

ANALISI STATISTICA DEGLI ERRORI

1. TIPI DI GRANDEZZE MISURATE

In generale possiamo trovare due tipi di grandezze: quelle discrete e quelle continue.

1.1 Grandezze discrete

Si dicono grandezze discrete quelle grandezze che si basano su eventi elementari che assumono valori distinti. Per esempio e' possibile avere 2 o 3 bambini in una famiglia ma non certo 2,5. Analogamente quando tiriamo un dado ci si aspetta un valore da 1 a 6, cioè non e' possibile avere un valore, ad esempio, di 4,5.

Il modo piu' semplice di presentare tali dati e' di ordinarli in ordine crescente e disegnarli con un diagramma di frequenza assoluta, con la grandezza misurata come ascissa e la frequenza con cui essa si presenta come ordinata. Alternativamente, l'ordinata del diagramma delle frequenze puo' essere resa adimensionale dividendo la frequenza dell'evento per il numero totale degli eventi. Questo diagramma e' detto diagramma delle frequenze relative (vedi FIG. 1).

1.2 Grandezze distribuite con continuita'

Le grandezze distribuite con continuita' sono quelle in cui gli eventi possono assumere qualsiasi valore tra limiti dati. Ossia, si puo' dire che e' sempre possibile avere un valore intermedio tra altri due valori, non importa come essi siano stati presi.

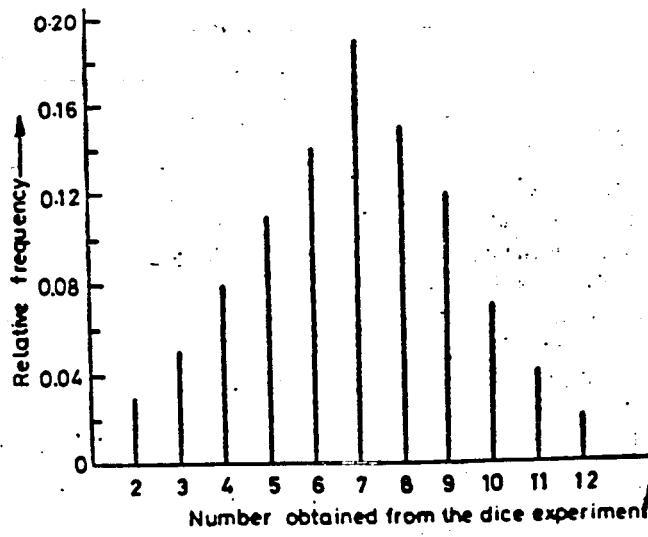
1.2.1 Istogramma

Un istogramma e' una rappresentazione grafica di dati misurati (che sono generalmente relativi a grandezze continue) in cui l'ascissa indica i valori misurati e l'ordinata la frequenza d'occorenza in un range specifico di valori misurati. Percio' e' importante nel costruire un istogramma suddividere le osservazioni in gruppi adeguati (intervalli, classi o celle) scegliendo i limiti adatti. Per prima cosa, quindi, determiniamo il "range" di valori R(x) che e' dato da:

$$R(x) = (X_{max} - X_{min}) \quad (1)$$

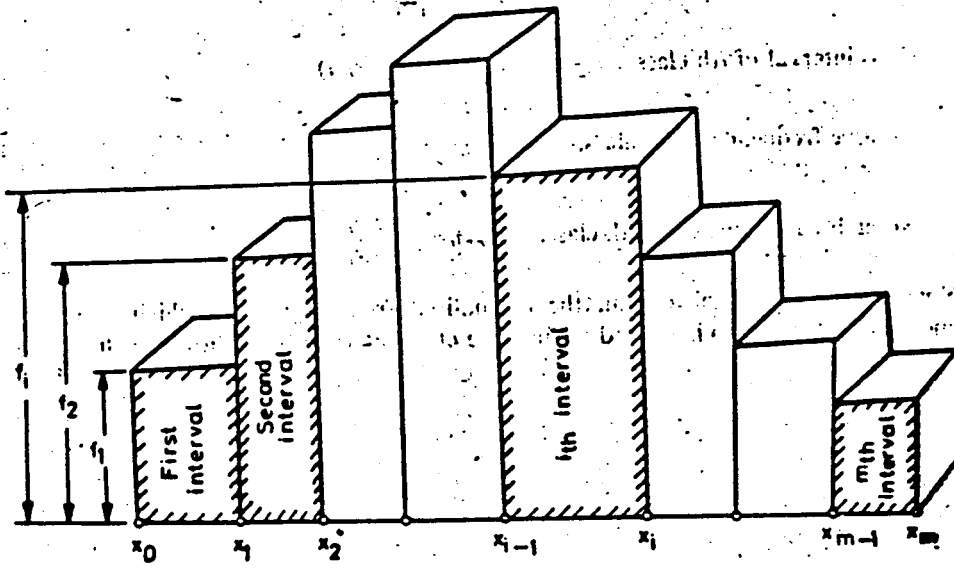
dove X_{max} e' il valore piu' elevato e X_{min} e' il piu' basso. Inoltre

