Prerequisiti:

- Conoscere adeguatamente il calcolo algebrico.
- Rappresentare punti e rette in un piano cartesiano.
- Possedere i primi elementi di probabilità e statistica.

OBIETTIVI DI APPRENDIMENTO

Una volta completata l'unità, gli allievi devono essere in grado di:

- *individuare situazioni che richiedono di rilevare lo stesso carattere su due soggetti o due caratteri diversi sullo stesso soggetto*
- *interpretare una tabella a doppia entrata*
- *costruire la distribuzione doppia delle frequenze di due variabili statistiche e rappresentarla graficamente anche con l'uso di uno strumento di calcolo automatico*
- *determinare le distribuzioni marginali di due variabili statistiche delle quali è nota la distribuzione doppia*
- *spiegare i concetti di dipendenza e indipendenza stocastica*
- *spiegare i concetti di connessione e correlazione di due variabili statistiche e rappresentare graficamente le due variabili correlate*
- *calcolare il coefficiente di correlazione di Bravais-Pearson con riferimento ad una situazione specifica*
- *spiegare il concetto di regressione di una variabile statistica su un'altra*
- *trovare la retta di regressione di una variabile statistica su un'altra*
- *spiegare quando la regressione è lineare*

L'unità è rivolta al 2° biennio di tutte le scuole superiori

Nozioni di statistica bivariata

Unità 55

55.1 CONSIDERAZIONI GENERALI

55.1.1 Riprendiamo alcuni concetti di statistica che già dovresti conoscere al fine di consolidarli e approfondirli.

- La **statistica descrittiva** è l'insieme dei procedimenti atti a raccogliere i dati – coerenti con l'obiettivo dell'indagine che si conduce – riguardanti tutti gli individui che compongono il collettivo o un opportuno campione rappresentativo. Tali dati sono chiamati ***dati statistici***.

Essi sono registrati in apposite tabelle – chiamate ***tabelle statistiche*** – ed eventualmente rappresentati con opportuni grafici (istogrammi, diagrammi cartesiani, diagrammi a torta, diagrammi a barre, eccetera). Sono quindi riassunti e descritti per mezzo di uno o più *valori di sintesi*, i cosiddetti ***indici di posizione*** e ***di dispersione***.

Tra gli indici di posizione, il valore più frequentemente usato è la ***media aritmetica***; tra quelli di dispersione ricordiamo la ***varianza*** e lo ***scarto quadratico medio*** (o ***deviazione standard***).

Supponiamo allora che i dati statistici siano quelli indicati dalla seguente successione di numeri:

[1] $$x_1, x_2, \ldots, x_n.$$

Si dicono, come noto, valori della **variabile statistica** X che descrive il fenomeno, cui i dati stessi si riferiscono.

La loro **media aritmetica**, indicata con $\mathbf{M(X)}$ o semplicemente con $\mathbf{M}$ o anche con $\boldsymbol{\mu}$, è tale che [(1)]:

$$M(X) = \frac{x_1 + x_2 + \cdots x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \ ;$$

mentre la **varianza** è il numero $\mathbf{Var(X)}$, indicato anche con $\boldsymbol{\sigma^2}$, tale che:

$$Var(X) = \sigma^2 = \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_n-\mu)^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \, ;$$

la radice quadrata di σ^2 è lo **scarto quadratico medio** (o **deviazione standard**), indicato con $\mathbf{dev(X)}$ o anche con $\boldsymbol{\sigma}$, cioè:

$$dev(X) = \sigma = \sqrt{\frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_n-\mu)^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

- La **statistica inferenziale** prende in esame solo qualche campione rappresentativo della collettività su cui verte l'indagine e ne ricava informazioni che possono estendersi all'intera popolazione. Per questo è detta anche **statistica induttiva**.

Le conclusioni cui essa giunge non sono certe ma soltanto probabili.

55.1.2 Oggigiorno l'analisi statistica riveste una notevole importanza in molti campi: scientifico, economico, sociale, politico, medico, eccetera. Alcuni esempi:

- Nelle scienze sperimentali (fisica, chimica, biologia, ...) si assume come misura più attendibile di una certa grandezza la media aritmetica di un certo numero di misure di quella grandezza, con un errore che viene calcolato ancora con considerazioni di tipo statistico.

---

(1) La sommatoria, cha a volte si scrive anche $\sum_{i=1}^{n} f(i)$, si legge "sommatoria per i che varia da 1 ad n di f(i)"; è un modo più compatto di indicare la somma f(1)+f(2)+...+f(n). Il simbolo **Σ** per indicare la "sommatoria" fu un'idea di Leonhard Euler (1707-1783).

- Le aziende produttrici ricorrono all'analisi statistica per valutare i gusti dei potenziali compratori.
- I partiti politici conducono sondaggi campionari per saggiare le tendenze dell'elettorato.
- Certe agenzie utilizzano i mezzi della statistica per le cosiddette "proiezioni", dopo il voto in una competizione elettorale.
- Le ditte farmaceutiche testano con metodi statistici l'efficacia di un farmaco, prima di immetterlo sul mercato.

55.1.3 Proponiamo alcune questioni con l'obiettivo di testare conoscenze e abilità che per la verità dovrebbero essere già state acquisite.

1. Sono assegnati 5 numeri. Sommandoli a 4 a 4 in tutti i modi possibili ma senza ripetizioni, si ottengono i seguenti numeri: 35, 38, 40, 43, 44. Quant' è la media aritmetica dei 5 numeri assegnati? [**R**. 10]
2. Sono assegnati 4 numeri. Sommando ciascuno di essi alla media aritmetica degli altri tre si ottengono i seguenti numeri: 19, 22, 32, 39. Quant'è la media aritmetica dei 4 numeri assegnati? [**R**. 14]
3. L'altezza media di un gruppo di giovani è 174 cm. Quella delle sole femmine del gruppo è 168 cm, mentre quella dei soli maschi è 176 cm. Se nel gruppo vi sono 4 femmine, quanti sono i maschi? [**R**. 12]
4. In un gruppo di amici fa la sua comparsa un nuovo venuto. La sua altezza supera di 5 cm l'altezza media degli amici del gruppo ma, dopo il suo arrivo, questa altezza media aumenta di 5 mm. Di quante persone è costituito il gruppo originario? [**R**. 9]

55.2 DISTRIBUZIONE STATISTICA DOPPIA

55.2.1 È probabile che la registrazione dei dati statistici mediante **tabelle a doppia entrata** ti sia già nota. Ci proponiamo comunque di approfondire l'argomento. Incominciamo con la descrizione di un esperimento.

Una sbarra S sia ottenuta incollando, una appresso all'altra, due sbarre S' ed S". Della sbarra S' si sono effettuate 20 misurazioni ed i valori ottenuti sono riassunti in apposita tabella (Tab. 1) e rappresentati graficamente (Fig. 1). Della sbarra S" si sono compiute 25 misurazioni ed i valori ottenuti sono riassunti in un'altra tabella (Tab. 2) e rappresentati con apposito istogramma (Fig. 2).

Tab. 1 – Misurazioni relative alla sbarra S'				
Misura (cm)	26,3	26,4	26,5	26,6
Frequenza assoluta	3	7	6	4

Tab. 2 – Misurazioni relative alla sbarra S"					
Misura (cm)	37,1	37,2	37,3	37,4	37,5
Frequenza assoluta	4	4	10	5	2

Osserviamo che le misure di S' ed S" si possono pensare come i valori assunti da due variabili statistiche che indichiamo rispettivamente con L' ed L". La tabella 1 sintetizza la distribuzione delle frequenze assolute di L'; la tabella 2 quella di L".

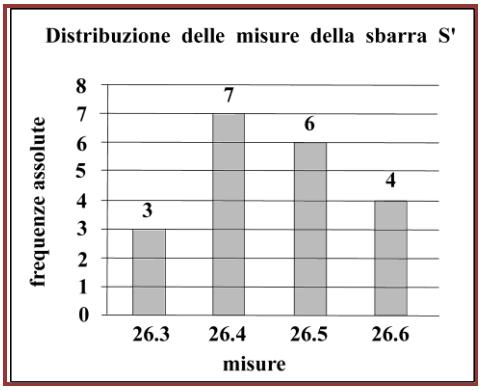

FIG. 1

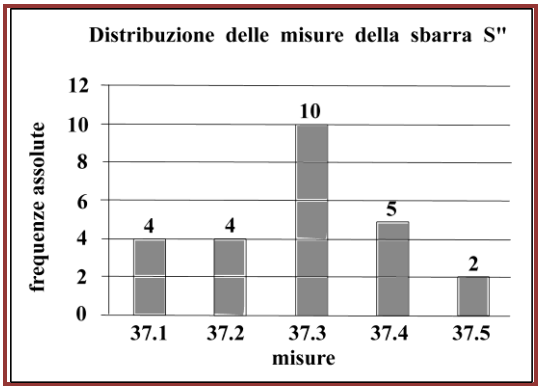

FIG. 2

Indicate con μ' e μ" le medie aritmetiche dei valori assunti rispettivamente dalle variabili L' ed L" e costatato che si tratta di medie ponderate, si ha:

$$\mu' = \frac{26{,}3 \cdot 3 + 26{,}4 \cdot 7 + 26{,}5 \cdot 6 + 26{,}6 \cdot 4}{3 + 7 + 6 + 4} \approx 26{,}455 \text{ (cm)};$$

$$\mu'' = \frac{37{,}1 \cdot 4 + 37{,}2 \cdot 4 + 37{,}3 \cdot 10 + 37{,}4 \cdot 5 + 37{,}5 \cdot 2}{4 + 4 + 10 + 5 + 2} \approx 37{,}288 \text{ (cm)}.$$

Se μ' e μ" sono assunte come le misure più attendibili di S' ed S" rispettivamente, sembra naturale assumere come misura più attendibile di S la media aritmetica μ delle misure di S.

Ma qual è questo valore μ, considerato che non sono state effettuate misurazioni dirette di S? È forse μ=μ'+μ"?

Proviamo a seguire questo ragionamento. La misura della sbarra S può essere pensata ottenuta dopo aver misurato separatamente i due pezzi, S' ed S", che la compongono. Per cui si può immaginare che una misura ottenuta per S sia 26,3+37,1.

Con quale frequenza?

Poiché 26,3 si presenta 3 volte come misura di S' e 37,1 si presenta 4 volte come misura di S", combinando ognuna delle volte in cui si presenta 26,3 con ciascuna delle volte in cui si presenta 37,1 possiamo concludere che la misura 26,6+37,1=63,4 di S si presenta 3×4=12 volte. Essa non si presenta con altre combinazioni.

Osserviamo invece che la misura 26,3+37,2=63,5 si ottiene anche sommando 26,4 e 37,1; quindi questa misura 63,5 si presenta con frequenza 3×4+7×4=40.

A sua volta, la misura 26,3+37,3=26,4+37,2=26,5+37,1=63,6 si presenta con frequenza 3×10+7×4+6×4=82.

