

ANALISI ESPLORATIVA DEI DATI

INTRODUZIONE

La fase iniziale di ogni valida analisi statistica dei dati deve essere necessariamente di tipo esplorativo:

**ELABORARE LE INFORMAZIONI A DISPOSIZIONE
AL FINE DI DESCRIVERE IN MODO SINTETICO
L'INSIEME DI DATI A DISPOSIZIONE**

Dal punto di vista dei metodi statistici, l'analisi esplorativa dei dati utilizza essenzialmente tecniche statistiche descrittive.

La finalità prevalente dell'analisi esplorativa è la descrizione delle strutture e delle relazioni presenti nei dati, per un eventuale successivo impiego in un modello statistico.

ANALISI ESPLORATIVA UNIVARIATA

Da questa semplice analisi esplorativa è possibile trarre importanti informazioni per le successive analisi multivariate e di modellazione.

I principali strumenti di analisi esplorativa univariata sono:

Le RAPPRESENTAZIONI GRAFICHE

- **Diagramma a barre e diagramma a torte (dati qualitativi nominali)**
- **Diagramma delle frequenze (caratteri qualitativi ordinali e quantitativi discreti)**
- **Istogramma (caratteri quantitativi continui)**

e una serie di INDICI SINTETICI

- **Indici di posizione (media, moda e mediana)**
- **Indici di variabilità (campo di variazione, differenza interquartilica, varianza, scarto quadratico medio)**
- **Indici di eterogeneità (entropia)**
- **Indici di asimmetria e di curtosi**

Misure di Posizione (o di Tendenza Centrale)

Nella maggior parte degli insiemi di dati, le osservazioni mostrano una tendenza a raggrupparsi attorno a un valore centrale.

Risulta in genere quindi possibile selezionare un valore tipico per descrivere un intero insieme di dati.

Tale valore descrittivo è una misura di posizione o di tendenza centrale.

Tipi di misure di posizione:

Medie

ANALITICHE

- aritmetica
- geometrica
- armonica
- quadratica

LASCHE

Valore centrale
Moda
Mediana
Quantili

LA MEDIA ARITMETICA

(invarianza dell'ammontare complessivo del carattere)

$$\mu = \frac{\sum_{i=1}^N x_i}{N} \quad \text{SEMPLICE}$$

$$\mu = \frac{\sum_{i=1}^s x_i n_i}{N} \quad \text{PONDERATA}$$

PROPRIETA' DELLA MEDIA ARITMETICA

- 1) La somma algebrica degli scarti è zero

$$\sum_{i=1}^N (x_i - \mu) = 0 \qquad \sum_{i=1}^S (x_i - \mu)n_i = 0$$

- 2) La somma dei quadrati degli scarti è un minimo

$$\sum_{i=1}^N (x_i - \mu)^2 = \min \qquad \sum_{i=1}^S (x_i - \mu)^2 n_i = \min$$

- 3) OMOGENEA
4) TRASLATIVA
5) ASSOCIATIVA
6) E' INTERNA

Misure di posizione: il Valore Centrale

Il **valore centrale (o midrange)** è dato dalla media tra la più piccola e la più grande delle osservazioni di un insieme di dati.

Il midrange

Il midrange si calcola sommando il valore più piccolo e quello più grande e dividendo per 2:

$$\text{Midrange} = \frac{X_{\text{più piccola}} + X_{\text{più grande}}}{2} \quad (3.3)$$

SOLUZIONE

La serie ordinata per questo insieme di dati è

10.0 20.6 28.6 28.6 29.4 29.5 29.9 30.1 30.5 30.5 32.1 32.2 32.4 33.0 35.2 37.1 38.0

Il midrange calcolato in base all'equazione (3.3) è

$$\begin{aligned} \text{Midrange} &= \frac{X_{\text{più piccola}} + X_{\text{più grande}}}{2} \\ &= \frac{10.0 + 38.0}{2} = 24.0 \end{aligned}$$

NOTA: il midrange deve essere usato con cautela. Infatti, poiché si basa solo sulla più grande e la più piccola delle osservazioni, il midrange è fortemente influenzato dalla presenza di valori anomali.

Misure di posizione: la Moda

La **moda** è il valore più frequente in un insieme di dati.

- A differenza della media, la moda non è influenzata dagli outlier.
- Tuttavia tale misura di posizione viene usata solo per scopi descrittivi, poiché è caratterizzata da maggiore variabilità rispetto alle altre misure di posizione (piccole variazioni in un insieme di dati possono far variare in modo consistente la moda).

