

# Capitolo 1

## Statistica descrittiva

Quando si raccolgono dei dati su una popolazione o su un campione, i valori ottenuti si presentano come un insieme di dati disordinati. I dati che non sono stati organizzati, sintetizzati o elaborati sono chiamati **dati grezzi**. A meno che il numero di dati sia piccolo, e' improbabile che i dati grezzi forniscano qualche informazione finche' non siano ordinati in qualche modo. Con il termine *statistica descrittiva* si intende una raccolta di metodi, strumenti matematici atti ad organizzare serie di dati per evidenziare in forma sintetica simmetrie, periodicit , leggi di altro genere ovvero di descrivere in maniera immediatamente comprensibile le informazioni dagli stessi dati.

Le tecniche di organizzazione dei dati variano in funzione dei modi di presentarsi delle caratteristiche degli elementi su cui e' svolta l'indagine. Si parla di caratteri qualitativi quando essi sono dei dati di natura non numerica, mentre si parla di caratteri quantitativi quando essi sono delle grandezze numeriche. I caratteri di tipo quantitativo si distinguono in *discreti* se assumono un numero limitato di valori e *continui* quando assumono qualsiasi valore reale in un certo intervallo.

## 1.1 Distribuzioni di frequenza

Sia  $E = \{x_1, x_2, \dots, x_n\}$  un insieme di dati di numerosità  $n$ .

### 1.1.1 Carattere numerico discreto

Quando il carattere è di tipo discreto, ed i valori assumibili dal carattere sono in numero limitato, è conveniente raggruppare i dati considerando l'insieme di tutti i valori assumibili, chiamati **modalità del carattere**. Sia  $N$  il numero dei valori assumibili dai dati dell'insieme  $E$  allora denoto con  $S = \{s_1, s_2, \dots, s_N\}$  l'insieme delle modalità. Definiamo :

i) **frequenza assoluta**

$$f_j = n^\circ \text{ elementi di } E \text{ con valore } S_j \quad j = 1, 2, \dots, N$$

ii) **frequenza relativa**

$$p_j = \frac{f_j}{n} \quad n = \text{numerosità di } E$$

iii) **Frequenza cumulata assoluta**

$$F_j = \sum_k f_k \quad \text{con } k \text{ tale che } S_k \leq S_j \quad j = 1, 2, \dots, N$$

iv) **Frequenza cumulata relativa**

$$P_j = \sum_k p_k \quad \text{con } k \text{ tale che } S_k \leq S_j \quad j = 1, 2, \dots, N$$

Le frequenze cumulate permettono di raggruppare i dati in modo da capire quando essi sono *minori di* o *maggiori di*.

Tabella 1.1: Numero stanze in un campione di appartamenti.

3	4	2	6	5	2	4	4
2	5	4	4	5	7	5	4
5	7	8	4	3	6	2	3
5	2	7	2	4	8	4	2
6	5	4	4	6	5	3	3
8	5	2	5	6	5	5	4
2	6	4	5	5	7	3	4
3	3	3	4	4	3	4	6
4	3	7	4	4	6	4	2
4	4	6	3	2	3	5	4

**Esempio 1.1** Nella tabella 1.1 sono elencati i dati relativi al numero di stanze possedute da 80 appartamenti scelti a caso tra tutti quelli che si trovano in un determinato quartiere di una data città.

**Esercizio 1.1** Dai dati della tabella 1.1 creare una tabella come la seguente

$n^\circ$ stanze	freq. ass.	freq. rel	freq. cum. ass.	freq. cum. rel.
1				
2				
3				
4				
5				
6				
7				
8				

Utilizzare Excel (funzione conta.se) e fare gli istogrammi.

### 1.1.2 Carattere numerico continuo

Quando il carattere da studiare e' di tipo continuo, o discreto con un numero elevato di modalita', allora conviene considerare dei sottoinsiemi di  $S$ . Chiamo **classe**  $C$  un sottoinsieme di  $S$ . Chiamo **partizione** di  $S$  ogni famiglia di classi tra loro disgiunte la cui unione sia  $S$  ovvero

$$C_i \subseteq S \quad i = 1, \dots, k \in N$$

$$C_i \cap C_j = \phi \quad \forall i \neq j$$

$$\bigcup_{i=1}^k C_i = S$$

Il modo di scegliere le classi non e' unico. Ad ogni classe si associano diverse grandezze che le caratterizzano:

- i *confini superiore e inferiore*, che sono i valori estremi della classe (aperti, chiusi)
- l'*ampiezza* che e' la differenza tra il confine superiore ed inferiore;
- il *valore centrale* che e' la media tra i due confini.

Di solito le classi hanno tutte la stessa ampiezza. Troppe classi rendono la tabella poco leggibile: il loro numero e' solitamente compreso tra 5 e 15. Diamo delle semplici regole pratiche

- scegliere un numero di classi  $k$ , approssimativamente uguale alla radice quadrata del numero dei dati, cioe'  $k \simeq \sqrt{n}$ .
- scegliere l'ampiezza delle classi  $a = R/k$ , dove  $R$  e' il campo di variazione dei dati, ovvero la differenza tra il valore massimo e quello minimo assunti dai dati.

Tabella 1.2: Costo mq di un campione di appartamenti.

2,11	3,08	2,35	3,54	0,44	2,24	4,60	1,88
2,08	1,90	2,15	5,11	3,69	0,88	2,56	4,00
3,15	3,67	3,15	4,09	4,57	1,06	2,05	2,34
4,17	4,10	4,75	1,90	2,36	0,90	2,07	3,23
4,21	2,12	1,21	2,10	4,05	5,42	0,85	4,80
2,11	5,08	2,78	4,88	1,11	1,83	1,85	2,87
2,23	3,20	2,80	2,19	1,88	2,16	2,74	2,45
1,19	3,79	1,24	3,06	2,11	3,70	2,91	1,80
3,48	4,10	3,13	0,90	3,07	4,10	1,66	2,88
2,11	1,90	1,18	0,75	1,60	3,85	1,45	2,00

**Esempio 1.2** Nella tabella 1.2 sono elencati i dati relativi al costo al metro quadro (in migliaia di Euro) di 80 appartamenti scelti a caso tra quelli che si trovano in un quartiere di una città italiana.

**Esercizio 1.2** Riferendosi alla tabella 1.2 raggruppare i dati in classi e calcolare le frequenze

Volendo raggruppare in classi i dati riportati in questa tabella, la prima cosa che occorre fare è osservare quali sono i valori minimo e massimo in essa riportati. Essendo questi 0,44 e 5,42, possiamo arbitrariamente pensare all'insieme  $S$  dei valori assumibili come al sottoinsieme  $[0.40, 5.50] \subset \mathbb{R}$ . Se vogliamo suddividere  $S$  in 5 classi, potremmo ad esempio scegliere le seguenti:

$$C_1 = (0.40, 1.50], \quad C_2 = (1.50, 2.30], \quad C_3 = (2.30, 3.00], \quad C_4 = (3.00, 4.00], \quad C_5 = (4.00, 5.50]$$

Quindi creare una matrice con

costo mq	freq. ass.	freq. rel	freq. cum. ass.	freq. cum. rel.
$C_1$				
$C_2$				
$C_3$				
$C_4$				
$C_5$				

Creare istogrammi con Excel.

