# II CENNI DI STATISTICA

## II.1 Generalità

La sfera d'influenza della statistica ha raggiunto ormai tutti i campi, dallo scientifico al sociologico, dal demografico al tecnico, dal politico all'economico, dal medico al psicologico ed a tanti altri.

Moltissime pubblicazioni sia nel campo medico che in quello tecnico fanno riferimento a principi di statistica; nei congressi, nelle conferenze vengono presentati dati e risultati di ricerche con precisi riferimenti statistici.

L'autore avverte la necessità che a tutti i livelli si debba possedere una sufficiente familiarità con le nozioni fondamentali della statistica e pertanto in questo capitolo cercherà di dare in rapida sintesi i principi generali della stessa, facendoli precedere da un paragrafo in cui sono esposti i più importanti concetti matematici usati. Nell'ordine vengono poi trattati i seguenti argomenti: la distribuzione di frequenze, le grandezze statistiche significative, la teoria elementare della probabilità e gli scopi della teoria dei campioni. Infine l'interpolazione ed il metodo dei minimi quadrati introducono la correlazione e la regressione tra due variabili.

## II.2 Elementi di aritmetica ed algebra

All'inizio di questa trattazione che si svolgerà in modo strettamente sintetico, diamo alcune definizioni che come vedremo sono sempre ricorrenti.

La *Statistica* comprende i metodi per raccogliere, ordinare, presentare ed analizzare i dati ed anche per dedurre conclusioni in modo da prendere ragionevoli decisioni fondate su queste basi.

Quando si vogliono raccogliere i dati che caratterizzano un fenomeno, una rilevazione, un gruppo di individui o di cose, se ci riferiamo a tutti i dati, questi vengono chiamati in forma significativa: *popolazione od universo*; quando se ne considera una parte che ne riassume le caratteristiche la chiameremo: *campione*.

Lo studio di un campione significativo dà sempre delle importanti notizie circa la popolazione a cui appartiene.

Diamo ora alcuni importanti concetti matematici riguardanti le variabili, le funzioni, i grafici, le equazioni, le diseguaglianze ed i logaritmi.

Si definisce *variabile* un simbolo letterale come X, Y, Z, x, y, z che può assumere un qualsiasi valore entro un insieme di valori. Quando la variabile assume un solo valore essa è detta *costante*, nel caso in cui

i dati descritti dalla variabile sono continui o discreti, la variabile è *continua o discreta.*

In genere le misurazioni sono continue e le enumerazioni sono discrete.

Nel caso in cui a ciascun valore di una variabile x corrispondono uno o più valori di un'altra variabile y, possiamo dire che y è *funzione* di x e scriveremo $y = f(x)$.

Se ad ogni valore di x (variabile indipendente) corrisponde un valore di y (variabile dipendente), si dice che y è funzione *univoca* di x, in caso contrario y funzione *non univoca* di x.

Si possono avere anche funzioni di due o più variabili ed in quel caso si scriverà $y = f(x, z, ...)$.

La dipendenza funzionale tra variabili può essere descritta con tabelle come è illustrato nel Cap. III, oppure con una *equazione.*

L'equazione è una espressione algebrica del tipo $y = ax + b$ oppure $y = ax^2 + bx + c$, ecc., dove ciò che si trova a sinistra del segno di uguaglianza è detto *membro di sinistra* e quello a destra *membro di destra.*

Nel Cap. III si vede come le variabili e le equazioni possono rappresentarsi con grafici e diagrammi.

Spesso vengono usati i simboli $<$ e $>$ che hanno il significato di *minore* e di *maggiore*; ad esempio: $4 < 5$ che si legge quattro minore di cinque oppure $6 > 3$ che si legge sei maggiore di tre.

I simboli $\leq$ e $\geq$ aggiungono al precedente significato anche quello dell'uguaglianza e cioè $4x \leq 5$ vuol significare che $4x$ può assumere tutti i valori inferiori od uguali a 5.

Si ricorda che si può effettuare contemporaneamente su entrambi i membri di una equazione la stessa operazione (esclusa la divisione per zero) che si ottiene un'espressione del tutto equivalente.

Cerchiamo ora di dare una definizione estremamente sintetica dei *logaritmi.*

Considerato un numero A positivo, questo può essere sempre espresso come una potenza del numero 10, vogliamo perciò trovare quel numero c tale che valga $A = 10^c$. Il numero c è detto logaritmo di A in base 10 e scriveremo brevemente $c = \log A$.