Procedendo allo stesso modo, si ottiene una tabella (Tab. 3) che registra la distribuzione di frequenze assolute delle misure di S; misure che possiamo considerare come i valori assunti da una terza variabile statistica, che indichiamo con L.

Tab. 3 – Misurazioni relative alla sbarra S

Misura (cm)	63,4	63,5	63,6	63,7	63,8	63,9	64,0	64,1
Frequenza assoluta	12	40	82	125	117	84	32	8

Prova a disegnare un istogramma che rappresenti graficamente questa distribuzione.

Calcolando la media aritmetica della variabile statistica L, si trova: $\mu \approx 63{,}743$.
Siccome: $63{,}743=26{,}455+37{,}288$ effettivamente: $\mu = \mu'+\mu''$.

55.2.2 La distribuzione delle frequenze assolute della variabile statistica L (Tab. 3) è detta **distribuzione doppia** delle frequenze delle variabili statistiche L' ed L" e può essere meglio specificata da una tabella a doppia entrata (Tab. 4), che in definitiva riassume le operazioni prima descritte per giungere alla distribuzione di L.

Lunghezza L' / Lunghezza L"	26,3	26,4	26,5	26,6	Somma frequenze L"
37,1	3×4	7×4	6×4	4×4	80
37,2	3×4	7×4	6×4	4×4	80
37,3	3×10	7×10	6×10	4×10	200
37,4	3×5	7×5	6×5	4×5	100
37,5	3×2	7×2	6×2	4×2	40
Somma frequenze L'	75	175	150	100	500

TAB. 4

Va aggiunto che, per il modo com'è stata costruita, la variabile statistica L si dice ***somma delle variabili*** L' ed L" e si scrive: L = L'+L" .
Indicata per comodità con M(Z) la media aritmetica di una generica variabile statistica Z, la teoria – confermata del resto dall'esempio precedente – mostra che si ha:

$$M(X+Y) = M(X)+M(Y) .$$

55.2.3 Invece della somma di due variabili statistiche X ed Y, si può prendere in considerazione il loro ***prodotto*** XY, costruito con lo stesso criterio seguito per la costruzione di X+Y e di cui abbiamo visto un esempio.

- In questo caso, se le variabili statistiche X ed Y sono *indipendenti* – cioè se i valori assunti da X non influenzano quelli assunti da Y e, viceversa, questi non influenzano quelli – allora la teoria mostra che risulta:

$$M(XY) = M(X) \cdot M(Y) .$$

Con riferimento all'esempio esaminato in 55.2.1, se S' ed S" sono due lati consecutivi di un rettangolo, il valore più attendibile dell'area del rettangolo è evidentemente M(L'L").

Ti invitiamo a calcolare questo valore dopo aver determinato la distribuzione di frequenze della variabile statistica L'L" ed a verificare che risulta:

$$M(L'L'') = M(L')\cdot M(L'') \approx 987 \text{ cm}^2.$$

- L'analisi statistica fornisce numerosi esempi di coppie di variabili statistiche non indipendenti e per le quali in genere non vale l'ultima relazione considerata. Di queste variabili ci occuperemo nelle prossime pagine, benché sotto altri punti di vista. Prima, però, vogliamo soffermarci su alcune considerazioni supplementari, le quali, quantunque condotte attraverso un esempio, hanno valore generale.

Del lato S di un quadrato sono state eseguite alcune misurazioni e si sono ottenuti i valori sintetizzati nella seguente tabella (Tab. 5) e pensati come i valori assunti da una variabile statistica L. Vogliamo

calcolare l'area del quadrato.
Intanto si trova la misura più attendibile per il suo lato: $M(L) = 15{,}0\ \text{cm}$.

Tab. 5 – Misurazioni relative alla lunghezza L			
Misura (cm)	14,8	15,0	15,1
Frequenza assoluta	1	2	2

Come misura dell'area si assume il valore: $[M(L)]^2 = 225{,}00\ \text{cm}^2$. È forse $[M(L)]^2 = M(L^2)$?
Precisiamo che, quando ci si riferisce alla variabile statistica L^2, nel caso specifico s'intende quella che assume i valori:

$$14{,}8^2 \quad 15{,}0^2 \quad 15{,}1^2$$

con le frequenze assolute rispettivamente:

$$1,\ \ 2,\ \ 2\,.$$

Per cui, operati i calcoli necessari: $M(L^2) = 225{,}01\ \text{cm}^2$. Dunque $M(L^2) \neq [M(L)]^2$, anche se la differenza si presenta trascurabile, almeno in questo caso.
D'altronde, se consideriamo la variabile statistica LL, ragionando come nel caso del rettangolo, si trova per i suoi valori la distribuzione di frequenze assolute sintetizzata nella seguente tabella (Tab. 6).

Tab. 6 – Misurazioni relative alla variabile statistica LL						
Misura (cm^2)	219,04	222,00	223,48	225,00	226,50	228,01
Frequenza assoluta	1	4	4	4	8	4

Insomma LL ed L^2 sono due variabili statistiche distinte.
Si ottiene, a conti fatti: $M(LL) = 225{,}00\ \text{cm}^2$. Ossia, concordemente con la conclusione del paragrafo precedente: $M(LL) = M(L) \cdot M(L) = [M(L)]^2$.

55.2.4 Il fatto che i due valori $[M(L)]^2$ ed $M(L^2)$ calcolati sopra differiscano di una quantità trascurabile può far pensare che le due grandezze siano in realtà uguali e che la differenza sia dovuta semplicemente ad un errore di approssimazione. Le cose non stanno così ed effettivamente, considerata una generica variabile statistica X, si dimostra che è in generale:

$$[M(X)]^2 \neq M(X^2) \ \text{ e } \ [M(X)]^2 = M(XX).$$

Lo facciamo vedere, però, solo in una situazione particolarmente semplice.
Sia allora la seguente variabile statistica:

$$X = \begin{bmatrix} a & b \\ 1 & 3 \end{bmatrix}$$

Si ha, evidentemente:

$$M(X) = \frac{a + 3b}{4}\,.$$

Da qui segue:

$$[M(X)]^2 = \left(\frac{a + 3b}{4}\right)^2 = \frac{a^2 + 9b^2 + 6ab}{16}\,.$$

Consideriamo, adesso, la variabile statistica X^2:

$$X^2 = \begin{bmatrix} a^2 & b^2 \\ 1 & 3 \end{bmatrix}$$

È chiaramente:

$$M(X^2) = \frac{a^2 + 3b^2}{4}.$$

È evidente, dunque, che si ha: $[M(X)]^2 \neq M(X^2)$.

Costruiamo, infine, la variabile statistica XX. Si trova abbastanza facilmente:

$$XX = \begin{bmatrix} a^2 & b^2 & ab \\ 1 & 9 & 6 \end{bmatrix}$$

Perciò:

$$M(XX) = \frac{a^2 + 9b^2 + 6ab}{16}$$

e di conseguenza: $[M(X)]^2 = M(XX)$.

55.2.5 La distribuzione doppia delle frequenze delle variabili statistiche L' ed L" (cfr. Tab. 4 in 55.2.2) è stata ottenuta sulla base delle distribuzioni assegnate di tali variabili. Più spesso la distribuzione doppia è assegnata direttamente.

Valga, per tutti, il seguente esempio, nel quale (Tab. 7) è data la distribuzione, per aree geografiche e per tipologia, delle scuole secondarie di 2° grado impegnate nella sperimentazione nell'anno scolastico 1986/87.

Tab. 7 – Distribuzione per area geografica e per tipologia delle scuole secondarie di 2° grado impegnate nella sperimentazione nell'anno scolastico 1986-87

Area geografica / Tipologia di scuola	NORD	CENTRO	SUD	TOTALI
Istruzione CLASSICA	151	94	55	300
Istruzione TECNICA	218	125	185	528
Istruzione PROFESSIONALE	108	63	52	223
Istruzione ARTISTICA	11	4	3	18
TOTALI	488	286	295	1069

Una rappresentazione grafica (Fig. 3) ben si presta ad evidenziare le due caratteristiche (distribuzione per area geografica e per tipologia) riferite allo stesso soggetto statistico (le scuole impegnate nella sperimentazione). Non sono evidenziati i totali.

Si potrebbe ricorrere ad altri tipi di grafici, ma riteniamo che la modalità illustrata sia quella più indicata per tutte le situazioni in cui bisogna rappresentare le due caratteristiche di uno stesso soggetto statistico o la stessa caratteristica di due soggetti statistici.

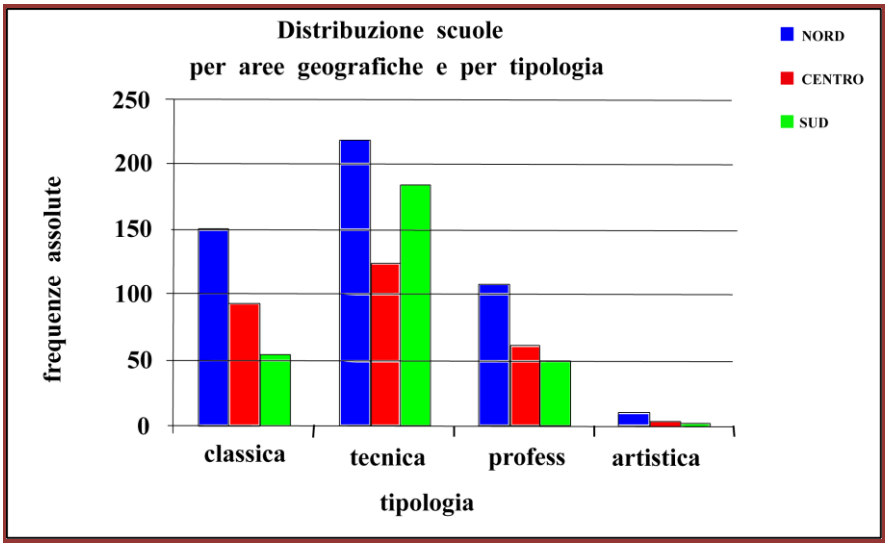

FIG. 3

55.2.6 La distribuzione doppia delle frequenze di due variabili X ed Y, suscettibili rispettivamente di m ed n determinazioni, è rappresentata in forma generale come nella tabella 8.
Questa *tabella a doppia entrata* è detta anche ***tabella di contingenza*** e la *distribuzione doppia* di frequenze che essa rappresenta è chiamata pure ***distribuzione congiunta*** delle variabili statistiche X ed Y.

