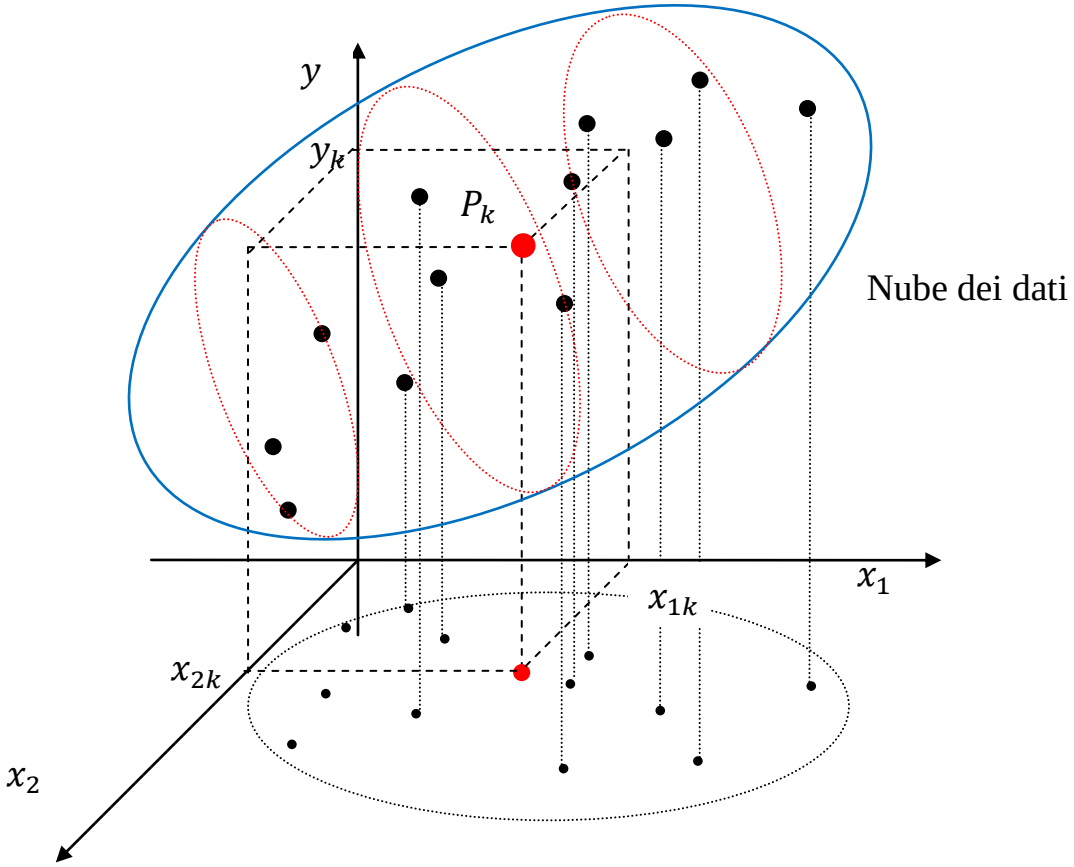
$$\begin{cases} a_{0,2} + 1{,}019a_{1,2} + 1{,}409a_{2,2} = 2{,}985 \\ 1{,}019a_{0,2} + 1{,}409a_{1,2} + 2{,}189a_{2,2} = 3{,}631 \\ 1{,}409a_{0,2} + 2{,}189a_{1,2} + 3{,}595a_{2,2} = 5{,}516 \end{cases}$$

I parametri del modello risultano: $\dot{a}_{0,2} = 0{,}617$; $\dot{a}_{1,2} = 3{,}872$; $\dot{a}_{2,2} = -1{,}121$; $Var(\dot{E}_2) = 0{,}082$; $Var(\dot{Y}_2) = 1{,}032$; $R_2^2 = 0{,}936$.

Il valore elevato dell'indice di determinazione per $r = 2$ fa ritenere sufficiente l'indagine condotta per quanto riguarda il modello polinomiale e questo è confermato anche dall'andamento dei valori $\dot{y}_2(x_k)$, riportati in tabella e dal corrispondente grafico.

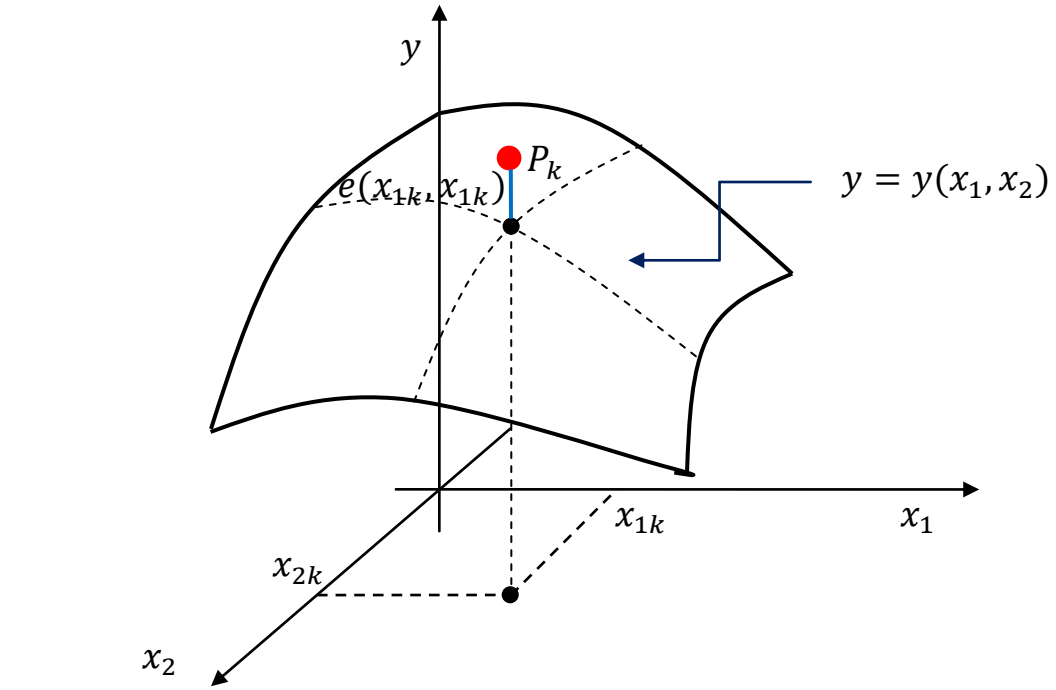
24. Cenni di analisi descrittiva congiunta di più grandezze quantitative: la regressione multipla