I logaritmi in base 10 si chiamano *decimali* o *volgari* o di *Briggs.* Quando si sceglie come base il numero $e = 2{,}718282$ (numero di Nepero) la scrittura convenzionale è $c = \ln A$ ed i logaritmi sono chiamati *naturali.*

Per i logaritmi esiste la condizione essenziale che A (detto anche *antilogaritmo* di c) sia sempre maggiore di zero.

I logaritmi presentano considerevoli vantaggi in alcuni tipi di calcoli, come nei casi appresso indicati:

$$\log (A \cdot B) = \log A + \log B$$

$$\log \frac{A}{B} = \log A - \log B$$

$$\log A^p = p \cdot \log A$$

per cui il logaritmo della seguente espressione può calcolarsi come segue:

$$\log\left(\frac{A \cdot B^p}{C}\right) = \log A + p \cdot \log B - \log C$$

Prima di concludere questo paragrafo è opportuno richiamare alcuni concetti sulle scritture numeriche concernenti dati, risultati e calcoli come l'arrotondamento e le cifre significative.

Dato il numero 23,72 *l'arrotondamento* alla 1ª cifra decimale è 23,7, mentre l'arrotondamento all'unità è 24, in quanto nel primo caso 23,72 è più prossimo a 23,7 che a 23,8, nel secondo invece 23,72 è più prossimo a 24 che a 23.

Nel caso in cui il numero termini per 5, l'arrotondamento viene fatto alla cifra pari che precede il 5, ad esempio 4,645 si arrotonda alla 2ª cifra decimale a 4,64 e 72,355 si arrotonda a 72,36.

Per quanto concerne le *cifre significative* occorre fissarne il significato con alcuni esempi.

Il numero 720,4 ha quattro cifre significative, il numero 127.000 ne ha sei, il numero 0,032 ne ha due ed infine 0,400 ne ha tre.

Quando si fanno dei calcoli come moltiplicazioni, divisioni ed estrazioni di radice, il risultato non può avere più cifre significative di quante ne abbia il numero con meno cifre significative, nel caso di addizione e sottrazione con numeri decimali, il risultato avrà dopo la virgola tante cifre quante ne ha il numero con meno cifre decimali.

## II.3 Distribuzioni di frequenze

I *dati grezzi* sono dati raccolti, ma non ordinati numericamente.

Una *serie* è un insieme di dati raccolti ed ordinati in modo crescente o decrescente.

Il *campo di variazione* dei dati raccolti è la differenza tra il più grande ed il più piccolo.

Quando si ha un numero notevole di dati grezzi è opportuno suddividerli in classi.

Le *classi* sono le suddivisioni nelle quali si intendono ripartire i dati grezzi e sono caratterizzate da due numeri che definiscono *l'intervallo* della classe e che sono il *limite inferiore* ed il *limite superiore* della stessa ed in genere la loro scrittura è del tipo: 120-140, 17,5 - 20,0. Quando manca un limite si dice che l'intervallo è *aperto*.

Il numero di dati compresi in una classe rappresenta la *frequenza* della classe.

Quando si prepara una tabella con dei dati suddivisi per classi, questa è chiamata *distribuzione di frequenze*.

Il *valore centrale* di una classe si ottiene sommando i limiti superiore ed inferiore e dividendo il risultato per due.

Graficamente i risultati possono essere ordinati e rappresentati in vari modi; i più comuni sono gli istogrammi, i poligoni e le curve di frequenza.

Un *istogramma* è formato da tanti rettangoli ordinati, quante sono le classi in cui sono suddivisi i dati, aventi per base l'ampiezza dell'intervallo della classe e centro sul valore centrale e per altezza la frequenza delle rispettive classi.

Il *poligono di frequenze* relativo allo stesso gruppo di dati è il grafico lineare che unisce i valori centrali delle stesse classi (vedi fig. II.3.1).