Tabella 1.3: Cause malfunzionamento macchina

fluttuazioni di tensione	6
instabilita' del sistema di controllo	22
errore dell'operatore	13
strumento usurato e non sostituito	2
altre cause	5
Totale	48

### 1.1.3 Carattere non numerico

Si cerca di raggruppare i dati in classi che non sono insiemi numerici e che formino una partizione dell'insieme.

**Esempio 1.3** *In uno stabilimento vengono registrati i casi di malfunzionamento di una macchina controllata dal computer, e le loro cause. I dati relativi ad un certo mese sono nella tabella 1.3*

**Esercizio 1.3** *Raggruppare i dati della tabella 1.3 (carattere non numerico) in classi.*

I dati della tabella 1.3 sono gia' raggruppati in classi dove

$$C_1 = \{\text{fluttuazioni tensione}\} \quad , \quad C_2 = \{\text{instabilita'}\} \quad , \quad \dots\dots\dots$$

*Quindi creare la matrice e graficare i dati con Excel.*

Classe	freq. ass.	freq. rel	freq. perc.
$C_1$			
$C_2$			
$C_3$			
$C_4$			
$C_5$			

### 1.1.4 Rappresentazioni grafiche

Hanno lo scopo di fornire immediatamente a chiunque le caratteristiche essenziali del fenomeno oggetto dell'indagine.

- Tali rappresentazioni sono basate essenzialmente su una proporzionalità fra frequenze (assolute o relative) e grandezze geometriche (aree o lunghezze) che vengono utilizzate per rappresentare il fenomeno.
- Non esistono regole fisse generali per la scelta della rappresentazione grafica con cui sintetizzare una distribuzione. L'importante è che venga assicurata l'immediata percezione del fenomeno in esame.
- Le rappresentazioni grafiche possono aiutare a scoprire relazioni fra le caratteristiche di distribuzioni.

#### Rappresentazioni grafiche di caratteri qualitativi

- Diagramma a settori circolari o a torta in cui, a ciascuna modalità  $x_i$  si associa un settore circolare avente area proporzionale alla sua frequenza  $f_i$ .
- Diagramma a barre o canne d'organo in cui a ciascuna modalità  $x_i$  si associa un rettangolo avente base costante ed un'altezza proporzionale alla frequenza  $f_i$ .
- Diagrammi figurativi in cui si utilizzano delle figure per rappresentare la distribuzione in esame: ciascuna figura rappresenta una modalità e la sua dimensione è proporzionale alla sua frequenza.

#### Rappresentazioni grafiche di caratteri quantitativi

- Istogramma (caratteri continui): plurirettangolo avente basi proporzionale all'ampiezza delle classi e aree proporzionali alla frequenza.



Nota: Poiche' nell'istogramma le aree dei singoli rettangoli sono proporzionali alle frequenze delle rispettive classi, l'altezza  $h_i$  del rettangolo della classe i-esima deve essere proporzionale al rapporto fra la frequenza della classe e la corrispondente ampiezza:

$$h_i \propto \frac{f_i}{x_{i+1} - x_i}$$

In particolare, la quantita'  $h_i$  e' la frequenza specifica della classe i-esima.

- Diagrammi a segmenti (caratteri discreti): grafico cartesiano in cui, in corrispondenza di ciascuna modalita'  $x_i$ , si riporta un segmento di altezza proporzionale alla corrispondente frequenza relativa ( $f_i$ ) oppure frequenza assoluta ( $n_i$ )
- Diagrammi cartesiani dove si esprime la dipendenza temporale (serie storiche) di un fenomeno quantitativo (asse delle ordinate) in funzione del tempo (asse delle ascisse).
- Diagrammi a radar: Questa rappresentazione grafica consiste di una sequenza di raggi che hanno origine da un centro e formano angoli uguali tra loro; ogni raggio rappresenta una delle variabili. La distanza dal centro del punto marcato sul raggio e' proporzionale al valore della variabile rispetto al valore massimo raggiungibile. I punti sui raggi vengono congiunti con segmenti, cosı che il grafico ha la forma di una stella o di una ragnatela.

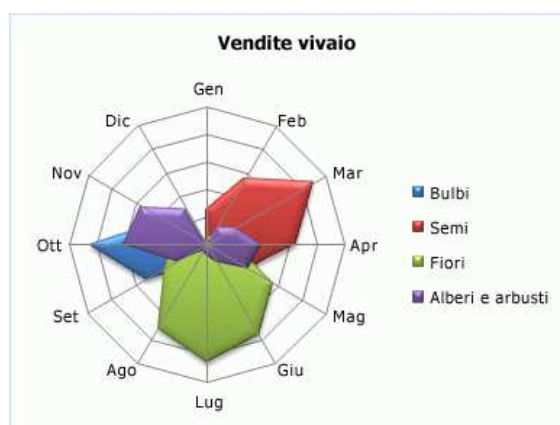


Figura 1.1: Diagramma a radar delle vendite di un vivaio

Supposto di aver ordinato i dati, possiamo ottenere dei grafici utilizzando dei software di tipo statistico. Il piu' comune e' il **foglio di calcolo elettronico**, disponibile nei pacchetti Office and OpenOffice. Si possono utilizzare istogrammi, diagrammi a torta, grafici di dispersione, radar.

**Esercizio 1.4** Supponiamo di aver rilevato il numero di incidenti avvenuti in una citta' durante una giornata e di avere ottenuto la tabella

ora	1	2	3	4	5	6	7	8	9	10	11	12
incidenti	3	2	1	0	1	3	5	10	8	4	4	6

ora	13	14	15	16	17	18	19	20	21	22	23	24
incidenti	7	3	3	2	3	5	6	5	4	6	4	3

Creare radar con Excel.

**Esercizio 1.5** Nella tabella che segue si riporta per ogni corso di Laurea il n. di studenti iscritti e quelli laureati

<i>corso di laurea</i>	<i>studenti iscritti</i>	<i>laureati</i>	<i>freq. percentuale</i>
<i>1 - scientifico</i>	<i>183300</i>	<i>15539</i>	
<i>2 - medicina</i>	<i>72107</i>	<i>7407</i>	
<i>3 - economia</i>	<i>457248</i>	<i>35272</i>	
<i>4 - scienze giuridiche</i>	<i>319068</i>	<i>18839</i>	
<i>5 - lettere</i>	<i>376446</i>	<i>27128</i>	
<i>6 - ingegneria</i>	<i>276345</i>	<i>7128</i>	

Creare diagramma a torta con Excel.