Esempio 3.5 *Il calcolo della moda*

Calcolate la moda dei rendimenti percentuali annui conseguiti dai fondi comuni azionari che prelevano le commissioni direttamente dalle attività del fondo utilizzando la serie ordinata nell'esempio 3.3.

SOLUZIONE

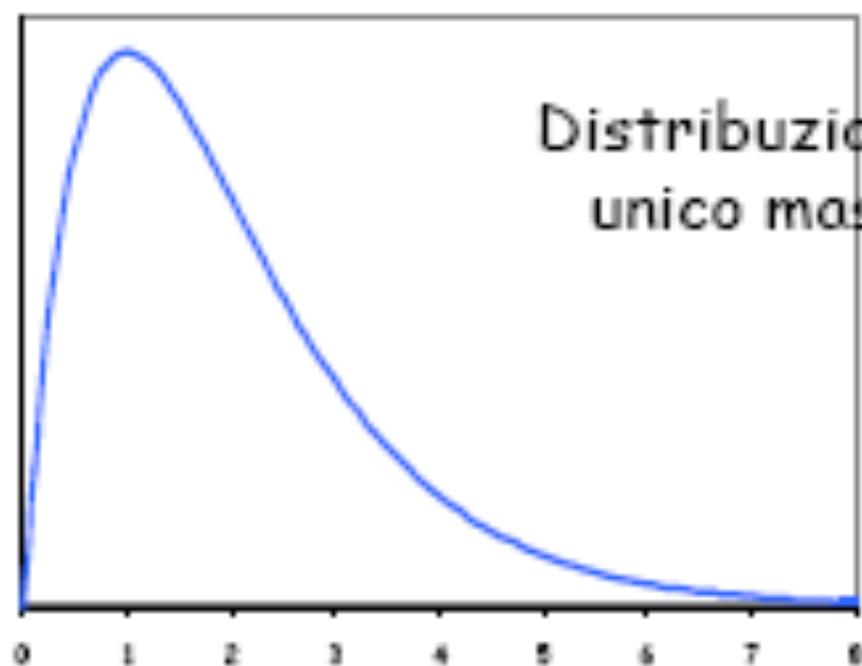
La serie ordinata per questi dati è la seguente:

10.0 20.6 28.6 28.6 29.4 29.5 29.9 30.1 30.5 30.5 32.1 32.2 32.4 33.0 35.2 37.1 38.0

Possiamo osservare che ci sono due valori “più tipici” o due mode: 28.6 e 30.5. Questo insieme di dati si dice *bimodale*.

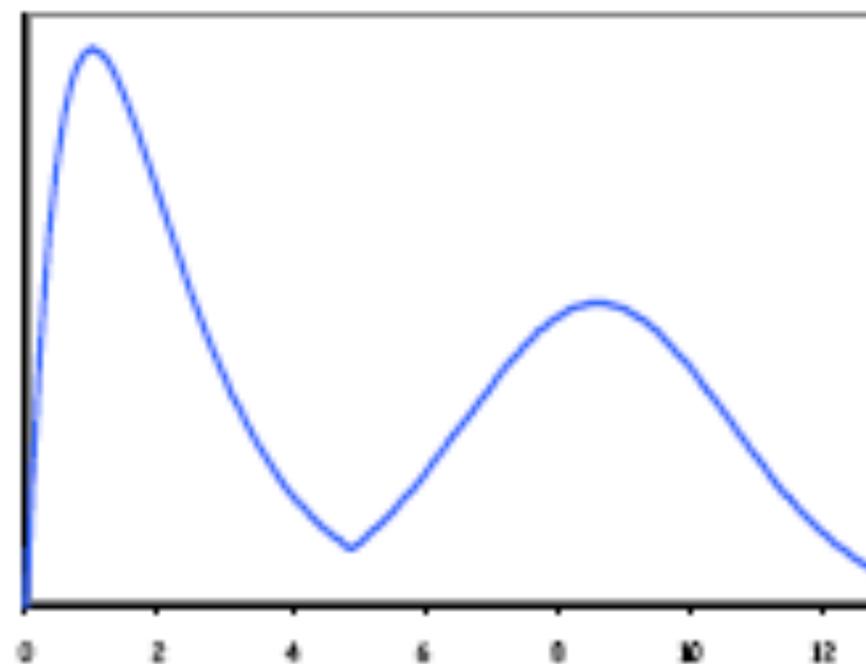
NOTA: un insieme di dati può non avere moda, se nessuno valore è “più tipico”.