$$R(x) = \text{numero di classi} * \text{ampiezza della classe.}$$

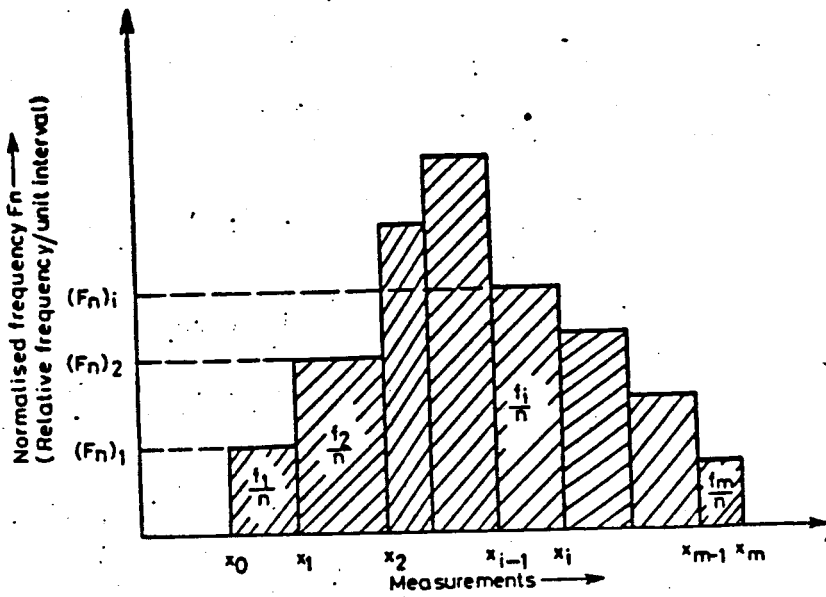


N	F	f
2	6	0,03
3	10	0,05
4	16	0,08
5	22	0,11
6	28	0,14
7	38	0,19
8	30	0,15
9	24	0,12
10	14	0,07
11	8	0,04
12	4	0,02
TOT.	300	1,00

- FIG. 1 -



- FIG 2 -



- FIG 3 -

Prima di decidere l'ampiezza, il numero e gli estremi delle classi, bisogna tener conto delle seguenti osservazioni:

- 1) Le classi devono contenere tutti i dati
- 2) Esse hanno usualmente almeno di 6 intervalli e non più di 16.
- 3) Ogni dato deve essere contenuto in uno ed in uno solo degli intervalli.

Per soddisfare i criteri su menzionati di solito si segue la seguente procedura. Il valore di $R(x)$ ottenuto dalla (1) e' diviso da un valore opportuno di ampiezza così che il numero delle classi stia tra 6 e 16. Se questo numero ottenuto non e' intero allora si prende il primo intero maggiore del numero per rappresentare il numero di classi. E' bene notare che non c'e' alcuna regola che obblighi a mantenere uniforme l'ampiezza delle classi. Percio' se la situazione richiede un numero fissato di classi, le loro ampiezze potranno essere disuguali per poterle adattare al range dato dei valori. Comunque, in pratica esse sono mantenute uguali poiche' questo facilita il confronto della frequenza tra classi diverse, eliminando la necessita' di tener conto delle differenze nella misura dell'intervallo di classe.

Dopo aver deciso l'intervallo di classe e il numero di classi, i punti medi della classe e gli estremi risultano fissati. Il punto medio o marker della classe e il valore della variabile che si trova a meta' tra l'estremo superiore e quello inferiore della classe stessa.

Nel determinare gli estremi della classe occorre fare in modo che nessuna misura cada sugli stessi così che senza difficoltà possiamo determinare a quale classe appartiene una particolare osservazione. Percio' normalmente fissiamo il valore di estrema' della classe con un numero di cifre significative maggiore di quello con cui effettuiamo le misure. Questo assicura che nessuno dei dati coincida con gli estremi.

Dopo di cio', i dati sperimentali vengono raggruppati nelle varie classi scelte. Il numero delle osservazioni che cade all'interno degli estremi di una particolare classe e' noto come la frequenza della classe. A questo punto possono essere disegnati gli istogrammi delle frequenze assolute, frequenze relative o frequenze percentuali. (vedi FIG. 2) ←

Nella pratica, tali distribuzioni di frequenze sono normalizzate rispetto all'ampiezza della classe e sono dunque rappresentate sotto forma di istogrammi normalizzati.

1.2.2 Istogrammi normalizzati

In un istogramma normalizzato, si utilizza come ordinata la frequenza normalizzata invece della frequenza assoluta o percentuale.

La frequenza normalizzata e' ottenuta come segue.

Avendo deciso di raggruppare i dati in m classi, siano X_0, X_1, \dots, X_m gli estremi delle classi e f_1, f_2, \dots, f_m le frequenze di occorrenza nelle varie classi.

Il numero totale delle misure e' $n = \sum_{i=1}^m f_i$

l'ampiezza della classe i -esima e' $(x_i - x_{i-1})$

la frequenza relativa della i -esima classe e' $(f_r)_i = f_i / \sum_{i=1}^m f_i$

la frequenza normalizzata della i -esima classe e' $(F_n)_i = \frac{(f_r)_i}{(x_i - x_{i-1})}$

Ora, in un istogramma normalizzato, la frequenza normalizzata $(F_n)_i$ (che e' la frequenza relativa per un intervallo unitario) e' tracciata come e' mostrato in Fig. 3.

L'area della i -esima banda in un istogramma normalizzato, A_i sarà pari a

$A_i = \text{ordinata} * \text{ampiezza della classe} =$

$$A_i = \frac{1}{(x_i - x_{i-1})} \frac{f_i}{\sum_{i=1}^m f_i} (x_i - x_{i-1}) = \frac{f_i}{\sum_{i=1}^m f_i} = \frac{f_i}{n} = (f_r)_i$$

Perciò, l'area assegnata ad ogni banda in un istogramma normalizzato rappresenta la frequenza di occorrenza relativa $(f_r)_i$ degli eventi che cadono all'interno dell'intervallo della classe data.