**Esercizio 1.6** Nella tabella che segue si riportano le aree dei continenti del mondo

<i>Continente</i>	<i>Area ( 1000 Km<sup>q</sup>)</i>
<i>Europa</i>	<i>10368</i>
<i>Asia</i>	<i>45078</i>
<i>Africa</i>	<i>30209</i>
<i>America sett. e centrale</i>	<i>24203</i>
<i>America merid.</i>	<i>17855</i>
<i>Oceania</i>	<i>8522</i>
<i>Antartide</i>	<i>14108</i>

Creare istogramma e diagramma a torta con Excel.

## 1.2 Indici di tendenza centrale

Gli indici di tendenza centrale sono delle misure in grado di sintetizzare con un solo valore numerico, i valori assunti dai dati.

### i) Media e Media pesata

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1.1)$$

$$\bar{x}_p = \frac{1}{p} \sum_{i=1}^n p_i x_i \quad , p_i \text{ sono i pesi} \quad p = \sum_{i=1}^n p_i \quad (1.2)$$

**Proposizione 1.1** *La media rende minima la funzione*

$$f(x) = \sum_{i=1}^n (x - x_i)^2$$

*ovvero la media e' quel punto che dista di meno da tutti i punti della serie di dati  $\{x_i\}$ .*

*Dim.*

$$f'(x) = 2 \sum_{i=1}^n (x - x_i) = 0 \quad \Rightarrow \quad \sum_{i=1}^n x - \sum_{i=1}^n x_i = 0 \quad \Rightarrow \quad n x = \sum_{i=1}^n x_i$$

da cui si ricava

$$x = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}.$$

Bisogna adesso dimostrare che in  $\bar{x}$   $f(x)$  è minima.

$$f''(x) = 2 \sum_{i=1}^n 1 = 2n > 0$$

allora  $\bar{x}$  è l'unico punto di minimo di  $f(x)$ . ■

## ii) Mediana ( $\hat{x}$ )

È quel numero che “sta nel mezzo”. Ovvero, ordinati i dati  $\{x_i\}$  in ordine crescente allora se  $n$  è dispari la mediana è l’elemento di posto

$$\frac{n+1}{2} .$$

Se  $n$  è pari la mediana non è univocamente determinata. Una possibile approssimazione si ottiene prendendo la **media aritmetica dei due valori centrali**, ovvero quelli di posto

$$\frac{n}{2} , \quad \frac{n}{2} + 1 .$$

Un’altra approssimazione si ottiene utilizzando un’interpolazione lineare a partire dai due valori centrali (vedi esempio).

*svantaggio*: risistemare i dati in ordine crescente (non nella media)

*vantaggio*: non dipende dai valori estremi

## iii) Moda

È quel valore che si ripete più volte nella serie di dati.

Dà un andamento qualitativo dei dati e non è garantito che sia un unico numero (distr. bi o multi-modale).

## iv) Quantili, percentili, Quartili

Supponiamo di avere un insieme di dati ordinati in modo crescente. Abbiamo già visto che la mediana è il *valore che sta nel mezzo*. In analogia possiamo definire

### Definizione 1.1

Si chiama quantile di ordine  $\alpha \in [0, 1]$ , e lo si indica con  $q_\alpha$ , un valore per cui alla sua sinistra compare almeno il  $100 \alpha\%$  delle osservazioni e alla sua destra almeno il  $100 (1 - \alpha)\%$  .

Alle volte, si usa il termine percentile, al posto di quantile, in questo caso  $\alpha$  è indicata come percentuale. Per esempio  $q_{0.95}$  è il novantacinquesimo percentile.

## Definizione 1.2

Si dicono primo quartile, secondo quartile e terzo quartile, e si indicano con  $Q_1, Q_2, Q_3$  i quantili, rispettivamente, di ordine 0.25, 0.5, 0.75 e quindi

$$Q_1 = q_{0.25} \quad , \quad Q_2 = q_{0.50} \quad , \quad Q_3 = q_{0.75}$$

Il secondo quartile coincide con la mediana.

### Calcolo dei quartili

Come già detto per la mediana, anche i quartili ed i percentili non sono univocamente determinati. Vediamo due modi per calcolarli:

#### 1. *media aritmetica*

In analogia con quanto già visto per il calcolo della mediana

- fissiamo  $\alpha = 0.25, 0.5, 0.75$  e calcoliamo  $\alpha(n+1)$
- se  $\alpha(n+1) = m \in \mathbb{N}$  allora  $Q_\alpha = x_m$
- se  $\alpha(n+1) \notin \mathbb{N}$  allora si prende la sua parte intera <sup>1</sup>, che è quel numero  $m \in \mathbb{N}$  tale che  $m < \alpha(n+1) < m+1$  e quindi, la media aritmetica

$$Q_\alpha = \frac{x_m + x_{m+1}}{2} \tag{1.3}$$

#### 2. *interpolazione lineare*

Supponiamo di avere due coppie di dati  $P_1(x_1, y_1)$ ,  $P_2(x_2, y_2)$  e di voler stimare il valore  $y^*$  che corrisponde a  $x^* \in ]x_1, x_2[$ . Con il metodo dell'interpolazione lineare si stima il valore  $y^*$  tramite la retta passante per i due punti  $P_1$  e  $P_2$  (detta retta interpolante), ovvero

$$y^* = y_1 + \frac{y_2 - y_1}{x_2 - x_1} (x^* - x_1) \quad . \tag{1.4}$$

Utilizziamo questo metodo per stimare la mediana:

---

<sup>1</sup>la parte intera di un numero  $x$  si indica con  $[x]$ , mentre la sua parte frazionaria è  $x - [x]$ .

- fissiamo  $\alpha = 0.25, 0.5, 0.75$  e calcoliamo  $\alpha(n+1)$
- se  $\alpha(n+1) = m \in \mathbb{N}$  allora  $Q_\alpha = x_m$
- se  $x^\star = \alpha(n+1) \notin \mathbb{N}$  allora utilizziamo la retta interpolante (1.4).

Sia  $m = [x^\star]$  (parte intera) e  $\beta = x^\star - [x^\star] = x^\star - m$  (parte frazionaria), allora  $P_1(m, x_m)$ ,  $P_2(m+1, x_{m+1})$  e quindi avremo

$$y^\star = Q_\alpha = x_m + \frac{x_{m+1} - x_m}{m+1 - m}(x^\star - m) = x_m + (x_{m+1} - x_m)\beta \quad (1.5)$$

In EXCEL la funzione QUARTILE, viene calcolata usando un' opportuna interpolazione lineare.

**Esempio 1.4** *Siano assegnati i seguenti dati, che rappresentano le età di un campione di persone*

$$E = \{x_i, i = 1 : 18\} = \{16, 18, 18, 19, 20, 20, 20, 20, 21, 21, 21, 22, 23, 25, 28, 30, 31, 37\}$$

La numerosità del campione è  $n=18$ . Si ottiene come media 18, mediana 21 e moda 20 (calcolare con EXCEL ed anche la distribuzione delle frequenze).

