1 Medie

La **statistica** consta di un insieme di metodi atti a elaborare e a sintetizzare i dati relativi alle caratteristiche di una fissata "popolazione", rilevati mediante osservazioni o esperimenti.
Col termine "popolazione" si usa designare un insieme (in genere piuttosto numeroso) di "individui": ma può trattarsi indifferentemente di una popolazione umana, di una colonia di batteri, delle molecole di un gas, degli esiti di un esperimento ripetuto un certo numero di volte, e cos via. Ovviamente, gli "individui" sono di volta in volta i singoli uomini, o i singoli batteri, o le singole molecole, o i singoli esiti dell'esperimento, ecc.
Il più delle volte i dati relativi ad una determinata caratteristica della popolazione sono di tipo quantitativo, ossia numerici. Nel caso di una popolazione umana, può trattarsi per es. delle misure delle altezze, o dei pesi, o delle età degli individui della popolazione. Ma non è esclusa l'eventualità che si abbia a che fare con dati di tipo qualitativo, ossia non numerici, quali per es. il tipo del gruppo sanguigno, oppure il colore degli occhi, ecc.
Supponiamo di aver fissato la nostra attenzione su una data caratteristica, per es. sull'altezza degli individui della popolazione in esame. Le singole misure saranno in generale diverse da un individuo all'altro. Il metodo più comunemente usato per estrarre dall'insieme dei dati numerici individuali qualche informazione globale, riferita al complesso della popolazione, consiste nel calcolo della media aritmetica. Esistono tuttavia anche vari altri tipi di medie, e il ricorso all'uno piuttosto che all'altro tipo dipende dalla natura dei dati e dall'utilizzazione che se ne vuole fare, come chiariremo meglio in seguito. Cominciamo col richiamare le principali definizioni.

Definizione 1.1. Dati n numeri, misure della grandezza in esame,

$$x_1, x_2, \ldots, x_n$$

la loro **media aritmetica** è il numero $\overline{x}$ dato dalla formula:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Mediante l'uso del simbolo di sommatoria, la stessa formula si può scrivere più sinteticamente come segue:

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

In luogo di $\overline{x}$ si usano anche i simboli M_a oppure μ (vedi in particolare il paragrafo 6).

Esempio 1.2. Dati i 5 numeri 176 181 168 176 172
la loro media aritmetica è:

$$\overline{x} = \frac{176 + 181 + 168 + 176 + 172}{5} = 174,6.$$

Esempio 1.3. Dati i 5 numeri: 145 187 151 165 225
la loro media aritmetica è ancora:

$$\overline{x} = \frac{145 + 187 + 151 + 165 + 225}{5} = 174,6.$$

Il confronto fra i due esempi fa vedere che, pur partendo da dati numerici piuttosto diversi tra loro, può capitare che la media aritmetica sia la stessa nei due casi.
Gli esempi 1.2 e 1.3 sono particolarmente semplici, ma proprio per questo motivo sono anche poco significativi: infatti una media calcolata su un insieme di soli 5 numeri fornisce scarse informazioni dal punto di vista statistico.
Ecco quindi un esempio più realistico:

Esempio 1.4. Nel rilevare le altezze di un gruppo di reclute, si ottenuta la seguente tabella delle frequenze:

Altezza (in cm)	F_{ass}
166	1
168	3
169	6
170	11
171	8
172	6
173	4
174	3
175	1
178	1

Per calcolare l' "altezza media" del gruppo di reclute, vale a dire la media aritmetica $\overline{x}$ delle altezze riscontrate, si tratta di sommare i numeri elencati nella prima colonna, *ciascuno considerato tante volte quant' è la sua frequenza, evidenziata nella seconda colonna*, e dividere questa somma per il numero totale n degli individui, della popolazione in esame. Naturalmente, invece di scrivere per tre volte l'addendo 168, conviene scrivere semplicemente $3 \cdot 168$, invece di scrivere per sei volte l'addendo 169, conviene

scrivere semplicemente $6 \cdot 169$, ecc. Quanto al calcolo del denominatore n, si tratta di sommare le frequenze elencate nella seconda colonna. In definitiva si ha:

$$\overline{x} = \frac{166 + 3 \cdot 168 + 6 \cdot 169 + \cdots + 175 + 178}{1 + 3 + 6 + \cdots + 1 + 1} = \frac{7521}{44} \simeq 170,9.$$

Si usa anche dire che $\overline{x}$ è la **media ponderata** dei numeri della prima colonna, considerati con le rispettive frequenze (o "molteplicità") riportate nella seconda colonna.

Passando dall'esempio specifico al caso generale, la media aritmetica ponderata di certi numeri x_i $(i = 1, 2, \ldots, k)$, ciascuno considerato con la sua frequenza assoluta f_i, è espressa dalla formula:

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i \cdot x_i}{\sum_{i=1}^{k} f_i} \tag{1}$$

Attenzione. Quando si ha a che fare con dati raggruppati, che quindi compaiono con determinate frequenze, la loro *media aritmetica* va intesa sempre ed esclusivamente nel senso di *media aritmetica ponderata* (anche quando l'aggettivo "ponderata" è sottinteso). Sarebbe profondamente sbagliato calcolare la media aritmetiaca dei soli dati, trascurando le rispettive frequenze

Ecco infine un'estensione della nozione di media aritmetica al caso di fenomeni che si sviluppano con continuità nel tempo. Per maggiore concretezza, consideriamo un esempio specifico: *la pressione arteriosa*. Com'è ben noto, la pressione arteriosa di un individuo (in condizioni normali) ha un andamento approssimativamente periodico $P = P(t)$ di periodo T (= alla durata di un battito cardiaco) e varia tra un valore minimo P_{min} (raggiunto nella fase diastolica) e un valore massimo P_{max} (raggiunto nella fase sistolica). Volendo introdurre la nozione di "pressione arteriosa media" questi soli due dati non sono sufficienti. Infatti occorre tenere conto dell'andamento complessivo del grafico di $P(t)$ in funzione del tempo t; in particolare, occorre tenere conto del fatto che la fase sistolica ha una durata generalmente inferiore alla durata della fase diastolica. Ecco come fare: si suddivide la durata del periodo T in un certo numero n di intervallini di durata $\frac{1}{n}T$; si misura il valore della pressione $P(t_i)$ all'istante centrale di ciascuno di questi intervallini; infine si calcola la media aritmetica degli n valori $P(t_i)$:

$$\frac{\sum_{i=1}^{n} P(t_i)}{n}. \tag{2}$$

Questa media aritmetica fornirà una misura tanto più precisa per quella che intendiamo chiamare "pressione arteriosa media", quanto maggiore sarà il numero n delle

suddivisioni considerate. Passando al limite, al tendere di n all'infinito, si definisce quindi la **pressione arteriosa media** come

$$P_{media} = \lim_{t\to\infty} \frac{\sum_{i=1}^{n} P(t_i)}{n}. \tag{3}$$

Chi ha presente la nozione di integrale definito non stenterà a riconoscere che la definizione precedente equivale a:

$$P_{media} = \frac{1}{T}\int_0^T P(t)dt \tag{4}$$

(il coefficiente correttivo $\frac{1}{T}$ serve a compensare il fatto che, quando si suddivide la durata T di un battito cardiaco in n intervallini, ciascuno di questi è caratterizzato da una durata $\frac{1}{n}T$ e non semplicemente dal numero $\frac{1}{n}$). È possibile anche un'interpretazione geometrica: l'area del sottografico di $P(t)$ sull'intervallo $[O,T]$ uguale all'area che, sullo stesso intervallo, avrebbe il sottografico di una pressione costante con intensità uguale a P_{media} (rettangolo di base $[O,T]$ e altezza P_{media}). Come già notato in varie altre occasioni, all'atto pratico non si effettuerà il passaggio al limite e ci si limiterà a calcolare P_{media} considerando una suddivisione di T in n intervallini, con n abbastanza grande da garantire una buona approssimazione dell'integrale.
Infine, una **regola pratica** che consente una valutazione (sia pure grossolana e non valida in condizioni di sforzo fisico) della pressione arteriosa media, sotto forma di media aritmetica (opportunamente *ponderata*) dei soli valori P_{min} e P_{max}:

$$P_{media} = \frac{1}{3}(2P_{min} + P_{max}). \tag{5}$$

2 La media geometrica

Dati n numeri, come sopra, con l'ulteriore condizione che essi siano tutti **positivi**, la loro **media geometrica** è il numero M_g dato dalla formula:

$$M_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}. \tag{6}$$

Se si introduce, in analogia col simbolo di sommatoria, il simbolo (detto **produttoria**) $\prod_{i=1}^{n} x_i$ per denotare il prodotto degli n numeri $x_1, x_2, \ldots, x_n$, la formula precedente si può scrivere più sinteticamente nella forma:

$$M_g = \sqrt[n]{\prod_{i=1}^{n} x_i}. \tag{7}$$

Esempio 2.1. Dati gli stessi 5 numeri dell'esempio 1.2, la loro media geometrica è:

$$M_g = \sqrt[5]{176 \cdot 181 \cdot 168 \cdot 176 \cdot 172} = \sqrt[5]{162\,009\,931\,776} \simeq 174,5. \tag{8}$$

Esempio 2.2. Dati gli stessi 5 numeri dell'esempio 1.3, la loro media geometrica è

$$M_g = \sqrt[5]{145 \cdot 187 \cdot 151 \cdot 165 \cdot 225} = \sqrt[5]{152\,003\,300\,625} \simeq 172,2. \tag{9}$$

Dal confronto fra gli esempi 1.2, 1.3 e 2.1, 2.2, si vede che la coincidenza delle medie aritmetiche non implica la coincidenza delle medie geometriche.