Nel caso in cui tutti gli intervalli delle classi sono uguali, l'ordinata di ogni intervallo di classe e' proporzionale alla frequenza relativa. L'area dell'istogramma e' data da:

$$\sum_{i=1}^m A_i = \sum_{i=1}^m \frac{f_i}{n} = \frac{n}{n} = 1$$

L'area di un istogramma normalizzato e' dunque unitaria.

Si può dimostrare che anche l'area di una distribuzione gaussiana e' unitaria. Perciò, se i dati distribuiti con continuita' devono essere comparati con una distribuzione Gaussiana, allora l'istogramma normalizzato diventa piu' appropriato degli istogrammi ordinari.

2. TENDENZA CENTRALE DEI DATI

Uno fra gli importanti parametri che descrivono le informazioni numeriche riguarda la locazione della tendenza centrale dei dati.

Ci sono svariati modi attraverso i quali e' possibile determinare la tendenza centrale dei dati (mediana, moda, media aritmetica). Comunque, puo' essere dimostrato che ciascuno dei modi precedentemente citati puo' essere piu' appropriato per descrivere un tipo di dati e completamente inadatto per altri.

2.1 Moda

Possiamo dire che la moda e' il valore che si presenta con maggiore frequenza. Ad esempio consideriamo un set di numeri contenente i seguenti valori discreti : 15, 18, 17, 14, 17, 13, 16. E' ovvio che in questo set di valori la frequenza del numero 17 e' la massima. Percio' puo' essere definito come valore modale.

Puo' essere notato dall'esempio precedente che, in un set di valori discreti, il valore modale puo' essere determinato con estrema facilità. Comunque nel caso di dati distribuiti con continuita', il valore modale si trova nel gruppo di dati che hanno la maggiore densita' di frequenza. In altre parole, esso si trova all'interno del piu' alto rettangolo dell'istogramma. Per determinare il valore modale consideriamo la parte di istogramma contenente il rettangolo di frequenza maggiore insieme con le classi adiacenti (vedi FIG. 4)

Noi definiamo la moda come l'ascissa X' del punto di intersezione P dei segmenti QS e RT. Puo' essere dimostrato che il valore della moda X' e' dato dalla seguente formula:

$$\text{Moda } X' = L_1 + \left[\frac{\Delta_1}{\Delta_2 + \Delta_1} \right] \cdot C$$

laddove L_1 e' l'estremo della classe piu' bassa

Δ_1 = eccesso della frequenza modale sulla frequenza della classe adiacente piu' bassa

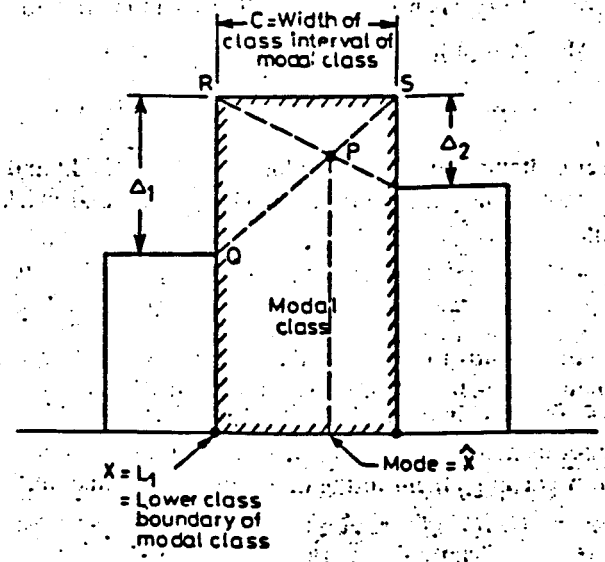
Δ_2 = eccesso della frequenza modale sulla frequenza della classe adiacente piu' alta

C = ampiezza della classe modale.

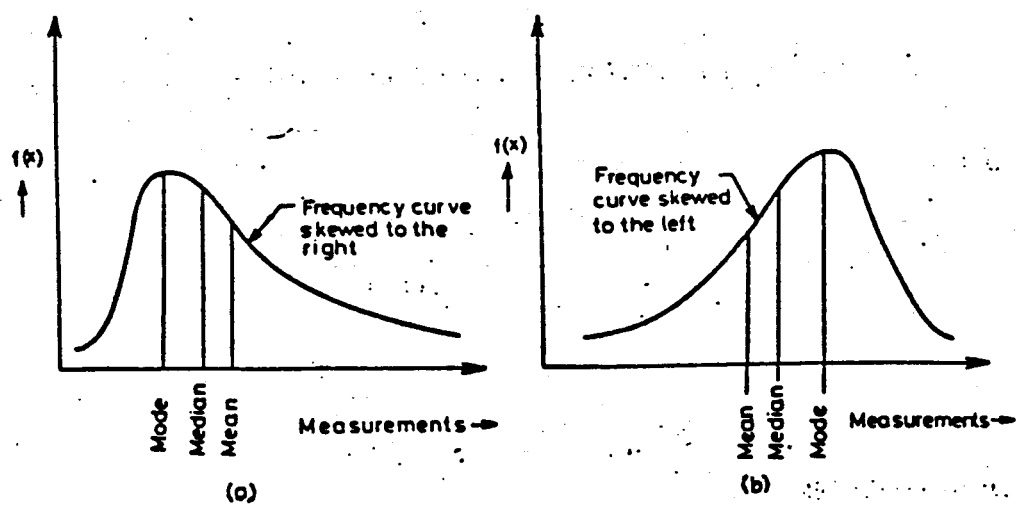
Occasionalmente una classe puo' essere bimodale, cioe' essa puo' contenere due distinti valori che possono occorrere molto piu' frequentemente degli altri valori della variabile. Tutto cio' sta usualmente ad indicare che il campione proviene da due popolazioni differenti.