Per il calcolo dei quartili si ha:

1. Per  $\alpha = 0.25$  (primo quartile)  $0.25(18+1) = 4.75$  quindi  $m = [4.75] = 4$  e la parte frazionaria  $\beta = 0.75$ . Usando la media aritmetica (1.3)

$$Q_{0.25} = \frac{x_4 + x_5}{2} = \frac{19 + 20}{2} = 19.5$$

mentre con l'interpolazione lineare (1.5)

$$Q_{0.25} = 19 + 0.75(20 - 19) = 19.75$$

2. Per  $\alpha = 0.5$  (secondo quartile o mediana)  $0.50(18+1) = 9.5$  quindi  $m = 9$  e  $\beta=0.5$ . Usando la (1.3)

$$Q_{0.5} = \frac{21 + 21}{2} = 21$$

che coincide con il valore ottenuto utilizzando la (1.5).

3. Per  $\alpha = 0.75$  (terzo quartile)  $0.75(18+1) = 14.25$  quindi  $m = 14$  e  $\beta=0.25$ . Usando la (1.3)

$$Q_{0.75} = \frac{25 + 28}{2} = 26,5$$

mentre con la (1.5) si ottiene

$$Q_{0.75} = 25 + 0.25(28 - 25) = 25,75$$

Da questa analisi possiamo concludere che il 25 % delle persone del campione hanno un'età minore o uguale a 19,5 anni, il 50 % minore o uguale a 21 e il 75 % minore o uguale a 26,5.



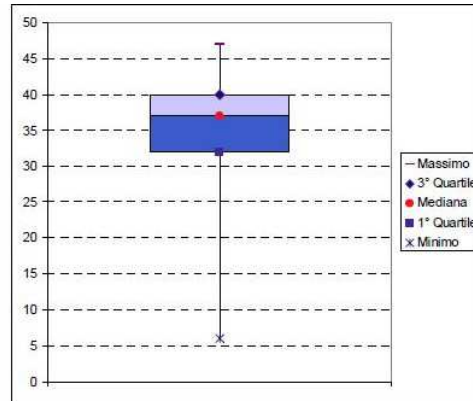


Figura 1.2: Il box plot

### 1.2.1 Il box plot

Il box plot o diagramma a scatola e baffi, e' un grafico, relativo a caratteri quantitativi, ottenuto a partire dai 5 numeri di sintesi

[minimo, I quartile (Q1), mediana, III quartile (Q3), massimo]

che descrive le caratteristiche salienti della distribuzione. Si ottiene riportando su un asse verticale (oppure orizzontale) i 5 numeri di sintesi. La scatola del box plot ha come estremi inferiore e superiore rispettivamente Q1 e Q3. La mediana divide la scatola in due parti. I baffi si ottengono congiungendo Q1 al minimo e Q3 al massimo. In alcuni grafici (ad esempio, quello ottenuto con il software SPSS) il baffo ha lunghezza pari a 1.5 volte l'altezza della scatola, data dalla distanza tra Q3 e Q1 - detto anche range interquartile; ovviamente è inferiore se il massimo valore osservato dista da Q3 meno di 1.5 volte il range interquartile. Confrontando tra loro le lunghezze dei due baffi (che rappresentano le distanze tra Q1 e il minimo e tra Q3 e il massimo) e le altezze dei due rettangoli che costituiscono la scatola (che rappresentano le distanze tra Q1 e mediana e tra mediana e Q3) si ottengono informazioni sulla simmetria della distribuzione: questa è tanto più simmetrica quanto le lunghezze dei baffi risultano simili tra loro e le altezze dei due rettangoli risultano simili tra loro. I baffi mettono inoltre in evidenza la presenza di eventuali outliers (osservazioni eccezionali).

## 1.3 Indici di variabilità

Può accadere che 2 serie di dati abbiano stessa media e/o mediana, ma le 2 serie sono molto diverse.

### Esempio 1.5

$$E_1 = \{0.5, 0.8, 2.0, 2.7, 4.0\} \quad \bar{x}_1 = \hat{x}_1 = 2$$

$$E_2 = \{1.4, 1.7, 2.0, 2.1, 2.8\} \quad \bar{x}_2 = \hat{x}_2 = 2$$

*Però i dati di  $E_2$  sono più omogenei (cioè vicini tra loro).*

Occorre pertanto definire indici che misurino il grado di variabilità o dispersione.

#### a) Varianza ( $s^2$ )

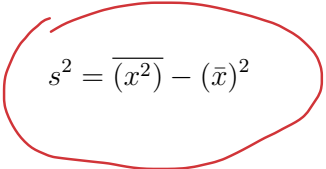
$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

ovviamente  $s^2$  è tanto più grande, tanto più i dati sono “distanti” dalla media.

### Esempio 1.6

$$s_{E_1}^2 = 1.64, \quad s_{E_2}^2 = 0.22 \quad (\text{omogeneo!})$$

### Proposizione 1.2


$$s^2 = \overline{(x^2)} - (\bar{x})^2$$

*Dim.*

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \tag{1.6}$$


$$= \frac{1}{n} \sum_{i=1}^n (x_i^2 + \bar{x}^2 - 2x_i\bar{x}) = \tag{1.7}$$

$$= \frac{1}{n} \left[ \sum_{i=1}^n x_i^2 + \sum_{i=1}^n \bar{x}^2 - 2 \sum_{i=1}^n x_i\bar{x} \right] = \tag{1.8}$$

$$= \frac{1}{n} \left[ \sum_{i=1}^n x_i^2 + n\bar{x}^2 - 2\bar{x} \sum_{i=1}^n x_i \right] = \tag{1.9}$$

$$= \overline{x^2} + (\bar{x}^2) - 2(\bar{x}^2) = \overline{x^2} - (\bar{x}^2) \quad \blacksquare \tag{1.10}$$

b) Scarto quadratico medio (Deviazione standard)

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$


c) Scarto medio assoluto

$$s.a. = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

Cosa fare se 2 serie di dati hanno la stessa varianza?

**Esempio 1.7**

$$E_3 = \{3, 4, 5, 6, 7\} \quad \bar{x}_3 = 5 \quad s_3^2 = 2.5$$

$$E_4 = \{13, 14, 15, 16, 17\} \quad \bar{x}_4 = 15 \quad s_4^2 = 2.5$$

*In entrambi i casi un dato rispetto al precedente varia di 1. Ma questa variazione è più importante nelle serie di dati  $E_3$  (che sono numeri più piccoli rispetto ad  $E_4$ ) che nella serie di dati  $E_4$ .*

È logico pensare che i dati di  $E_3$  siano più dispersi di quelli di  $E_4$ , anche perchè  $\bar{x}_4 > \bar{x}_3$ . Per questo motivo si definisce un **coefficiente di variazione** (c.v.)

$$c.v. = \frac{s}{\bar{x}}$$

Per  $E_3$   $c.v. = \frac{\sqrt{2.5}}{5}$  per  $E_4$   $c.v. = \frac{\sqrt{2.5}}{15}$ . A valori maggiori del c.v. corrisponde una maggiore variabilità dei dati.