Moda utile per distribuzioni unimodali



Distribuzione unimodale
unico massimo locale

Distribuzione bimodale
più di un massimo locale



Misure di posizione: la Mediana

1/2

La **mediana** è il valore centrale in una successione ordinata di dati.

La mediana

La mediana è l'osservazione che, nella serie ordinata dei dati, si lascia alla destra il 50% delle osservazioni e a sinistra il 50% delle osservazioni. Quindi, il 50% delle osservazioni risulteranno maggiori della mediana e il 50% risulteranno minori della mediana.

$$\text{Mediana} = \text{osservazione di posto } \frac{n+1}{2} \text{ nella serie ordinata} \quad (3.2)$$

Commento: La mediana non è influenzata dalle osservazioni estreme di un insieme di dati: nel caso di osservazioni estreme è quindi opportuno descrivere l'insieme di dati con la mediana piuttosto che con la media.

REGOLA 1. Se l'ampiezza del collettivo è un numero **dispari**, la mediana coincide con il valore centrale, vale a dire con l'osservazione che occupa la posizione $(n+1)/2$ nella serie ordinata delle osservazioni.

REGOLA 2. Se l'ampiezza del collettivo è un numero **pari**, la mediana allora coincide con la media dei valori corrispondenti alle due osservazioni centrali.

Calcolo della **MEDIANA** nel caso di una v.s. divisa in intervalli

Supponendo che le frequenze si distribuiscono in maniera uniforme all'interno delle classi, si utilizza la relazione:

$$Me = x_i + \frac{x_{i+1} - x_i}{n_i} \left(\frac{N}{2} - N_{i-1} \right)$$

dove

- x_i** è l'estremo inferiore della classe che contiene la **Me**
- x_{i+1}** è l'estremo superiore “ “ “ “ “ “
- n_i** è la frequenza assoluta “ “ “ “ “ “
- $N/2$** il posto occupato dalla **Me**
- N_{i-1}** la frequenza accumulata della classe precedente la **Me**

INTRODUZIONE

La VARIABILITA' è l'attitudine di un fenomeno quantitativo ad assumere diverse modalità:

- **DISPERSIONE** maggiore o minore addensamento delle osservazioni intorno alla media prestabilita;
- **DISUGUAGLIANZA** diversità delle varie osservazioni tra loro

Proprietà essenziali di una misura della Variabilità:

- 1. Essere nulla quando e solo quando tutti i termini della distribuzione sono uguali tra loro;**
- 2. Crescere all'aumentare della disuguaglianza tra i termini, nel senso che la grandezza della variabilità cresca ogni qual volta cresce almeno una delle quantità assunte per la misurazione della disuguaglianza tra due termini.**

Misure di variabilità (DISPERSIONE)

Una seconda caratteristica importante di un insieme di dati è la **variabilità**: la quantità di dispersione o di disuguaglianza presente nei dati.

Due insiemi di dati possono differire o nella posizione o nella variabilità oppure sia nella posizione che nella variabilità.

FIGURA 3.3

Due distribuzioni simmetriche a forma campanulare che differiscono solo nella posizione

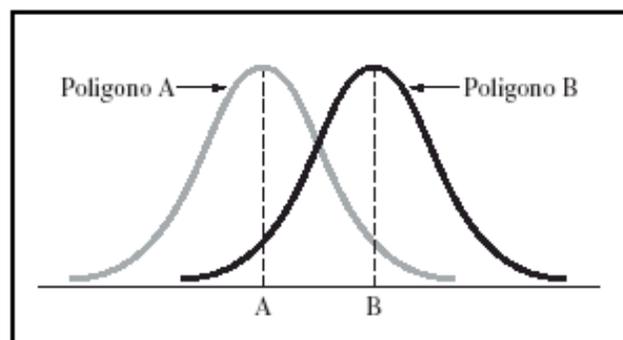
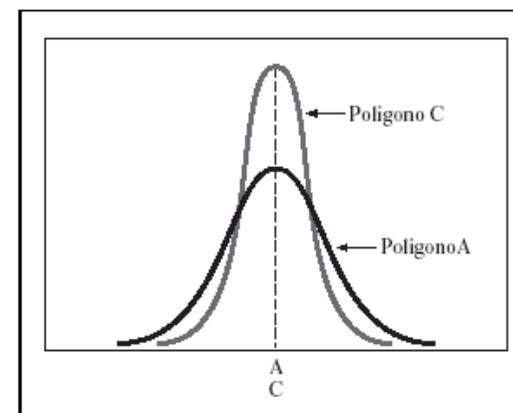