Tab. 8 – Distribuzione doppia di frequenze

variabile Y / variabile X	y_1	y_2	y_3	...	y_n	distribuzione marginale di X
x_1	f_{11}	f_{12}	f_{13}	...	f_{1n}	S_1
x_2	f_{21}	f_{22}	f_{23}	...	f_{2n}	S_2
x_3	f_{31}	f_{32}	f_{33}	...	f_{3n}	S_3
...	...	...	...	...	...	...
x_m	f_{m1}	f_{m2}	f_{m3}	...	f_{mn}	S_m
distribuzione marginale di Y	T_1	T_2	T_3	...	T_n	TOTALE

In questa tabella, con riferimento ai valori dell'ultima colonna, si ha:

$$S_i = f_{i1}+f_{12}+f_{13}+...+f_{in} \quad (i=1,2,3,...,m),$$

mentre, con riferimento a quelli dell'ultima riga, è:

$$T_i = f_{1j}+f_{2j}+f_{3j}+...+f_{mj} \quad (j=1,2,3,...n).$$

L'ultima riga e l'ultima colonna forniscono poi le cosiddette ***distribuzioni marginali*** delle variabili statistiche X ed Y. Non sono altro che le distribuzioni delle frequenze dei due caratteri osservati singolarmente ed i valori di tali frequenze sono le somme dei valori delle righe o colonne corrispondenti.
Naturalmente il "TOTALE" dell'ultima casella in basso a destra è lo stesso sia calcolato per riga sia calcolato per colonna.
Ovviamente le distribuzioni marginali di X ed Y possono anche essere rappresentate autonomamente nei modi indicati rispettivamente nelle tabelle 9 e 10.

Tab. 9 – Distribuzione marginale di X				
x_1	x_2	x_3	...	x_m
S_1	S_2	S_3	...	S_m

Tab. 10 – Distribuzione marginale di T				
y_1	y_2	y_3	...	y_n
T_1	T_2	T_3	...	T_n

Osservazione. Data la distribuzione doppia delle frequenze di due variabili statistiche, è facile, addirittura banale, ottenere le distribuzioni marginali delle due variabili. Non è vero il contrario. Vale a dire che la conoscenza delle distribuzioni marginali non consente di risalire alla distribuzione doppia. Questo perlomeno in generale. Ci sono infatti circostanze particolarissime in cui ciò è possibile.

55.2.7 In una tabella doppia di m righe ed n colonne si segnalano m ***distribuzioni condizionate di riga*** ed n ***distribuzioni condizionate di colonna***. Dove il termine "condizionate" dipende dal fatto che la distribuzione di riga o di colonna che si considera è subordinata alla scelta del valore rispettivamente della colonna o della riga.
Con riferimento alla tabella 8, la i-esima distribuzione condizionata di riga, vale a dire la distribuzione condizionata di X dato $Y=y_i$, è rappresentata nella tabella 11, mentre la k-esima distribuzione condizionata di colonna, vale a dire la distribuzione condizionata di Y dato $X=x_k$, è rappresentata nella tabella 12.

Tab.11-Distribuzione condizionata di X dato $Y=y_i$ (riferita alla tabella doppia 8)				
x_1	x_2	x_3	...	x_m
f_{1i}	f_{2i}	f_{3i}	...	f_{mi}

Tab.12-Distribuzione condizionata di Y dato $X=x_k$ (riferita alla tabella doppia 8)				
y_1	y_2	y_3	...	y_n
f_{k1}	f_{k2}	f_{k3}	...	f_{kn}

In altre parole, se X ed Y sono due variabili statistiche, la cui distribuzione doppia è costituita da m righe ed n colonne, la ***i-esima distribuzione condizionata di riga***, vale a dire la ***distribuzione condizionata di* X *dato* $Y=y_i$**, è la distribuzione di X limitata ai soggetti che presentano la modalità y_i di Y, mentre la ***k-esima distribuzione condizionata di colonna***, vale a dire la ***distribuzione condizionata di* Y *dato* $X=x_k$**, è la distribuzione di Y limitata ai soggetti che presentano la modalità x_k di X.
Per esempio, nella tabella 7 vi sono 4 distribuzioni condizionate di riga e 3 distribuzioni condizionate di colonna. In particolare nella sottostante tabella 13 è rappresentata la distribuzione condizionata di X dato $Y=y_2$ (= Istruzione Tecnica), mentre nella tabella 14 è rappresentata la distribuzione condizionata di Y dato $X=x_3$ (= Sud).

Tab. 13 – Distribuzione condizionata di X dato $Y = y_2$ = Istruzione Tecnica (riferita alla tabella doppia 7)			
NORD	CENTRO	SUD	TOTALE
218	125	185	528

Tab. 14 – Distribuzione condizionata di Y dato $X = x_3$ = Sud (riferita alla tabella doppia 7)				
Istruzione CLASSICA	Istruzione TECNICA	Istruzione PROFESSIONALE	Istruzione ARTISTICA	TOTALE
55	185	52	3	295

In realtà, ai fini pratici, interessano più le distribuzioni condizionate relative, vale a dire quelle che si

ottengono sostituendo al valore di ogni linea (riga o colonna) il suo rapporto rispetto al totale della linea corrispondente. Con riferimento alle due precedenti distribuzioni, le distribuzioni relative sono rappresentate nelle tabelle 15 e 16 sottostanti.

Tab. 15 – Distribuzione condizionata relativa di X dato $Y = y_2 =$ Istruzione Tecnica (riferita alla tabella doppia 7)			
NORD	CENTRO	SUD	TOTALE
0,41	0,24	0,35	1

Tab. 16 – Distribuzione condizionata relativa di Y dato $X = x_3 =$ Sud (riferita alla tabella doppia 7)				
Istruzione CLASSICA	Istruzione TECNICA	Istruzione PROFESSIONALE	Istruzione ARTISTICA	TOTALE
0,19	0,63	0,17	0,01	1

55.3 CORRELAZIONE

55.3.1 Quando si considerano due fenomeni collettivi distinti o due aspetti di uno stesso fenomeno, può accadere che uno di essi influenzi l'altro, come può darsi che ciò non avvenga. Nel primo caso i due fenomeni si dicono ***stocasticamente indipendenti*** (o, più semplicemente: ***indipendenti***) nel secondo si dicono ***stocasticamente dipendenti*** (o, più semplicemente: ***dipendenti***).

- Per esempio, negli ultimi 50 anni l'altezza media degli italiani è aumentata; nello stesso tempo è migliorata l'alimentazione (più proteine, più vitamine, eccetera). È legittimo supporre che il miglioramento dell'alimentazione influenzi l'altezza media degli italiani.
- Altro esempio: la percentuale di persone affette da cancro ai polmoni è più elevata se rilevata in un campione di fumatori rispetto a quella rilevata in un campione di non fumatori. È legittimo supporre che il fumo sia una delle cause del cancro ai polmoni.

L'indagine relativa a due fenomeni sotto osservazione può riguardare aspetti qualitativi per entrambi (esempio: il colore dei capelli di un gruppo di persone e la nazione di provenienza; oppure: il colore degli occhi e le preferenze in campo sportivo; eccetera), aspetti quantitativi per entrambi (esempio: l'altezza di un gruppo di persone e i loro pesi) oppure aspetti qualitativi per un fenomeno e aspetti quantitativi per l'altro.

Ricordiamo che l'insieme delle modalità di un carattere osservato e delle rispettive frequenze si chiama ***variabile statistica***. A volte questa denominazione è riservata alle modalità di tipo quantitativo mentre se esse sono di tipo qualitativo si parla più propriamente di ***mutabile statistica***. Per questo possiamo dire che l'indagine su due fenomeni può riguardare due mutabili statistiche o due variabili statistiche o una mutabile ed una variabile.

In ogni caso, la dipendenza di una variabile dall'altra si chiama **correlazione** o **connessione** [(2)].

Quel settore della statistica che si occupa delle relazioni che intercorrono fra due fenomeni collettivi o fra due caratteri di uno stesso fenomeno si chiama **statistica bivariata**.

55.3.2 Un modo per stabilire se due variabili statistiche sono o no dipendenti è quello di ricorrere alla

[2] Alcuni autori distinguono fra *connessione* e *correlazione*, chiamando correlazione il legame fra due variabili e connessione il legame tra due mutabili o fra una mutabile ed una variabile.

distribuzione congiunta delle due variabili che descrivono i fenomeni.
Ebbene, si può affermare che il carattere Y è ***indipendente*** da X se, per tutte le modalità X, le distribuzioni condizionate relative di Y sono uguali fra loro e sono uguali alla distribuzione relativa marginale di Y.
Se ciò non accade Y è ***dipendente*** da X.
Per esempio, al fine di stabilire se il peso delle persone (variabile Y) dipende dall'altezza (variabile X) sono stati messi sotto osservazione l'altezza e il peso di un gruppo di persone ed i dati sono registrati nella tabella sottostante (Tab. 17), precisando che per ogni classe è compreso il primo estremo ed è escluso il secondo (per esempio, nella classe dei pesi 50-60 è incluso 50 ed è escluso 60, in quella delle altezze 180-190 è incluso 180 ed escluso 190).

Y=Peso (kg) X=Altezza (cm)	50-60	60-70	70-80	80-90	90-100	100-110	TOTALI
150-160	12	13	6	1	0	0	32
160-170	6	15	18	5	2	2	48
170-180	1	10	21	15	4	5	56
180-190	0	3	13	19	15	16	66
190-200	0	0	3	8	14	16	41
TOTALI	19	41	61	48	35	39	243

TAB. 17

Da questa tabella, dividendo i valori di ogni colonna per i corrispettivi totali di colonna, si ottengono le distribuzioni condizionate relative di colonna, compresa la distribuzione relativa marginale di colonna (Tab. 18). Si constata che tali distribuzioni non hanno le medesime frequenze per le diverse modalità delle altezze. Ne consegue che il peso dipende dall'altezza. Non ci voleva molto ad intuirlo, ma l'esempio ci è servito per chiarire il concetto precedente.

Y=Peso (kg) X=Altezza (cm)	50-60	60-70	70-80	80-90	90-100	100-110	TOTALI
150-160	0,63	0,32	0,10	0,02	0,00	0,00	0,13
160-170	0,32	0,37	0,30	0,10	0,06	0,05	0,20
170-180	0,05	0,24	0,34	0,31	0,11	0,13	0,23
180-190	0,00	0,07	0,21	0,40	0,43	0,41	0,27
190-200	0,00	0,00	0,05	0,17	0,40	0,41	0,17
TOTALI	1	1	1	1	1	1	1,00

TAB. 18

Ma c'è di più. La tabella consente infatti di valutare come, per una data fascia di pesi, questi dipendono dalle altezze. Così, ad esempio, si può constatare che, per la fascia di pesi 80–90 kg, sono di più le persone di altezza appartenente alla fascia 180–190 cm, mentre sono di meno quelle di altezza appartenente alla fascia 150–160 cm. Cosa che per la verità si poteva constatare anche dalla tabella 18 delle frequenze assolute.

ESERCIZIO. Prendi in esame le seguenti tabelle doppie (Tab. 19 e Tab. 20), che riassumono dati relativi alle due variabili statistiche X ed Y, determina per entrambe la tabella delle distribuzioni condizionate relative di colonna e stabilisci se e come X influenza Y. Ti consigliamo l'uso di un foglio elettronico.