**Esempio 1.8** Nella seguente tabella vengono riportati il prezzo (in EURO/LITRO) di un particolare combustibile e la frequenza con cui esso viene venduto giornalmente in un distributore.

<i>prezzi</i>	14.5	16.8	12.3	10.7	11.4	18.1	20.6	13.8
<i>frequenze</i>	7	5	8	12	10	6	4	11

Calcolare media, varianza, moda e mediana.

In questo caso:

- la media e' quella pesata (1.2) con  $x_i$  i *prezzi* ed i pesi  $s_i$  dati dalle *frequenze*.
- La varianza si calcola

$$s^2 = \frac{\sum_{i=1}^k (x_i - \bar{x}_p)^2 f_i}{\sum_{i=1}^k f_i}$$

dove  $k = 8$ .

- Per prima cosa si devono ordinare i dati in modo crescente nel prezzo. Sia  $N$  la somma delle frequenze. Essendo  $N = 63$  (dispari) allora la mediana e' elemento di posto  $64/2=32$  e quindi pari a 13.8

Tabella 1.4: Dati

classe	ampiezza della classe	freq. assoluta	freq. percentuale	altezza istogr.
110 - 130	20	20	10 %	0.5
130 - 150	20	40	20 %	1
150 - 170	20	60	30 %	1.5
170 - 210	40	80	40 %	1
	Tot.	200	100 %	

### 1.3.1 Dati raggruppati per classi

Supponiamo che gli  $n$  dati siano raggruppati in classi  $C_i$  ( $i = 1, \dots, k$ ), come nell'esempio 1.2. Ricordiamo che l'*ampiezza della classe* e' la differenza tra il valore massimo e minimo nella classe; il *valore centrale della classe* e' la semisomma tra il massimo ed il minimo nella classe; la *frequenza assoluta* e' il numero di elementi che appartengono alla classe; la *frequenza relativa* e' la percentuale di elementi che appartengono alla classe.

Per graficare questa tabella possiamo costruire un **istogramma normalizzato**. Esso consiste in un insieme di rettangoli adiacenti (ognuno relativo ad una classe) aventi come base sull'asse  $x$  con punto medio nel valore centrale della classe e altezza proporzionale alla frequenza della classe, in modo che l'area del rettangolo sia pari alla frequenza assoluta o percentuale della classe. Quindi l'altezza del rettangolo si ottiene dividendo la frequenza relativa o percentuale per l'ampiezza della classe. In questo modo se si sommano tutte le aree dei rettangoli si otterra' 1 o 100 %.

**Esempio 1.9** Nella tabella 1.4 sono riportati dei dati. In questo caso ho 4 classi, le prime 3 di ampiezza 20 e la quarta di ampiezza 40. Considerando le frequenze percentuali, l'istogramma della prima classe avra' altezza  $10/20 = 0.5$ , della seconda  $20/20 = 1$ , della terza  $30/20 = 1.5$  e della quarta  $40/40 = 1$ . Il relativo istogramma normalizzato e' graficato nella figura 1.3.

### Media e varianza

Supponiamo che gli  $n$  dati sono raggruppati in classi  $C_i$  ( $i = 1, \dots, k$ ), come nell'esempio 1.2.

Detto  $m_i$  il valore centrale della classe  $C_i$ , allora la media e la varianza sono così definite

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k m_i f_i \quad , \quad s^2 = \frac{1}{n} \sum_{i=1}^k (m_i - \bar{x})^2 f_i \quad (1.11)$$

dove  $f_i$  è la frequenza assoluta della classe.

**Esempio 1.10** *Calcolare la media e la varianza dell'esempio 1.2.*

In questo caso il numero di dati è  $n = 80$ , il numero di classi  $k = 5$  e

$$\begin{aligned} C_1 &= (0.40, 1.50] \quad , \quad m_1 = \frac{0.4 + 1.5}{2} = \dots \\ C_2 &= (1.50, 2.30] \quad , \quad m_2 = \frac{1.5 + 2.3}{2} = \dots \\ C_3 &= (2.30, 3.00] \quad , \quad m_3 = \frac{2.3 + 3.0}{2} = \dots \\ C_4 &= (3.00, 4.00] \quad , \quad m_4 = \frac{3 + 4}{2} = \dots \\ C_5 &= (4.00, 5.50] \quad , \quad m_5 = \frac{4 + 5.5}{2} = \dots \end{aligned}$$

applicando la (1.11) si ottiene ....

### Mediana

In questo caso, si può calcolare la mediana attraverso l'istogramma normalizzato. Occorre trovare quel valore sull'asse  $x$  tale che divida esattamente a metà l'area delimitata dall'istogramma. Ricordiamo che, per come viene costruito l'istogramma normalizzato, l'area totale sottesa ha un valore fissato: vale 1 se si stanno utilizzando le frequenze relative, 100 % se si stanno utilizzando le frequenze percentuali. Chiariamo anche qui con un esempio.

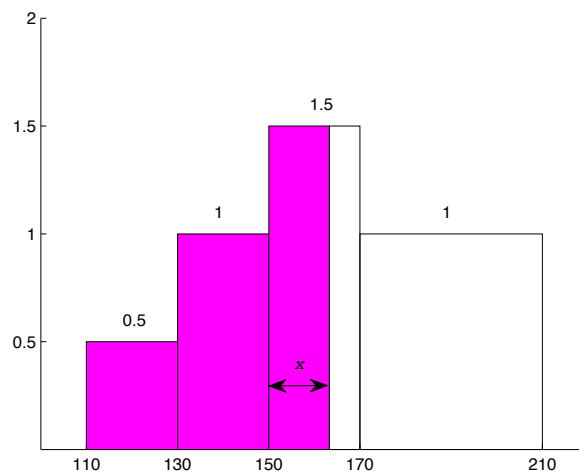


Figura 1.3: Istogramma normalizzato e calcolo della mediana

**Esempio 1.11** Consideriamo i dati della tabella 1.4. L'istogramma normalizzato, costruito con le frequenze percentuali, ha area pari a 100. La mediana è quel valore che ripartisce l'area dell'istogramma normalizzato in due parti uguali. Nel nostro caso 50 la prima (colorata) e 50 la seconda (vedi figura 1.3). Dobbiamo calcolare :

- area del primo rettangolo :  $20$  (base)  $\times$   $0.5$  (altezza)  $= 10$
- area del secondo rettangolo :  $20$  (base)  $\times$   $1$  (altezza)  $= 20$
- area di parte del terzo rettangolo :  $x$  (base)  $\times$   $1.5$  (altezza)  $= 1.5 x$

quindi dobbiamo imporre che quest'area sia uguale a 50

$$10 + 20 + 1.5 x = 50 \quad \rightarrow \quad x = 13.3$$

pertanto la mediana  $Q_2$  sarà

$$Q_2 = 150 + 13.3 = 163.3 \quad .$$

In modo analogo si possono calcolare gli altri quartili.