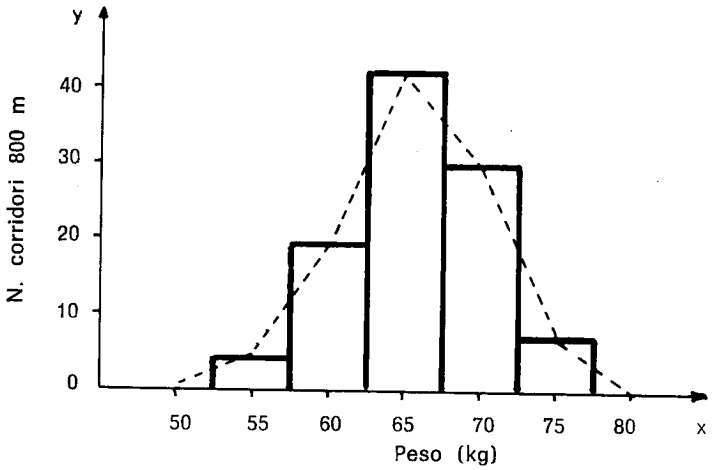



*Fig. II.3.1*

Quando i dati raccolti sono molti e l'ampiezza degli intervalli delle classi è molto piccola, si vede come il poligono di frequenze tenda ad un grafico continuo e non più ad una spezzata; la curva così formata viene chiamata *curva di frequenze*.

I valori delle frequenze degli istogrammi, dei poligoni e delle curve sopra definiti possono essere dati in forma percentuale, riferendoli al numero totale dei dati raccolti. Anche in questo caso possono costruirsi i grafici suddetti, che avranno gli stessi nomi, soltanto che le frequenze saranno indicate come *frequenze percentuali*.

## II.4 Grandezze statistiche significative

Ricordiamo alcune convenzioni di scrittura.

Considerato un insieme N di dati i cui valori indicheremo con $x_1$, $x_2$, . . ., $x_N$, si chiamano *indici* i numeri 1, 2, . . ., N e con i quello generico, per cui il dato generico sarà indicato con $x_i$.

Per indicare la somma di tutti i dati $x_i$, con l'indice i che varia da 1 a N, scriviamo:

$$\sum_{i=1}^{N} x_i = x_1 + x_2 + \ldots + x_N$$

Quando non vi è possibilità di equivoco si usano anche forme abbreviate come

$$\Sigma\, x,\quad \Sigma\, x_i,\quad \Sigma_i x_i.$$

Il valore più rappresentativo di un gruppo di dati è la *media*.

Esistono vari tipi di media, ciascuno più significativo dell'altro in casi specifici; ne diamo le definizioni.

1. Si definisce *media aritmetica* o *media* di un gruppo N di dati la relazione:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

2. Nel caso in cui i dati $x_i$ compaiono ciascuno con la frequenza (o peso) $f_i$, si definisce *media aritmetica ponderata* la relazione:

$$\bar{x} = \frac{\sum_{i=1}^{N} f_i \cdot x_i}{\sum_{i=1}^{N} f_i} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_N x_N}{f_1 + f_2 + \ldots + f_N}$$

3. La *mediana* di un gruppo di dati è il valore di mezzo degli stessi e si indica con $\tilde{x}$.

4. La *moda* di un insieme di dati è il valore che ha la frequenza più alta e si indica con $\dot{x}$.

5. La *media geometrica* $x_g$ di un gruppo N di numeri è data dalla relazione:

$$x_g = \sqrt[N]{x_1 x_2 \ldots x_N}$$

e si calcola usando i logaritmi, in quanto più semplicemente si ha:

$$\log x_g = \frac{1}{N}(\log x_1 + \log x_2 + \ldots + \log x_N)$$

6. La *media quadratica* $x_q$ di un insieme N di dati si ottiene dalla relazione:

$$x_q = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}} = \sqrt{\bar{x}^2}$$

Definite così le principali medie, diamo cenno, in forma altrettanto sintetica, di altre grandezze statistiche significative.

Si definisce *dispersione* o *variazione* di un gruppo di dati l'attitudine di questi a disporsi intorno ad un valore medio.

La *dispersione massima* è la differenza tra il valore massimo ed il minimo del gruppo di dati N.

Si chiama *scarto* la differenza dell'i-esimo dato $x_i$ e la media aritmetica $\bar{x}$ e si indica con:

$$\xi = x_i - \bar{x}$$

Una proprietà degli scarti è che la loro somma è nulla, cioè:

$$\sum_{i=1}^{N} \xi_i = 0$$

Si può allora definire lo *scarto medio assoluto dalla media aritmetica* la relazione:

$$\frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N} = \frac{\sum_{i=1}^{N} |\xi_i|}{N}$$

dove il simbolo $|\ldots|$ indica il valore assoluto dello scarto, cioè il suo valore numerico privo di segno.