Y / X	y_1	y_2	y_3	y_4	TOTALI
x_1	81	92	35	88	296
x_2	45	51	20	49	165
x_3	54	61	23	59	197
TOTALI	180	204	78	196	658

TAB. 19

Y / X	y_1	y_2	y_3	y_4	TOTALI
x_1	93	86	23	45	247
x_2	29	51	53	16	149
x_3	123	67	80	135	404
TOTALI	245	204	156	195	800

TAB. 20

55.3.3 Quando l'indagine verte su aspetti quantitativi per entrambi i fenomeni indagati, la correlazione può essere di tipologie diverse. In particolare essa può essere:

- ***diretta***, se a valori crescenti di una variabile corrispondono mediamente valori crescenti dell'altra. Una correlazione diretta si dice pure ***concordanza***;
- ***inversa***, se a valori crescenti di una variabile corrispondono mediamente valori decrescenti dell'altra. Una correlazione inversa si chiama pure ***discordanza.***

È possibile conoscere il grado della correlazione, mediante il calcolo di appositi numeri, chiamati **coefficienti** (o **indici**) **di correlazione**. Sono espressi da formule basate sulle medie aritmetiche delle variabili statistiche che caratterizzano i due fenomeni e sulle deviazioni standard. Non ci occuperemo della dimostrazione di tali formule, anzi accenneremo ad uno soltanto degli indici che esse esprimono.
Siano, allora, X ed Y due variabili statistiche, espressioni di altrettanti fenomeni collettivi, suscettibili rispettivamente dei seguenti valori:

$$x_1, x_2, \dots, x_n; \qquad y_1, y_2, \dots, y_n.$$

Quale che sia l'indice "i", al valore x_i della variabile X è associato il valore y_i assunto da Y. Per cui: ad x_1 resta associato y_1, ad x_2 resta associato y_2, e così via.
Per esempio: le x_i sono le altezze dei padri e le y_i quelle dei rispettivi figli.
Oppure: le x_i sono le altezze di alcune persone e le y_i i loro rispettivi pesi.
Indicate con m_x ed m_y le medie aritmetiche delle due variabili e con σ_x e σ_y le loro deviazioni standard e posto:

$$p = \sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \dots + x_n y_n,$$

un coefficiente di correlazione particolarmente usato è il numero r dato dalla seguente formula:

$$\mathbf{r = \frac{p - n\, m_x\, m_y}{n\, \sigma_x\, \sigma_y}}.$$

Si chiama **coefficiente di correlazione lineare di Bravais-Pearson** [3].

[3] **Bravais**, August; scienziato francese, 1811-1863. **Pearson**, Karl; matematico e statistico inglese, 1857-1936.

Si tratta di un numero compreso fra –1 e 1. Precisamente:

- quando $0<r\leq 1$ la correlazione è diretta (*concordanza*);
- quando $-1\leq r<0$ la correlazione è inversa (*discordanza*);
- quando $r=0$ la correlazione è *nulla*.

Naturalmente, quanto più r è vicino a 0 tanto meno i due fenomeni sono correlati e, di conseguenza, tanto maggiore è la dispersione. Al contrario, quanto più r è vicino a ± 1 tanto più essi sono correlati e, di conseguenza, tanto minore è la dispersione.

Nei casi particolari in cui $r=\pm 1$, i punti (x_1,y_1), (x_2,y_2), ..., (x_n,y_n) si distribuiscono lungo una retta. Si parla allora di **correlazione lineare perfetta**: diretta (se $r=1$) o inversa (se $r=-1$).

55.3.4 Vediamo un paio di esempi.

- ESEMPIO 1. Nella tabella 21 sono indicate le altezze X (misurate in centimetri) di un gruppo di 15 persone ed i loro rispettivi pesi Y (misurati in chilogrammi).

Tab. 21 – Altezze e pesi di un gruppo di persone

numero	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Altezza X (cm)	173	165	125	182	178	162	175	181	164	172	170	150	158	173	176
Peso Y (kg)	70	54	24	72	92	70	78	69	60	70	71	45	63	68	72

Con un po' di pazienza, ma anche con l'ausilio di uno strumento di calcolo automatico (è sufficiente un foglio elettronico), si trova:

$$m_x \approx 166{,}933 \text{ cm};\ \ m_y \approx 65{,}200 \text{ kg};\ \ \sigma_x \approx 14{,}553 \text{ cm};\ \ \sigma_y \approx 14{,}498 \text{ kg};\ \ p \approx 166069 \text{ kg}\cdot\text{cm}\,.$$

Pertanto il coefficiente di correlazione di Bravais-Pearson è:

$$r = \frac{166069 - 15 \cdot 166{,}933 \cdot 65{,}200}{15 \cdot 14{,}533 \cdot 14{,}498} \approx 0{,}887.$$

La dispersione è scarsa e la correlazione è diretta. Anzi non è molto lontana dalla correlazione lineare perfetta.

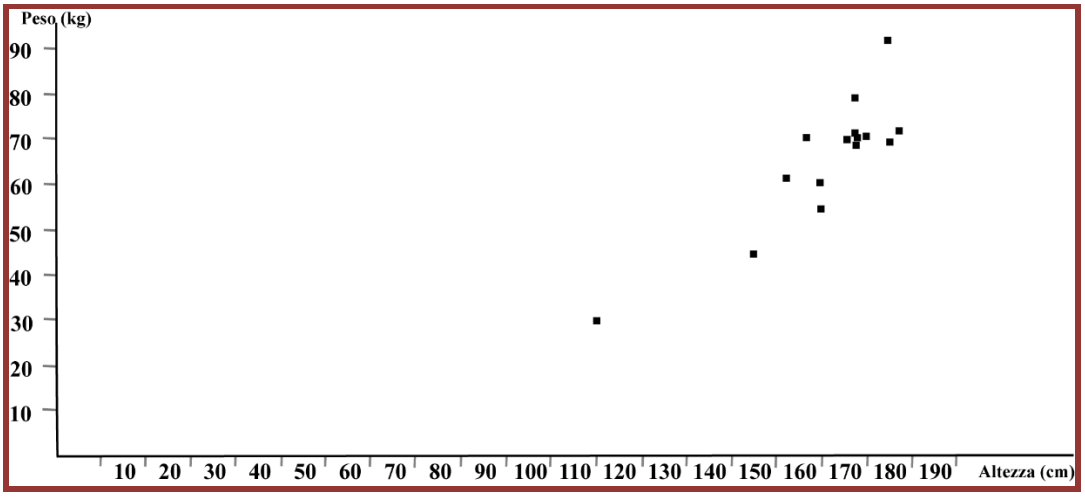

FIG. 4

Il cosiddetto ***diagramma a dispersione***, vale a dire la rappresentazione grafica della correlazione fra le variabili X ed Y, si ottiene (Fig. 4) disegnando in un piano cartesiano ortogonale (Oxy) i punti (x_i,y_i). Da tale diagramma si può intuire come, mediamente, al crescere delle altezze X crescono i pesi Y delle persone.

- ESEMPIO 2. Nella tabella 22 sono indicati, per le 20 Regioni d'Italia ed in riferimento ad un certo anno, le percentuali X di persone che lavoravano nell'industria ed i redditi pro-capite Y di quell'anno espressi in migliaia di euro.

Regioni	X	Y
Piemonte	42	10,5
Valle d'Aosta	28	12,5
Lombardia	47	11,0
Trentino Alto Adige	26	9,6
Veneto	42	8,8
Friuli Venezia Giulia	33	9,8
Liguria	25	10,8
Emilia Romagna	36	10,7
Toscana	39	9,6
Umbria	35	8,8
Marche	41	8,3
Lazio	21	9,0
Abruzzo	28	7,1
Molise	24	6,4
Campania	24	6,2
Puglia	25	6,3
Basilicata	25	6,3
Calabria	20	5,6
Sicilia	23	6,1
Sardegna	25	6,5

TAB. 22

Calcoliamo i valori dei parametri che figurano nella formula di Bravais-Pearson:

$$m_x \approx 30{,}450;\quad m_y \approx 8{,}495;\quad \sigma_x \approx 8{,}179;\quad \sigma_y \approx 2{,}058;\quad p \approx 5348{,}500.$$

Pertanto il coefficiente di correlazione di Bravais-Pearson è:

$$r = \frac{5348{,}500 - 20 \cdot 30{,}450 \cdot 8{,}495}{20 \cdot 8{,}179 \cdot 2{,}058} \approx 0{,}52.$$

C'è una correlazione diretta tra le due variabili statistiche ma non è molto alta, per cui si ha una discreta dispersione. Di nuovo, il diagramma a dispersione evidenzia la "nuvola" dei punti (x_i,y_i) (Fig. 5) e fa intuire come il reddito pro-capite cresca mediamente al crescere della percentuale di persone che lavorano nell'industria.

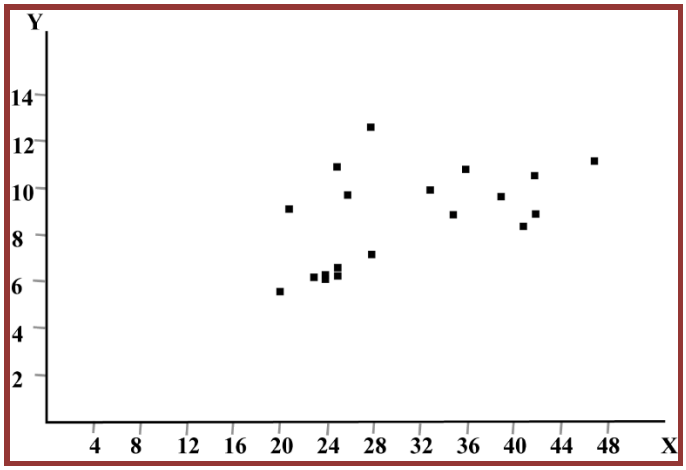

FIG. 5

55.3.5 Nei due esempi precedenti abbiamo osservato il diagramma a dispersione di due particolari correlazioni, entrambe dirette. Mostriamo adesso alcune figure che hanno lo scopo di dare un'idea d'insieme della rappresentazione di due generiche variabili statistiche: - correlate direttamente (Fig. 6), - correlate inversamente (Fig. 7), - non correlate (Fig. 8).