2.2 Mediana

La mediana e' definita come il valore centrale di un gruppo di dati raggruppati in ordine numerico che puo' essere crescente



— FIG 4 —
moda



— FIG 5 —

Posizione relativa di media, moda e mediana

(dal piu' basso valore di osservazione al piu' alto) o viceversa. Nel caso in cui la dimensione del campione e' pari, noi determiniamo la mediana operando la media dei due valori centrali.

La mediana e' anche detta 50-simo percentile e si denota con P_{50} . 50-esimo percentile significa che approssimativamente il 50% delle misure nella distribuzione sono minori di P_{50} ed il 50% delle misure nella distribuzione sono maggiori di P_{50} . Supponiamo che N osservazioni dei dati siano ordinate in senso crescente o decrescente da 1 a N . Allora possiamo esprimere matematicamente il valore della mediana come:

$$\text{Mediana} = \{X\}_{(n+1)/2} \quad \text{quando } N \text{ è dispari}$$

$$\text{Mediana} = \frac{1}{2} [\{X\}_{n/2} + \{X\}_{(n/2)+1}] \quad \text{quando } N \text{ è pari}$$

La mediana e' una quantita' posizionale, nel senso che il suo valore e' determinato dal valore di osservazione che occupa una particolare posizione nella distribuzione. Puo' essere notato che la mediana non e' influenzata dai valori estremi dei dati.

In taluni casi, la distribuzione dei dati puo' non essere simmetrica. In tali casi la mediana e' considerata come la locazione centrale dei dati.

Geometricamente, il valore della mediana e' ottenuto dalla intersezione dell'ascissa dell'istogramma con una retta verticale che divide l'istogramma stesso in due parti aventi aree uguali.

Comunque, si puo' precisare che il valore ottenuto per la mediana puo' non essere rappresentativo se i dati non si raggruppano al centro della distribuzione.

2.3 Media Aritmetica

La Media Aritmetica e' il piu' comune ed il piu' usato tra i parametri di tendenza centrale dei dati. Per definizione, la media aritmetica di un gruppo di dati e' uguale alla somma dei valori del gruppo diviso per il numero di valori contenuti nello stesso. E' possibile determinare sia la media di un insieme di valori misurati che la media di una popolazione. Il simbolo \bar{X} tradizionalmente denota il primo (la media di un insieme di dati o campione) mentre \overline{X} denota il secondo. Si intende per popolazione il totale dei valori tra cui il campione e' preso. Laddove noi facciamo riferimento ad un campione, noi non intendiamo solo una porzione di popolazione, ma una porzione selezionata in modo tale che i valori osservati non siano affetti da bias. Tale processo e' generalmente definito come campionatura casuale o random.

Supponiamo che un particolare campione abbia un totale di n osservazioni in cui i veri valori osservati sono:

X_1 occorso f_1 volte, X_2 occorso f_2 volte, X_m occorso f_m volte. Ora la somma delle frequenze di occorrenza dei valori veri ci da' ovviamente il numero totale di osservazioni del campione.

La media \bar{X} e' data da:

$$\bar{X} = \frac{\sum_{i=1}^m f_i \cdot x_i}{\sum_{i=1}^m f_i}$$

La stessa espressione puo' anche essere applicata per dati distribuiti con continuita' laddove X_i rappresenta il valore del punto medio della i -esima classe avente la frequenza f_i . Similarmente e' possibile calcolare la media di una popolazione. Consideriamo la popolazione costituita da N osservazioni e siano X_1, X_2, \dots, X_k le medie relative a k campioni. \dots , .. siano le medie dei valori di K campioni. La media della popolazione, \bar{X} , e' data da:

$$\bar{X} = \frac{\sum_{j=1}^k X_j}{N}$$

2.4 Media Aritmetica Pesata

A volte, noi associamo ai numeri X_1, X_2, \dots, X_m certi coefficienti pesati o pesi W_1, W_2, \dots, W_m rispettivamente, a seconda del significato o dell'importanza attribuita al singolo numero. In tal caso, la media aritmetica pesata e' definita come:

$$\bar{X} = \frac{W_1 X_1 + W_2 X_2 + \dots + W_n X_n}{W_1 + W_2 + \dots + W_n} = \frac{\sum_{i=1}^n W_i \cdot X_i}{\sum_{i=1}^n W_i}$$

2.5 Migliore stima del valore vero di una misura

Noi abbiamo fino ad ora descritto tre parametri chiamati moda, mediana e media aritmetica, che possono essere usati per descrivere la tendenza centrale dei dati. Per curve che sono moderatamente asimmetriche, vi e' una relazione empirica che lega questi parametri:

$$\text{Media} - \text{moda} = 3 (\text{media} - \text{mediana})$$

La FIG. 5 mostra le relative posizioni di moda, mediana e media per curve di frequenza spostate a destra o a sinistra. Comunque, per curve simmetriche, moda, mediana e media coincidono.

Nell'analisi di dati sperimentali, ci si trova usualmente di fronte al problema della determinazione della migliore stima del valore vero della misura. Supponiamo che X_1, X_2, \dots, X_m siano misure di una quantita' il cui valore vero e' X . Allora la misura potrebbe avere i seguenti errori:

$$\begin{aligned} \text{Errore in } X_1, e_1 &= X_1 - X \\ &\dots \dots \dots \end{aligned}$$

Errore in X_n , $e_n = X_n - X$

Alcuni di questi errori potrebbero essere positivi mentre altri potrebbero essere negativi. Ora, se noi diamo eguale importanza agli errori sia negativi che positivi, allora la strada migliore e' quella di elevare al quadrato tutti gli errori.