Esempio 2.3. Data la stessa tabella di numeri dell'esempio 1.4, un calcolo diretto della loro media geometrica non sarebbe agevole. Conviene quindi passare ai logaritmi (in una base qualsiasi, per es. in base 10) e osservare che il logaritmo della media geometrica viene ad essere semplicemente la media aritmetica dei logaritmi dei dati considerati. Nel nostro caso:

$$Log M_g = Log(166 \cdot 168^3 \cdot 169^6 \cdot 170^{11} \cdot 171^8 \cdot 172^6 \cdot 173^4 \cdot 174^3 \cdot 175 \cdot 178)^{\frac{1}{44}} \tag{10}$$

$$= \frac{1}{44}(Log166 + 3Log168 + \cdots + Log178) \simeq 2,2324$$

A questo punto basta ritornare dai logaritmi ai numeri, e si ottiene:

$$M_g \simeq 170,8.$$

3 La mediana

Dati sempre n numeri, si comincia col riordinarli per valore crescente (dal più piccolo al più grande):

$$x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n$$

e si chiama **mediana** (o **valore mediano**) l'elemento M_e che in questa successione occupa il posto centrale. Per la precisione, se n è dispari, l'indice che individua il posto centrale è $i = \frac{n+1}{2}$, mentre se n è pari non esiste un elemento di posto centrale; si considerano allora i due elementi più prossimi al posto centrale, individuati dagli indici $i_1 = \frac{n}{2}$ ed $i_2 = \frac{n}{2} + 1$ e se ne fa la semisomma. In conclusione la definizione di mediana, data sopra, va perfezionata distinguendo i due casi e ponendo rispettivamente:

$$M_e = \begin{cases} x_{\frac{n+1}{2}} & \text{se n è dispari} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} & \text{se n è pari} \end{cases} \tag{11}$$

Esempio 3.1. Gli stessi numeri dell'esempio 1.2, riordinati per valore crescente, sono:

$$168 \quad 172 \quad 176 \quad 176 \quad 181.$$

In questa successione, l'elemento di posto centrale è il terzo, ossia il valore mediano è:

$$M_e = 176.$$

Esempio 3.2. Gli stessi numeri dell'esempio 1.3, riordinati per valore crescente, sono:

$$145 \quad 151 \quad 165 \quad 187 \quad 225.$$

Quindi in questo caso il valore mediano è

$$M_e = 165.$$

Esempio 3.3. I numeri della tabella dell'esempio 1.4 sono già ordinati per valore crescente; tenendo conto delle rispettive frequenze, si tratta di una successione di 44 numeri. Quindi il valore mediano è dato dalla semisomma dei numeri che occupano il ventiduesimo e il ventitreesimo posto. In questo caso x_{22} coincide con x_{23} (si tratta del numero 171) e quindi anche

$$M_e = 171.$$

Dal confronto fra gli esempi 1.2, 1.3, 3.1 e 3.2 risulta che la coincidenza delle medie aritmetiche non implica la coincidenza delle mediane.

4 La moda

I vari tipi di medie considerati finora si riferivano esclusivamente alle misure *numeriche* di una data grandezza. Quando capita invece di dover considerare variabili di tipo *non numerico*, come per es. il gruppo sanguigno degli individui di una data popolazione, è naturale ripartire la popolazione stessa in classi (nel caso del gruppo sanguigno le classi saranno "O" "A", "B", "AB") e stabilire qual è la classe più numerosa. Tale classe prende il nome di **classe modale** o **moda**.

Naturalmente, nulla vieta di parlare di moda anche nel caso di una grandezza numerica: se i valori della grandezza sono distribuiti in un numero finito di classi, si chiama *classe modale*, o *moda*, la classe alla quale appartiene il maggior numero di misure. Infine, non è escluso che esistano due o più classi ugualmente numerose, nel qual caso si parla di *classi modali*, al plurale. Ovviamente, il ricorso alla nozione di moda è significativo solo se la concentrazione delle frequenze nella classe modale è abbastanza pronunciata. Così sarebbe per es. del tutto fuori luogo parlare di moda a proposito delle situazioni considerate negli esempi 1.2 e 1.3.

Esempio 4.1. Nella situazione ipotizzata nell'esempio 1.4, la classe modale è quella che corrisponde all'altezza di 170 cm.

Esempio 4.2. I ricoveri in un reparto ospedaliero nel corso di una settimana hanno avuto il seguente andamento:

Giorno	Numero ricoveri
Lunedì	18
Martedì	9
Mercoledì	18
Giovedì	10
Venerdì	6

In questo caso si hanno due classi modali, corrispondenti al Lunedì e al Mercoledì.

Regola pratica. Si preferisce usare la media aritmetica in tutte quelle situazioni dove le singole misure della grandezza in esame risultano distribuite in modo abbastanza simmetrico a sinistra e a destra di $\overline{x}$, con un addensamento dei valori in prossimità di $\overline{x}$. (Ciò capita per es. per le altezze delle reclute, considerate nell'esempio 1.4). Nelle situazioni in cui questa condizione di "simmetria" non è soddisfatta, si può provare a passare ai logaritmi delle misure della grandezza in esame: se i logaritmi si distribuiscono in modo abbastanza simmetrico a sinistra e a destra del logaritmo di M_g, è opportuno usare la media geometrica. Per es. i dati numerici dell'esempio 1.3 sono distribuiti in modo vistosamente asimmetrico rispetto alla loro media aritmetica; si constata invece che i corrispondenti logaritmi sono distribuiti abbastanza simmetricamente a sinistra e a destra del logaritmo della media geometrica; quindi in questo caso è preferibile usare la media geometrica, piuttosto che quella aritmetica. In medicina, l'uso della media geometrica è particolarmente indicato quando si ha a che fare con misure relative a fenomeni caratterizzati da leggi di tipo esponenziale, come per es. conteggi di una popolazione di batteri o titoli di un anticorpo. Si preferisce usare la mediana in quelle situazioni dove non interessano tanto i valori numerici delle grandezze in esame, quanto piuttosto il loro ordinamento. Per es. nel caso dei voti assegnati agli elaborati dei partecipanti ad un concorso, la maggiore o minore severit nel metro di giudizio è relativamente irrilevante. Ciò che conta è esclusivamente la posizione di ciascun candidato nella graduatoria: la mediana consente di separare il 50% dei candidati "peggiori" dal 50% dei candidati "migliori". Si usano le classi modali sia in quei casi in cui non sarebbe possibile usare altri tipi di medie, perché i dati raccolti sono di tipo non numerico (per es. il tipo del gruppo sanguigno) sia quando, avendo a che fare con

dati numerici, interessa considerare un campione dal comportamento "tipico" o "normale", identificabile col valore più frequente, non influenzato dall'eventuale presenza di elementi "spurii" o "anomali".

Esercizio 4.3. Un'indagine effettuata su un campione di 50 famiglie ha dato il seguente risultato:

numero dei figli per famiglia	F_{ass}
0	6
1	12
2	16
3	9
4	4
5	1
6	2

Calcolate il numero medio di figli per famiglia.

Esercizio 4.4. Schematizzate la situazione dell'esercizio precedente, pensando la popolazione campione costituita dai 50 "capifamiglia" e dai rispettivi "figli" (gli altri eventuali familiari non hanno rilevanza ai fini dell'indagine). È ragionevole presumere che, per raccogliere i dati, poi riassunti nella tabella, a tutti i "capifamiglia" sia stata rivolta una domanda del tipo:

"Quanti figli vi sono nella sua famiglia?" In alternativa, si sarebbe potuta rivolgere a tutti i "figli" una domanda del tipo: "In quanti fratelli siete nella vostra famiglia?" La tabella, e i corrispondenti valori medi, sarebbero risultati gli stessi nei due casi? E, in caso di risposta negativa, in che modo si sarebbero modificati?

Suggerimento. Le due domande sono solo apparentemente equivalenti: per ogni famiglia la prima domanda viene posta una volta sola; la seconda domanda viene posta invece tante volte quanti sono i "figli" di quella famiglia. Cambia quindi addirittura la "popolazione" sottoposta all'indagine, cambia la tabella delle frequenze, e cambiano i valori medi.

Esercizio 4.5. Ecco una "tavola di mortalità" relativa ad una popolazione (fittizia) di 100 individui:

Per semplicità di calcolo, supponiamo che i decessi ipotizzati nel corso di ciascun decennio avvengano tutti esattamente alla metà del decennio (vale a dire al compimento

Fascia di età	Numero di decessi
$0 \leq x < 10$	2
$10 \leq x < 20$	1
$20 \leq x < 30$	1
$30 \leq x < 40$	2
$40 \leq x < 50$	4
$50 \leq x < 60$	7
$60 \leq x < 70$	15
$70 \leq x < 80$	27
$80 \leq x < 90$	36
$90 \leq x < 100$	5

del 5°, del 15°, del 25° ,... anno di età). Ricordiamo infine la seguente definizione: si chiama *attesa di vita all' età* X la media aritmetica degli anni che restano ancora da vivere agli individui che hanno raggiunto l'età X. Ciò premesso, calcolate, per la popolazione ipotizzata nella tabella, l'attesa di vita:

- al momento della nascita
- all'età di 50 anni
- all'età di 70 anni.

Esercizio 4.6. Considerate la tabella

Tabella 1: cause di morte registrate n Italia dal 1982 al 1989

Anno	Sistema circ	Tumori	App. Resp.	App. diger.	Altre cause	Totale
1982	251811	127333	34335	30621	90835	534935
1983	266885	131499	40010	31955	93981	564330
1984	243396	130143	34658	31322	87046	526565
1985	245690	134384	36878	31693	92527	541172
1986	245611	137179	38724	30797	92183	544489
1987	239287	141494	33932	29001	89057	532771
1988	232609	143350	34064	29387	92876	532286
1989	231577	145583	33266	29647	91780	531853

(a) È appropriato sintetizzare mediante qualche tipo di media i dati riportati nelle singole righe di tale tabella? E i dati riportati nelle singole colonne?