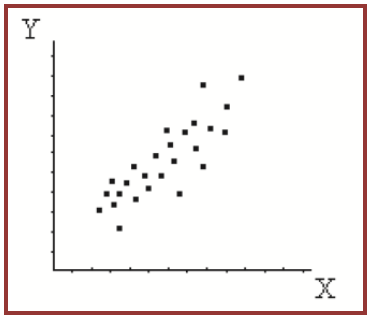

FIG. 6

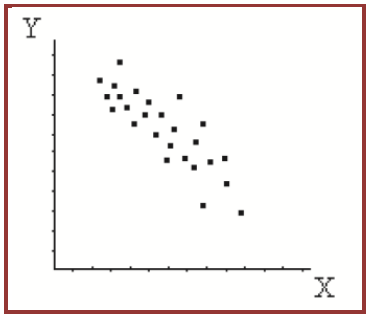

FIG. 7

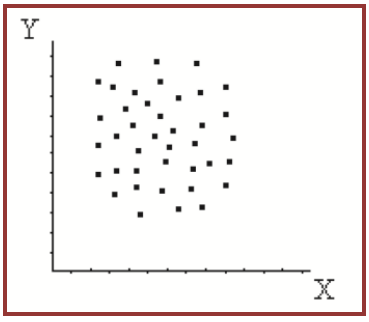

FIG. 8

55.3.6 La correlazione fra due variabili statistiche è una base eccellente per lo studio dei fenomeni in vari campi: dalla fisica all'economia, dalle assicurazioni alla pubblicità, dalla medicina alle scienze in genere. Quello che abbiamo esposto è però solo una piccola parte di ciò che è possibile fare, giusto per dare un'idea. Bisogna comunque evitare di trarre in maniera affrettata conclusioni di causa-effetto fra i due fenomeni studiati, perché altrimenti si rischia di prendere spiacevoli cantonate. In realtà, le relazioni statistiche spesso non hanno nulla a che fare con una relazione di causa-effetto. Valgono più di ogni spiegazione un paio di esempi, ancorché stravaganti [(4)].

- Le statistiche dimostrano che la maggior parte degli incidenti d'auto avvengono con automobili che viaggiano a velocità moderata e che si verificano pochissimi incidenti a velocità superiori a 150 km/h.

Questo significa che è più sicuro viaggiare a velocità elevate?

Neanche per idea. Il fatto è che la stragrande maggioranza delle persone guida a velocità moderate e pochissime vanno a 150 km/h o più. È quindi naturale che la maggior parte degli incidenti avvenga a velocità moderate.

[4] Cfr.: Martin Gardner, *Ah! Ci sono! Paradossi stimolanti e divertenti*, RBA Italia, 2008, pag. 165 e segg..

Ciò che potrebbe essere utile per qualche conclusione sensata è il confronto fra la percentuale di incidenti che si verificano fra le persone che guidano a velocità moderate e quella fra le persone che vanno a velocità di 150 km/h ed oltre.

- Una ricerca rivelò che, in una certa città, contemporaneamente ad un forte aumento demografico si era verificata una notevole crescita del numero dei nidi di cicogna.

Questo conferma la credenza che i neonati siano portati dalle cicogne?

Ovviamente NO. Significa semplicemente che con l'aumento del numero degli edifici, resosi necessario in seguito all'incremento demografico, le cicogne disponevano di più posti in cui potevano fare il nido.

55.4 REGRESSIONE

55.4.1 Una volta disegnato il diagramma a dispersione relativo a due variabili statistiche X ed Y, vale a dire la rappresentazione grafica dei punti (x_i,y_i), può essere utile conoscere una funzione $y=f(x)$ verificata da tutte le coppie ordinate (x_i,y_i). In tal caso si disporrebbe di una curva passante per tali punti. A volte ciò è possibile ma nei casi che stiamo esaminando, data la distribuzione molto irregolare di questi punti, bisogna accontentarsi di una funzione con caratteristiche diverse. In particolare essa, non potendo essere soddisfatta dalle coppie ordinate (x_i,y_i), deve essere tale che i punti che rappresentano tali coppie si addensino nel miglior modo possibile intorno al suo grafico. Il metodo di ricerca di una funzione siffatta è chiamato *interpolazione statistica* e la funzione è detta *funzione interpolatrice*.

Tra i metodi di interpolazione statistica quello più usato è la **regressione**, nel qual caso la funzione interpolatrice si chiama più propriamente ***funzione di regressione*** di Y su X.

Le funzioni di regressione possono essere di vario tipo: lineari, quadratiche, cubiche, iperboliche, eccetera. Noi ci occuperemo solamente del modello lineare. In tal caso, il grafico della funzione è una retta, che è chiamata ***retta di regressione*** di Y su X. Tale retta ha un'equazione del tipo:

$$y = ax+b$$

e la teoria mostra che i coefficienti a, b sono tali da soddisfare le seguenti condizioni:

[2] $$a = \frac{p - n\, m_x\, m_y}{n\, \sigma_x^2}\,, \qquad b = m_y - a\, m_x$$

dove i simboli presenti hanno lo stesso significato chiarito in precedenza.

- ESEMPIO 1. Riprendiamo la precedente tabella 21. Si calcola facilmente (naturalmente con l'uso di uno strumento di calcolo automatico): $a \approx 0{,}884$; $b \approx -82{,}37$.

Cosicché la retta di regressione di Y (peso delle persone) su X (altezze delle persone) ha la seguente equazione: $y = 0{,}884\, x - 82{,}37$.

La sua rappresentazione grafica (Fig. 9) completa il diagramma a dispersione della correlazione fra le due variabili (Fig. 4) e mostra come, effettivamente, tali punti si addensino intorno a questa retta.

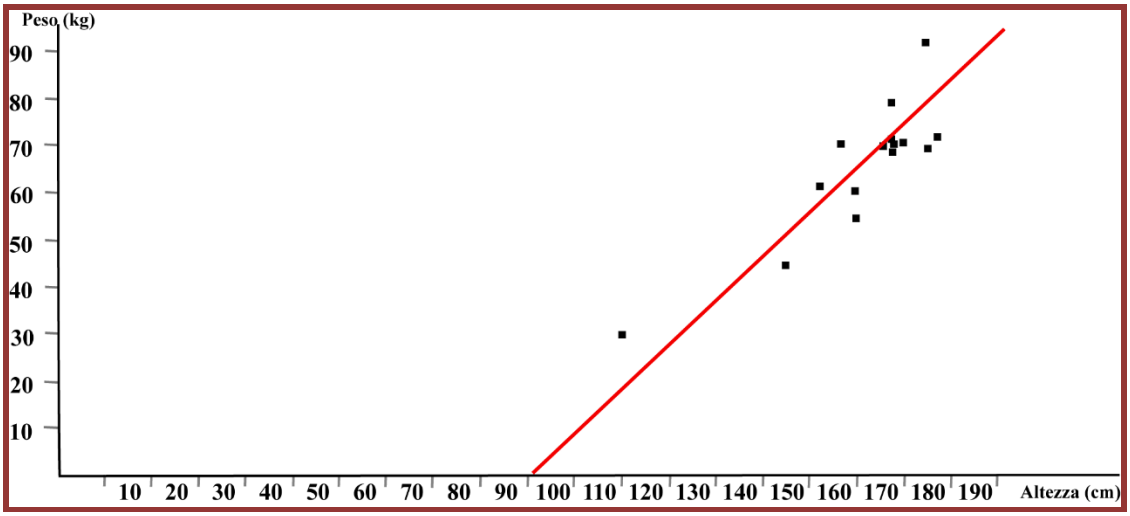

FIG. 9

- ESEMPIO 2. Con riferimento alla tabella 22 si calcola che: a≈0,13; b≈4,5.

Per cui la retta di regressione di Y (reddito pro-capite) su X (percentuale di persone che lavorano nell'industria) ha la seguente equazione: $y = 0{,}13x+4{,}5$ ed è rappresentata in figura 10, ottenuta completando la figura 5. Mostra come, effettivamente, i punti (x_i, y_i) si addensino intorno a questa retta, anche se meno intensamente rispetto all'esempio precedente.

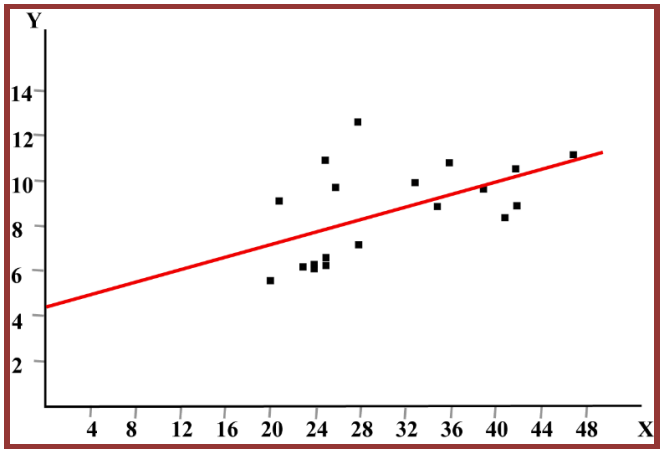

FIG. 10

55.4.2 A volte, accanto alla retta di regressione di Y su X, è utile conoscere una retta che esprima un'approssimazione della dipendenza di X da Y: si chiama **retta di regressione** di X su Y. È rappresentata da un'equazione del tipo:

$$x = a'y+b',$$

dove i valori dei coefficienti a' e b', con il solito significato per i simboli usati, sono dati dalle formule seguenti:

[3] $$a' = \frac{p - n\, m_x\, m_y}{n\, \sigma_y^2}, \qquad b' = m_x - a' m_y\,.$$

Approfondiamo ritornando sui due esempi descritti poco sopra.

1) La retta di regressione di X (altezze delle persone) su Y (pesi) è espressa dalla seguente equazione: $x = 0{,}780\ y + 116{,}14$ e quindi, esprimendo y in funzione di x:

$$y = 1{,}282\ x - 148{,}90.$$

La figura 9, dove oltre al diagramma a dispersione è rappresentata la retta r di regressione di Y su X, può allora essere integrata con il disegno di questa nuova retta s (Fig. 11).

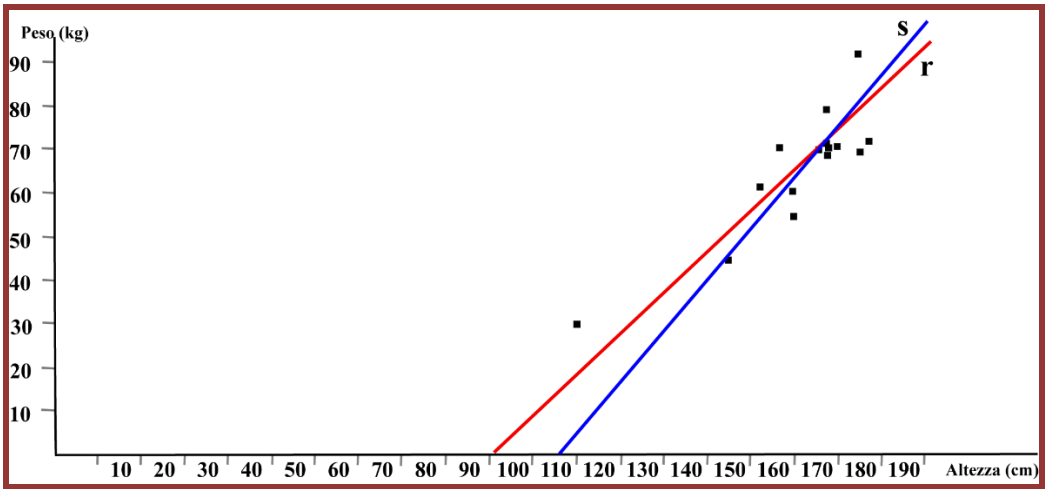

FIG. 11

2) La retta di regressione di X (percentuale di persone che lavorano nell'industria) su Y (reddito pro-capite) è espressa dalla seguente equazione:

$$x = 2{,}066\ y + 12{,}899$$

e quindi:

$$y = 0{,}484\ x - 6{,}243.$$

La figura 10, dove oltre al diagramma a dispersione è rappresentata la retta r di regressione di Y su X, può allora essere integrata con il disegno di questa nuova retta s (Fig. 12).