## 1.4 Indici di forma

Due indici statistici numerici che tengono conto della forma di una distribuzione di una serie di dati sono:

- a) Asimmetria:** E' una misura dello scostamento di una distribuzione dalla simmetria. Se la curva di frequenza di una distribuzione ha una coda piu' lunga a destra del massimo centrale, piuttosto che a sinistra, la distribuzione si dice *positivamente asimmetrica*. Se e' vero il contrario si dice *negativamente asimmetrica*.

### Definizione 1.3

Date  $n$  osservazioni  $x_1, x_2, \dots, x_n$  e' detto indice di asimmetria (o skewness), la quantita'

$$sk = \frac{m_3}{\sqrt{m_2^3}} \quad , \quad m_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k \quad k = 2, 3, 4, \dots$$

dove  $m_k$  sono i momenti centrati di ordine  $k$ .

Questo indice indica se la distribuzione del campione e' simmetrica rispetto alla media ( $sk=0$ ): se  $sk > 0$  la distribuzione sara' piu' concentrata a sinistra, con una coda piu' lunga a destra, il contrario se  $sk < 0$ .



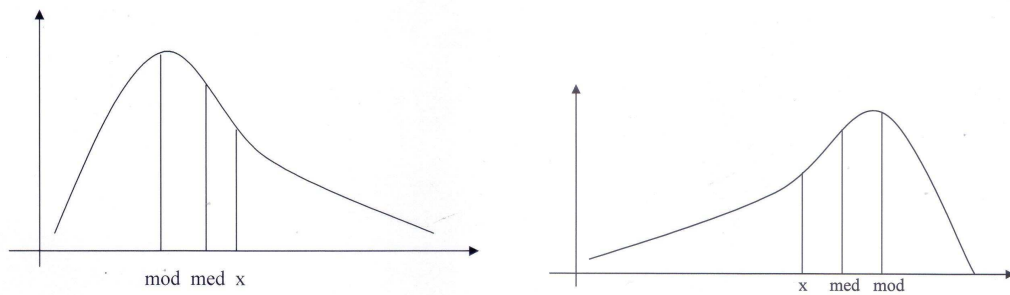


Figura 1.4: Figura sinistra: indice di asimmetria:  $sk > 0$  (asimmetria a destra). Figura destra: indice di asimmetria:  $sk < 0$  (asimmetria a sinistra)

**a) Curtosi:** E' una misura dell'appiattimento di una distribuzione di dati rispetto alla distribuzione normale (gaussiana).

#### Definizione 1.4

Date  $n$  osservazioni  $x_1, x_2, \dots, x_n$  e' detta Curtosi, la quantita'

$$\kappa = \frac{m_4}{m_2^2}$$

Si prova che, se  $\kappa > 3$  allora la distribuzione (detta leptocurtica) e' piu' appuntita rispetto alla normale (con code piu' grandi), se  $\kappa < 3$  (platicurtica) e' piu' appiattita (con code piu' piccole), infine se  $\kappa = 3$  ha la stessa altezza di una normale.

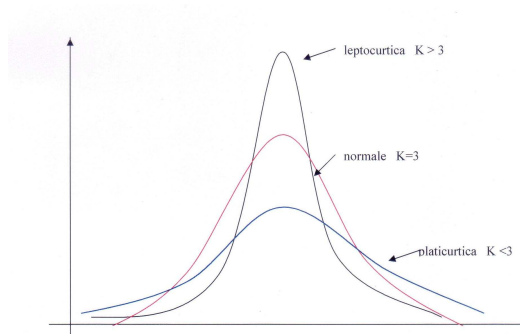


Figura 1.5: La curtosi

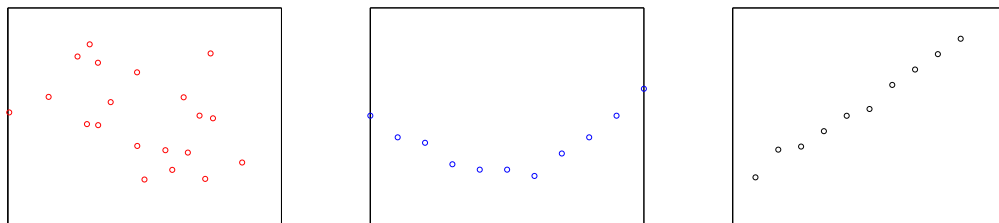


Figura 1.6: Scatterplot di dati

## 1.5 La correlazione tra due serie di dati

Talvolta piu' caratteri vengono misurati per ogni individuo come peso, altezza, reddito, ecc. Si vuole capire se c'e' una qualche relazione tra essi. Consideriamo di avere due caratteri quantitativi  $X$  e  $Y$  di una popolazione e supponiamo che i dati siano sotto forma di coppie  $\{x_i\}$ ,  $\{y_i\}$  di numerosità  $n$ , in cui la prima coordinata rappresenta il primo carattere  $X$  e la seconda quello  $Y$ . Ogni coppia di dati e' relativa allo stesso individuo. In un primo approccio grafico si possono disegnare sul piano tutti punti  $(x_i, y_i)$  e vedere se essi tendono a disporsi secondo un andamento regolare (*scatterplot*). Nella figura 1.6 sono riportati degli scatterplot di tre serie di dati: nel primo da sinistra sembra che non ci sia alcuna relazione tra i dati ovvero sono *indipendenti*. Nel secondo e terzo, invece, si vede una tendenza a forma di parabola e retta rispettivamente. Ci chiediamo se esiste una certa relazione tra questi dati ovvero se sono tra loro indipendenti. Per rispondere a questa domanda si puo' confrontare le variazioni delle

coppie di dati rispetto ai rispettivi valori medi.

$$x_i - \bar{x} \quad y_i - \bar{y}.$$

È ovvio supporre che esista una relazione di dipendenza tra  $\{x_i\}$  e  $\{y_i\}$  se  $x_i - \bar{x}$ ,  $y_i - \bar{y}$  hanno lo stesso segno.

Quindi, tanto più i prodotti  $(x_i - \bar{x})(y_i - \bar{y})$  hanno concordanza di segno, tanto più i dati considerati hanno forte dipendenza. Anche nel caso in cui a valori positivi di  $(x_i - \bar{x})$  corrispondono valori negativi di  $(y_i - \bar{y})$  o viceversa, denota una forte dipendenza tra i dati considerati.

Invece, se tutti i prodotti  $(x_i - \bar{x})(y_i - \bar{y})$  hanno segni diversi, la loro somma risulta piccola in valore assoluto (ovvero tende a zero) e potrebbe esserci indipendenza tra le due serie di dati.