Ora, il criterio piu' plausibile per la migliore scelta di X sarebbe quello grazie al quale gli errori commessi nelle varie misure risultano minimi. Alternativamente la somma dei quadrati degli errori possa essere minima.

La somma dei quadrati degli errori puo' essere scritta come:

$$S_e = e_1^2 + e_2^2 + \dots + e_i^2 + \dots + e_n^2 = (x_1 - X)^2 + (x_2 - X)^2 + \dots + (x_i - X)^2 + \dots + (x_n - X)^2$$

La somma dei quadrati degli errori e' una funzione di X che consente la migliore stima del valore vero. Ora, al fine di determinare il valore di X che renda minima S_e si procede come segue. Detto $X = X_0$ il valore di X che minimizza S_e si ha:

$$\frac{dS_e}{dX} = -2(x_1 - X) - 2(x_2 - X) \dots - 2(x_i - X) \dots - 2(x_n - X) = 0$$
$$\Rightarrow X_0 = (x_1 + x_2 + \dots + x_i + \dots + x_n) / n = \bar{X}_n$$

In generale, quindi, la media aritmetica e' considerata la migliore stima del valore vero di un insieme di valori misurati.

2.5 Esempio

La tabella mostra la distribuzione di frequenza della resistenza di un resistore fabbricato da un pro

Resistenza [Ω]	Frequenza
93 - 95	4
96 - 98	15
99 - 101	33
102 - 104	21
105 - 107	7

Determinare: a) la media aritmetica, b) la mediana, c) la m d , d) usare la formula empirica per calcolare il valore modale e comparare tale valore con quello precedentemente ottenuto.

a) Per trovare la media aritmetica, noi possiamo supporre che 4 resistori hanno una resistenza di 94Ω, 15 di 97Ω ecc.

La media aritmetica diviene:

$$\frac{4 \times 94 + 15 \times 97 + 3 \times 100 + 21 \times 103 + 7 \times 106}{4 + 15 + 33 + 21 + 7} = \frac{8036}{80} = 100,45 \Omega$$

- b) il valore mediano puo' essere determinato attraverso il metodo di interpolazione o attraverso la definizione. Per definizione la mediana e' quel valore tale che la meta' della frequenza totale ($80/2 = 40$) giace prima di esso e l'altra meta' dopo.

Ora la somma delle frequenze delle due prime classi e' $4+15=19$. Così per avere i desiderati 40, abbiamo bisogno di 21 dei 33 campioni della terza classe. Il valore mediano dovrebbe stare ai $21/33$ dell'ampiezza dell'intervallo tra 99 e 101.

$$\text{Mediana} = L_1 + \left\{ \frac{\frac{n}{2} - (\sum p_i)}{f_{\text{mediana}}} \right\} \cdot C = 99 + \frac{40 - (4+15)}{33} (101-99) = 100,273 \Omega$$

- c) Il valore modale per dati distribuito con continuita' e':

$$L_1 = 99$$

$$\Delta_1 = 33 - 15 = 18$$

$$\Delta_2 = 33 - 21 = 12$$

$$\hat{X} = L_1 + \left\{ \frac{\Delta_1}{\Delta_1 + \Delta_2} \right\} C = 99 + \left(\frac{18}{12+18} \right) (101-99) = 100,2 \Omega$$

- d) Usando la formula empirica

$$\text{Moda} = \text{media} - 3 (\text{media} - \text{mediana})$$

si ottiene:

$$\text{Moda} = 99,92 \Omega$$

Si puo' notare che c'e' un buon accordo tra il valore modale calcolato per definizione e quello ottenuto usando la relazione empirica per distribuzioni moderatamente asimmetriche.

3. MISURE DI DISPERSIONE

Le sole misure di tendenza centrale usualmente non danno un'adeguata descrizione dei dati sperimentali. Infatti si dovrebbe prendere in considerazione anche la variabilita' dei dati. Per esempio se una persona mette la sua testa in un frigorifero ed i suoi piedi in un forno, allora mediamente si puo' dire che si senta bene. Ma in pratica la persona vorrebbe star meglio, a causa dell'estrema differenza tra le due temperature. Quindi un altro importante parametro dei dati sperimentali e' l'entita' della dispersione o il grado di compattezza dei dati. Nel seguito si esaminano alcune misure di dispersione dei dati.

3.1 Range

Il modo piu' semplice di rappresentare la dispersione e' il range, che non e' altro che la differenza tra i valore massimo

o cioè: $99 + \frac{21}{33} (101-99) = 100,273 \Omega$ - Utilizzando la definizione:

e minimo assunti dai dati. Il maggiore svantaggio associato al range come misura di dispersione e' il fatto che esso e' basato esclusivamente sulla dispersione dei valori estremi, mentre non fornisce informazioni sulla compattezza, o meno, dei valori compresi tra gli estremi stessi. Quindi e' azzardato considerare il range come misura di dispersione. Comunque si puo' utilizzare il range per avere una idea approssimativa circa l'estensione dei dati disponibili e della compattezza dei dati.

3.2 La deviazione media

La deviazione media e' un miglioramento sul range in quanto nella sua valutazione si tiene conto di tutti i valori delle misure effettuate.