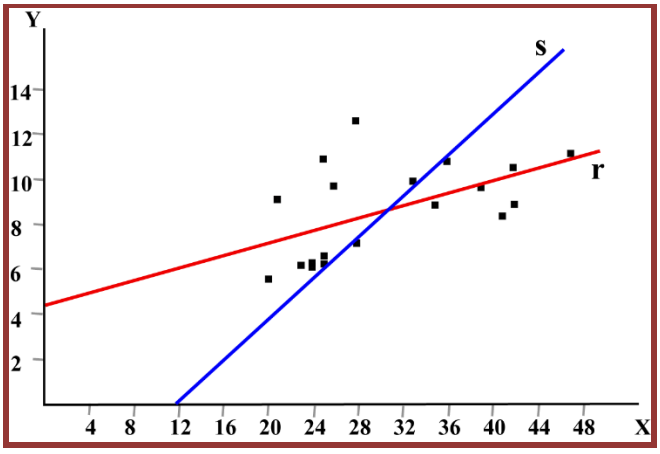

FIG. 12

55.4.3 Si possono notare alcuni fatti interessanti.

1) In entrambi i casi presi in esame la retta s di regressione di X su Y ha una pendenza maggiore della retta r di regressione di Y su X. Questo fatto si verifica sempre.

2) Nella situazione di figura 11, nella quale si ha una dispersione minore di quella di figura 12, l'angolo formato dalle due rette di regressione è minore rispetto a quello di figura 12. Anche questo vale in generale. Precisamente, quanto minore è la dispersione tanto minore è l'angolo delle due rette di regressione e quanto maggiore è la dispersione tanto maggiore è l'angolo delle due rette di regressione.
3) Il punto in cui le due rette di regressione si intersecano si chiama ***baricentro*** della nuvola di punti. Non è detto che sia necessariamente uno dei punti della nuvola. Nel primo dei due esempi descritti esso è il punto di coordinate $(166{,}93;\ 65{,}16)$, nel secondo esempio è il punto di coordinate $(30{,}35;\ 8{,}45)$. Cosa che puoi dimostrare da solo.
4) Dall'analisi dei risultati precedenti, al netto degli errori di approssimazione, sembra che il baricentro della nuvola di punti (x_i,y_i) coincida con il punto di coordinate (m_x,m_y). Non è un caso ma la regola. Cosa che si può dimostrare (in maniera noiosa per le lungaggini nei calcoli, se "fatti a mano", ma abbastanza rapidamente con l'ausilio di un idoneo software matematico) risolvendo il sistema delle due rette di regressione, vale a dire il sistema delle equazioni $(y=ax+b,\ x=a'y+b')$, e sostituendo ai coefficienti a, b, a', b' le loro espressioni date dalle [2] e dalle [3]. Si trova per l'appunto: $x=m_x$, $y=m_y$.

55.4.4 Quando i punti (x_1,y_1), (x_2,y_2), ..., (x_n,y_n) sono situati sulla retta di regressione o perlomeno si discostano da essa in misura trascurabile allora si parla di **regressione lineare**. In questo caso la retta di regressione di Y su X e quella di X su Y tendono a sovrapporsi.
Con riferimento ai due esempi precedenti, mentre la regressione rappresentata in figura 11 non è molto distante da una regressione lineare, non si può dire altrettanto di quella rappresentata in figura 12.
Vediamo adesso un paio di esempi di regressione lineare vera e propria.

- ESEMPIO 1. Un gas è riscaldato a pressione costante a partire da una data temperatura. L'aumento Δt_i di temperatura, riferito alla temperatura iniziale, e il corrispondente aumento di volume ΔV_i sono indicati nella tabella sottostante (Tab. 23).

Tab. 23 – Aumento del volume di un gas rispetto all'aumento della temperatura.

Δt_i (°C)	10	20	30	40	50	60	70	80	90	100
ΔV_i (dm^3)	200	350	500	670	850	1000	1150	1350	1500	1670

Considerata la retta di regressione della variabile ΔV sulla variabile Δt:

$$\Delta V = a\,\Delta t + b,$$

si tratta di determinare i coefficienti a, b in base alle formule [2]. Si trova:

$$a \approx 14{,}79;\ b \approx 110{,}4.$$

Cosicché la retta di regressione ha la seguente equazione:

$$\Delta V = 14{,}79\,\Delta t + 110{,}4.$$

Essa è rappresentata in figura 13, assieme al diagramma a dispersione, vale a dire alla rappresentazione grafica dei punti $(\Delta t_i,\Delta V_i)$. Si nota come questi punti si discostino pochissimo dalla retta di regressione: si tratta pertanto di regressione lineare.
Prova a determinare la retta di regressione di Δt su ΔV e a disegnarla completando il grafico di fig. 13.

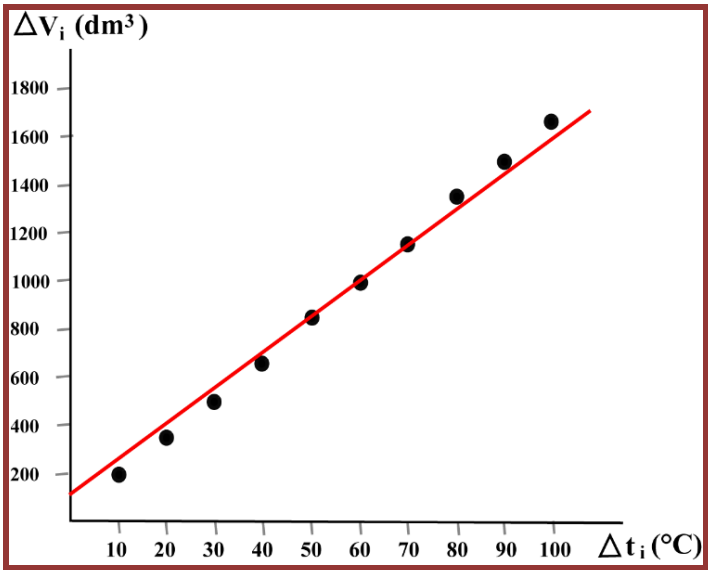

FIG. 13

- ESEMPIO 2. Di un secondo esempio di regressione lineare vera e propria ci limitiamo a fornire il risultato. Si tratta di una ricerca condotta verso la fine del secolo XIX da Karl Pearson sulla correlazione fra l'altezza H di una persona di sesso maschile e la lunghezza L del suo femore. Egli, in seguito a numerose misurazioni, oltre un centinaio, trovò la seguente retta di regressione lineare (tutto è espresso in centimetri):

$$H = 1{,}88\ L + 81{,}31\ .$$

Con questa formula si riesce a stimare la statura, tanto per fare un esempio, dell'uomo di Neanderthal. Sapendo infatti che il suo femore è lungo circa 44,52 cm, si trova che la sua altezza H, espressa in centimetri, è:

$$H = 1{,}88 \times 44{,}52 + 81{,}31 \approx 165\ .$$

Ribadiamo che si tratta di una stima e non dell'altezza esatta.

55.4.5 L'esempio precedente mostra un fatto interessante. Date due variabili statistiche X ed Y, la conoscenza della retta di regressione di Y su X, in particolare se si tratta di una regressione lineare (o quasi lineare), permette di "stimare" quale valore può assumere Y noto un determinato valore di X, a condizione che questo valore sia interno all'intervallo dei valori osservati o, se esterno, non sia molto discosto da tale intervallo.

Oltre al precedente esempio 2, di cui abbiamo detto, anche l'esempio 1 permette di trarre qualche conclusione esemplificativa:

- Se la temperatura del gas aumenta di $\Delta t = 45$ °C, si può stimare che il suo volume aumenti di $\Delta V = 14{,}79 \times 45 + 110{,}4 \approx 776$ (dm^3).
- Se la temperatura del gas aumenta di $\Delta t = 103$ °C, si può stimare che il suo volume aumenti di $\Delta V = 14{,}79 \times 103 + 110{,}4 \approx 164$ (dm^3).

Consideriamo un'altra situazione, per la quale chiediamo la tua collaborazione.

- ESERCIZIO. La tabella sottostante (Tab. 24) registra la popolazione residente in Italia (Y) in alcuni anni (X) nei quali è stato fatto un censimento (dati ISTAT). I valori di Y sono espressi in migliaia.

Si chiede di disegnare il diagramma a dispersione, trovare la retta di regressione di Y su X e stimare quale popolazione era residente in Italia negli anni 1941 (in quell'anno non fu fatto il censimento a causa della guerra), 1936 e 1991.

X = anno	1921	1931	1941	1951	1961	1971	1981
Y = popolazione (in migliaia)	39943	41651	?	47515	50623	54136	56556

TAB. 24

RISOLUZIONE (traccia). Il calcolo del coefficiente di correlazione di Bravais-Pearson ($r \approx 0{,}831$) mostra che si tratta di una correlazione non molto lontana da una correlazione lineare perfetta. Il grafico della retta di regressione di Y su X, completando il diagramma a dispersione, conferma poi che si ha a che fare con una regressione quasi lineare. La retta di regressione ha la seguente equazione:

$$y = 239{,}217\,x - 418707.$$

Fatti i calcoli opportuni, il modello lineare ipotizzato permette di trarre delle conclusioni:

- La popolazione residente in Italia nell'anno 1941 può essere stimata in circa 45 milioni e 600 mila persone residenti. Come detto, non sappiamo quanto fosse esattamente questa popolazione dal momento che in quell'anno non è stato fatto alcun censimento della popolazione.
- Nell'anno 1936 la popolazione residente può essere stimata in circa 44 milioni e 400 mila. In realtà, nel 1936 un censimento fu fatto e si rilevò una popolazione di quasi 43 milioni di persone residenti.
- La proiezione della popolazione residente nel 1991 dà un valore di circa 57 milioni e mezzo. In realtà, anche questo valore è noto ed è di 56 milioni 778 mila.

VERIFICHE

Tabelle a doppia entrata. Distribuzioni statistiche (nn. 1-11).