**Definizione 1.5** Si definisce *covarianza*

$$c_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

**Teorema 1.1** Si prova facilmente che  $c_{xy} = \overline{xy} - \bar{x} \bar{y}$

**Definizione 1.6** Due serie di dati  $x_i, y_i$  sono *statisticamente incorrelate* se:

$$c_{xy} = 0 \quad .$$

Da quanto detto questo indice è tale che:

- i)  $c_{xy} \in \mathbb{R}$
- ii) se  $\{x_i\}, \{y_i\}$  sono fortemente correlati  $c_{xy}$  è grande in valore assoluto;
- iii) Se  $c_{xy} > 0$  i due set di dati si dicono *correlati positivamente*, e questo significa che si muovono nella stessa direzione (all'aumentare dell'uno l'altro aumenta e viceversa). Viceversa se  $c_{xy} < 0$  i due set di dati si dicono *correlati negativamente*.
- iv) se  $\{x_i\}, \{y_i\}$  sono *statisticamente incorrelati*, dal teorema precedente, si ha che  $\overline{xy} \simeq \bar{x} \bar{y}$  ovvero la media del prodotto delle due serie di dati ( $\overline{xy}$ ) è uguale al prodotto delle medie delle singole serie di dati ( $\bar{x} \bar{y}$ ). Ma questo non ci assicura che non ci sia una dipendenza tra le due serie di dati. Vedremo più avanti un esempio in cui le due serie di dati hanno covarianza piccola senza essere per questo indipendenti.

## 1.6 Analisi di regressione per una serie di dati

Assegnato un insieme  $E$  di coppie di dati  $\{x_i\}, \{y_i\}$  di numerosita'  $n$ , ci domandiamo se esiste un legame funzionale del tipo

$$y = f(x)$$

che descriva bene la relazione tra i dati.

Un'analisi di questo tipo si chiama **analisi di regressione**.

A questo punto come si fa a determinare la  $f$ , che al suo interno contiene dei parametri in modo che questo legame sia buono?

### 1.6.1 Metodo dei minimi quadrati

Si cerca  $f(x)$  tale che sia minima la funzione *residuo*

$$g(f) = \sum_{i=1}^n [f(x_i) - y_i]^2$$

Questa funzione rappresenta la somma dei quadrati delle distanze tra i dati sperimentali ( $y_i$ ) e quelli calcolati con funzione  $f(x_i)$ , ovvero la somma degli "errori".

### 1.6.2 Regressione lineare

In questo caso la funzione  $f$  e' una retta

$$f(x) = mx + q$$

$$\rightarrow g(m, q) = \sum_{i=1}^n [mx_i + q - y_i]^2$$

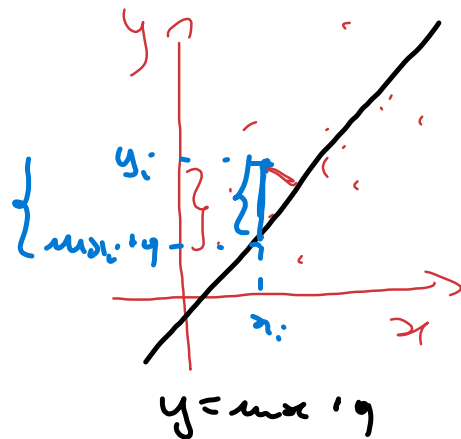
(incognite  $m, q$ !)

Poiché  $g(m, q)$  è una funzione di due variabili:

*condizione necessaria e sufficiente affinché  $P_* = (m_*, q_*)$  sia minimo relativo è che*

i)

$$\left. \frac{\partial g}{\partial m} \right|_{P_*} = 0 \quad , \quad \left. \frac{\partial g}{\partial q} \right|_{P_*} = 0$$



ii)

$$H(m_*, q_*) = g_{mm}g_{qq} - (g_{mq})^2 > 0$$

iii)

$$g_{qq}(m_*, q_*) > 0$$

dove

$$g_m = \frac{\partial g}{\partial m} = \lim_{h \rightarrow 0} \frac{g(m+h, q) - g(m, q)}{h}$$

$$g_{mm} = \frac{\partial^2 g}{\partial m^2} \quad , \quad g_{mq} = \frac{\partial^2 g}{\partial m \partial q}$$

$$\frac{\partial g}{\partial m} = 2 \sum_{i=1}^n (mx_i + q - y_i)x_i$$

$$\frac{\partial g}{\partial q} = 2 \sum_{i=1}^n (mx_i + q - y_i)$$

$$\frac{\partial^2 g}{\partial m^2} = 2 \sum_{i=1}^n x_i^2 > 0$$

$$\frac{\partial^2 g}{\partial q^2} = 2n > 0$$

$$\frac{\partial^2 g}{\partial m \partial q} = 2 \sum_{i=1}^n x_i > 0 \quad \text{Provare che } H(m, q) > 0!$$

Risolviamo ora il sistema:

$$\begin{cases} \sum_i (mx_i + q - y_i)x_i = 0 \\ \sum_i (mx_i + q - y_i) = 0 \end{cases}$$

$$\begin{cases} \sum_i mx_i^2 + \sum_i qx_i - \sum_i x_i y_i = 0 \\ \sum_i mx_i + nq - \sum_i y_i = 0 \end{cases}$$

dividendo per n si ottiene:

$$\begin{cases} \frac{m}{n} \sum_i x_i^2 + q\bar{x} - \frac{1}{n} \sum_i x_i y_i = 0 \\ m\bar{x} + q - \bar{y} = 0 \Rightarrow q = \bar{y} - m\bar{x} \end{cases}$$

$$m \frac{\sum_i x_i^2}{n} + \bar{x} \bar{y} - m(\bar{x})^2 - \frac{1}{n} \sum_i x_i y_i = 0 \quad , \quad m = \frac{\frac{1}{n} \sum_i x_i y_i - \bar{x} \bar{y}}{\frac{\sum_i x_i^2}{n} - (\bar{x})^2}$$

Da cui ricordando la covarianza

$$c_{xy} = \frac{1}{n} \sum_i x_i y_i - \bar{x} \bar{y}$$

e la varianza per la variabile x

$$s_x^2 = \frac{\sum_i x_i^2}{n} - (\bar{x})^2$$

da cui

$$m = \frac{c_{xy}}{s_x^2} \quad , \quad q = \bar{y} - \frac{c_{xy}}{s_x^2} \bar{x}$$

Il metodo applicato fornisce la retta che meglio approssima i dati, ma non il grado di approssimazione. Per questo motivo si introduce il **coefficiente di correlazione lineare (o di Pearson)**

$$r_{xy} = \frac{c_{xy}}{s_x s_y}$$

### Proposizione 1.3

- i)  $r_{xy} \in [-1, 1]$
- ii) se  $r_{xy} = \pm 1$  i dati  $(x_i, y_i)$  sono perfettamente allineati con la retta di regressione
- ii) se  $r_{xy} > 0$  la retta è ascendente

NB: nella pratica, se  $|r_{xy}| < 0.9$ , i dati si allontanano dall'andamento rettilineo.