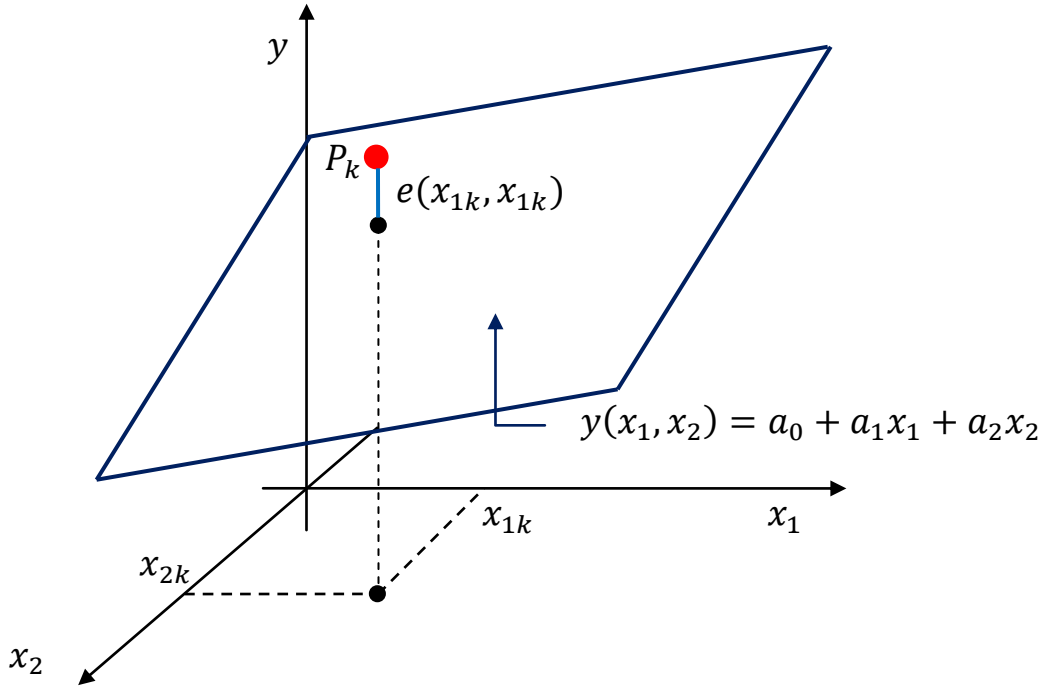
Lo studio di fenomeni complessi richiede la raccolta di dati e l'interpretazione di più di due variabili di cui una, indicata con *Y*, è di particolare interesse, mentre le altre sono variabili esplicative, $(X_1, X_2, \cdots, X_r)$, di cui si vuole conoscere l'influenza sulla variabile *Y*. Per semplicità si considera $r = 2$ potendosi generalizzare i risultati alle situazioni con un maggior numero di variabili esplicative. In corrispondenza di ogni unità statistica u_k osservata, con $k = 1,2, \cdots, n$, si dispone di un punto $P_k \equiv (x_{1k}, x_{2k}, y_k)$. L'insieme dei punti P_k, in un grafico cartesiano a tre dimensioni, rappresenta l'intera popolazione che si concretizza come la "nube dei dati".



In relazione alle conoscenze "a priori" del fenomeno si sceglie la funzione $y = y(x_1, x_2)$, in particolare, spesso si adotta una funzione lineare nelle variabili

$$y(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2$$





Esprimendo la variabile regressa Y come $Y = y(X_1, X_2) + E$ che evidenzia le due componenti: strutturale e casuale, si ha

$$E = Y - y(X_1, X_2) = Y - (a_0 + a_1 X_1 + a_2 X_2)$$

La componente casuale dipende, oltre che dai dati $(X_1, X_2, Y) \rightarrow \{(x_{1k}, x_{2k}, y_k); k = 1,2, \cdots, n\}$, anche dai parametri $\{a_s; s = 0,1,2\}$.

Similmente a quanto fatto per la regressione polinomiale la scelta dei valori da assumere per $\{a_s; s = 0,1,2\}$ è demandata alla minimizzazione di una funzione di perdita media di tipo quadratico che corrisponde al criterio di accostamento dei "minimi quadrati"

$$\min_{a_s;s=0,1,2} \Psi\big(a_s; s = 0,1,2; (X_1, X_2, Y)\big) = \min_{a_s;s=0,1,2} M(E^2)$$

$$\min_{a_s;s=0,1,2} \Psi = \min_{a_s;s=0,1,2} M(E^2) = \min_{a_s;s=0,1,2} M([Y - y(X_1, X_2)]^2)$$

$$\min_{a_s;s=0,1,2} \Psi = \min_{a_s;s=0,1,2} M([Y - (a_0 + a_1 X_1 + a_2 X_2)]^2)$$

La condizione di minimo è soddisfatta dall'uguaglianza a zero delle derivate parziali di Ψ rispetto ai parametri a_s per $s = 0,1,2$. In particolare, per $s = 0$ si ha

$$\frac{\partial \Psi}{\partial a_0} = (-2)M(E) = 0 \Rightarrow M(E) = 0$$

$$M\big(Y - (a_0 + a_1 X_1 + a_2 X_2)\big) = 0$$

$$M\{y(X_1, X_2)\} = a_0 + a_1 M(X_1) + a_2 M(X_2) = M(Y)$$

da cui si ottiene

$$a_0 = M(Y) - a_1 M(X_1) - a_2 M(X_2) \quad (*)$$

e sostituendo nell'espressione da minimizzare abbiamo

$$\min_{a_s;s=1,2} \Psi = \min_{a_s;,1,2} M\Big(\big[\big(Y - M(Y)\big) - a_1\big(X_1 - M(X_1)\big) - a_2\big(X_2 - M(X_2)\big)\big]^2\Big)$$

Al posto delle variabili (X_1, X_2, Y) si possono introdurre le variabili "scarto" dalla rispettiva media (U_1, U_2, Z):

$$U_1 = X_1 - M(X_1),\ U_2 = X_2 - (X_2), Z = Y - M(Y)$$

si ha la seguente funzione da minimizzare

$$\min_{a_s;s=1,2} \Psi = \min_{a_s;s=1,2} M(E^2) = \min_{a_s;,1,2} M([Z - a_1U_1 - a_2U_2]^2)$$

Derivando Ψ rispetto a_1 e a_2 si ottiene un sistema lineare di 2 equazioni in 2 incognite:

$$\begin{cases} \partial\Psi/\partial a_1 = (-2)M(EU_1) = 0 \\ \partial\Psi/\partial a_1 = (-2)M(EU_2) = 0 \end{cases} \rightarrow \begin{cases} M(EU_1) = 0 \\ M(EU_2) = 0 \end{cases}$$

$$\begin{cases} M\big((Z - a_1U_1 - a_2U_2)U_1\big) = 0 \\ M\big((Z - a_1U_1 - a_2U_2)U_2\big) = 0 \end{cases}$$

$$\begin{cases} a_1M(U_1^2) + a_2M(U_1U_2) = M(U_1Z) \\ a_1M(U_1U_2) + a_2M(U_2^2) = M(U_2Z) \end{cases} \quad (**)$$

La "matrice dei coefficienti" è data da medie di potenze degli "scarti" di X_1 e X_2, ossia

$$M(U_1^2) = M\left\{\big(X_1 - M(X_1)\big)^2\right\} = \overline{m}_{20} = Var\{X_1\}$$

$$M(U_2^2) = M\left\{\big(X_2 - M(X_2)\big)^2\right\} = \overline{m}_{02} = Var\{X_2\}$$

$$M(U_1U_2) = M\big\{\big(X_1 - M(X_1)\big)\big(X_2 - M(X_2)\big)\big\} = Cov\{X_1, X_2\}$$

mentre il vettore dei "termini noti" è dato da medie degli "scarti" di Y per quelli di X_1 e X_2, rispettivamente:

$$M(U_1Z) = M\big\{\big(X_1 - M(X_1)\big)\big(Y - M(Y)\big)\big\} = Cov\{X_1, Y\}$$

$$M(U_2Z) = M\big\{\big(X_2 - M(X_2)\big)\big(Y - M(Y)\big)\big\} = Cov\{X_2, Y\}$$