1. Le due variabili statistiche indipendenti, X ed Y, sono distribuite come nelle tabelle sottostanti:

Variabile X	1	3	5	7		Variabile Y	2	4	6
Frequenza assoluta	2	3	4	1		Frequenza assoluta	3	2	1

Calcolare M(X) ed M(Y). Determinare le distribuzioni delle frequenze assolute di X+Y e di XY e calcolare M(X+Y) ed M(XY). [**R**. $M(X) = 2{,}7$; $M(Y) = 2{,}5$; $M(X+Y) = 5{,}2$; $M(XY) = 6{,}75$]

2. In un'urna vi sono 4 palline contrassegnate coi numeri: $-1, 0, 1, 2$. Dopo 40 estrazioni, ovviamente con reinserimento, le frequenze di estrazione sono risultate rispettivamente:

$$7, 11, 10, 12.$$

Detta X la variabile statistica che prende i valori contrassegnati sulle palline con le suddette frequenze, determinare le distribuzioni delle frequenze assolute delle variabili XX ed X^2 e calcolarne le medie aritmetiche.

[**R**. XX assume i valori $-2, -1, 0, 1, 2, 4$ rispettivamente con le frequenze $168, 140, 759, 149, 240, 144$; X^2 prende i valori $0, 1, 4$ rispettivamente con le frequenze $11, 17, 12$; ...]

3. La variabile statistica A prende i valori 1, 2, 3, 4 rispettivamente con le frequenze 6, 5, 4, 7; la variabile statistica B prende gli stessi valori, ma con le frequenze 5, 4, 4, 3. Determinare le distribuzioni delle frequenze assolute delle variabili statistiche:

$$X = \max(A,B) \text{ e } Y = |A - B|,$$

dove max(A,B) indica la variabile statistica che assume come valore il massimo dei valori a, b assunti rispettivamente dalle variabili A, B e |A – B| indica la variabile statistica che assume i valori |a – b|. Calcolare quindi M(X) ed M(Y).

[**R.** X: valori 1, 2, 3, 4 con frequenze 30, 69, 96, 157;
Y: valori 0, 1, 2, 3 con frequenze 87, 125, 87, 53; ...]

4. Si estrae un numero della tombola. La variabile statistica A assume il valore –1 se esso è divisibile per 3, il valore 1 se è divisibile per 4 ma non per 3 ed il valore 0 in ogni altro caso. Le frequenze relative si assumono uguali alle rispettive probabilità. Determinare la distribuzione delle frequenze relative di A. Determinare poi le distribuzioni delle frequenze relative delle variabili:

$$X = A^2, \quad Y = \max(A,X), \quad Z = |A - X|.$$

$\left[\textbf{R.}\ \text{A: valori} -1, 0, 1 \text{ con frequenze relative } \frac{1}{3}, \frac{1}{2}, \frac{1}{6}; \ \ldots\right]$

5. Si lanciano due dadi con le facce numerate da 1 a 6. La variabile statistica A assume il valore –2 se la somma dei due numeri usciti è minore di 5, il valore 2 se è maggiore di 8 ed il valore 0 in ogni altro caso. Le frequenze relative si assumono uguali alle rispettive probabilità. Determinare la distribuzione delle frequenze relative di A. Determinare poi le distribuzioni delle frequenze relative delle variabili:

$$X = A^2, \quad Y = X + A, \quad Z = \min(X,A).$$

6. Considerate le due variabili statistiche X ed Y, di cui all'esercizio n. 1, costruire la loro distribuzione congiunta e fornire una sua rappresentazione grafica. Determinare quindi le distribuzioni marginali di X ed Y. [**R.** ... ; d.m. di X: 12, 18, 24, 6; d.m. di Y: 30, 20, 10]

7. Considerate le due variabili statistiche XX ed X^2, di cui all'esercizio n. 2, costruire la loro distribuzione congiunta e fornire una sua rappresentazione grafica. Determinare quindi le distribuzioni marginali di XX ed X^2.

[**R.** ... ; d.m. di XX: 6720, 5600, 30360, 5960, 9600, 5760; d.m. di X^2: 17600, 27200, 19200; ...]

8. Considerate le due variabili statistiche X ed Y, di cui all'esercizio n. 3, costruire la loro distribuzione congiunta e fornire una sua rappresentazione grafica. Determinare quindi le distribuzioni marginali di X ed Y. Rappresentare inoltre graficamente la terza distribuzione condizionata di riga e la seconda distribuzione condizionata di colonna.

[**R.** ... ; d.m. di X: 30624, 44000, 30624, 18656; d.m. di Y: 10560, 24288, 33792, 55264; ...]

9. Considerate le due variabili statistiche X ed Y, di cui all'esercizio n. 1, indicare con X' ed Y' le distribuzioni delle frequenze relative di X ed Y rispettivamente. Quindi costruire la distribuzione congiunta di X' ed Y' e fornire una sua rappresentazione grafica. Determinare quindi le distribuzioni marginali di X' ed Y'.
Relativamente a tali distribuzioni marginali si nota qualche particolarità?

10. Risolvere lo stesso esercizio precedente con riferimento, questa volta, alle due variabili statistiche X ed Y di cui all'esercizio n. 3.

11. Le istituzioni scolastiche di 2° grado erano distribuite per area geografica e per tipologia di istituto, nell'anno scolastico 2000/01, secondo la seguente tabella di contingenza:

	Nord	Centro	Sud e isole
Licei classici	204	149	297
Licei pedagogici	178	84	262

Licei scientifici	418	219	412
Licei linguistici	102	36	72
Licei artistici	59	18	45
Istituti d'arte	50	48	83
Istituti professionali	601	261	647
Istituti tecnici	923	436	1009

Rappresentare graficamente la seconda distribuzione condizionata di riga e la prima distribuzione condizionata di colonna. Determinare inoltre le distribuzioni marginali e rappresentare graficamente la situazione complessiva.

Correlazione (nn. 12-21).

12. LABORATORIO DI MATEMATICA. Dopo aver misurato le altezze H ed i pesi P degli studenti della tua classe e dopo aver raccolto i dati su un'apposita tabella, trova il coefficiente di correlazione di Bravais-Pearson. Trai quindi qualche conclusione circa la dipendenza di P da H.

13. Nella tabella sottostante sono indicate le altezze H e le circonferenze toraciche T – entrambe espresse in centimetri ed approssimate a meno di 1 cm – di un gruppo di 20 militari di leva. Rappresentare graficamente la "nuvola" di punti che descrive il fenomeno. Determinare quindi il coefficiente di correlazione di Bravais-Pearson e trarre qualche conclusione circa la dipendenza di T da H.

H	T		H	T		H	T		H	T		H	T
171	92		165	89		173	94		168	95		178	92
175	101		167	88		165	88		179	92		169	90
168	98		178	102		167	80		177	101		172	91
180	102		172	82		173	102		180	104		167	89

14. La tabella sottostante sintetizza come sono distribuite l'una rispetto all'altra le due variabili statistiche X ed Y relative rispettivamente alle altezze ed alle circonferenze toraciche di 200 giovani alla visita di leva (entrambe espresse in centimetri). Calcolare il coefficiente di correlazione di Bravais-Pearson e trarre qualche conclusione circa la dipendenza di Y da X.

Y \ X	73	75	76	78	79	80	83	84	85	87	88	89	90	91	92	94	95	96	97	98	99	100	101	102	104	106
165	1			1	1			1			1	1														
166					1		1	2	3				2													
167		1	1			1	1			2	2			1	1	1										
168				1	1	1		1	1	2		3			2		1			1						
169							1	1	2		1	2	3	1		1		1	1		1	1				
170								1		1	2		2	3	3	1	1			1	1		1			
171										1		2	1	4	4	3	2		1	1		1				
172											1	1	2	2	4	5	3	2	1		1					
173							1			1		1		2	3	3	2	2		1		2		1		

174										1	1		1	1	2	2	2	2	1		1		1	1		
175											1	1			1	1	2	1	2	1			1		1	
176						1				1			1	1		1		2	1		1	1		1		
177									1		1				1		1			2			1		1	1
178																1		1	1			1		2		1
179												1							1		1				1	
180																				1	1		1		1	
181																						1				1
182																							1			

15. LABORATORIO DI MATEMATICA. Conduci, assieme ai tuoi compagni di classe, una ricerca su tutti gli studenti che frequentano la tua stessa scuola, volta a stabilire come sono distribuite l'una rispetto all'altra le due variabili statistiche X ed Y, relative rispettivamente ai voti di matematica e di italiano riportati in pagella nell'orale dai vari alunni alla fine del 1° quadrimestre. Dopo aver compilato la relativa tabella del tipo di quella dell'esercizio precedente, calcola il coefficiente di correlazione di Bravais-Pearson e trai qualche conclusione circa la dipendenza di Y da X.

16. Per studiare la dilatazione lineare dei corpi, la classe è stata ripartita in 6 gruppi. Ognuno dei 6 gruppi, ai quali sono state assegnate altrettante sbarre dello stesso metallo ma non di uguale lunghezza, fa 5 misurazioni riscaldando la sbarra sempre dello stesso intervallo termico. I risultati ottenuti sono riportati nella tabella sottostante, dove ΔL (espresso in millimetri) rappresenta l'allungamento della sbarra ed L_i la sua lunghezza iniziale (espressa in centimetri). Calcolare il coefficiente di correlazione di Bravais-Pearson e trarre qualche conclusione circa la dipendenza di ΔL da L.

ΔL / L_i	0,29	0,30	0,31	0,32	0,59	0,60	0,61	0,62	0,88	0,89	0,90	0,91	1,19	1,20	1,21
50	2	1	1	1											
100					3	5	1	1							
150									1	2	6	1			
200													1	3	1

17. Per studiare la caduta dei gravi, un corpo è fatto cadere da 10 altezze diverse e da ogni altezza è fatto cadere 5 volte. Ogni volta è misurato il tempo di caduta e, dopo le 5 cadute dalla stessa altezza, è calcolato il tempo medio di caduta (vale a dire la media aritmetica dei tempi trovati). I risultati sono riassunti nella tabella sottostante, dove le altezze sono espresse in centimetri ed i tempi in secondi. Rappresentare graficamente la situazione e trovare il coefficiente di correlazione di Bravais-Pearson. Trarre qualche conclusione circa la dipendenza del tempo di caduta del grave dall'altezza da cui cade.

Altezza (cm)	150	160	170	180	190	200	210	220	230	240
Tempo (s)	0,54	0,56	0,58	0,61	0,62	0,63	0,65	0,67	0,68	0,70

18. Le pressioni p (in atmosfere) segnate da un manometro alle profondità h (in metri) rispetto alla superficie libera di un liquido sono raccolte nella seguente tabella:

h (m)	20	40	60	80	100	120
p (atm)	2,9	4,7	6,8	8,8	10,5	12,6

Dopo aver rappresentato i dati in un piano cartesiano, determinare il coefficiente di correlazione di Bravais-Pearson e trarre qualche conclusione circa la dipendenza di p da h..