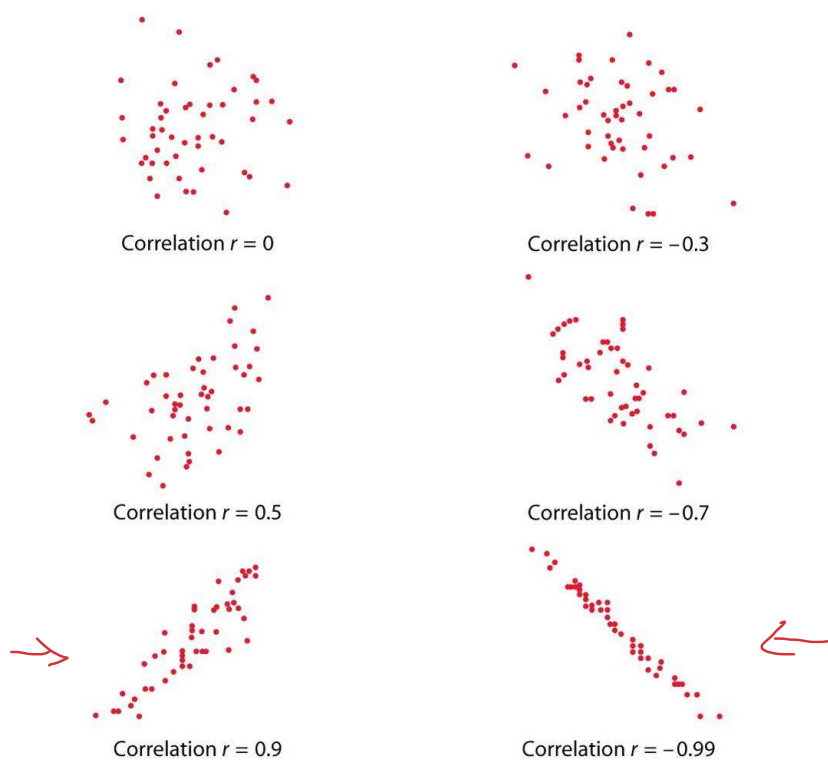


Figura 1.7: Indice di correlazione  $r = r_{xy}$

**Esempio 1.12** *Tabella carichi di rottura* Si deve controllare la resistenza di un campione di 15 travi di cemento, tutte ottenute dalla stessa gittata, misurando sia i carichi di prima lesione  $X_i$  che quelli di rottura finale  $Y_i$  (in Kg), come in tabella 1.5.

Con EXCEL calcolare:  $s_x$ ,  $s_y$ ,  $c_{xy}$ ,  $r_{xy}$ ,  $\bar{x}$ ,  $\bar{y}$ , retta di regressione e grafico dei dati x,y con retta di regressione.

**Esempio 1.13** *Assegnati i set di dati in tabella 1.6, calcolare la covarianza ed il coefficiente di Pearson. Si osservi che la covarianza tende a zero ma non il coefficiente di Pearson !*

Tabella 1.5: Carichi di rottura.

$I^a$ lesione	rottura
2550	4650
2900	4650
3000	4700
3000	4750
3000	4775
3000	4775
3250	4800
3250	4950
3250	5050
3600	5100
4225	5100
4650	5150
4750	5175
5175	5250
5300	5300

Tabella 1.6: Esempio sulla covarianza ed indipendenza

x	0.185	0.22	0.233	0.247	0.255	0.2745
y	0.049	0.053	0.054	0.0565	0.058	0.0605



### 1.6.3 Parabola dei minimi quadrati

#### Regressione non lineare

In questo caso

$$f(x) = a + bx + cx^2$$

a questo punto devo rendere minimo

$$g(a, b, c) = \sum_{i=1}^n [a + bx_i + cx_i^2 - y_i]^2$$

allora

$$\frac{\partial g}{\partial a} = 0 \quad , \quad \frac{\partial g}{\partial b} = 0 \quad , \quad \frac{\partial g}{\partial c} = 0$$

$$\left\{ \begin{array}{l} \sum_i y_i = an + b \sum_i x_i + c \sum_i x_i^2 \\ \sum_i x_i y_i = a \sum_i x_i + b \sum_i x_i^2 + c \sum_i x_i^3 \\ \sum_i x_i^2 y_i = a \sum_i x_i^2 + b \sum_i x_i^3 + c \sum_i x_i^4 \end{array} \right.$$

Sistema di 3 equazioni in 3 incognite

**Esempio 1.14** Trovare la soluzione del sistema lineare 3x3 con EXCEL

$$AX = B$$

Se  $\det A \neq 0$  (teorema di Cramer) allora

$$\exists A^{-1} \quad t.c. \quad A^{-1}AX = A^{-1}B$$

quindi la soluzione e'

$$X = A^{-1}B$$

NB:

$$A \rightarrow n \times n, X \rightarrow n \times 1, B \rightarrow n \times 1$$

**Esempio 1.15** Utilizzando la tabella dei carichi di rottura 1.5, calcolare la parabola dei minimi quadrati.

Un modo grossolano per vedere la bontà dell'approssimazione è calcolare il *residuo*:

$$g(\bar{a}, \bar{b}, \bar{c}) = \sum_i [\bar{a} + \bar{b}x_i + \bar{c}x_i^2 - y_i]^2$$

- Confrontare il valore ottenuto con la retta di regressione
- Mostrare il grafico dei dati e la parabola di regressione

**Esempio 1.16** *Popolazione USA: in tabella 1.7 sono assegnati per alcuni anni la popolazione degli Stati Uniti d'America.*

Si vogliono fare delle stime di crescita di questa popolazione. In particolare:

- Calcolare la retta di regressione e  $r_{xy}$ ;
- Calcolare la parabola di regressione;
- Approssimare i dati con la curva esponenziale

$$y = ae^{bx}$$

In questo caso ci possiamo ricondurre al caso della regressione lineare :

$$\ln y = \ln(ae^{bx})$$

$$\ln y = \ln a + \ln e^{bx} = \ln a + bx$$

con un cambiamento di variabili, si ha

$$\tilde{y} = \alpha + bx$$

- Approssimare i dati con una legge esponenziale  $y = b m^x$  (utilizzare REGR.LOG)
- Grafico dei dati sperimentali con quelli di regressione.

Tabella 1.7: Popolazione USA

Anno	Popol(mln)
1840	17.1
1850	23.2
1860	31.4
1870	39.8
1880	50.2
1890	62.9
1900	76.0
1910	92.0
1920	105.7
1930	122.8
1940	131.7
1950	151.1
1960	179.3

Suggerimento: scegliere  $x$  in modo che:

.....

anno 1890  $\rightarrow x=-1$

anno 1900  $\rightarrow x=0$

anno 1910  $\rightarrow x=1$

anno 1920  $\rightarrow x=2$

.....