Tutti i coefficienti del sistema si ottengono dai dati osservati $\{(x_{1k}, x_{2k}, y_k); k = 1,2,\cdots,n\}$, in particolare, le covarianze si ottengono, ad esempio per $Cov\{X_1, Y\}$, come:

$$Cov\{X_1, Y\} = \frac{1}{n}\sum_{k=1}^{n}\big(x_{1k} - M(X_1)\big)\big(y_k - M(Y)\big)$$
$$= \frac{1}{n}\sum_{k=1}^{n} x_{1k}y_k - M(X_1)M(Y)$$

Gli elementi della matrice dei coefficienti e del vettore dei termini noti fanno parte della classe dei "momenti" (centrali, cioè calcolati rispetto al valor medio) della variabile tridimensionale $\{(X_1, X_2, Y)\}$. Si osservi che l'operatore "covarianza" assume valori positivi, nulli e negativi; inoltre, si dimostra, ad esempio, che:

$$|Cov\{X_1, Y\}| \leq [Var\{X_1\}Var\{Y\}]^{1/2}$$

Risolto il sistema lineare (**) si ottengono i valori dei parametri del modello di regressione multipla che rispettano il criterio di accostamento, specificatamente indicati: $\dot{a}_1, \dot{a}_2$ che sostituiti nella (*) determinano anche l'intercetta $\dot{a}_0$. Disponendo di tali parametri è possibile definire il modello:

$$\dot{y}(x_1, x_2) = \dot{a}_0 + \dot{a}_1 x_1 + \dot{a}_2 x_2$$

e, in particolare, determinare i valori della variabile Y corrispondenti alle osservazioni di (X_1, X_2):

$$\{X_1, X_2, \dot{y}(X_1, X_2)\} \equiv \{x_{1k}, x_{2k}, \dot{y}_k = \dot{a}_0 + \dot{a}_1 x_{1k} + \dot{a}_2 x_{2k}\}$$

e i valori della componente accidentale

$$\dot{E} \equiv Y - \dot{y}(X_1, X_2) = Z - (\dot{a}_1 U_1 + \dot{a}_2 U_2)$$
$$\rightarrow \{e_k = y_k - (\dot{a}_0 + \dot{a}_1 x_{1k} + \dot{a}_2 x_{2k})\}$$
$$= z_k - (\dot{a}_1 u_{1k} + \dot{a}_2 u_{2k})$$

La media aritmetica di $\dot{E}$, abbiamo già visto è pari a zero: $M(\dot{E}) = 0$.

Come misura dell'accostamento si impiega la varianza dei "residui":

$$Var\left(\dot{E}\right) = M\left(\dot{E}^2\right) = \frac{1}{n}\sum_{k=1}^{n}(y_k - \dot{y}_k)^2 \leq Var\ (Y)$$

da cui si ottiene un indice "standardizzato", che è detto "indice di determinazione":

$$R^2 = 1 - \frac{Var\left(\dot{E}_r\right)}{Var\ (Y)} \quad \text{con} \quad 0 \leq R^2 \leq 1$$

Oltre alla varianza dei residui si considera anche la varianza "spiegata" che misura la variabilità dei valori ottenuti dal modello $\{\dot{y}_k\}$:

$$\begin{aligned} Var\{\dot{y}(X_1, X_2)\} &= M\{[\dot{y}(X_1, X_2) - M(Y)]^2\} \\ &= \frac{1}{n}\sum_{k=1}^{n}(\dot{y}_k - M(Y))^2 \leq Var\ (Y) \end{aligned}$$

La varianza spiegata può anche ottenersi come

$$\begin{aligned} Var\{\dot{y}(X_1, X_2)\} &= M\{(\dot{a}_1 U_1 + \dot{a}_2 U_2)^2\} = M\{(\dot{a}_1 U_1 + \dot{a}_2 U_2)^2\} \\ &= M\{(\dot{a}_1 U_1 + \dot{a}_2 U_2)(\dot{a}_1 U_1 + \dot{a}_2 U_2)\} \\ &= M\{\dot{a}_1(\dot{a}_1 U_1^2 + \dot{a}_2 U_1 U_2) + \dot{a}_2(\dot{a}_1 U_1 U_2 + \dot{a}_2 U_2^2)\} \\ &= \dot{a}_1(\dot{a}_1 M\{U_1^2\} + \dot{a}_2 M\{U_1 U_2\}) \\ &+ \dot{a}_2(\dot{a}_1 M\{U_1 U_2\} + \dot{a}_2 M\{U_2^2\}) \end{aligned}$$

essendo per il sistema (**) $(\dot{a}_1 M\{U_1^2\} + \dot{a}_2 M\{U_1 U_2\}) = M(U_1 Z)$ e $\dot{a}_1 M(U_1 U_2) + \dot{a}_2 M(U_2^2) = M(U_2 Z)$, si ha

$$Var\{\dot{y}(X_1, X_2)\} = \dot{a}_1 Cov\{X_1, Y\} + \dot{a}_2 Cov\{X_2, Y\}$$

dove $Cov\{X_1, Y\}$ e $Cov\{X_2, Y\}$ sono i"termini noti" del sistema lineare dato da (**).

Potendosi ancora dimostrare che

$$Var\{\dot{y}(X_1, X_2)\} + Var\left(\dot{E} = Y - \dot{y}(X_1, X_2)\right) = Var\ (Y)$$

L'indice di determinazione è dato anche da:

$$R^2 = \frac{Var\{\dot{y}(X_1, X_2)\}}{Var\ (Y)} \quad \text{con} \quad 0 \leq R^2 \leq 1$$

Esercizio 18

Si voglia determinare un modello di regressione lineare che esprima la grandezza *Y prodotto interno lordo (PIL)* degli USA (in milioni di $) sulla base delle seguenti grandezze:

- X_1 *quantità di lavoro* (in milioni di uomini/anno);
- X_2 *capitale investito* (in milioni di $).

Si disponga dei seguenti $n = 10$ rilievi, relativi agli anni dal 1946 al 1955 (fonte: Goldberg), posto l'anno 1946→ $k = 1$.

k	X_1	X_2	Y	X_1^2	X_2^2	Y^2	X_1X_2	X_1Y	X_2Y
1	51	9	209	2601	81	43681	459	10659	1881
2	53	25	214	2809	625	45796	1325	11342	5350
3	53	39	225	2809	1521	50625	2067	11925	8775
4	50	51	221	2500	2601	48841	2550	11050	11271
5	52	62	243	2704	3844	59049	3224	12636	15066
6	54	75	257	2916	5625	66049	4050	13878	19275
7	54	94	265	2916	8836	70225	5076	14310	24910
8	55	108	276	3025	11664	76176	5940	15180	29808
9	52	118	271	2704	13924	73441	6136	14092	31978
n=10	54	124	291	2916	15376	84681	6696	15714	36084
Σ	528	705	2472	27900	64097	618564	37523	130786	184398
Σ/n	52,8	70,5	247,2	2790	6409,7	61856,4	3752,3	13078,6	18439,8
				2,16	1439,45	748,56	29,90	26,44	1012,20

Nelle colonne successive sono riportati i valori necessari per il calcolo dei momenti dall'origine che interessano i cui risultati sono raccolti nella penultima riga della tabella precedente. Nell'ultima riga sono riportati i valori delle varianze e covarianze.