(b) In caso di risposte affermative alle domande del punto precedente, precisate quali tipi di media riterreste opportuno calcolare. Quindi effettuate i relativi calcoli, specificando di volta in volta ciò che i risultati ottenuti rappresentano.

Esercizio 4.7. Determinate il "peso medio" degli individui considerati nella tabella seguente.

Peso p (in kg)	F_{ass}
$40 \leq p < 45$	2
$45 \leq p < 50$	12
$50 \leq p < 55$	21
$55 \leq p < 60$	17
$60 \leq p < 65$	18
$65 \leq p < 70$	22
$70 \leq p < 75$	18
$75 \leq p < 80$	7
$80 \leq p < 90$	3

Suggerimento. Alla locuzione "peso medio" si può attribuire sia il significato di media aritmetica, sia quello di mediana, sia infine quello di classe modale (o di classi modali). Discutete vantaggi e svantaggi dell'uso di questi diversi tipi di medie. Quando, come in questo caso, le misure della grandezza in esame (nel nostro esempio, il peso) sono raggruppate in fasce di una certa ampiezza, conviene supporre ai fini del calcolo della media aritmetica che tutte le misure che cadono entro una detenninata fascia coincidano col valore centrale della fascia stessa (per es. i 18 pesi della fascia compresa tra 60 kg e 65 kg si supporranno convenzionalmente tutti uguali a 62,5 kg). Ai fini del calcolo della mediana è invece più appropriato supporre che le misure che cadono entro una determinata fascia siano equidistribuite nella fascia stessa (secondo questa convenzione, per es. i 18 pesi della fascia compresa tra 60 kg e 65 kg si supporranno ordinati in una successione crescente del tipo 60 kg, 60,275 kg, 60,55 kg ,..., 64,95 kg).

Esercizio 4.8. Determinate la "durata media" dei ricoveri ospedalieri considerati nella tabella seguente.

9	6	7	6	13	12	15	7	8	9	7	11	10	18	8
14	11	6	7	15	3	13	8	13	7	8	13	11	9	5
6	12	13	8	14	13	16	11	20	4	12	9	12	4	10
13	14	6	7	17	1	10	7	11	6	7	10	13	14	11
10	6	14	12	7	13	8	13	3	13	10	8	12	2	5
14	7	12	4	11	7	14	9	20	15	10	13	8	14	12
9	13	12	10	8	9	11	8	15	9	15	6	11	11	3
12	14	1	7	13	7	10	13	13	10	12	16	12	8	10

Esercizio 4.9. (a) Date un esempio di una dozzina di dati numerici (non tutti uguali tra loro), tali che la loro media aritmetica coincida con la mediana.

(b) Date un esempio di una dozzina di dati numerici, tali che la loro media aritmetica risulti minore della mediana.

(c) Date un esempio di una dozzina di dati numerici, tali che la loro media aritmetica risulti maggiore della mediana.

Esercizio 4.10. Analizzate la seguente situazione, ricorrendo eventualmente ad un'esemplificazione numerica: a causa di un errore strumentale, o di trascrizione, in un insieme di dati numerici piuttosto accurati è stato inserito un dato "sballato" (molto più grande o molto più piccolo degli altri dati). Il dato "sballato" influenza maggiormente la media aritmetica o la mediana?

Esercizio 4.11. Un'indagine statistica relativa a due popolazioni disgiunte A, B ha dato i seguenti risultati: età media della popolazione A: 42,5 anni; età media della popolazione B: 48,3 anni. Calcolate l'età media della popolazione complessiva, sapendo che la popolazione A è costituita da 47500 individui e che la popolazione B è costituita da 68 350 individui.

Esercizio 4.12. Supponete di sapere che l'altezza media dei giovani di leva (classe del 1970) sia stata di: 174,7 cm nell'Italia del Nord;
173,5 cm nell'Italia Centrale;
171,8 cm nell'Italia del Sud;
170,3 cm nell'Italia Insulare.
Queste informazioni sono sufficienti per calcolare l'altezza media di tutti i giovani di leva italiani (classe del 1970)? In caso di risposta affennativa, qual è questa altezza media? In caso di risposta negativa, quali sono i dati mancanti?

Esercizio 4.13. Dati n numeri x_i, verificate che l'espressione nella variabile x:

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

si annulla per $x = \overline{x} =$ *media aritmetica* dei valori x_i.

Esercizio 4.14. Dati n numeri x_i, verificate che l'espressione nella variabile x:

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

assume il suo valore minimo per $x = \overline{x} =$ *media aritmetica* dei valori x_i.

Esercizio 4.15. Dati n numeri x_i, verificate che l'espressione nella variabile x:

$$\sum_{i=1}^{n}|x_i - x|$$

assume il suo valore minimo (oppure uno dei suoi valori minimi) per $x = M_e =$ *mediana* dei valori x_i.

5 Dispersione

Le varie medie considerate nel Par. prec. sono dette anche *indici di posizione*, in quanto ogni media rappresenta appunto una particolare "posizione" sulla scala delle grandezze del tipo considerato. Ma la sola conoscenza di una media (sia essa la media aritmetica, o quella geometrica, o la mediana, o la moda) non è sufficiente per descrivere in che modo i dati di partenza risultano distribuiti intorno a quel valore medio. Infatti gli esempi 1.2 e 1.3 del Par. prec. fanno vedere che una medesima media aritmetica può scaturire da insiemi di dati molto dissimili tra loro: mentre i numeri dell'esempio 1.2 sono tutti piuttosto "addensati" vicino alla media aritmetica $\overline{x}$, i numeri dell'esempio 1.3 sono assai pi "dispersi". Ovviamente, considerazioni analoghe si possono ripetere anche a proposito degli altri tipi di medie. Per misurare questo grado di dispersione, si introducono degli ulteriori indicatori numerici, detti appunto *indici di dispersione*. Ecco le principali definizioni:

Definizione 5.1. Si chiama **intervallo di variazione** I_V di un insieme di dati (in inglese: *range*) la differenza

$$I_V = x_{max} - x_{min}$$

dove x_{max} e x_{min} denotano rispettivamente il più grande e il più piccolo tra i valori della serie di misure in esame.

Così nell'esempio 1.2 risulta $x_{max} = 181$; $x_{min} = 168$; $I_V = 13$. Nell'esempio 1.3 risulta invece $x_{max} = 225$; $x_{min} = 145$; $I_V = 80$.

La nozione di *intervallo di variazione* presenta un grave inconveniente: la sua ampiezza dipende in maniera determinante dalla presenza anche di un solo è valore molto diverso dagli altri, valore che il più delle volte è scarsamente significativo ai fini statistici (per es. può essere frutto della lettura errata di uno strumento, o di un errore di trascrizione, o simili). Ciò giustifica l'introduzione di altri indici di dispersione, meno influenzati dai valori estremi. Ecco una prima idea: a partire dagli n numeri:

$$x_1 \quad x_2 \quad \cdots x_n$$

e dalla loro media aritmetica $\overline{x}$, si calcolano i cosiddetti scarti, ossia le differenze:

$$x_1 - \overline{x} \quad x_2 - \overline{x} \quad \cdots x_n - \overline{x}$$

. A questo punto si sarebbe tentati di calcolare la *media aritmetica* degli scarti. Ma si constata che così facendo si ottiene sempre 0, in quanto gli scarti positivi compensano esattamente quelli negativi (cfr. l'esercizio 4.13). Occorre dunque introdurre qualche ulteriore correttivo, in modo tale da rendere positivi anche gli scarti negativi. Una possibilità consiste nel sostituire gli scarti con i rispettivi valori assoluti; un'altra possibilità consiste nel sostituirli con i rispettivi quadrati. Poiché i valori assoluti sono poco maneggevoli, si preferisce ricorrere ai quadrati. In definitiva, si dà la seguente definizione.

Definizione 5.2. Si chiama **varianza** di un insieme di dati statistici, in simboli *Var*, la media aritmetica dei quadrati degli scarti:

$$Var = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} \tag{12}$$

Ma neppure la varianza è esente da inconvenienti. Infatti dal punto di vista "dimensionale" essa non è omogenea con i dati di partenza (se per es. gli x_i sono lunghezze, la varianza rappresenta una lunghezza al quadrato; se gli x_i sono temperature, o pressioni, ecc. la varianza rappresenta una temperatura al quadrato, una pressione al quadrato, ecc.). Con un'ulteriore modifica si passa allora ad un nuovo indice, che di solito risulta preferibile alla varianza: la modifica consiste nell'annullare l'effetto degli elevamenti al quadrato mediante un'estrazione di radice quadrata. Ecco la definizione.

Definizione 5.3. Si chiama **scarto quadratico medio** o **deviazione standard** (in inglese: *standard deviation*), e si denota abitualmente con s oppure con σ (vedi in particolare il Par. succ.), la radice quadrata della varianza. In simboli:

$$s = \sqrt{Varianza} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}} \tag{13}$$

Nota. Tenuto conto del legame tra varianza e scarto quadratico medio, si scrive spesso s^2 (rispettivamente σ^2) in luogo di Var.