In essa si considera il valore assoluto della deviazioni di ogni dato dalla media aritmetica, e viene effettuata la media. Quindi:

$$\text{Deviazione Media (DM)} = \frac{\sum_{i=1}^n |x_i - \bar{X}|}{n} = \frac{\sum_{i=1}^n |d_i|}{n} = \overline{|(x_i - \bar{X})|} \quad (2)$$

Dove d_i e' la deviazione dalla media della i -esima osservazione. Nell'equazione (2) $|x_i - \bar{X}|$ rappresenta il valore assoluto dello scostamento di x_i dalla media.

Se x_1, \dots, x_n occorrono con frequenze f_1, \dots, f_n , allora la deviazione media puo' essere scritta come:

$$\text{DM} = \frac{\sum_{i=1}^n f_i |x_i - \bar{X}|}{\sum_{i=1}^n f_i} = \frac{\sum_{i=1}^n f_i |d_i|}{N} = \frac{\sum_{i=1}^n f_i |(x_i - \bar{X})|}{N} = \overline{f_i (x_i - \bar{X})} \quad (3)$$

dove $\sum_{i=1}^n f_i = N$, numero totale di osservazione effettuate.

L'equazione (3) e' anche utile per dati raggruppati, dove x_i rappresenta l'identificativo della classe e f_i la frequenza della classe corrispondente.

Occasionalmente la DM puo' anche essere definita in termini di deviazioni assolute dalla mediana o altre medie. Una proprieta' interessante a tale proposito e' che la $\sum_{i=1}^n |(x_i - a)|$ assume un valore minimo per a pari alla mediana.

Si puo' notare che sarebbe piu' appropriato usare la terminologia "deviazione media assoluta" piuttosto che "deviazione media." In effetti si puo' dimostrare che la media delle deviazioni reali dai valori medi e' zero:

$$\frac{\sum_{i=1}^n d_i}{n} = \frac{\sum_{i=1}^n (x_i - \bar{X})}{n} = \frac{\sum_{i=1}^n x_i}{n} - \frac{n\bar{X}}{n} = \bar{X} - \bar{X} = 0 \quad (4)$$

Nel calcolo della DM si riscontrano alcune deviazioni positive, le rimanenti negative. Possiamo notare nell'equazione (4) che esse si bilanciano l'un l'altra intorno alla media. Pertanto per ogni deviazione deve essere calcolato il valore assoluto che e' piuttosto scomodo ai fini pratici. Percio' la deviazione media e' meno comunemente impiegata come misura di dispersione dei dati, rispetto ai parametri che verranno introdotti nel seguito.

3.3 Varianza

Ora invece di considerare la media delle deviazioni assolute intorno al valore medio, determineremo la deviazione media quadratica intorno al valore medio. L'operazione di elevazione al quadrato evita il problema inerente ai valori assoluti, poiche' tutti i quadrati sono in segno positivi. La deviazione media quadratica anche definita come varianza e' molto usata come misura della variabilita' dei dati. Infatti, i suoi vantaggi sono:

- 1) impiega tutti i valori dei dati ed e' sensibile al cambiamento di uno qualunque di essi;
- 2) e' indipendente dalla tendenza centrale, cioe' il valor medio, perche' usa le deviazioni dal valore medio;
- 3) il suo computo e' ragionevolmente semplice perche' i valori quadrati di ogni deviazione dalla media sono positivi.

La varianza e' generalmente indicata con il simbolo σ^2 ed e' cosi' calcolata:

$$\sigma^2 \triangleq \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2$$

svolgendo il quadrato:

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

Quindi potremmo esprimere la varianza come differenza tra la media dei valori al quadrato meno il quadrato del valore medio dell'insieme dei dati. $= (\bar{X}^2) - (\bar{X})^2$

3.4 Deviazione standard

Molto spesso viene utilizzata la radice quadrata della varianza, denotata con σ e chiamata deviazione standard o radice quadrata dello scarto quadratico medio. Il vantaggio principale della deviazione standard rispetto alla varianza e' che la sua unita' di misura e' la stessa della quantita' misurata. Essa e' definita come:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n d_i^2}{n}} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n}} \quad (5)$$

Nell'eventualita' di dati raggruppati si potrebbe scrivere:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{X})^2}{n}}$$

Dove $\sum_{i=1}^n f_i = n$ e' il numero totale di osservazioni. Si puo' notare che per distribuzioni di frequenza non eccessivamente asimmetriche, la relazione empirica tra la deviazione media e la deviazione standard e' data da:

$$\text{Deviazione media} = \frac{4}{5} \cdot (\text{deviazione standard})$$

$$\text{i.e. } DM = \frac{4}{5} \sigma$$

3.5 Fattore di dispersione

Il fattore di dispersione e' un numero adimensionale che rappresenta il grado di dispersione dei dati rispetto al valore medio. E' ottenuto dividendo la deviazione standard per il valor medio.

$$C.V. = \frac{\sigma}{\bar{X}}$$

E' generalmente espresso in percentuale ed e' chiamato anche fattore di dispersione percentuale. Si noti che il fattore di dispersione cade in difetto quando \bar{X} e' prossimo a zero.

3.6 Deviazione standard corretta

La misura di dispersione σ , cade in difetto quando il numero di osservazioni e' piccolo. Per esempio quando si considera una sola osservazione di una grandezza fisica il valore di σ e' zero. Cio' implica che la misura non ha dispersione e conseguentemente non soggetta a nessun altro errore di misura. Al contrario il risultato e' indefinibile qualora si consideri una sola osservazione. Comunque l'affidabilita' di σ cresce al crescere del numero di osservazioni. Percio' il valore della deviazione quadratica media σ viene corretto in maniera tale da fornire una stima piu' realistica della precisione delle misure, anche per un numero ridotto di osservazioni. Cio' si ottiene dividendo la somma dei quadrati delle deviazioni dalla media per $(n-1)$ che rappresenta il grado di liberta', anzichè per n che e' il numero totale di osservazioni.