FIGURA 3.4

Due distribuzioni simmetriche a forma campanulare che differiscono solo nella variabilità



Tipi di misure di variabilità:

- **Campo di variazione (Range)**
- **Differenza interquartilica (Range Interquartile)**
- **Scarto Semplice Medio**
- **Scarto Quadratico Medio**
- **Varianza**
- **Devianza**

Misure di variabilità: il Range

Il **range** (o campo di variazione) è la differenza tra l'osservazione più grande e quella più piccola in un insieme di dati.

Il range

Il range è uguale all'osservazione più grande meno quella più piccola.

È importante sottolineare che il range deve assumere sempre valori maggiori di zero.

Quindi se la quantità $(X_{\text{più grande}} - X_{\text{più piccola}})$ risulta minore di zero, dobbiamo considerarne l'opposto, $-(X_{\text{più grande}} - X_{\text{più piccola}})$. In definitiva quindi:

$$\text{Range} = |X_{\text{più grande}} - X_{\text{più piccola}}| \quad (3.7)$$

(Quando inseriamo un certo valore tra le due barrette, $| \cdot |$, significa che stiamo considerando il *valore assoluto* della quantità considerata; ad esempio, $|3| = 3$ e $|-3| = 3$).

La serie ordinata è

10.0 20.6 28.6 28.6 29.4 29.5 29.9 30.1 30.5 30.5 32.1 32.2 32.4 33.0 35.2 37.1 38.0

Per questi dati il *range* è $38.0 - 10.0 = 28.0$.

NOTA: un limite del range consiste nel fatto che non tiene conto di come i dati si distribuiscono effettivamente tra il valore più piccolo e quello più grande.

Per questo motivo, in presenza di osservazioni estreme, risulta una misura inadeguata della variabilità.

Misure di variabilità: il Range Interquartile

Il **range (o differenza) interquartile** è la differenza tra il terzo e il primo quartile in un insieme di dati.

Il range interquartile

Il range interquartile si ottiene sottraendo al terzo quartile il primo quartile. Anche il range interquartile deve essere sempre maggiore di zero. Quindi:

$$\text{Range interquartile} = | Q_3 - Q_1 | \quad (3.8)$$

NOTA: Questa misura di variabilità sintetizza la dispersione del 50% delle osservazioni che occupano le posizioni centrali, e non è pertanto influenzata da valori estremi.

La serie ordinata è

10.0 20.6 28.6 28.6 29.4 29.5 29.9 30.1 30.5 30.5 32.1 32.2 32.4 33.0 35.2 37.1 38.0

Per questi dati sappiamo già dall'esempio 3.8 che $Q_1 = 29.0$ e $Q_3 = 32.7$. In base all'equazione (3.8) abbiamo:

$$\text{Range interquartile} = 32.7 - 29.0 = 3.7$$

L'intervallo compreso tra i due quartili 29 e 32.7 racchiude il 50% delle osservazioni centrali. L'ampiezza di tale intervallo, 3.7, racchiude i rendimenti percentuali annui conseguiti dal *gruppo centrale* dei 17 fondi comuni azionari che prelevano le commissioni dalle attività del fondo

Misure di variabilità: lo Scarto Semplice medio

Sebbene il range sia una misura della dispersione totale e il range interquartile della dispersione centrale, nessuna di queste due misure tiene conto di come le osservazioni si distribuiscano o si concentrino intorno a una misura di tendenza centrale, come ad esempio la media.

Lo SCARTO SEMPLICE MEDIO considera gli scarti in valore assoluto.

$$\delta = \frac{\sum_{i=1}^N |x_i - \mu|}{N} \quad \text{semplice}$$

$$\delta = \frac{\sum_{i=1}^s |x_i - \mu| n_i}{N} \quad \text{ponderata}$$

Misure di variabilità:

Scarto quadratico medio, Varianza e Devianza.

Varianza, la sua radice quadrata (*Scarto Quadratico Medio*) e il suo numeratore (*Devianza*) sintetizzano la dispersione dei valori osservati attorno alla loro media.