Esempi. Con riferimento ai dati numerici dell'esempio 1.2, e tenuto presente che $\overline{x} = 174,6$, si calcola:

i	x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
1	176	1,4	1,96
2	181	6,4	40,96
3	168	-6,6	43,56
4	176	1,4	1,96
5	172	-2,6	6,76
	Somma		95,6

da cui:

$$s^2 = \frac{95,6}{5} \simeq 19,1$$

e infine:

$$s = \sqrt{19,1} \simeq 4,37.$$

Con lo stesso procedimento, a partire dai dati numerici dell'esempio 1.3 del Par prec. si calcola:

$$s^2 = \frac{4219,2}{5} \simeq 844$$

e dunque:

$$s \simeq 29,05.$$

Lo scarto quadratico medio di questo secondo esempio è sensibilmente più grande di quello del primo esempio, a conferma del fatto che s misura la maggiore o minore dispersione dei singoli valori rispetto alla media.

Naturalmente, se nel calcolo della varianza o della deviazione standard gli scarti $x_i - \overline{x}$ compaiono con determinate frequenze, occorre tenere conto di tali frequenze (esattamente come nel caso delle medie ponderate di cui al Par. prec.). Più esplicitamente,

siano dati certi numeri x_i (i = 1,2, ..., k), ciascuno con frequenza assoluta f_i e sia $\overline{x}$ la loro media aritmetica (sempre ponderata, s'intende). La varianza e lo scarto quadratico medio sono espressi allora rispettivamente dalle formule:

$$Var = \frac{\sum_{i=1}^{k} f_i \cdot (x_i - \overline{x})^2}{\sum_{i=1}^{k} f_i} \tag{14}$$

$$s = \sqrt{Varianza} = \sqrt{\frac{\sum_{i=1}^{k} f_i \cdot (x_i - \overline{x})^2}{\sum_{i=1}^{k} f_i}} \tag{15}$$

Per es., partendo dai dati numerici dell'esempio 1.4 del Par. prec., si calcola:

$$s^2 = \frac{200,84}{44} \simeq 4,56$$

e

$$s \simeq 2,14.$$

Spesso le tecniche statistiche qui esposte vengono applicate non all'intera popolazione, ma solo ad un suo campione. Si cerca poi di stimare nel miglior modo possibile le caratteristiche dell'intera popolazione a partire dalle informazioni desunte dal campione. Quando si opera in questo modo, conviene modificare leggermente le formule 12 e 13, ponendo a denominatore il numero $n - l$ in luogo del numero n:

$$Var = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n - 1} \tag{16}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n - 1}} \tag{17}$$

Per evitare fraintendimenti, si parla allora di **varianza stimata** (formula 16) e di **scarto quadratico medio stimato** o di **deviazione standard stimata** (formula 16). Il motivo di questa modifica trova la sua giustificazione sulla base del computo dei cosiddetti "gradi di libertà", concetto importante che però in questa sede non approfondiremo. Va da sé che, per n abbastanza grande, la diversità tra *varianza* e *varianza stimata*, come pure tra *scarto quadratico medio* e *scarto quadratico medio stimato*, diventa trascurabile.

Esempio 5.4. A partire dai dati numerici dell'esempio 1.4 del si calcola:

$$\text{Varianza stimata} = \frac{200,84}{43} \simeq 4,67$$

e quindi

$$\text{Scarto quadratico medio stimato} \simeq 2,16$$

Dunque lo scarto quadratico medio stimato ($\simeq 2,16$) differisce dallo scarto quadratico medio ($\simeq 2,14$) appena nella seconda cifra dopo la virgola.

Distanza interquartile

Ecco infine la definizione di un altro indice di dispersione, che si ricollega alla nozione di mediana. Ricordiamo preliminannente che, dopo avere riordinato gli n numeri x_i per valori crescenti, la mediana M_e suddivide questo insieme di numeri in due parti ugualmente numerose. Nulla vieta di suddividere lo stesso insieme ordinato di numeri in quattro parti ugualmente numerose. Se per es. $n = 27$, si comincia col determinare la mediana: $M_e =$ elemento di posto centrale nell'insieme ordinato dei 27 valori x_i, ossia x_{14}. Si determina poi l'elemento di posto centrale nel sottoinsieme ordinato, formato dai 13 valori x_i che precedono M_e, ossia x_7; analogamente si determina l'elemento di posto centrale nel sottoinsieme ordinato, formato dai 13 valori x_i che seguono M_e, ossia x_{21}. I tre valori così ottenuti:

$$q_l = x_7 \quad q_2 = M_e = x_{14} \quad q_3 = x_{21}$$

vengono detti **quartili** e più precisamente, nell'ordine, primo, secondo, terzo quartile. Naturalmente, se si applica il procedimento or ora descritto ad un insieme ordinato costituito da un numero qualsiasi n di valori x_i, può capitare che qualcuno dei sottoinsiemi da suddividere in due parti ugualmente numerose sia formato da un numero pari di elementi; in tal caso, come valore del corrispondente quartile si assume, al solito, la semisomma dei due valori più prossimi al posto centrale.

Con queste notazioni, si considera come ulteriore indice di dispersione la cosiddetta **distanza interquartile**, definita da $\Delta = q_3 - q_l$. Per definizione, quindi, la distanza interquartile "taglia via" il 25% dei valori più bassi e il 25% dei valori più alti.

Esempio 5.5. Nel caso della tabella dell'esempio 1.4 del Par. prec. il primo quartile è dato dalla semisomma dell'11^a e della 12^a altezza; quindi $q_l = 170$. Il terzo quartile è dato dalla semisomma della 32^a e della 33^a altezza; quindi $q_3 = 172$. La distanza interquartile è dunque $\Delta = q3 - ql = 172 - 170 = 2$.

Graficamente, la dispersione di una serie di misure può essere visualizzata efficacemente con uno schema del tipo di quello disegnato in fig.:

il rettangolo centrale racchiude le misure comprese tra il primo e il terzo quartile. Il segmento sulla sinistra rappresenta l'intervallo entro cui variano le misure inferiori

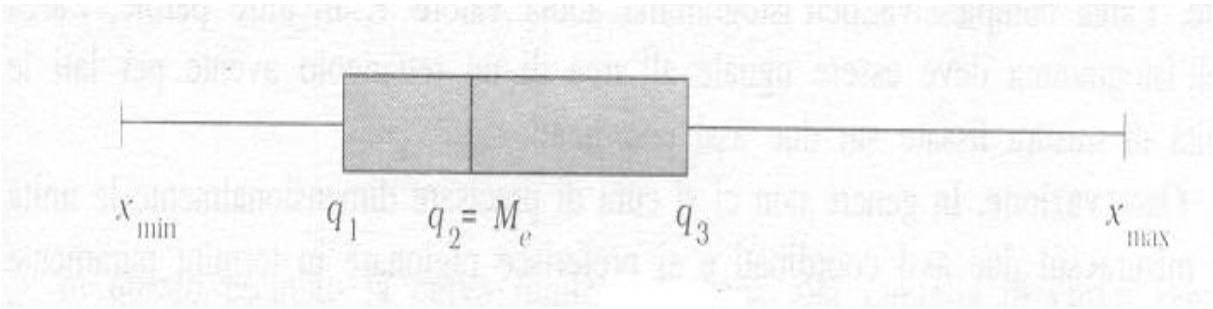

Figura 1:

al primo quartile; il segmento sulla destra rappresenta l'intervallo entro cui variano le misure superiori al terzo quartile.

Esercizio 5.6. A partire dai dati riportati nella tabella dell'esercizio 4.3 calcolate lo scarto quadratico medio e la distanza interquartile.

Esercizio 5.7. Calcolate lo scarto quadratico medio e la distanza interquartile per i pesi degli individui considerati nella tabella dell'esercizio 4.7.

Esercizio 5.8. Calcolate lo scarto quadratico medio e la distanza interquartile per le durate dei ricoveri ospedalieri considerati nella tabella dell'esercizio 4.8.

Esercizio 5.9. (a) Scegliete una dozzina di dati numerici (non tutti uguali tra loro) tali che la loro media aritmetica sia il numero 25. Calcolate il corrispondente scarto quadratico medio.

(b) Scegliete un'altra dozzina di dati numerici, facendo in modo che la loro media aritmetica sia ancora il numero 25, mentre lo scarto quadratico medio sia il doppio di quello calcolato in (a).

Esercizio 5.10. Prendendo spunto dal famoso sonetto "La statistica" di Trilussa, supponete di dover ripartire 800 polli tra 1600 individui; Ecco 4 possibili criteri di suddivisione:

(a) Si dà mezzo pollo a ciascun individuo.

(b) Si dà un pollo a 800 individui ("fortunati" o "raccomandati") e nulla ai restanti 800 individui.

(c) Si danno due polli a 400 individui (particolarmente "fortunati" o "raccomandati") e nulla ai restanti 1200 individui.

(d) Si danno tutti gli 800 polli ad un unico individuo ("super-fortunato" o "super-raccomandato") e nulla ai restanti 1599 individui.

Per ciascuna delle quattro ripartizioni, calcolate la media aritmetica e lo scarto quadratico medio.

6 La distribuzione normale

Per costruire l'istogramma delle frequenze di un insieme di misure di una grandezza che può variare con continuità, si suddivide l'intero intervallo delle misure in un numero finito n di intervallini (di solito tutti della stessa ampiezza). Si assume poi ciascun intervallino come base di una "canna d'organo" dell'istogramma, facendo in modo che la corrispondente area risulti proporzionale al numero delle misure che cadono entro l'intervallino considerato. Per evitare problemi di scala, conviene inoltre fare una volta per tutte la convenzione che, indipendentemente dal numero delle misure considerate, l'area complessiva dell'istogramma abbia valore 1. In altre parole, l'area dell'istogramma deve essere uguale all'area di un rettangolo avente per lati le unità di misura fissate sui due assi coordinati.
Osservazione. In genere non ci si cura di precisare dimensionalmente le unità di misura sui due assi coordinati e si preferisce ragionare in termini puramente numerici. Ciò dipende da una certa difficoltà ad interpretare intuitivamente il significato della grandezza che va posta sull'asse y. Fortunatamente, in questo contesto, l'interpretazione dimensionale delle grandezze in gioco è abbastanza irrilevante, in quanto tutto è riconducibile ad un confronto di aree.