Il numero di gradi di liberta' si riferisce al numero di informazioni indipendenti originate dai dati disponibili. Esso puo' essere determinato usando il ragionamento induttivo che segue.

Un caso semplice costituito da un insieme di 2 misure fornisce solo un input utile rispetto alla dispersione valutata intorno alla popolazione media perche' il set di campioni non fornisce solo informazioni sulla dispersione ma viene anche utilizzato per valutare la variabile rispetto alla quale la dispersione e' misurata, vale a dire X . Così un insieme di 2 misure fornisce solo una osservazione indipendente riguardo la dispersione. Analogamente in un insieme di 10 misure noi possiamo valutare 10 deviazioni rispetto al valor medio. Di conseguenza si puo' dire che in realtà ci sono 9 deviazioni indipendenti e la decima potrebbe essere determinata dal fatto che la somma delle deviazioni e' uguale a 0. Quindi in generale un insieme di n misure fornisce $n-1$ osservazioni indipendenti con riferimento alla deviazione standard della popolazione. Infatti l'espressione generale per il grado di

libertà e' $(n-m)$ dove n e' il numero di osservazioni ed m e' il numero di costanti da valutare dai dati di partenza, allo scopo di generare valori dai quali le deviazioni devono essere calcolate. Nel caso in cui $m=1$ e cioè in cui solo una costante deve essere calcolata, la media aritmetica, la deviazione standard corretta (denotata con S) detta anche "unbiased" o la migliore stima della precisione di un apparato può essere scritta:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{(n-1)}} = \sqrt{\frac{\sum_{i=1}^n d_i^2}{(n-1)}} \quad (6)$$

Confrontando le equazioni (5) e (6) la relazione tra S e s può essere scritta

$$s = \sqrt{\frac{n\sigma}{(n-1)}}$$

La quantità $\sqrt{\frac{n}{n-1}}$ e' conosciuta come fattore di correzione di Bessel. L'equazione mostra che all'aumentare di n diminuisce la differenza tra S e s . Quindi l'aumentare del numero di campioni la deviazione standard diventa una stima di precisione più affidabile. In pratica per $n > 25$ il fattore di Bessel diviene tanto vicino all'unità da poter essere trascurato.

4. DEVIATION STANDARD DELLA MEDIA.

Il termine deviazione standard corretta o migliore stima di precisione dell'apparato fornisce un'informazione sugli errori casuali che vengono commessi in ogni misura. In altre parole noi possiamo dire che:

$$x = x_{\text{misurata}} \pm s$$

Analogamente, la precisione di un insieme di misure vorrebbe indicare come e' vicina la media delle misure effettuate al valore vero. In altre parole, l'errore standard intrinseco (denotato s_n da alcuni autori) o la migliore stima di incertezza U_n e' data da $\sigma(\bar{X}_n)$.

Quindi,

$$X = (\bar{X}_n)_{\text{misurata}} \pm \sigma(\bar{X}_n) = (\bar{X}_n)_{\text{misurata}} \pm U_n$$

4.1 Migliore stima dell'incertezza di un insieme di misure

Come prima e' stato spiegato, la migliore stima di incertezza rappresenta la estensione dell'errore casuale nei valori misurati. Essa viene determinata considerando m insiemi di misure di n campioni in una popolazione di $(n \times m)$ osservazioni. La media di un insieme di misure varrà dunque:

$$\bar{X}_i = \frac{\sum_{j=1}^n x_{ij}}{n}$$

dove i rappresenta l' i -esimo insieme di misure e j l'osservazione nell' i -esimo insieme.
Comunque la media delle medie anche conosciuta come media della popolazione puo' essere definita come:

$$\bar{\bar{X}} = \frac{\sum_i^m \bar{X}_i}{m} \approx \frac{m\bar{X}}{m} \approx \bar{X}$$

dove $\bar{X} \approx \bar{X}_1 \dots \approx \bar{X}_m = \text{costanti}$

Se la deviazione in un insieme e' denotata con d_{ij} mentre quella tra le medie dei diversi insiemi e' denotata con D_i , allora

$$\sigma^2(x) = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n d_{ij}^2$$

e

$$\sigma^2(\bar{X}_n) = \frac{1}{m} \sum_{i=1}^m D_i^2 \quad (7)$$

dove $\sigma^2(x)$ e $\sigma^2(\bar{X}_n)$ rappresentano rispettivamente le variazioni delle singole misure e quella delle medie dei gruppi di misure. Si ha: $D_i = (\bar{X}_i - \bar{X}) \approx \frac{1}{n} \sum_{j=1}^n x_{ij} - \frac{n\bar{X}}{n}$ ($\because \bar{X} \approx \bar{X}_i$)

$$\approx \frac{1}{n} \sum_{j=1}^n (x_{ij} - \bar{X}) \approx \frac{1}{n} \sum_{j=1}^n d_{ij}$$

(8)

Sostituendo il valore di D_i dall'equazione (8) nella (7) si ottiene:

$$\sigma^2(\bar{X}_n) = \frac{1}{m} \sum_{i=1}^m \left\{ \frac{1}{n} \sum_{j=1}^n d_{ij} \right\}^2 = \frac{1}{mn^2} \sum_{i=1}^m \left\{ \sum_{j=1}^n d_{ij} \right\}^2 \quad (9)$$