VARIANZA

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} \quad \text{semplice} \qquad \sigma^2 = \frac{\sum_{i=1}^s (x_i - \mu)^2 n_i}{N} \quad \text{ponderata}$$

SCARTO QUADRATICO MEDIO

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} \quad \text{semplice} \qquad \sigma = \sqrt{\frac{\sum_{i=1}^s (x_i - \mu)^2 n_i}{N}} \quad \text{ponderata}$$

DEVIANZA

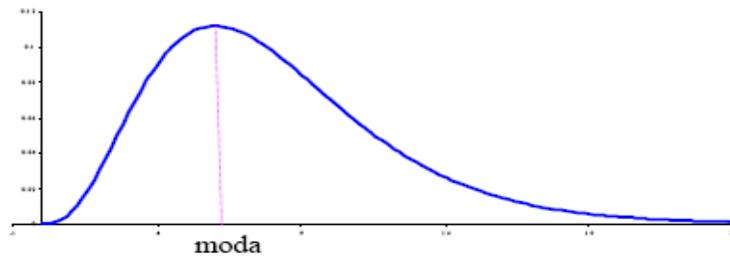
$$Dev(x) = \sum_{i=1}^N (x_i - \mu)^2 \quad \text{semplice} \qquad Dev(x) = \sum_{i=1}^s (x_i - \mu)^2 n_i \quad \text{ponderata}$$

ASIMMETRIA

- **La terza caratteristica dei dati che prendiamo in considerazione è la forma della loro distribuzione, cioè il modo in cui si distribuiscono.**
- **La distribuzione dei dati può essere simmetrica o meno.**
- **Se la distribuzione dei dati non è simmetrica, si dice asimmetrica oppure obliqua.**

- **Tipi di misure di forma:**
- **Asimmetria**

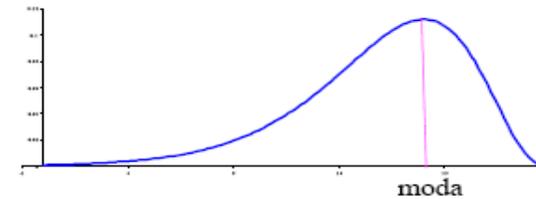
Asimmetria positiva



Moda \leq mediana \leq media

Addensamento delle frequenze nei valori più **bassi** di X

Asimmetria negativa

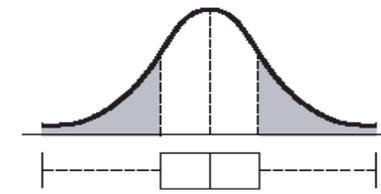
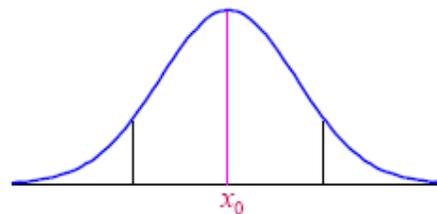


media \leq Moda \leq mediana

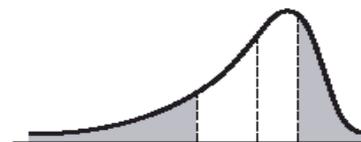
Addensamento delle frequenze nei valori più **alti** di X

Distribuzione simmetrica unimodale

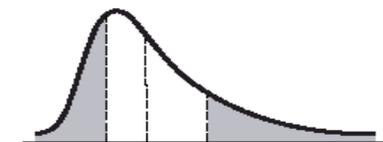
1. Moda = x_0
 2. Media = x_0
 3. Mediana = x_0
- } Moda = Media = Mediana



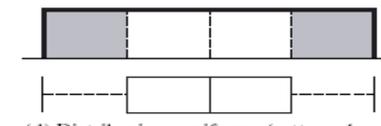
(a) Distribuzione simmetrica a campana



(b) Distribuzione obliqua a sinistra



(c) Distribuzione obliqua a destra



(d) Distribuzione uniforme (rettangolare)

MISURE DELL'ASIMMETRIA

Indice di Asimmetria S_k di Pearson

$$S_k = \frac{\mu - Mo}{\sigma} \quad \text{oppure}$$

$S_K > 0$ asimmetria positiva

$$S_k = \frac{3(\mu - Me)}{\sigma}$$

$S_K < 0$ asimmetria negativa

Coefficiente di Asimmetria

$$\gamma_1 = \frac{\sum_{i=1}^N (x_i - \mu)^3}{N\sigma^3} \quad \text{nel caso di una serie}$$

$$\gamma_1 = \frac{\sum_{i=1}^s (x_i - \mu)^3 n_i}{N\sigma^3} \quad \text{nel caso di una distribuzione}$$