Facciamo ora l'ulteriore ipotesi, che la popolazione considerata sia molto numerosa (costituita da una quantità praticamente illimitata di individui). In tal caso il numero n degli intervallini può essere aumentato a piacere, diminuendone corrispondentemente le ampiezze. Si ottengono "canne d'organo" via via più sottili e istogrammi via via più regolari, che in genere tendono a stabilizzarsi intorno ad una forma limite, approssimabile con una curva continua, detta curva di distribuzione delle frequenze, come illustrato in fig 2.

In questo esempio la curva limite appartiene alla famiglia di curve aventi equazioni del tipo:

$$y = Ae^{-B(x-C)^2}$$

con A, B, C parametri opportuni, dei quali ci occuperemo tra un momento.

Una siffatta distribuzione delle frequenze si chiama **distribuzione normale** o **distribuzione gaussiana**. Non sempre un insieme di misure tende a disporsi secondo

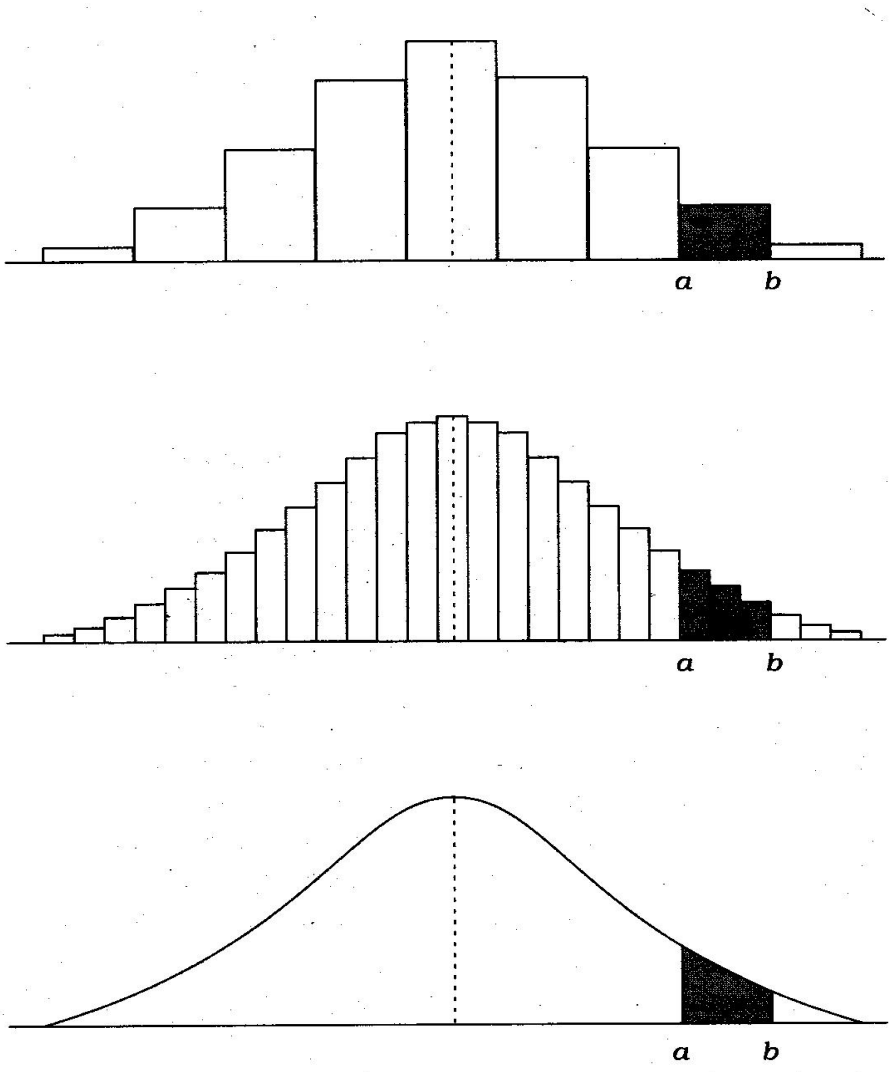

Figura 2:

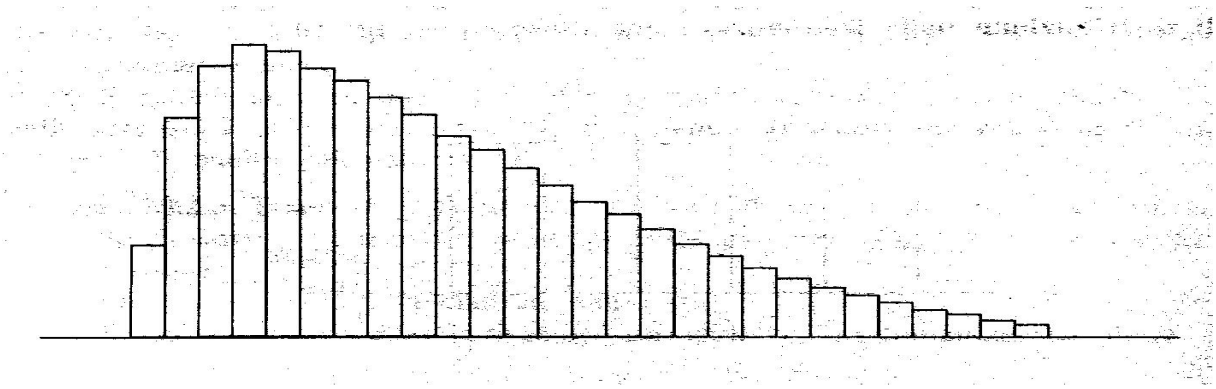

Figura 3:

una distribuzione gaussiana. Per es. in fig. 3 è visualizzata una distribuzione di dati chiaramente non gaussiana.

La constatazione se un insieme di misure sperimentali sia approssimabile o meno con una distribuzione gaussiana è un fatto di natura sperimentale. Tuttavia in certi casi si può prevedere anche sulla base di considerazioni teoriche che un certo insieme di dati sperimentali avrà un andamento gaussiano. Per es. è ben noto che se uno stesso sperimentatore, o sperimentatori diversi, ripetono più volte la misura di una medesima grandezza (sia essa il numero dei leucociti/mm^3 nel sangue di un ammalato oppure la distanza Terra-Sole) i risultati delle singole misure in generale non coincidono tra loro, per effetto della presenza di numerosi piccoli errori casuali. Le misure tendono però ad addensarsi in prossimità di un valore centrale, identificabile con la loro media aritmetica, dando luogo ad una distribuzione di tipo gaussiano. Se le misure non sono affette da errori sistematici (dovuti per es. ad un'errata taratura degli strumenti) è ragionevole assumere tale valore centrale come misura "vera" della grandezza in esame. Quanto ai parametri A, B, C che caratterizzano la curva gaussiana "limite degli istogrammi desunti da un certo insieme di dati sperimentali, si potrebbe pensare di determinarli per tentativi, modificandoli uno alla volta e tracciando le corrispondenti curve, fino ad ottenere una buona approssimazione degli istogrammi desunti dai dati sperimentali. Tuttavia, *se si sa già che la distribuzione è di tipo gaussiano*, la determinazione dei valori numerici di A, B, C può essere ricondotta al solo calcolo della media aritmetica, che in questo contesto si denota tradizionalmente con μ, e dello scarto quadratico medio, che in questo contesto si denota tradizionalmente con σ. Risulta infatti:

$$A = \frac{1}{\sigma\sqrt{2\pi}} \quad B = \frac{1}{2\sigma^2} \quad C = \mu$$

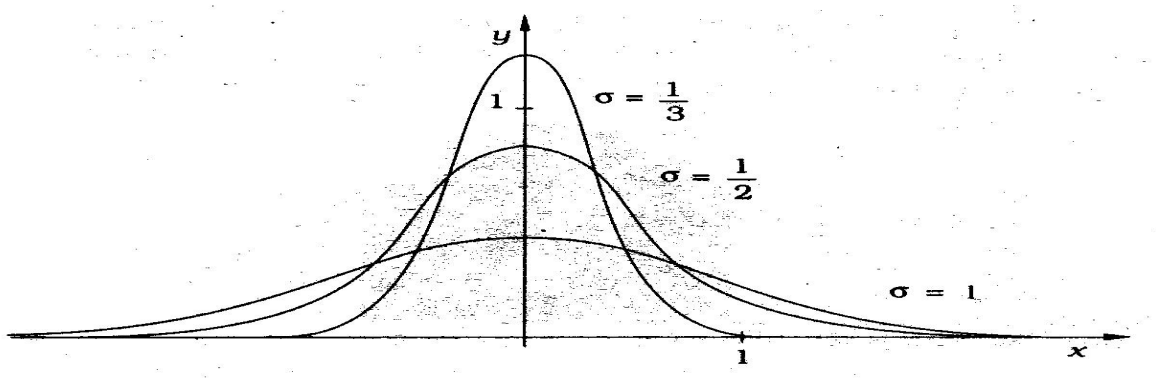

Figura 4:

In altre parole, se si ha a che fare con una distribuzione gaussiana di cui si conosce la media aritmetica, μ e lo scarto quadratico medio σ, la corrispondente gaussiana è il grafico della funzione:

$$y = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{18}$$

Cerchiamo ora di interpretare il significato dei tre parametri A, B, C. Il valore di C si spiega facilmente: la distribuzione gaussiana è simmetrica e i valori delle singole misure si addensano intorno alla loro media aritmetica. Quindi la curva gaussiana teorica deve avere un massimo proprio in corrispondenza al valore $C = \mu$. Il valore $\frac{1}{2\sigma^2}$ assunto da B determina la maggiore o minore "ripidità" della curva gaussiana, e dipende quindi dalla maggiore o minore dispersione dei dati: quanto più σ è piccolo, tanto più la curva è "ripida", quanto più σ è grande, tanto più la curva è "piatta" (vedi fig. 4).