Altro valore molto significativo della dispersione di un insieme di dati è lo *scarto quadratico medio* definito dalla:

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

Quando $N > 30$, si può semplificare nella:

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

Lo scarto quadratico medio, tra le proprietà che possiede, ne ha una che statisticamente riveste grande importanza; infatti in una distribuzione normale di dati (dove la distribuzione di frequenze è simmetrica rispetto alla media) il 68,27% dei dati $x_i$ è compreso tra $\bar{x} - s$ e $\bar{x} + s$, il 95,45% tra $\bar{x} - 2s$ e $\bar{x} + 2s$ ed il 99,73% tra $\bar{x} - 3s$ e $\bar{x} + 3s$.

Infine la *varianza* è definita come il quadrato dello scarto quadratico medio e cioè:

$$\sigma = s^2$$

Nello studio delle distribuzioni di frequenze sovente, invece di usare come variabile la x, si preferisce sostituirla con una *variabile standardizzata* adimensionale, definita dalla:

$$z = \frac{x - \bar{x}}{s}$$

dove $\bar{x}$ ed s hanno i significati sopra esposti.

Per concludere questo paragrafo rappresentiamo gli andamenti delle curve di frequenze più comuni con la fig. II.4.1.

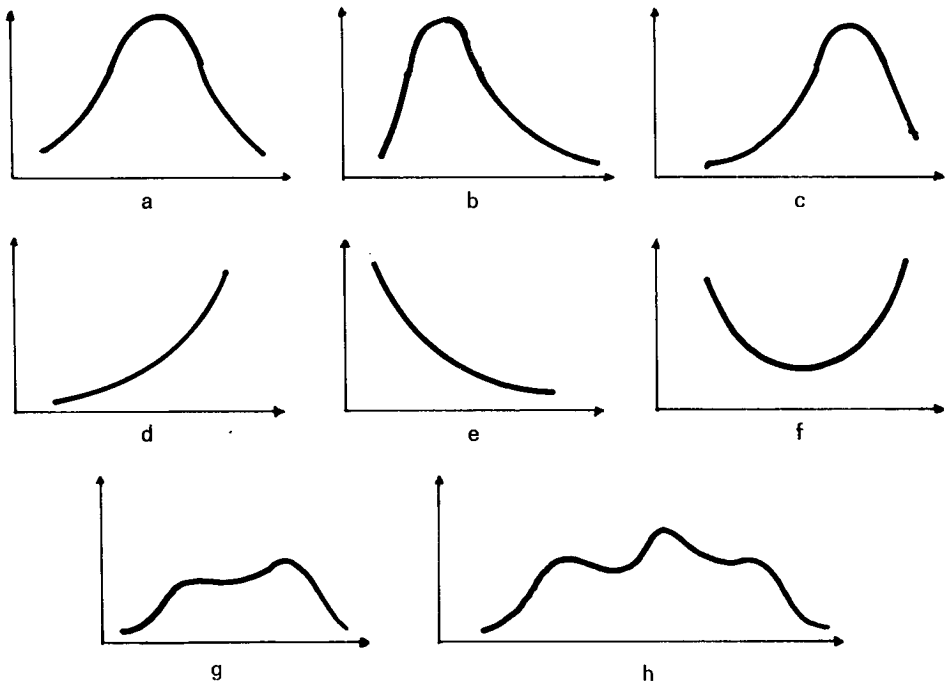
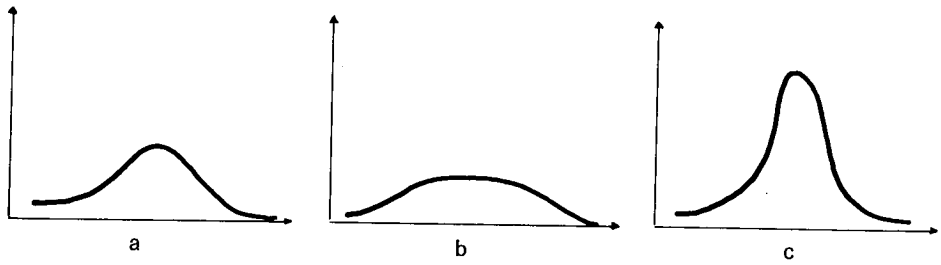