Si sono considerati i seguenti modelli di regressione.

- $y(x_1) = a_0 + a_1x_1$, i cui parametri risultano: $\dot{a}_1 = \frac{26,44}{70,5} = 12,24;\ \dot{a}_0 = 247,2 - 12,24 \times 52,8 = -399,11;\ R^2 = 0,4324.$

- $y(x_2) = b_0 + b_1 x_2$, i cui parametri risultano: $\dot{b}_1 = \frac{1012,2}{1439,45} = 0,703$; $\dot{a}_0 = 247,2 - 0,703 \times 70,5 = 197,63$; $R^2 = 0,9508$.
- $y(x_1, x_2) = c_0 + c_1 x_1 + c_2 x_2$, i cui parametri risultano sono la soluzione del sistema:

$$\begin{cases} 2,16c_1 + 29,9c_2 = 26,44 \\ 29,9c_1 + 1439,45c_2 = 1012,2 \end{cases}; \ \dot{c}_1 = 3,519; \dot{c}_2 = 0,630;$$

$$R^2 = \frac{3,519 \times 26,44 + 0,630 \times 1012,2}{748,56} = 0,9763$$

Si lascia al lettore ogni commento sui risultati ottenuti.

25. I momenti di variabili statistiche unidimensionali e bidimensionali

Una classe di indicatori sintetici di variabili unidimensionali e bidimensionali tali da comprendere indici di posizione e di variabilità, ma anche altri indicatori che evidenziano specificità del modo di distribuirsi variabili (asimmetria, curtosi, correlazione, ecc.), è data dai "momenti".

A. Nel caso di una generica variabile unidimensionale *X*, si definisce il momento di ordine *r* da un "polo" θ la seguente media:

$$_{\theta}m_r = M\{(X-\theta)^r\} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \theta)^r = \sum_{i=1}^{p}(x_i - \theta)^r f_i$$

per $r = 0, 1, 2, \cdots$, dove i valori della variabile *X* sono estesi all'intero asse reale.

Per quanto riguarda i valori di θ ci si limita a considerare le due situazioni di interesse: $\theta = 0$ e $\theta = m = M(X)$, avendo le due classi momenti: "dall'origine" e "centrali", rispettivamente:

$$m_r = M\{X^r\} \quad \text{per} \quad r = 0, 1, 2, \cdots$$

$$\bar{m}_r = M\{(X-m)^r\} \quad \text{per} \quad r = 0, 1, 2, \cdots$$

I principali momenti che hanno rilevanza in campo statistico sono quelli di ordine inferiore o pari a 4, in particolare abbiamo

$r = 0 \quad m_0 = 1 \qquad\qquad \bar{m}_0 = 1$

$r = 1 \quad m_1 = M\{X\} = m \quad \bar{m}_1 = M\{X-m\} = 0$

$r = 2 \quad m_2 = M\{X^2\} \qquad \bar{m}_2 = M\{(X-m)^2\} = Var(X) = s^2$

Tra i momenti dall'origine e quelli centrali, applicando lo sviluppo del binomio di Newton, esistono le relazioni:

$$\bar{m}_r = \sum_{j=0}^{r} \binom{r}{j} (-1)^j m^j m_{r-j} \quad \text{e} \quad m_r = \sum_{j=0}^{r} \binom{r}{j} m^j \bar{m}_{r-j}$$

che permettono di ottenere i momenti centrali conoscendo i momenti dall'origine di ordine inferiore o uguale e viceversa di ottenere quelli dall'origine conoscendo la media aritmetica e i momenti centrali di ordine inferiore o uguale.

Osservazioni

- C'è equivalenza nel descrivere una variabile statistica X mediante la distribuzione di frequenza e la sequenza dei momenti:

$$X \to \{x_i, f_i\} \begin{matrix} \nearrow \{m, m_2, m_3, \cdots\} \\ \searrow \{m, \bar{m}_2, \bar{m}_3, \cdots\} \end{matrix}$$

- I momenti possono intendersi come una sequenza di parametri che definiscono la variabile statistica X.
- Due variabili che hanno uguale distribuzione hanno anche gli stessi momenti.
- Due o più variabili sono tanto più "somiglianti" quanto più uguali sono i momenti, a partire da quelli di ordine inferiore. C'è, quindi, una "gerarchia" tra i momenti in sequenza inversa rispetto all'ordine r.
- Nei momenti si ritrovano indici sintetici della variabile X:

$$m_1 = m \Longrightarrow "media\ aritmetica" \Longrightarrow "indice\ di\ posizione"$$

$$\bar{m}_2 = s^2 \Longrightarrow "varianza" \Longrightarrow "indice\ di\ dispersione"$$

$$\bar{m}_3 \Longrightarrow \gamma_1 = \bar{m}_3/s^3 \Longrightarrow "indice\ di\ asimmetria"$$

$$\bar{m}_4 \Longrightarrow \gamma_2 = \bar{m}_4/s^4 \Longrightarrow "indice\ di\ forma - curtosi"$$

B. Per una variabile statistica bidimensionale (X, Y) si definiscono le classi di momenti "dall'origine" e "centrali" medie dei prodotti delle potenze di ordine (*r*,*s*), in modo analogo a quello impiegato per i momenti di variabile unidimensionale:

$$m_{rs} = M\{X^r Y^s\} = \frac{1}{n}\sum_{k=1}^{n} x_k^r\, y_k^s \quad \text{per} \quad r, s = 0,1,2,\cdots$$

$$\bar{m}_{rs} = M\left\{\big(X - M(X)\big)^r \big(Y - M(Y)\big)^s\right\}$$
$$= \frac{1}{n}\big(x_k - M(X)\big)^r \big(y_k - M(Y)\big)^s \quad \text{per} \quad r, s = 0,1,2,\cdots$$

Osservazioni

- Se $r = 0$ (oppure, $s = 0$) si ottengono i momenti corrispondenti della sola componente *Y* (oppure, *X*), quindi, ad es., $m_{01} = M(Y)$, $\bar{m}_{20} = Var(X)$.
- I momenti con sia $r \neq 0$ sia $s \neq 0$, denominati "momenti misti", sono quelli che evidenziano caratteristiche congiunte delle variabili componenti e presentano un interesse specifico nello studio multivariato.
- Come per i momenti di variabili unidimensionali risultano più importanti o maggiormente rappresentativi del fenomeno bivariato quelli di ordine "complessivo" $(r + s)$ minori.

$$\text{Per } r + s = 0 \Longrightarrow r = s = 0 \Longrightarrow m_{00} = \bar{m}_{00} = 1$$