Infine, il valore $\frac{1}{\sigma\sqrt{2\pi}}$ attribuito ad A serve a fare sì che l'area complessiva racchiusa tra la curva gaussiana e l'asse delle ascisse abbia misura unitaria, secondo quanto convenuto all'inizio di questo paragrafo.

Esempio 6.1. Esempio. Se $\mu = 0$ e $\sigma = 1$, l'equazione 18 assume una forma particolarmente semplice, detta **curva normale standardizzata**:

$$y = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}} \tag{19}$$

La convenzione di denotare la media aritmetica e lo scarto quadratico medio con le lettere greche μ e σ facilita una distinzione tra questi valori, riferiti all'intera popolazione (che, come già detto, si deve supporre costituita da una quantità "illimitata" di

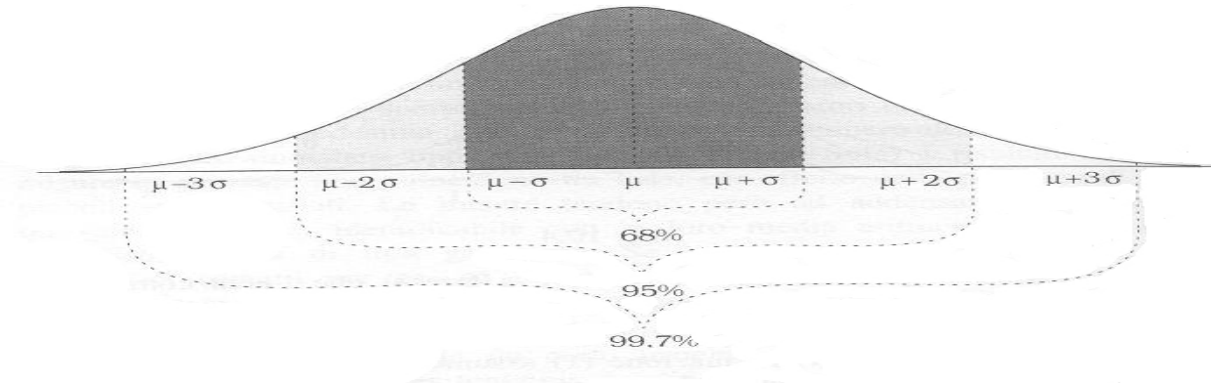

Figura 5:

individui) ed i valori $\overline{x}$ ed s, relativi ad un sottoinsieme della popolazione scelto come "campione". Mentre μ e σ sono valori teoricamente ben individuati, ma in genere sconosciuti per l'impossibilità pratica di effettuare le misure su tutta la totalità degli individui della popolazione, $\overline{x}$ ed s rappresentano solo delle stime di tali valori. Queste stime presentano il vantaggio di poter essere facilmente calcolate a partire dalle misure effettuate sugli individui del campione prescelto, ma presentano al tempo stesso l'inconveniente di dipendere di volta in volta dal particolare campione esaminato.
Ancora un'osservazione. Nella curva limite sono ormai scomparse le "canne d'organo" degli istogrammi da cui eravamo partiti. Nondimeno, fissati due valori qualsiasi a e b sull'asse delle ascisse, possiamo considerare l'area del corrispondente sottografico (regione tratteggiata fig. 10.2).
Questa area rappresenta la porzione delle misure che hanno un valore compreso fra a e b. Se per es. l'area costituisce l'$8,6\%$ dell'area totale, vorrà dire che circa l'$8,6\%$ delle misure della grandezza considerata cade entro l'intervallo $[a, b]$. Orbene, indipendentemente dal fatto che la curva gaussiana sia più o meno ripida, si dimostra il seguente importante risultato (vedi fig. 5):
nell'intervallo $[\mu - \sigma, \mu + \sigma]$ cade circa il 68% delle misure;
nell 'intervallo $[\mu - 2\sigma, \mu + 2\sigma]$ cade circa il 95% delle misure;
nell'intervallo $[\mu - 3\sigma, \mu + 3\sigma]$ cade circa il $99,7\%$ delle misure.

Volendo conoscere le misure delle aree che cadono entro intervalli di ampiezze diverse da quelle or ora segnalate, si consultano le apposite tavole, calcolate una volta per tutte (cfr. tab. 6). Se ne desume per es. che circa il 50% delle misure risulta compreso nell'intervallo $[\mu - 0.7\sigma, \mu + 0,7\sigma]$. Per comodità di consultazione, la tabella 6 è

Tabella 2: VALORI DELLE AREE SOTTESE DALLA CURVA GAUSSIANA

Valori di μ	Nell'intervallo $[\mu - u\sigma, \mu + u\sigma]$	Fuori dell'intervallo $[\mu - u\sigma, \mu + u\sigma]$	Nell'intervallo $[\mu + u\sigma, +\infty)$
0	0	1	0,5
0,2	0,1586	0,8414	0,4207
0,4	0,3108	0,6892	0,3446
0,6	0,4514	0,5486	0,2743
0,8	0,5762	0,4238	0,2119
1	0,6826	0,3174	0,1587
1,2	0,7698	0,2302	0,1151
1,4	0,8384	0,1616	0,0808
1,6	0,8904	0,1096	0,0548
1,8	0,9282	0,0718	0,0359
2	0,9544	0,0456	0,0228
2,2	0,9722	0,0278	0,0139
2,4	0,9836	0,0164	0,0082
2,6	0,9906	0,0094	0,0047
2,8	0,9950	0,0050	0,0025
3	0,9974	0,0026	0,0013
3,2	0,9986	0,0014	0,0007

articolata su varie colonne. In realtà la conoscenza dei valori di una colonna consente di dedurre facilmente i corrispondenti valori delle altre colonne. Per es. i valori scritti nella seconda e nella terza colonna hanno sempre somma 1 (area totale sottesa dalla curva gaussiana). Analogamente, i valori della terza colonna sono sempre doppi dei valori della quarta colonna (per simmetria).

Un'ulteriore nozione di uso frequente è il cosiddetto *errore standard della media*. Ecco di cosa si tratta.

Se le misure di una certa grandezza (si pensi per es. alla solita altezza delle reclute) vengono effettuate su un campione formato da n individui estratti casualmente dall'intera popolazione, nasce il problema di stabilire entro quali limiti di precisione l'altezza media $\overline{x}$ calcolata per il solo campione (e quindi nota) può essere assunta come stima per l'altezza media (ma sconosciuta) relativa all'intera popolazione. A tal fine, detto s lo scarto quadratico medio riscontrato sul campione, si introduce la nozione di **errore standard della media**, in simboli $e.s.m$, ponendo per definizione:

$$e.s.m. = \frac{s}{\sqrt{n}}$$

Con considerazioni che si riallacciano alle proprietà delle distribuzioni gaussiane, ma che sono valide anche per distribuzioni non gaussiane dei dati di partenza, si dimostra che la media (sconosciuta) sull'intera popolazione cade entro l'intervallo $[\overline{x} - e.s.m., \overline{x} + e.s.m.]$ nel 68% dei casi, cade entro l'intervallo $[\overline{x} - 2e.s.m., \overline{x} + 2e.s.m.]$ nel 95% dei casi, e cade entro l'intervallo $[\overline{x} - 3e.s.m., \overline{x} + 3e.s.m.]$ nel $99,7\%$ dei casi. Attenzione a non confondere l'*errore standard della media* con lo *scarto quadratico medio* Lo scarto quadratico medio è una caratteristica propria della popolazione (misura quanto questa è dispersa rispetto alla media). L'errore standard della media è invece un indice di quanto bene una media "campionaria" riesce ad approssimare la media "globale" sull'intera popolazione. Fissata la numerosità n del campione, l'errore standard della media è tanto più piccolo (e quindi l'approssimazione è tanto migliore) quanto minore è la dispersione della popolazione. In ogni caso, però, l'errore standard della media può essere reso arbitrariamente piccolo pur di aumentare la numerosità n del campione. Utilizzando la nozione di *errore standard della media* siamo ora in grado di precisare meglio le convenzioni in uso per caratterizzare l'entit dell'*errore* dal quale si ritiene che possano essere affette le misure sperimentali di una grandezza fisica. Una scrittura del tipo $a \pm \Delta a$ significa che la misura della grandezza in questione è compresa nell'intervallo $[a - \Delta a, a + \Delta a]$. Orbene, mentre questa convenzione esprime una certezza quando si parla delle scritture decimali troncate (o arrotondate) di un fissato numero reale (per es. $3,141592 < \pi < 3,141593$), nel caso delle misure sperimentali non è possibile pervenire ad un'analoga certezza. Poiché però, in assenza di errori sistematici, la media aritmetica di una serie di misure fisiche della stessa grandezza tende ad approssimare la misura "vera", appare ragionevole ripiegare sulla seguente **convenzione**. In una scrittura del tipo $a \pm \Delta a$, il valore a rappresenta la *media aritmetica* della serie delle misure sperimentali effettuate e Δa rappresenta l'ampiezza del corrispondente *errore standard della media*. Anticipando una terminologia che sarà precisata successivamente, possiamo dunque concludere che, in base a questa convenzione, la misura "vera" (ma sconosciuta) della grandezza in questione sarà contenuta nell'intervallo $[a - \Delta a, a + \Delta a]$ con una probabilità del 68% circa. L'intervallo $[a - \Delta a, a + \Delta a]$ viene detto **intervallo di confidenza** della misura.