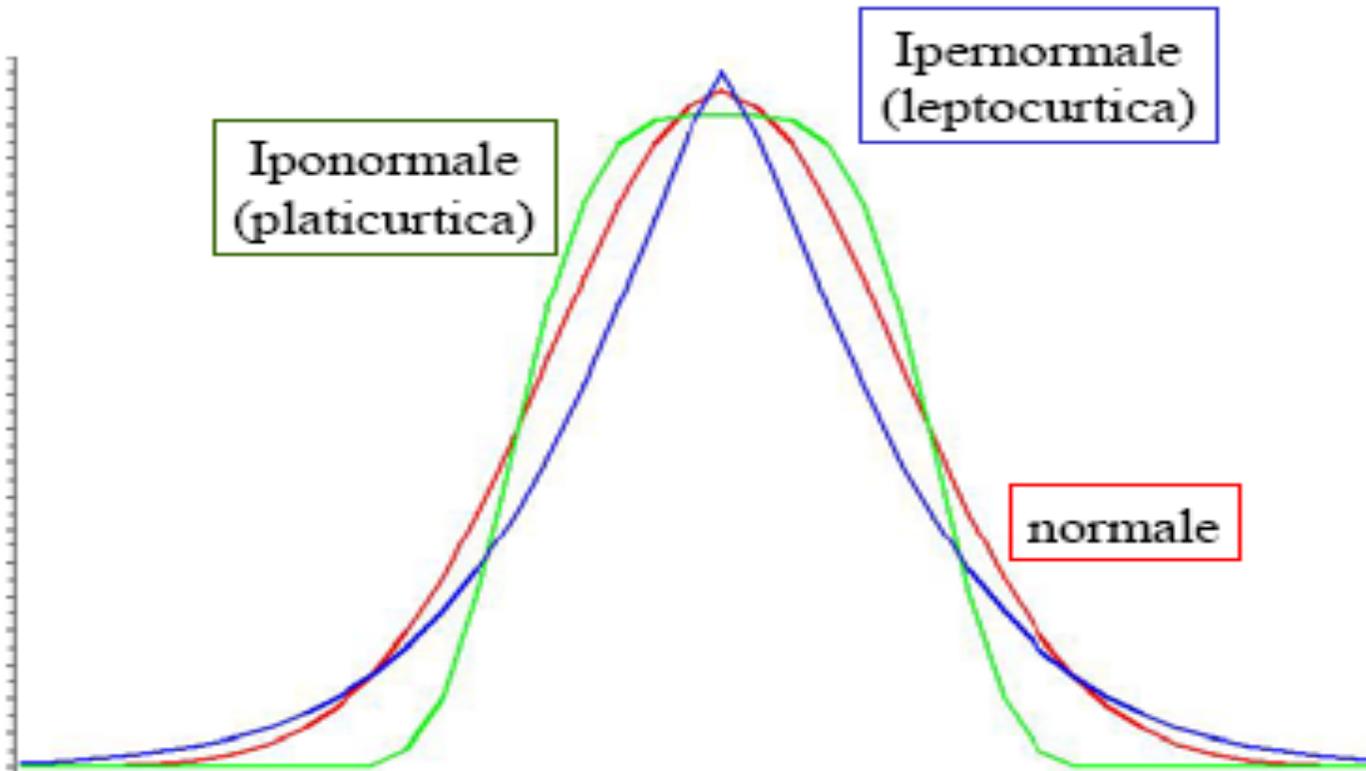
$\gamma_1 > 0$ asimmetria positiva

$\gamma_1 < 0$ asimmetria negativa

DISNORMALITA' O CURTOSI

CURTOSI

Per distribuzioni simmetriche, valuta il peso nelle code rispetto alla distribuzione normale con medesima media e ds



MISURE DELLA DISNORMALITA'

COEFFICIENTE DI CURTOSI DI PEARSON

$$\gamma_2 = \frac{\sum_{i=1}^N (x_i - \mu)^4}{N\sigma^4} - 3$$

nel caso di una serie

$$\gamma_2 = \frac{\sum_{i=1}^s (x_i - \mu)^4 n_i}{N\sigma^4} - 3$$

nel caso di una distribuzione

$\gamma_2=0$ distribuzione normale

$\gamma_2>0$ distribuzione ipernormale

$\gamma_2<0$ distribuzione iponormale

INDICE DI DISNORMALITA' DI GINI

$$I = \frac{2\sigma^2}{\delta^2} - \pi$$

$I=0$ distribuzione normale

$I>0$ distribuzione ipernormale

$I<0$ distribuzione iponormale

Indipendenza in Generale(1)

- In matematica si dice che la variabile y non dipende dalla variabile x quando essa rimane costante al variare dei valori assunti da x .
- Nel caso contrario si dice che la y dipende ed è funzione di x .