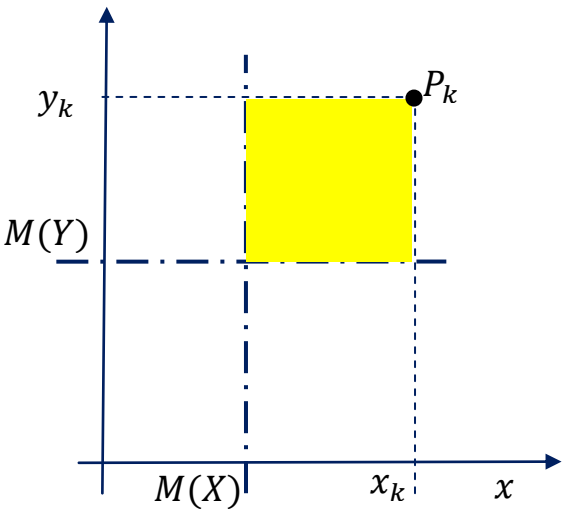
$$\text{Per } r+s=1 \Longrightarrow \begin{cases} r=1, s=0 \Longrightarrow \begin{cases} m_{10} = M(X) \\ \bar{m}_{10} = M(X - m_{10}) = 0 \end{cases} \\ r=0, s=1 \Longrightarrow \begin{cases} m_{01} = M(Y) \\ \bar{m}_{01} = M(Y - m_{01}) = 0 \end{cases} \end{cases}$$

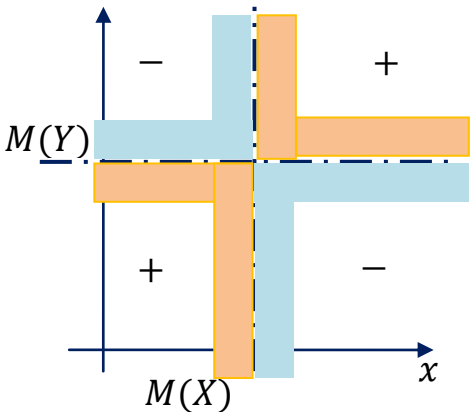
$$\text{Per } r+s=2$$

$$\Longrightarrow \begin{cases} r=2, s=0 \Longrightarrow \begin{cases} m_{20} = M(X^2) \\ \bar{m}_{20} = M\{(X - m_{10})^2\} = Var(X) \end{cases} \\ r=0, s=2 \Longrightarrow \begin{cases} m_{02} = M(Y^2) \\ \bar{m}_{02} = M\{(Y - m_{01})^2\} = Var(Y) \end{cases} \\ r=1, s=1 \Longrightarrow \begin{cases} m_{11} = M(XY) \\ \bar{m}_{11} = M\{(X - m_{10})(Y - m_{01})\} = Cov(X,Y) \end{cases} \end{cases}$$

26. La covarianza, significato e proprietà

La "covarianza" tra le componenti della variabile statistica (X, Y) è la media dei prodotti degli scarti delle variabili X e Y dalle rispettive medie aritmetiche $M(X)$ e $M(Y)$.

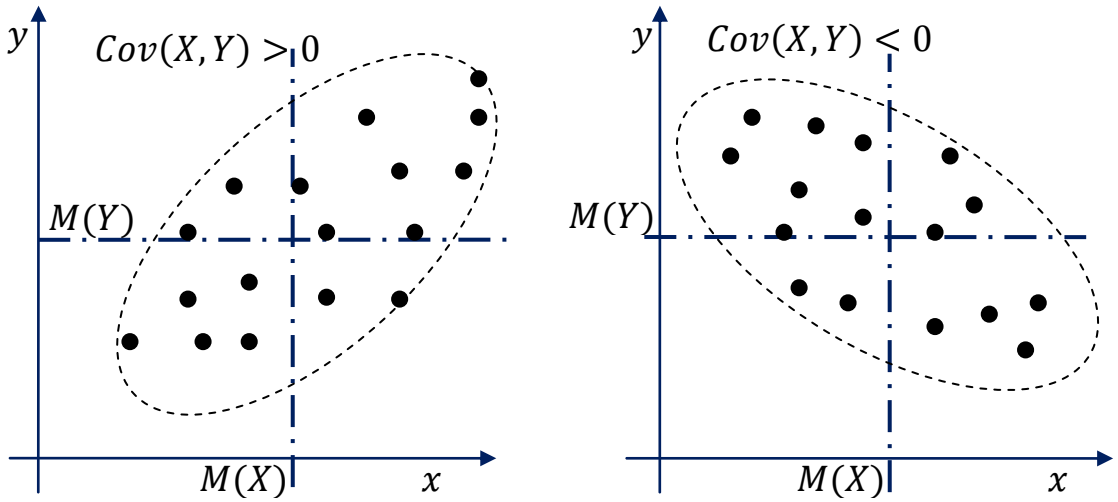




$$Cov(X,Y) = \bar{m}_{11} = M\{(X - m_{10})(Y - m_{01})\}$$
$$= \frac{1}{n}\sum_{k=1}^{n}\big(x_k - M(X)\big)\big(y_k - M(Y)\big)$$

La covarianza presenta valori positivi, negativi e nulli a seconda del prevalere dei punti $P_k \equiv (x_k, y_k)$ rispetto alle linee delle medie, nel 1° e 3° quadrante $(+)$; nel 2° e 4° quadrante $(-)$.

Per queste proprietà la covarianza è una misura del "legame diretto" $(+)$ oppure "inverso" $(-)$ tra le variabili X e Y, più precisamente della "correlazione lineare" tra le variabili.



Determinazione della covarianza mediante i momenti dall'origine

Consideriamo la variabile bidimensionale (X, Y), dalla definizione della covarianza, si ottiene

$$\begin{aligned} Cov(X,Y) &= M\{(X - m_{10})(Y - m_{01})\} \\ &= M\{X(Y - m_{01}) - m_{10}(Y - m_{01})\} \\ &= M\{XY - m_{01}X\} - m_{10}M(Y - m_{01}) \\ &= M\{XY\} - m_{01}M\{X\} = M\{XY\} - M\{X\}M\{Y\} \\ &= m_{11} - m_{10}m_{01} \end{aligned}$$

relazione che permette di ottenere $Cov(X,Y)$ mediante i momenti dall'origine che può impiegarsi come formula di calcolo.