*Fig. II.4.1*

a curva simmetrica od a forma campanulare
b curva obliqua sinistra con inclinazione positiva
c curca obliqua destra con inclinazione negativa
d curva ascendente od omegamodale
e curva discendente od alfamodale
f curva ad U
g curva bimodale
h curva plurimodale

Relativamente alle curve di frequenza indicate si richiama il concetto di *asimmetria*, che rappresenta il grado di scostamento dalla simmetria e per le curve di tipo normale, anche quello di *curtosi*, che è il grado di altezza raggiunto; si distinguono tre casi di curtosi che sono rappresentati nella figura II.4.2.




a mesocurtosi, b platicurtosi, c leptocurtosi.

*Fig. II.4.2*

## II.5 Teoria elementare della probabilità e sue applicazioni

Cercheremo di dare i fondamenti della teoria elementare della probabilità evidenziando le analogie con le distribuzioni di frequenze sopra riassunte.

Supponiamo che di una grandezza x, legata ad un fenomeno fisico oppure ad un qualsiasi evento, sia stato rilevato un gruppo di valori $x_1, x_2, \ldots, x_N$ ognuno dei quali abbia rispettivamente la probabilità $p_1, p_2, \ldots, p_N$ di verificarsi, come prima cosa possiamo dire che la somma delle probabilità è evidentemente: $p_1 + p_2 + \ldots + p_N = 1$ (oppure = 100%).

Pensiamo ora di voler rappresentare in forma analitica o grafica la funzione che rappresenta la probabilità p della grandezza x, detta *funzione di probabilità* di x; dalla rappresentazione appare immediata l'analogia tra la funzione di probabilità e la distribuzione di frequenze precedentemente illustrata.

Per chiarire con un esempio, pensiamo alla distribuzione di probabilità come alla distribuzione della popolazione (cioè all'insieme di tutti i valori) ed alla distribuzione di frequenze come alla distribuzione di campioni estratti dalla popolazione stessa.

Se i campioni considerati tendono, come ampiezza, all'universo dei dati, la curva di frequenze che noi conosciamo sarà valida come distribuzione della popolazione, da ciò la nostra affermazione che la funzione di probabilità ha lo stesso andamento delle curve di frequenze.

Tra le tante proprietà che una distribuzione di probabilità possiede ne ricordiamo qualcuna, pertanto, premesso che la *probabilità* $p_i$ che un evento $x_i$ accada è uguale al rapporto tra il numero m degli eventi favorevoli ed il numero N di tutti gli eventi, abbiamo che:

1. chiamate $p_1$ e $p_2$ le probabilità di due eventi distinti, la probabilità che avvenga "sia l'uno che l'altro evento" è uguale al prodotto di $p_1$ per $p_2$.
2. chiamate $p_1$ e $p_2$ le probabilità di due eventi distinti, la probabilità che avvenga "o l'uno o l'altro evento" è uguale alla somma di $p_1$ più $p_2$.

Le due proprietà sopra ricordate valgono anche nel caso di più probabilità.

Per quanto riguarda le distribuzioni di probabilità possiamo dire che ne esistono varie: la distribuzione binomiale, quella multinomiale, quella normale o Gaussiana e su quest'ultima noi fisseremo la nostra attenzione, rimandando ad altri testi coloro che volessero avere maggiori notizie.

Una *distribuzione normale o Gaussiana* è definita dall'equazione:

$$y = \frac{1}{s \cdot \sqrt{2\pi}} e^{-\frac{(x-\bar{x})^2}{2s^2}}$$

ed in termini di variabile standard da:

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Quando la variabile x viene espressa in termini di variabile standardizzata z diciamo che la distribuzione di z avviene con media $\bar{x} = 0$ e scarto quadratico medio $s = 1$.

La fig. II.5.1. rappresenta la curva normale standardizzata con area uguale ad uno.