Indipendenza in Generale (2)

- **Partendo da questa definizione, è immediato stabilire se c'è INDIPENDENZA tra due V.S. espresse da una serie di coppie di valori.**
- **Nel caso di TABELLE A DOPPIA ENTRATA, perché ci sia indipendenza si deve verificare:**

$$\eta_{ih} = \frac{\eta_{io} \cdot \eta_{oh}}{N}$$

**Per tutte le caselle della tabella.
L'INDIPENDENZA E' RECIPROCA**

TABELLA A DOPPIA ENTRATA

Variabile X	Variabile Y						Totale
	y₁	y₂	..	y_h	..	y_t	
x₁	n₁₁	n₁₂	..	n_{1h}	..	n_{1t}	N₁₀
x₂	n₂₁	n₂₂	..	n_{2h}	..	n_{2t}	N₂₀
..
x_i	n_{i1}	n_{i2}	..	n_{ih}	..	n_{it}	N_{i0}
..
x_s	n_{s1}	n_{s2}	..	n_{sh}	..	n_{st}	N_{s0}
Totale	N₀₁	N₀₂	..	N_{0h}	..	N_{0t}	N

Indipendenza in media (1)

- **La moderna metodologia ha grandemente valorizzato la sostituzione delle distribuzioni parziali con i corrispondenti valori medi.**
- **Quando ricorrono le condizioni, si sarebbe indotti a ritenere che la relazione tra x e y farebbe corrispondere ad x il valore medio \bar{y}_i delle y_h**

dato da
$$\bar{y} = \frac{\sum_{h=1}^t y_h \eta_{ih}}{\eta_{io}} \quad i = 1, 2, \dots, s$$

...

Indipendenza in media (2)

...quindi se $\bar{y}_1 = \bar{y}_2 = \dots = \bar{y}_s$

si ha INDIPENDENZA IN MEDIA DI Y DA X.

Così anche considerando le distribuzioni parziali di x

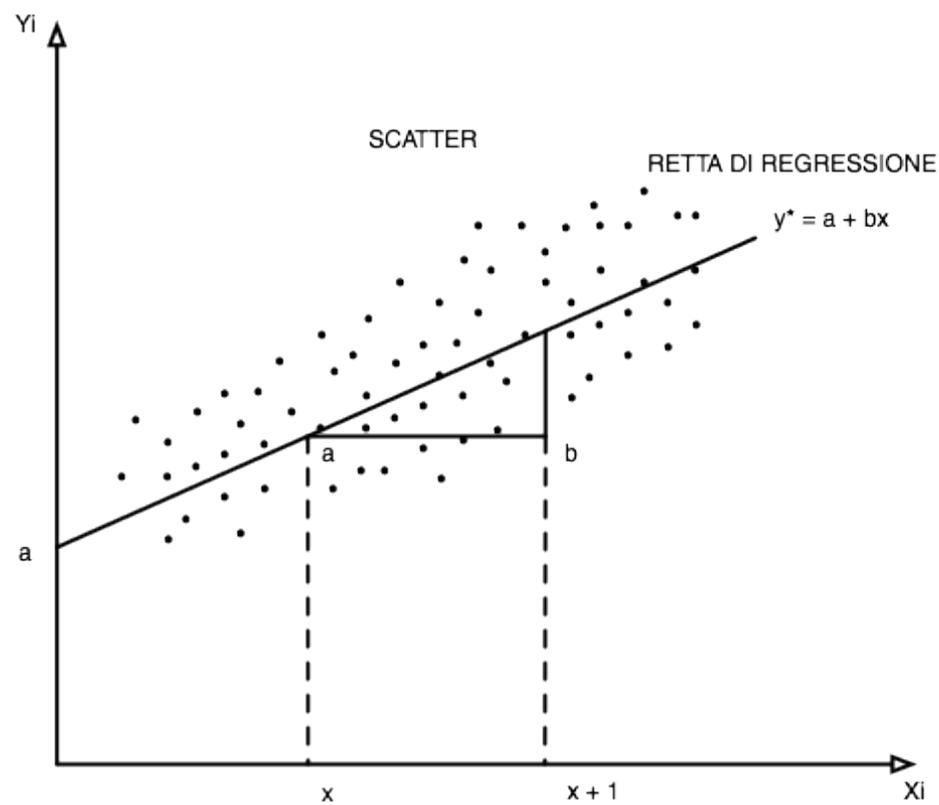
dati da $\bar{x}_h = \frac{\sum_{i=1}^s x_i \eta_{ih}}{\eta_{oh}} \quad h = 1, 2, \dots, t$