Nota. Taluni autori preferiscono assumere come valore di Δa il *doppio* dell'errore standard della media. Ovviamente, se si adotta quest'altra convenzione, la probabilità che la misura "vera" della grandezza sia contenuta nell'intervallo $[a - \Delta a, a + \Delta a]$ sale al 95% circa. L'uso, in ambito sperimentale, della nozione di *intervallo di confidenza* impone anche una modifica alle regole di propagazione degli errori. Un teorema assicura

che la somma (o la differenza) di due distribuzioni gaussiane è ancora una distribuzione gaussiana, che ha come *media* la somma (o la differenza) delle medie e come *varianza* (in entrambi i casi) la somma delle varianze. Pertanto, date due grandezze fisiche indipendenti $a \pm \Delta a$, $b \pm \Delta b$, la loro somma e la loro differenza avranno rispettivamente espressioni della forma $s \pm \Delta s$, $d \pm \Delta d$, con:

$$s = a + b \qquad d = a - b$$

e

$$(\Delta s)^2 = (\Delta a)^2 + (\Delta b)^2 \qquad (\Delta d)^2 = (\Delta a)^2 + (\Delta b)^2$$

da cui, infine, estraendo le radici quadrate:

$$\Delta s = \sqrt{(\Delta a)^2 + (\Delta b)^2} \qquad \Delta d = \sqrt{(\Delta a)^2 + (\Delta b)^2}$$

Con ragionamenti analoghi si prova che per il *prodotto* $p \pm \Delta p$ e per il *quoziente* $q \pm \Delta q$ sussistono le espressioni

$$p = a \cdot b \qquad q = a/b$$

e

$$\left(\frac{\Delta p}{p}\right)^2 = \left(\frac{\Delta a}{a}\right)^2 + \left(\frac{\Delta b}{b}\right)^2 \qquad \left(\frac{\Delta q}{q}\right)^2 = \left(\frac{\Delta a}{a}\right)^2 + \left(\frac{\Delta b}{b}\right)^2$$

da cui infine, estraendo le radici quadrate:

$$\frac{\Delta p}{p} = \sqrt{\left(\frac{\Delta a}{a}\right)^2 + \left(\frac{\Delta b}{b}\right)^2} \qquad \frac{\Delta q}{q} = \sqrt{\left(\frac{\Delta a}{a}\right)^2 + \left(\frac{\Delta b}{b}\right)^2}$$

Nota. Le ampiezze degli intervalli di confidenza di somme, differenze, prodotti e quozienti, espresse da queste fonnule, sono sempre minori o uguali di quelle usate nell'ambito delle regole di propagazione degli errori. La discordanza si spiega in termini probabilistici, in quanto l'eventualità di compensazioni tra errori di segno opposto rende meno frequente il verificarsi del caso più sfavorevole (somma di errori di ugual segno).

Esercizio 6.2. Scrivete l'equazione della gaussiana relativa alle altezze del gruppo di reclute considerato nell'esempio 1.4 del Par. 1.

Esercizio 6.3. Supponete che la distribuzione dei pesi degli individui di una popolazione abbia una distribuzione gaussiana con media $\mu = 61$ kg e scarto quadratico medio $\sigma = 5$ kg.

(a) Scrivete l'equazione della gaussiana relativa ai pesi di tale popolazione e tracciatene il grafico.

(b) Calcolate la percentuale degli individui di quella popolazione, il cui peso è:

1. inferiore a 56 kg
2. superiore a 66 kg
3. inferiore a 53 kg
4. superiore a 69 kg
5. compreso tra 59 e 63 kg.

e interpretate i risultati sul grafico della gaussiana.

Esercizio 6.4. Le altezze di un certo gruppo di reclute sono distribuite con buona approssimazione secondo una curva gaussiana con media $\mu = 170$ cm e scarto quadratico medio $\sigma = 5$ cm. Le divise sono disponibili in 5 taglie:
A: per individui di altezza ≤ 161 cm
B: per individui di altezza compresa tra 161 e 167 cm
C: per individui di altezza compresa tra 167 e 173 cm
D: per individui di altezza compresa tra 173 e 179 cm
E: per individui di altezza ≥ 179 cm.
Stimate il numero di divise delle varie taglie occorrenti per 750 reclute.

Esercizio 6.5. A partire dalla distribuzione statistica delle altezze x delle reclute considerata nell'esercizio 6.4, e nell'ipotesi che tutte le reclute siano individui "ben proporzionati", utilizzate la formula $y = 13,2x^3$ per calcolare i pesi y corrispondenti alle altezze:

$$\mu - \sigma = 1,65m \quad \mu = 1,70m \quad \mu + \sigma = 1,75m.$$

Notate qualche asimmetria tra la distribuzione delle altezze e quella dei pesi?

Esercizio 6.6. Considerate l'equazione della generica curva gaussiana (con media μ e scarto quadratico medio σ). Verificate, mediante derivazione:

(a) Che la gaussiana assume il suo valore massimo nel punto μ.

(b) Che la gaussiana presenta due flessi, rispettivamente nei punti $\mu - \sigma$ e $\mu + \sigma$.

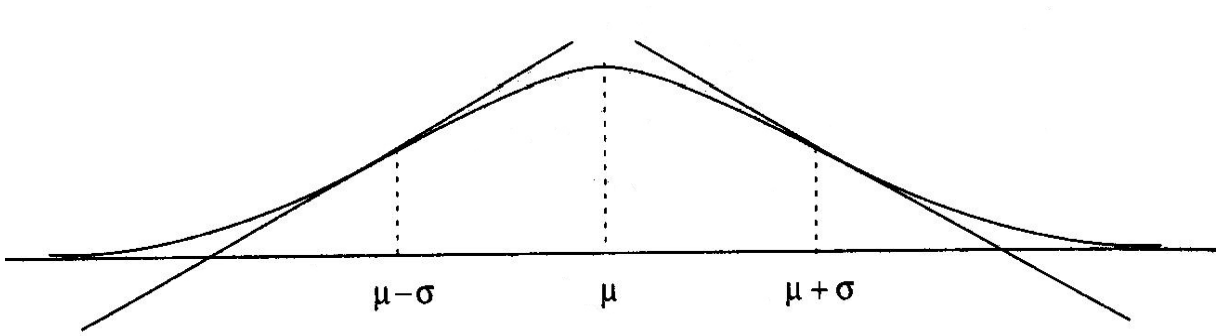

Figura 6:

7 Distribuzioni a due caratteri

Le situazioni prese in esame nei paragrafi precedenti si riferivano sempre ad un'unica caratteristica della popolazione in esame. Spesso interessa invece considerare simultaneamente due caratteristiche quantitative degli individui di una stessa popolazione, per stabilire se esiste una qualche relazione tra l'una e l'altra. Per es., nel caso di una popolazione di individui adulti, supponiamo di voler cercare una eventuale relazione tra pressione arteriosa ed età (ma nulla vieta di cercare eventuali relazioni tra altre coppie di grandezze, per es. tra pressione arteriosa e peso, oppure tra pressione arteriosa e numero di sigarette fumate mediamente al giorno). Numeriamo gli individui della popolazione da 1 ad n e associamo all'$i-$esimo individuo la coppia ordinata di numeri $(x_i; y_i)$, dove x_i denota la sua età (misurata per es. in anni) e y_i denota la sua pressione arteriosa (misurata per es. in mm di Hg). In un sistema di coordinate cartesiane del piano, ogni coppia $(x_i; y_i)$ individua un punto P_i, e il complesso degli n punti forma una specie di "nube". Orbene, a seconda delle coppie di grandezze prese in esame, questa nube può presentare delle regolarità più o meno appariscenti.

Se la nube è del tipo visualizzato in fig. 7 a, si intuisce che al crescere dei valori di x anche i corrispondenti valori di y tendono a crescere (si parla allora di una *concordanza* o di una *correlazione positiva*); se invece la nube è del tipo visualizzato in fig. 10.6 b, si intuisce che al crescere dei valori di x i corrispondenti valori di y tendono a diminuire (si parla allora di una *discordanza* o di una *correlazione negativa*); se la nube è del tipo visualizzato in fig. 7 c, si intuisce che al crescere dei valori di x i valori di y si mantengono sostanzialmente costanti (si parla allora di *indifferenza* della y rispetto ad x). Infine, se la nube del tipo visualizzato in fig. 10.6 d, si deve concludere che i dati a disposizione *non evidenziano alcuna correlazione* tra le due grandezze considerate.

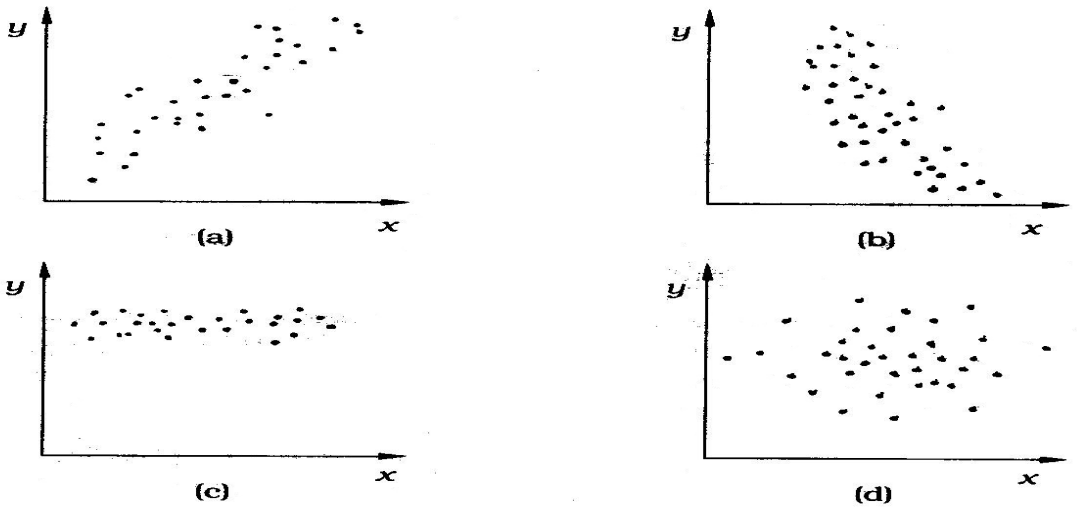

Figura 7:

Ritornando all'esempio specifico dell' età e della pressione arteriosa, supponiamo di disporre della seguente tabella di dati, su un campione che per ragioni di semplicità riterremo costituito da soli 7 individui.