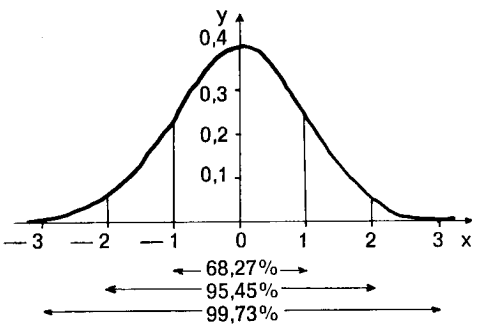



*Fig. II.5.1*

Dalla figura si vede come le aree comprese tra $z = -1$ e $z = 1$, tra $z = -2$ e $z = 2$, tra $z = -3$ e $z = 3$ valgono rispettivamente il 68,27%, il 95,45% ed il 99,73% dell'area totale unitaria.

## II.6 Simboli convenzionali

Elenchiamo i simboli normalmente adoperati per indicare alcune grandezze statistiche:

$\bar{x}$ media campionaria

$s$ scarto quadratico medio campionario

$s^2$ varianza campionaria

$\hat{s} = s\sqrt{\frac{N}{N-1}}$

$\mu$ media della popolazione

$\sigma$ scarto quadratico medio della popolazione

$\sigma^2$ varianza della popolazione

## II.7 Scopi della teoria dei campioni

Nel caso si volessero conoscere i parametri di una popolazione, come la media, la varianza, ecc. è di grande aiuto la teoria dei campioni, la quale non è altro che lo studio delle relazioni che esistono tra una popolazione ed i campioni estratti dalla stessa. La teoria dei campioni ci aiuta anche a confrontare le differenze osservate tra due campioni della stessa popolazione ed a individuare se sono significative.

Questo avviene normalmente con l'uso dei test di significatività o di ipotesi.

Affinché un test sia buono esso deve essere formulato in modo da minimizzare gli errori di decisione; questi possono essere di I tipo quando si scarta un'ipotesi che dovrebbe essere accettata o di II tipo nel caso contrario.

Se riusciamo a limitare il I tipo di errore aumentiamo il *livello di significatività* di un test (si scelgono livelli del 5% o dell'1%), se minimizziamo il II tipo di errore aumentiamo la *potenza* del test.

I campioni si dividono in grandi campioni con $N > 30$ ed in piccoli campioni per $N < 30$.

Per i primi in genere le distribuzioni campionarie sono del tipo normale e l'approssimazione a questa va migliorando al crescere di N, per i secondi invece essa è meno buona e peggiora al diminuire di N.

Per i piccoli campioni in particolare si sono studiate due distribuzioni, quella del t di Student e quella del chi-quadrato ($\chi^2$), le quali sono valide anche per i grandi campioni.

Vediamo in sintesi il loro significato.

1. *Distribuzione del t di Student*

Data una popolazione, consideriamo alcuni campioni estratti da essa e per ciascuno si consideri il valore statistico

$$t = \frac{\bar{x} - \mu}{s/\sqrt{N}} \qquad \text{(II.7.1)}$$

Nel caso in cui i campioni estratti siano tutti di ampiezza N si può costruire la distribuzione:

$$y = \frac{A}{1 + \frac{t^2}{\nu}^{(\nu+1)/2}}$$

dove A è una costante che dipende da N e $\nu = N - 1$.

La distribuzione y, detta distribuzione t di Student, è molto simile a quella normale, specialmente per $N > 30$ e quindi i test di significatività possono essere estesi anche ai piccoli campioni con la sola differenza che il valore z della variabile standardizzata della distribuzione normale è sostituito da un opportuno valore di t.

2. *Distribuzione chi-quadrato* ($\chi^2$)

Analogamente a quanto visto per la distribuzione precedente si consideri il valore statistico:

$$\chi^2 = \frac{Ns^2}{\sigma^2} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2}{\sigma^2}$$

se estraiamo dei campioni di ampiezza N e per ciascuno di essi calcoliamo il $\chi^2$, possiamo costruire la distribuzione del $\chi^2$ data da:

$$y = A(\chi)^{\nu-2} e^{-\frac{1}{2}\chi^2}$$

dove A è una costante che dipende da $\nu = N - 1$ che è il numero di gradi di libertà della statistica che si sta osservando.

Anche in questo caso si possono fissare i livelli di significatività dei test.

Per poter calcolare i valori statistici (II.7.1) e (II.7.2) si vede che è necessario conoscere dei parametri appartenenti sia al campione che alla popolazione, ma se questi sono sconosciuti essi vengono ricavati dal campione stesso.