se $\bar{x}_1 = \bar{x}_2 = \dots = \bar{x}_z$ si ha INDIPENDENZA IN MEDIA di X da Y.

Analisi della Dipendenza

- Nel caso in cui tra i due caratteri esiste una relazione unidirezionale nel senso che X indica il CARATTERE ANTECEDENTE o INDIPENDENTE e Y il CARATTERE CONSEGUENTE o DIPENDENTE si parla di DIPENDENZA DEL CARATTERE Y DAL CARATTERE X.

Retta di REGRESSIONE



a = valore teorico del carattere y allorché $x = 0$
b = COEFFICIENTE DI REGRESSIONE

Minimi quadrati

Il calcolo dei parametri della RETTA DI REGRESSIONE si effettua con il metodo dei MINIMI QUADRATI; in altre parole si sceglie la retta per la quale la somma dei quadrati degli scostamenti tra i valori teorici e quelli osservati del carattere y sia minima:

$$\sum_{i=1}^N (y_i^* - y_i)^2 = \text{MINIMO}$$

ossia:
$$\sum_{i=1}^N (a + bx_i - y_i)^2 = \text{MINIMO}$$

...

Parametri della Regressione

risolvendo abbiamo:

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

$$b = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2} = \frac{CODEV(xy)}{DEV(x)}$$

IL COEFFICIENTE DI REGRESSIONE ESPRIME DI QUANTO, IN MEDIA, VARIA IL CARATTERE DIPENDENTE ALL'AUMENTARE DI UNA UNITA' DEL CARATTERE INDIPENDENTE.

Indice di Determinazione

$$DEV(y) = \sum_{i=1}^N (y_i - \bar{y})^2 \quad \text{DEVIANZA TOTALE}$$

$$DEV(R) = \sum_{i=1}^N (y_i^* - \bar{y})^2 \quad \text{DEVIANZA DI REGRESSIONE}$$

$$DEV(E) = \sum_{i=1}^N (y_i - y_i^*)^2 \quad \text{DEVIANZA RESIDUA}$$

$$R^2 = \frac{DEV(R)}{DEV(Y)} = \frac{[CODEV(x, y)]^2}{DEV(x)DEV(y)} \quad 0 \leq R^2 \leq 1$$

L'INDICE DI DETERMINAZIONE, esprime QUANTA PARTE DELLA DEVIANZA TOTALE DI Y E' DETERMINATA o SPIEGATA DALLA RETTA DI REGRESSIONE

Assume valore 0 quando la Devianza di Regressione è nulla ($b=0$)

Assume valore 1 quando la Devianza Residua è nulla ossia quando i punti sono allineati e giacciono sulla retta di Regressione.

Analisi della Interdipendenza (1)

Nei casi in cui non è possibile individuare qual è il carattere antecedente e qual è il carattere conseguente o perché i due caratteri si influenzano reciprocamente si studia la relazione simmetrica ossia l'INTERDIPENDENZA mediante il:

COEFFICIENTE DI CORRELAZIONE LINEARE (Bravais - Pearson).

$$r = \frac{CODEV(X,Y)}{\sqrt{DEV(x) \cdot DEV(y)}} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2 \cdot \sum_{i=1}^N (y_i - \bar{y})^2}}$$

$$-1 \leq r \leq 1$$

...

Analisi della Interdipendenza (2)

**$r = -1$ PERFETTA RELAZIONE LINEARE DECRESCENTE
(DISCORDANZA PERFETTA)**

**$r = +1$ PERFETTA RELAZIONE LINEARE CRESCENTE
(CONCORDANZA PERFETTA)**

$r = 0$ INDIFFERENZA

Diagramma di dispersione e indice di correlazione