Dalla relazione $Cov(X,Y) = M\{XY\} - M\{X\}M\{Y\}$ si hanno le seguenti condizioni riguardanti il segno che presenta la covarianza:

$$\begin{cases} Cov(X,Y) > 0 \Longrightarrow M\{XY\} > M\{X\}M\{Y\} \\ Cov(X,Y) = 0 \Longrightarrow M\{XY\} = M\{X\}M\{Y\} \\ Cov(X,Y) < 0 \Longrightarrow M\{XY\} < M\{X\}M\{Y\} \end{cases}$$

Campo di esistenza della covarianza

Si consideri una variabile (X,Y) con componenti "non degenere", cioè tali che $Var(X) > 0$ e $Var(Y) > 0$, si definisca la variabile statistica Z funzione di (X,Y) pari a:

$$Z = a\big(X - M(X)\big) + (Y - M(Y)) \text{ con } a \in \Re, a \neq 0$$

La media e la varianza di Z risultano

$$\begin{aligned} M(Z) &= M\big\{a\big(X - M(X)\big) + (Y - M(Y))\big\} \\ &= aM\big(X - M(X)\big) + M(Y - M(Y)) = 0 \end{aligned}$$

essendo $M\big(X - M(X)\big) = M\big(Y - M(Y)\big) = 0$, per la proprietà della media aritmetica.

$$\begin{aligned} Var\{Z\} &= M\{(Z - M(Z))^2\} = M\{Z^2\} \\ &= M\left\{\left[a\big(X - M(X)\big) + (Y - M(Y))\right]^2\right\} \\ &= M\left\{a^2\big(X - M(X)\big)^2 + 2a\big(X - M(X)\big)\big(Y - M(Y)\big)\right. \\ &\left. + (Y - M(Y))^2\right\} \\ &= a^2 M\left\{\big(X - M(X)\big)^2\right\} \\ &+ 2aM\left\{\big(X - M(X)\big)\big(Y - M(Y)\big) + M\{(Y - M(Y))^2\}\right\} \\ &= a^2 Var(X) + 2aCov(X,Y) + Var(Y) \\ &= a^2\overline{m}_{20} + 2a\overline{m}_{11} + \overline{m}_{02} \end{aligned}$$

La varianza di Z è una funzione quadratica di a: $Var\{Z\} = \xi(a)$, dovendo, come varianza essere $\xi(a) \geq 0$ quindi le radici dell'equazione di secondo grado $\xi(a) = 0 \Rightarrow\ a^2\overline{m}_{20} + 2a\overline{m}_{11} +$

$\bar{m}_{02} = 0$ deve presentare radici reali coincidenti o complesse e quindi presentare il discriminante $\Delta \leq 0$, da cui consegue

$$\frac{\Delta}{4} = [Cov(X,Y)]^2 - Var(X)Var(Y) \leq 0$$

da cui

$$[Cov(X,Y)]^2 \leq Var(X)Var(Y) \Rightarrow |Cov(X,Y)| \leq s(X)s(Y)$$

avendo indicato con $s(X) = \sqrt{Var(X)}$ e $s(Y) = \sqrt{Var(Y)}$ gli s.q.m. rispettivamente delle variabili X e Y si ottiene il campo di esistenza della covarianza:

$$-s(X)s(Y) \leq Cov(X,Y) \leq s(X)s(Y)$$

Normalizzazione della covarianza, il coefficiente di correlazione lineare

Dalla relazione precedente, si ha

$$-\frac{s(X)s(Y)}{s(X)s(Y)} \leq \frac{Cov(X,Y)}{s(X)s(Y)} \leq \frac{s(X)s(Y)}{s(X)s(Y)}$$

$$-1 \leq \frac{Cov(X,Y)}{s(X)s(Y)} \leq 1$$

Il rapporto $Cov(X,Y)/s(X)s(Y)$ è detto "coefficiente di correlazione di Bravais-Pearson" e misura, in forma standardizzata il legame lineare tra le due variabili X e Y, e viene indicato con $r_{X,Y} = r(X,Y)$

$$r_{X,Y} = \frac{Cov(X,Y)}{s(X)s(Y)}, -1 \leq r_{X,Y} \leq 1, |r_{X,Y}| \leq 1, r_{X,Y}^2 \leq 1, r_{X,Y} = r_{Y,X}$$

Si tratta di una standardizzazione "impropria" in quanto pur eliminando l'aspetto dimensionale mette in luce, con il segno $(+ \text{ o } -)$, la natura lineare del legame tra le variabili *X* e *Y*.

I valori "estremi" (± 1) si verificano se la variabile (X, Y) presenta un "perfetto legame lineare tra le due componenti *X* e *Y* ".

Posto infatti che $Y = a + bX \rightarrow y_k = a + bx_k \quad \text{per} \quad k = 1, 2, \cdots, n$ e inoltre sia $b \neq 0$, si ha:

$$M(Y) = M(a + bX) = a + bM(X)$$

$$Y - M(Y) = (a + bX) - (a + bM(X) = b(X - M(X))$$

$$\begin{aligned} Var(Y) &= M\{(Y - M(Y))^2\} = M\{b^2(X - M(X))^2\} \\ &= b^2 M\{(X - M(X))^2\} = b^2 Var(X) \end{aligned}$$

da cui

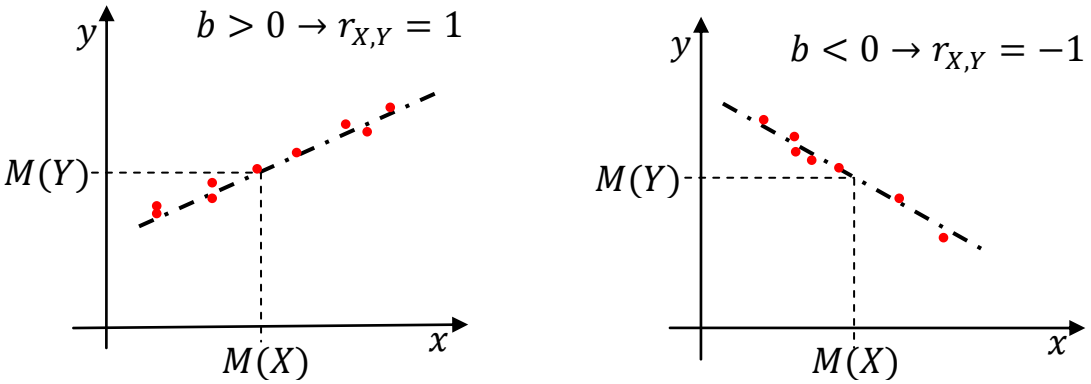
$$s(Y) = +\sqrt{Var(Y)} = |b| s(X)$$

La covarianza $Cov(X, Y)$ risulta:

$$\begin{aligned} Cov(X, Y) &= M\{\big(X - M(X)\big)\big(Y - M(Y)\big)\} = M\{b(X - M(X))^2\} \\ &= bM\{(X - M(X))^2\} = bVar(X) = b[s(X)]^2 \end{aligned}$$

Il coefficiente di correlazione lineare $r_{X,Y} = r(X, Y)$ risulta:

$$r_{X,Y} = \frac{Cov(X, Y)}{s(X)s(Y)} = \frac{b[s(X)]^2}{s(X)[|b| s(X)]} = \frac{b}{|b|} = \begin{matrix} \nearrow +1 \ \ se \ \ b > 0 \\ \searrow -1 \ \ se \ \ b < 0 \end{matrix}$$



Le rette di regressione

La covarianza e il coefficiente di correlazione lineare giocano un ruolo importante nello studio del legame lineare dato dalle due rette di regressione $\dot{y}(x) = \dot{a}_0 + \dot{a}_1 x$ e $\dot{x}(y) = \dot{b}_0 + \dot{b}_1 y$, in cui i parametri sono determinati mediante il criterio di accostamento dei minimi quadrati.