Rimandiamo ad altre pubblicazioni i lettori che volessero avere maggiori notizie al riguardo, ricordando però che il calcolo del $\chi^2$ permette di avere una rapida visione della discrepanza esistente tra le frequenze osservate di una distribuzione e quelle attese o teoriche del tipo normale, binominale, ecc.

## II.8 L'interpolazione ed il metodo dei minimi quadrati come introduzione alla regressione ed alla correlazione di due variabili

L'oggetto di questo paragrafo riguarda la ricerca del legame che esiste tra due variabili, ad esempio: il peso degli atleti seniores lanciatori di disco e la loro statura.

Una volta raccolti i dati in una tabella e poi rappresentati su di un grafico a coordinate cartesiane (vedi III parte), ci troveremo senz'altro di fronte ad un *diagramma a dispersione* (fig. II.8.1).

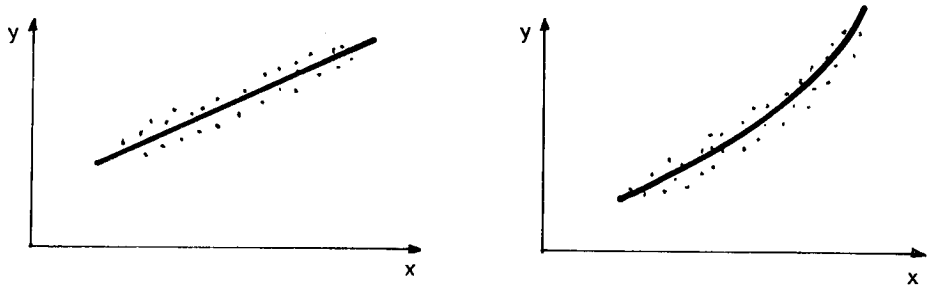



*Fig. II.8.1*

Vogliamo trovare una curva, detta interpolante, che approssimi nel miglior modo possibile tali dati, tale operazione viene chiamata *interpolazione.*

E' il ricercatore, che dopo aver preso visione della dispersione dei dati, sceglie il tipo di funzione che meglio possa approssimare la distribuzione.

Le funzioni usate sono molte, a partire dalla retta fino alle curve di n-esimo grado, alla iperbole, alla curva esponenziale o logaritmica, ecc.

Per evitare il giudizio soggettivo nel ricavare la curva interpolante è adottato comunemente il *metodo dei minimi quadrati.*

Presa una distribuzione a dispersione come in fig. II.8.2, dove i $d_i$

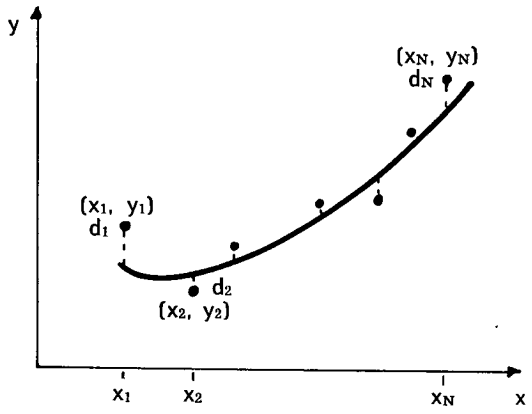



*Fig. II.8.2*

sono le distanze dei punti sperimentali dalla curva interpolante, la migliore di queste è quella per cui è minima la somma: $d_1^2 + d_2^2 + \ldots + d_N^2$.

Se la curva è una retta, questa è la retta dei minimi quadrati, se è una parabola, questa è la parabola dei minimi quadrati, ecc. Pertanto data una distribuzione a dispersione e trovata col metodo dei minimi quadrati la curva interpolante, questa è la *curva di regressione* della grandezza y rispetto alla x e sarà chiamata negativa o positiva a seconda che y diminuisca od aumenti al crescere di x.

Si ricorda che in genere la curva di regressione di y rispetto ad x non è uguale alla curva di regressione di x rispetto alla y.

La *correlazione*, strettamente legata alla regressione, rappresenta il grado della relazione tra le variabili e quando una funzione descrive bene il legame tra due variabili si potrà dire che queste sono ben correlate.

Un indice della bontà delle correlazioni esistenti tra variabili è determinata dal coefficiente di correlazione r, il cui valore varia tra $-1$ e $+1$.

Si rimanda ad altri testi l'approfondimento del significato matematico di detto coefficiente.