Età	Pressione
25	120
30	125
42	135
55	140
55	145
63	140
70	160

Siamo dunque nel caso della fig. 7 a.
Quando si presume che tra due variabili x, y possa sussistere una relazione di dipendenza della y dalla x schematizzabile in termini matematici mediante una funzione lineare, si usa tracciare la cosiddetta *retta di regressione*, cioè la retta che meglio approssima la nube dei dati. Occorre però precisare ancora cosa si debba intendere per "migliore

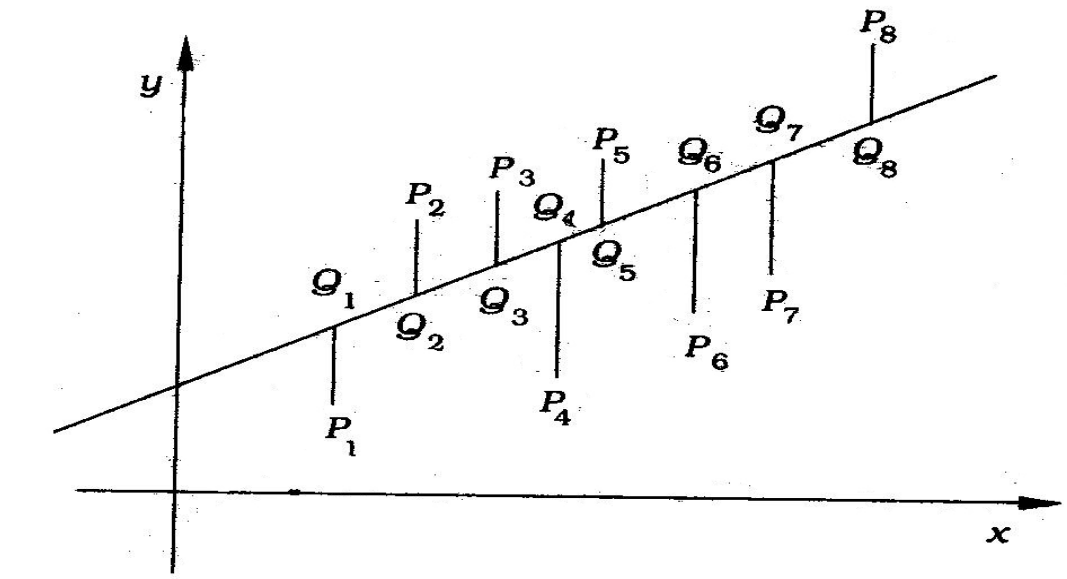

Figura 8:

approssimazione".
Ecco la definizione precisa: data una retta generica s, si tracciano i segmenti paralleli all'asse y che congiungono i punti dati $P_i = (x_i; y_i)$ con i punti $Q_i = (x_i; y_i')$ di uguale ascissa, posti sulla retta s (vedi figura). Si calcolano quindi i quadrati delle lunghezze di tali segmenti e infine se ne fa la somma:

$$\sum_{i=1}^{n}(y_i - y_i')^2$$

Fermi restando i punti P_i, il valore di questa espressione dipende evidentemente da s. Orbene, si dimostra che c'è una posizione di s, che rende minimo tale valore. La retta così individuata è la **retta di regressione** relativa alla nube di punti P_i.

Per determinare l'equazione cartesiana

$$y = a + bx$$

della retta di regressione conviene calcolare in primo luogo la media aritmetica $\overline{x}$ delle n ascisse x_i e la media aritmetica $\overline{y}$ delle n ordinate y_i. Dopodiché si dimostra che il

valore numerico di b è dato da:

$$b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Quanto al valore numerico di a, si dimostra che: $a = \overline{y} - b\overline{x}$ (b essendo sempre il numero del quale abbiamo fornito or ora l'espressione). Nel caso dell'esempio numerico da cui eravamo partiti (età e pressione in un insieme di 7 individui) a conti fatti e a seguito di opportuni arrotondamenti otteniamo la seguente equazione per la retta di regressione:

$$y = 0,73x + 102.$$

In sostanza, abbiamo espresso la pressione arteriosa y come funzione lineare dell' età x. Naturalmente si tratta solo di una schematizzazione matematica e non di un legame funzionale vero e proprio tra le due grandezze in esame: se avessimo preso in esame i dati di un altro campione, saremmo pervenuti in generale ad una retta di regressione diversa. Del resto anche i punti rappresentativi dei singoli individui del campione considerato sono pi o meno discosti dalla loro retta di regressione. Nonostante queste limitazioni, la conoscenza della retta di regressione utile per es. per stabilire se la pressione di un determinato individuo molto superiore o molto inferiore a quella che ci si aspetterebbe in base alla sua et. Concludiamo il paragrafo, riportando la formula del cosiddetto **coefficiente di correlazione** (di Pearson):

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

I valori che questo numero può assumere sono tutti compresi nell'intervallo [-1, 1]. Quando $r = -1$, si ha una correlazione negativa perfetta (vale a dire i punti P_i sono tutti perfettamente allineati su una retta con coefficiente angolare negativo); quando $r = l$, si ha una correlazione positiva perfetta (vale a dire i punti P_i sono tutti perfettamente allineati su una retta con coefficiente angolare positivo). Quanto più r si discosta dai valori -l e l, tanto meno preciso risulta l'allineamento dei punti P_i: per r prossimo a 0, non sussiste alcuna correlazione lineare fra le due variabili in esame (può darsi nondimeno che sussista qualche altro tipo di correlazione, di natura più complessa e quindi non esprimibile mediante funzioni lineari).

Esercizio 7.1. (a) Calcolate l'equazione della retta di regressione relativa alla tabella seguente

assumendo come variabile x l'anno dell'olimpiade e come variabile y il tempo stabilito nella gara dei 400 m stile libero, espresso in secondi.

Anni	Tempi (in min e s)
1908	5 37
1912	5 24
1920	5 27
1924	5 04
1928	5 02
1932	4 48
1936	4 44
1948	4 41
1952	4 31

(b) Sulla base dell'equazione della retta di regressione, quale tempo si può congetturare sia stato stabilito alle olimpiadi del 1984? E a quelle del 1988?
(c) Sempre sulla base dell'equazione della retta di regressione, quale tempo si può presumere che verr stabilito alle olimpiadi del 2088? E a quelle del 2188?

Esercizio 7.2. Considerate nuovamente la tabella dell'esercizio 4.6. È significativo calcolare le rette di regressione relative a certe coppie di colonne di tale tabella? In caso di risposte affermative, svolgete i calcoli. Confrontate quindi le informazioni desumibili dalle equazioni delle rette di regressione con quelle ottenute nell'esercizio 4.6.

Esercizio 7.3. In un gruppo di 5 adulti, la somministrazione di dosi diverse di un farmaco ha comportato le seguenti diminuzioni della pressione arteriosa:

Dose (in mg)	Diminuz. dela pressione (in $mmHg$)
7	10
12	18
15	20
20	25
22	25

(a) Scrivete l'equazione della retta di regressione.
(b) Calcolate la dose ottimale per ottenere una diminuzione della pressione pari a 15 $mmHg$.

Esercizio 7.4. . Si dispone dei seguenti dati, relativi alle altezze di un gruppo di reclute, e ai punteggi conseguiti dalle stesse reclute ad un test attitudinale:

Altezza	Punteggio
168	12
176	25
170	10
178	20
167	24
175	18

Calcolate l'equazione della retta di regressione e il coefficiente di correlazione. Quindi interpretate opportunamente i risultati ottenuti.

8 Spunti per ulteriori approfondimenti

Abbiamo affrontato solo una minima parte degli argomenti che costituiscono oggetto di studio per la statistica (e in particolare per la statistica medica). Per es., non abbiamo neppure accennato ad altre distribuzioni di probabilità -diverse dalla distribuzione normale -che pure sono molto importanti in svariati contesti applicativi. Ci siamo poi limitati al caso della regressione lineare, trascurando tutti gli altri tipi di regressione. Soprattutto non abbiamo parlato di un tema fondamentale, qual è l'*inferenza statistica*: uso di tecniche appropriate (test t, test χ^2, ecc.) per analizzare eventuali differenze tra due o più gruppi, in vista di confrontare per es. l'efficacia di diverse terapie. Per tutte queste tematiche rinviamo agli appositi trattati, e in particolare al libro di **S. A. Glantz**: Statistica per discipline bio-mediche (McGraw-Hill Libri Italia, 1997), a quello di *C. Rossi-G. Serio*: La metodologia statistica nelle applicazioni biomediche (Springer, 1990), e al libro di **P. Armitage, G. Berry**: Statistica medica (McGraw-HiII Libri Italia, 1996). Segnaliamo inoltre il capitolo di **G. Gallus e S. Milani**: Elementi di metodologia statistica con applicazioni, in *Medicamenta, voi II*, Cooperativa Farmaceutica Ed., Milano 1993, pagg. 689-835.