Nel caso della prima retta, dal sistema di equazioni normali abbiamo

$$\begin{cases} a_0 + a_1 M(X) = M(Y) \\ a_0 M(X) + a_1 M(X^2) = M(XY) \end{cases}$$

$$\begin{cases} a_0 = M(Y) - a_1 M(X) \\ a_0 M(X) + a_1 M(X^2) = M(XY) \end{cases}$$

$$\begin{cases} a_0 = M(Y) - a_1 M(X) \\ [M(Y) - a_1 M(X)]M(X) + a_1 M(X^2) = M(XY) \end{cases}$$

$$\begin{cases} a_0 = M(Y) - a_1 M(X) \\ a_1 [M(X^2) - M^2(X)] = M(XY) - M(X)M(Y) \end{cases}$$

$$\begin{cases} a_0 = M(Y) - a_1 M(X) \\ a_1 Var(X) = Cov(X,Y) \end{cases}$$

$$\begin{cases} a_0 = M(Y) - a_1 M(X) \\ \dot{a}_1 = Cov(X,Y)/Var(X) \end{cases}$$

Essendo $Cov(X,Y) = r_{X,Y}s(X)s(Y)$ e $Var(X) = s^2(X)$ risulta

$$\dot{a}_1 = r_{X,Y}\frac{s(Y)}{s(X)}$$

e

$$\dot{a}_0 = M(Y) - \left[r_{X,Y}\frac{s(Y)}{s(X)}\right]M(X)$$

La funzione di regressione può scriversi come:

$$\big(\dot{y} - M(Y)\big) = \dot{a}_1\big(x - M(X)\big) = \left[r_{X,Y}\frac{s(Y)}{s(X)}\right]\big(x - M(X)\big)$$

La varianza "residua", inoltre, risulta

$$\begin{aligned} Var\{(Y - \dot{Y})\} &= M\left\{\left(Y - \dot{Y}\right)^2\right\} \\ &= M\left\{\left[(Y - M(Y)) - \dot{a}_1\big(X - M(X)\big)\right]^2\right\} \\ &= Var(Y) - 2\dot{a}_1 Cov(X,Y) + \dot{a}_1^2 Var(X) \\ &= Var(Y) - 2r_{X,Y}\frac{s(Y)}{s(X)}r_{X,Y}s(X)s(Y) \\ &+ \left(r_{X,Y}\frac{s(Y)}{s(X)}\right)^2 Var(X) = \big(1 - r_{X,Y}^2\big)Var(Y) \end{aligned}$$

La varianza "spiegata" da tale modello è

$$Var\{\dot{Y}\} = M\left\{\left(\dot{Y} - M(Y)\right)^2\right\} = M\left\{\left[\left(r_{X,Y}\frac{s(Y)}{s(X)}\right)\big(X - M(X)\big)\right]^2\right\}$$
$$= \left(r_{X,Y}\frac{s(Y)}{s(X)}\right)^2 Var(X) = r_{X,Y}^2 Var(Y)$$

L'indice di determinazione, infine, è

$$R_y^2 = Var\{\dot{Y}\}/Var(Y) = r_{X,Y}^2 = \frac{[Cov(X,Y)]^2}{Var(X)Var(Y)}$$

Analogamente per la seconda retta di regressione si ottiene

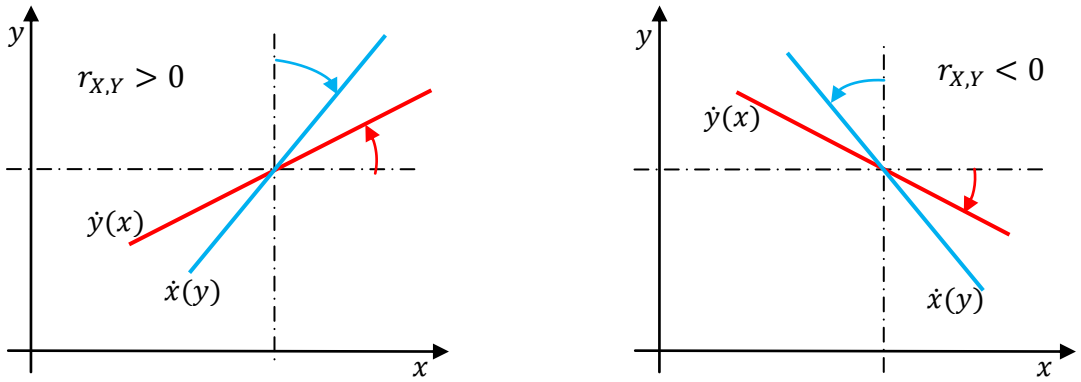
$$\dot{b}_1 = r_{X,Y}\frac{s(X)}{s(Y)} \quad \text{e} \quad \dot{b}_0 = M(X) - \left[r_{X,Y}\frac{s(X)}{s(Y)}\right]M(X)$$

Le due rette di regressione hanno in comune il punto "baricentrico" $(M(X), M(Y))$, le inclinazioni (coefficienti angolari) dello stesso segno (quello della covarianza), lo stesso indice di determinazione: pari al quadrato del coefficiente di correlazione lineare.

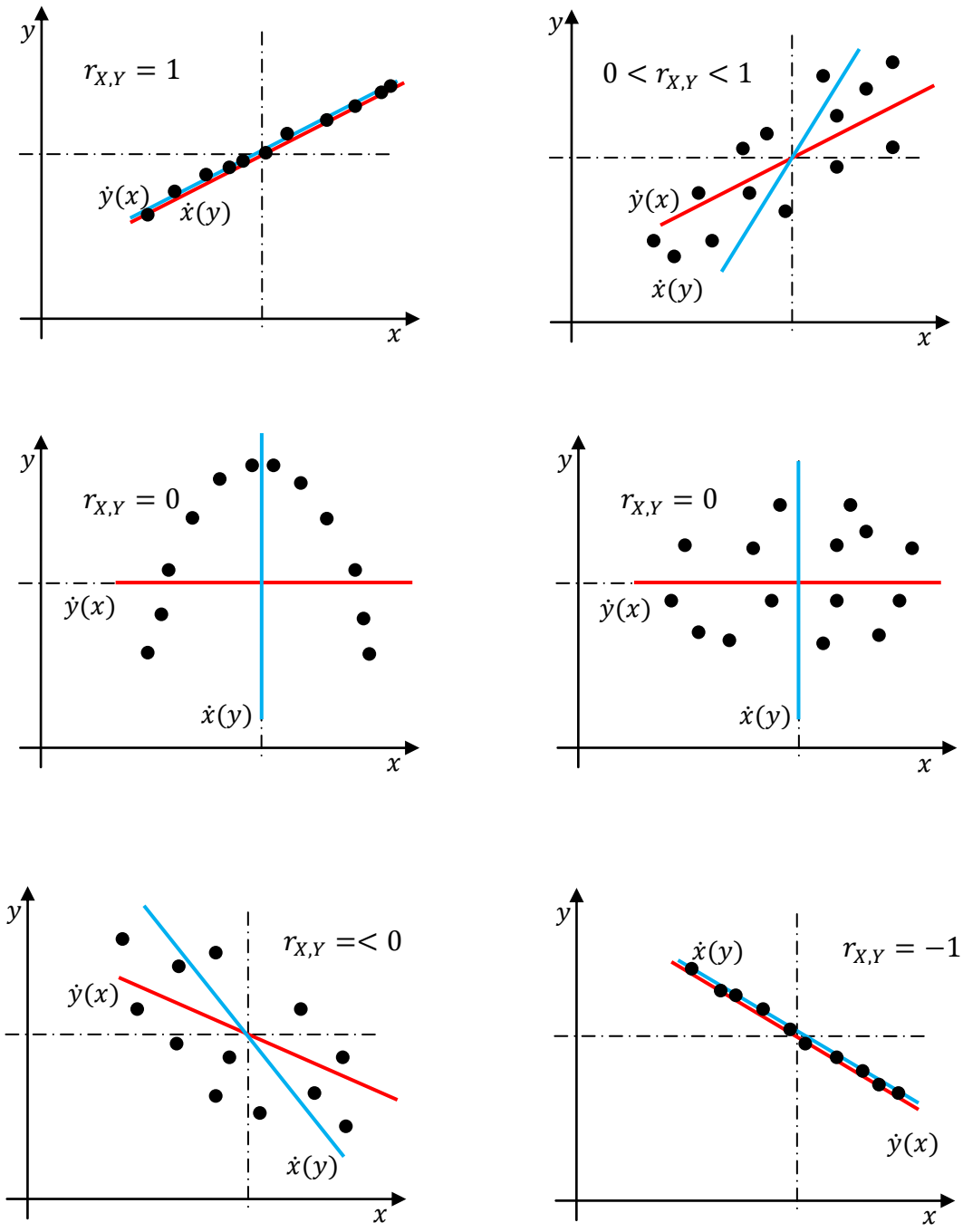
Tra i coefficienti di regressione $\dot{a}_1$ e $\dot{b}_1$ è valida la relazione:

$$\dot{a}_1 \cdot \dot{b}_1 = r_{X,Y}^2 \rightarrow r_{X,Y} = \pm\sqrt{\dot{a}_1 \cdot \dot{b}_1}$$

dove il segno è quello comune a covarianza, coefficiente di correlazione e ai parametri di regressione $\dot{a}_1$, $\dot{b}_1$.



Nella successiva figura sono presentate alcune situazioni tipiche di correlazione.



27. Attualità della statistica descrittiva

L'attuale tendenza che si manifesta nell'analisi dei dati è orientata alla scelta del modello interpretativo del fenomeno, effettuata in forma automatica solo sulla base delle informazioni raccolte. Un importante contributo è dato dall'Informatica, con la predisposizione di appropriati software.

Infatti, in molte ricerche osservative o sperimentali, sia scientifiche sia sociali, si dispone di un'elevata numerosità di dati (al contrario di quanto avveniva fino a pochi decenni fa) che sono spesso eterogenei per provenienza, per tempi di raccolta e per modalità di organizzazione, ma hanno la qualità di risultare disponibili a costi relativamente contenuti.

In particolare, in ambito gestionale, bio-medico, sanitario e fisico-tecnologico, la dimensione sia del numero di unità osservate sia della numerosità delle variabili/grandezze rilevate è sempre più ampia. Anche, ad esempio, medie o piccole aziende possono disporre di "data base" a basso costo e che, supportati da sistemi informatici adeguati, permettono di condurre analisi statistiche adeguate di carattere descrittivo.

Tali analisi presentano problemi peculiari, legati al modo difforme in cui speso avviene la raccolta dei dati, comportando metodologie particolari, una di queste è denominata "Data Mining".

Si deve, infatti, operare come in presenza di "giacimenti" in campo minerario o petrolifero, essendo le informazioni spesso eterogenee, avvalendosi, oltre che della metodologia statistica, dell'informatica, delle intelligenze artificiali, delle reti neurali e altro ancora. Particolare attenzione si dovrà dedicare alle innovazioni tecnologiche che permettono l'utilizzo di strumenti di hardware e di software sempre più adeguati.

L'attuale ricorso al "Data Mining" per la costruzione di modelli di realtà molto complesse consente di avvalersi di strumenti metodologici, anche semplici, propri della statistica descrittiva, quali grafici, indicatori sintetici, correlazioni e regressioni a due variabili, ovviamente replicate per tutte le combinazioni delle grandezze osservate per le quali si hanno a disposizione masse di dati numericamente rilevanti. Si osserva inoltre che spesso si opera in

assenza, o quasi, di assunzioni e anche talora di uno scopo/obiettivo preciso.

Quali esempi di ambienti adatti all'impiego dell'analisi "Data Mining" abbiamo

- *scontrini dei supermercati, carte "fidelity";*
- *dati utenze registrate da società telefoniche;*
- *sistema WWW, Internet, Google, ecc.;*
- *ricerche sulla struttura del DNA, microarrays;*
- *clima, meteorologia, rilevazioni aereo-spaziali.*

La disponibilità di dati e la velocità della loro elaborazione, fornite dagli hardware e dai software oggi disponibili, rende possibile concordare con quanto afferma R.H. Coase (1910-), Premio Nobel 1991 per l'economia:

"Se torturate i dati abbastanza a lungo, infine la natura confesserà i suoi segreti"

Si noti la diversità di questo approccio rispetto a quanto verrà illustrato in seguito e a cui si è accennato nel Paragrafo 1, dove era presentata la centralità del "modello" rappresentativo del fenomeno allo studio, nelle sue due componenti: strutturale o relazionale e casuale o aleatoria.

Riferimenti bibliografici

Di Ciaccio A., Borra S., (1996) *Introduzione alla Statistica Descrittiva*, McGraw-Hill Italia, Milano.

Landenna G., (1994) *Fondamenti di Statistica Descrittiva*, il Mulino, Bologna.

Leti G., (1983) *Statistica Descrittiva*, il Mulino, Bologna.

Zanella A., (1995) *Elementi di statistica descrittiva. Una presentazione sintetica*, CULS, Milano.

Sommario

0. Premessa 1

1. Ricerca di una definizione della disciplina Statistica 3

2. La "Statistica Descrittiva" 12

3. Analisi descrittiva di un carattere unidimensionale 17

4. Rappresentazioni grafiche 25

5. Rappresentazioni alternative di una variabile quantitativa **X** 28

6. Rappresentazione sintetica di una variabile quantitativa **X** 29

7. Sintesi di una variabile quantitativa unidimensionale 30

8. Proprietà degli indici di posizione 31

9. La media aritmetica 32

10. Altri tipi di indici di posizione 40

11. Moda o valore modale 46

12. Mediana o valore mediano 48

13. Valori quantili 54

14. La scelta degli indici di posizione 55

15. Minimizzazione della funzione di perdita 56

16. Media "obiettivo" secondo Chisini 60

17. Principali tipi di "medie obiettivo" 60

18. Concetto e misure di variabilità 65

19. Tipologie di indici di dispersione 67

20. Principali indici di dispersione rispetto a un centro 68

21. Alcuni indici di dispersione “globali” 70

22. Indici di dispersione “assoluti” e “relativi” 72

23. Analisi descrittiva congiunta di due grandezze quantitative: la regressione polinomiale 73

24. Cenni di analisi descrittiva congiunta di più grandezze quantitative: la regressione multipla 81

25. I momenti di variabili statistiche unidimensionali e bidimensionali 89

26. La covarianza, significato e proprietà 92

27. Attualità della statistica descrittiva 100