19. Una sbarra metallica è riscaldata a partire da una data temperatura. L'aumento ΔL della sua lunghezza (in millimetri) in funzione dell'aumento Δt di temperatura (in gradi centigradi) è fornito dalla tabella seguente:

Δt (°C)	50	100	150	200	250	300
ΔL (mm)	2	4	7	9	10	13

Dopo aver rappresentato i dati in un piano cartesiano, determinare il coefficiente di correlazione di Bravais-Pearson e trarre qualche conclusione circa la dipendenza di ΔL da Δt.

20. Le posizioni x (in millimetri) occupate da un punto materiale che si muove su una retta (sulla quale è stato fissato un riferimento cartesiano OU) in funzione del tempo t (in secondi) sono fornite dalla seguente tabella:

t (s)	0	5	10	15	20
x (mm)	5	128	254	390	500

Dopo aver rappresentato i dati in un piano cartesiano, determinare il coefficiente di correlazione di Bravais-Pearson e trarre qualche conclusione circa la dipendenza di x da t.

21. Per valutare come varia la temperatura di ebollizione dell'acqua in funzione della pressione ambientale sono stati effettuati alcuni rilevamenti, riportati nella tabella sottostante, dove la pressione è misurata in millimetri di mercurio (mm_{Hg}) e la temperatura in gradi centigradi (°C).

Pressione (mm_{Hg})	25	50	75	100	150	200	250	300	400	500	600	700	800	900
Temperatura (°C)	26	38	46	51	59	68	73	76	82	87	93	98	102	105

Dopo aver rappresentato i dati in un piano cartesiano, determinare il coefficiente di correlazione di Bravais-Pearson e trarre qualche conclusione circa la dipendenza della temperatura di ebollizione dell'acqua dalla pressione ambientale. Valutare, in particolare, quale temperatura corrisponde alla pressione di 760 mm_{Hg}.

Regressione.

22. Con riferimento alle coppie di variabili statistiche considerate nell'esercizio numero:

a) 12; b) 14; c) 16; d) 17; e) 18; f) 19; g) 20; h) 21,

determinare le due rette di regressione e dire se si tratta di regressione lineare o no.

UNA BREVE SINTESI PER DOMANDE E RISPOSTE

DOMANDE.

1. Considerate due variabili statistiche X ed Y, è vero che $M(X+Y) = M(X)+M(Y)$?
2. Se la variabile statistica B rappresenta le misure della base di un rettangolo e la variabile statistica H rappresenta le misure dell'altezza dello stesso rettangolo, indicata con S la variabile statistica che rappresenta le misure dell'area del rettangolo, è vero che risulta $M(S) = M(B)\cdot M(H)$?
3. Se la variabile statistica L rappresenta le misure del lato di un quadrato, indicata con S la variabile statistica che rappresenta le misure dell'area del quadrato, è vero che risulta $M(S) = M(L^2)$?
4. È vero che dalla distribuzione doppia delle frequenze di due variabili statistiche si possono ottenere le distribuzioni marginali delle due variabili?
5. È vero che dalle distribuzioni marginali di due variabili statistiche si può risalire alla distribuzione doppia delle frequenze delle due variabili?
6. È vero che in una correlazione diretta l'indice di correlazione di Bravais-Pearson è maggiore di 1, mentre in una correlazione inversa è minore di 1?
7. Come si può definire la regressione?
8. Ammesso che la retta di equazione $y = ax+b$ sia la retta di regressione della variabile statistica Y sulla variabile statistica X, è vero che la retta di regressione di X su Y ha come equazione quella che si ottiene dalla precedente esprimendo x in funzione di y?
9. Se x_i ed y_i (con $i=1,2,...,n$) sono le n determinazioni delle due variabili statistiche X ed Y rispettivamente, cos'è il baricentro della "nuvola" di punti (x_i, y_i) rappresentati in un piano cartesiano? Ha qualche legame con gli indici di posizione delle variabili?

RISPOSTE.

1. Sì.
2. Sì, dal momento che le due variabili B e H sono indipendenti.
3. No. La misura corretta dell'area del quadrato è espressa dalla media della variabile statistica LL, che è diversa dalla variabile statistica L^2.
4. Sì, addirittura in maniera banale.
5. No.
6. No. L'indice di correlazione di Bravais-Pearson ha sempre valore assoluto non maggiore di 1. In una correlazione diretta esso è positivo mentre in una correlazione inversa è negativo. Quand'è nullo non c'è alcuna correlazione fra le due variabili statistiche prese in esame. Tanto più esso è prossimo a zero tanto meno correlate risultano le due variabili. Tanto più l'indice ha valore assoluto prossimo ad 1 tanto più le due variabili sono correlate. Quando l'indice è uguale a ± 1 la correlazione (diretta o inversa) è perfetta.
7. La regressione è il metodo che permette di trovare una funzione $y=f(x)$ idonea ad esprimere la dipendenza della variabile statistica Y dalla variabile statistica X. Questa funzione si chiama funzione di regressione di Y su X.
8. No. In effetti, se x_i ed y_i (dove $i = 1,2,...,n$) sono le n determinazioni delle variabili statistiche X ed Y rispettivamente, m_x ed m_y le medie di tali variabili, σ_x e σ_y le relative deviazioni standard e inoltre $p=\sum_{i=1}^{n} x_i y_i$, i coefficienti a e b della prima equazione sono espressi dalle formule seguenti:

$$a = \frac{p - n\, m_x\, m_y}{n\, \sigma_x^2}\,, \qquad b = m_y - a\, m_x\,,$$

mentre l'equazione della retta di regressione di X su Y è x = a'y+b', dove si ha:

$$a' = \frac{p - n\, m_x\, m_y}{n\, \sigma_y^2}\,, \qquad b' = m_x - a' m_y\,.$$

9. Il baricentro della nuvola di punti è il punto in cui si secano le rette di regressione di Y su X e di X su Y. Si dimostra che le sue coordinate cartesiane sono (m_x, m_y).

Complementi: Il metodo dei minimi quadrati.

1. Assegnati n punti (x_i, y_i), con $i = 1,2,\dots,n$, di una distribuzione statistica, si pone il problema di trovare la linea di equazione y=f(x) che meglio approssima la distribuzione. In particolare, ove ne ricorrano le condizioni, la retta che meglio risponde ai requisiti richiesti.
La tecnica per la ricerca e la determinazione della funzione y=f(x) è nota come **metodo dei minimi quadrati**. Lo descriviamo per grandi linee.
Si vuole trovare allora la retta di equazione:

$$y = a\,x + b$$

che meglio approssima la distribuzione presa in esame.
Si considera al riguardo il quadrato della distanza di ogni punto $P_i(x_i y_i)$ della distribuzione dal punto Q_i della retta avente la medesima ascissa, vale a dire $Q_i(x_i,\ ax_i + b)$; dunque:

$$\overline{P_i Q_i}^2 = \big(y_i - (ax_i + b)\big)^2;$$

si trova quindi l'espressione della somma degli n valori $\overline{P_i Q_i}^2$, cioè:

$$\sum_{i=1}^{n} \big(y_i - (ax_i + b)\big)^2.$$

Il metodo dei minimi quadrati prevede di rendere minima questa somma, vale dire di trovare per quali valori di a, b ciò avviene. Ebbene, la teoria (che però non possiamo sviluppare) mostra che ciò accade per i valori di a, b forniti dalle formule [2] in 55.4.1.
Questi due valori, la *pendenza* a e l'*ordinata all'origine* (o *intercetta*) b sono spesso chiamati ***stimatori OLS***, dove la sigla *OLS* sta per *Ordinary Least Squares*, che è l'espressione inglese per *Metodo dei minimi quadrati*.

È il caso di far presente che la regressione lineare va bene se effettivamente i dati sperimentali sono distribuiti in modo da non discostarsi molto da un andamento lineare. Ma se questo andamento è molto discosto da quello lineare e fa pensare di più ad una linea curva allora è preferibile approssimare l'andamento dei dati sperimentali con la linea che meglio si adatta alla situazione (Fig. 14) e che può essere una parabola, un'iperbole o altra curva. Anche in questo caso il metodo dei minimi quadrati permette di risolvere la questione. Ma di questo non possiamo occuparci.

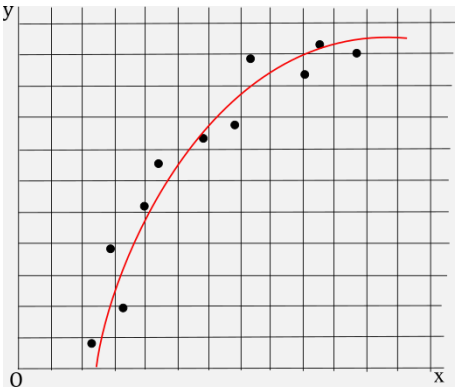

FIG. 14

2. La creazione del metodo dei minimi quadrati fece registrare a suo tempo, cioè agli inizi del secolo XIX, un'aspra polemica fra i due matematici che se ne attribuivano la paternità, vale a dire: il tedesco Carl Friedrich Gauss (1777-1855) e il francese Adrien Marie Legendre (1752-1833).

Oggi sappiamo che fu Gauss a creare e utilizzare per primo il "metodo", nel 1802, al fine di determinare la traiettoria di Cerere, un asteroide all'interno del Sistema Solare, scoperto da poco, precisamente il 1° gennaio 1801, dall'astronomo italiano Giuseppe Piazzi (1746-1826). Ma Gauss pubblicò i suoi lavori circa 7 anni dopo, nel 1809. Solo che alcuni anni prima, nel 1806, Legendre aveva pubblicato una sua opera in cui descriveva la tecnica per la quale egli stesso coniò la denominazione *metodo dei minimi quadrati*.

Fu per questo motivo che, dopo la pubblicazione di Gauss, Legendre gli scrisse per rivendicare la priorità nella creazione del "metodo". Ovviamente Gauss non gliela riconobbe.

Siccome la cosa si era già ripetuta per altre scoperte matematiche che avevano visto implicati ancora Gauss e Legendre, la polemica, soprattutto da parte di Legendre nei confronti di Gauss, che egli accusava di plagio oltreché di scorrettezza, non solo non si placò, ma si inasprì.

Oggi, dopo la pubblicazione postuma dei lavori di Gauss, e in particolare di un suo diario dove annotava scrupolosamente i suoi risultati, sappiamo senza alcun dubbio che Gauss precedette sempre Legendre.

Detto per completezza, i lavori di Gauss e di Legendre che riportano il metodo dei minimi quadrati sono rispettivamente i seguenti:

- *Theoria motus corporum coelestium in sectionibus conicis solem ambientium (Teoria del moto di corpi celesti (che si muovono) secondo sezioni coniche intorno al Sole);*
- *Nouvelles méthodes pour la détermination des orbites des comètes (Nuovi metodi per la determinazione delle orbite delle comete).*