La doppia somma dell'equazione (9) contiene due differenti tipi di termini. Ci sono alcuni termini in cui d_{ij} e' elevato al quadrato mentre altri contengono il prodotto di due differenti d_{ij} . Poiche' c'e' una eguale probabilita' che i termini siano positivi e negativi, al limite, quando si considera un elevato numero di osservazioni $n \times m$, la somma dei termini con il prodotto di due differenti d_{ij} tende a zero.

$$\sum_{i=1}^m \left\{ \sum_{j=1}^n d_{ij} \right\}^2 \approx \sum_{i=1}^m \sum_{j=1}^n d_{ij}^2$$

La (9) diviene dunque

$$\sigma^2(\bar{X}_n) = \frac{1}{mn^2} \sum_{i=1}^m \sum_{j=1}^n d_{ij}^2 = \frac{1}{n} \left\{ \frac{1}{m \cdot n} \sum_{i=1}^m \sum_{j=1}^n d_{ij}^2 \right\} = \frac{\sigma^2(x)}{n}$$

5. VALUTAZIONE DELLA MEDIA E DELLA DEVIAZIONE STANDARD MEDIANTE CODIFICHE INTERMEDIE

5.1 Metodo del cambio dell'origine dei dati

La procedura per operare mediante il cambiamento dell'origine dei dati e' anche chiamata "metodo della media assunta". Talvolta intuitivamente noi utilizziamo questo metodo di calcolo.

Si considerino, ad esempio 4 osservazioni:

100,20 - 100,30 - 100,15 - 100,25
 e si desidera determinare la media. Invece di fare il totale 400,80, quindi dividere per 4 ottenendo 100,20, e' piu facile fare la media di 0,20 - 0,30 - 0,15 - 0,25, ottenendo 0,20, quindi sommare 100. In effetti stiamo operando su valori X' dove $X' = X - 100$; in altre parole abbiamo spostato l'origine dei dati ad $X_A = 100$.

In generale, noi possiamo scrivere i valori osservati nella seguente forma

$$X_i = X'_i + X_A \tag{10}$$

dove X_A e' la media assunta (pari ad un valore costante). La media valutata come nell'equazione precedente puo' essere ottenuta facendo la somma e dividendo per il numero totale di osservazioni:

$$\frac{\sum X_i}{n} = \frac{1}{n} \sum (X'_i + X_A) \quad \text{oppure} \quad \bar{X}_n = \bar{X}'_n + X_A \tag{11}$$

dove \bar{X}'_n e' il valore medio dei valori ottenuti spostando l'origine dei dati. Similmente la varianza puo' essere ottenuta attraverso la sostituzione dei valori di x_i ed \bar{X}_n dalle equazioni (10) e (11) nella seguente equazione:

$$\sigma^2(x) = \frac{1}{n} \sum (x_i - \bar{X}_n)^2 = \frac{1}{n} \sum \{(x'_i + X_A) - (\bar{X}'_n + X_A)\}^2 = \frac{1}{n} \sum (x'_i - \bar{X}'_n)^2 = \sigma^2(x')$$

ovvero; $\sigma(x) = \sigma(x')$

5.2 Metodo del cambiamento dell'ordine di grandezza dei dati

Talvolta, un cambio nell'ordine di grandezza facilita il calcolo riducendo il numero di digits richiesto per le manipolazioni aritmetiche.

Ad esempio i valori di 4 osservazioni sono 400, 600, 200, 500 e si vuole determinare la loro media. Si potrebbero trascurare gli zeri e calcolare la media di 4,6 2 e 5 che e' 4,25, dopo di cio' si dovrebbero rimettere gli zeri e presentare la media come 425. In questo esempio e' stato eseguito un cambiamento dell'ordine di grandezza dei dati originali attraverso un appropriato fattore di moltiplicazione che nel caso suindicato e' 1/100. In tal modo possiamo scrivere i valori assegnati nella seguente forma generale

$$X_i = a \cdot X'_i,$$

dove a e' un appropriato fattore di moltiplicazione. La media e la deviazione standard possono essere trovate in maniera simile a quella utilizzata nel paragrafo precedente, ed i risultati sono i seguenti:

$$\bar{X}_n = a \cdot \bar{X}'_n \quad \sigma_n(x) = a \cdot \sigma_n(x')$$

6. LA DISTRIBUZIONE NORMALE

6.1 Introduzione

In generale agli strumenti di misura sono associati un numero di fattori che sono causa di errori casuali (random errors). Comunque, l'ampiezza degli errori, presi individualmente, e' usualmente piccola. Ora, se le misure fatte sono molte, i dati mostrano una distribuzione continua; questa potrebbe essere facilmente rappresentata mediante un istogramma normalizzato che presenta la frequenza relativa per ogni intervallo come ordinate, ed i valori misurati come ascissa. Se l'ampiezza dell'intervallo e' piccola e le ordinate dei vari punti medi delle classi sono congiunte da una curva regolare, la distribuzione risultante e' chiamata limite della distribuzione di frequenza. Tale distribuzione e' comunemente denominata "distribuzione normale o gaussiana". Tale distribuzione e' rappresentata matematicamente dalla funzione :

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\bar{X})^2}{2\sigma^2}\right\}$$

dove: $p(x)$ e' la funzione densita' di probabilita' che, per un dato intervallo, rappresenta la frequenza relativa di occorrenza del valore misurato in quell'intervallo;

σ e' la deviazione standard dei valori misurati;

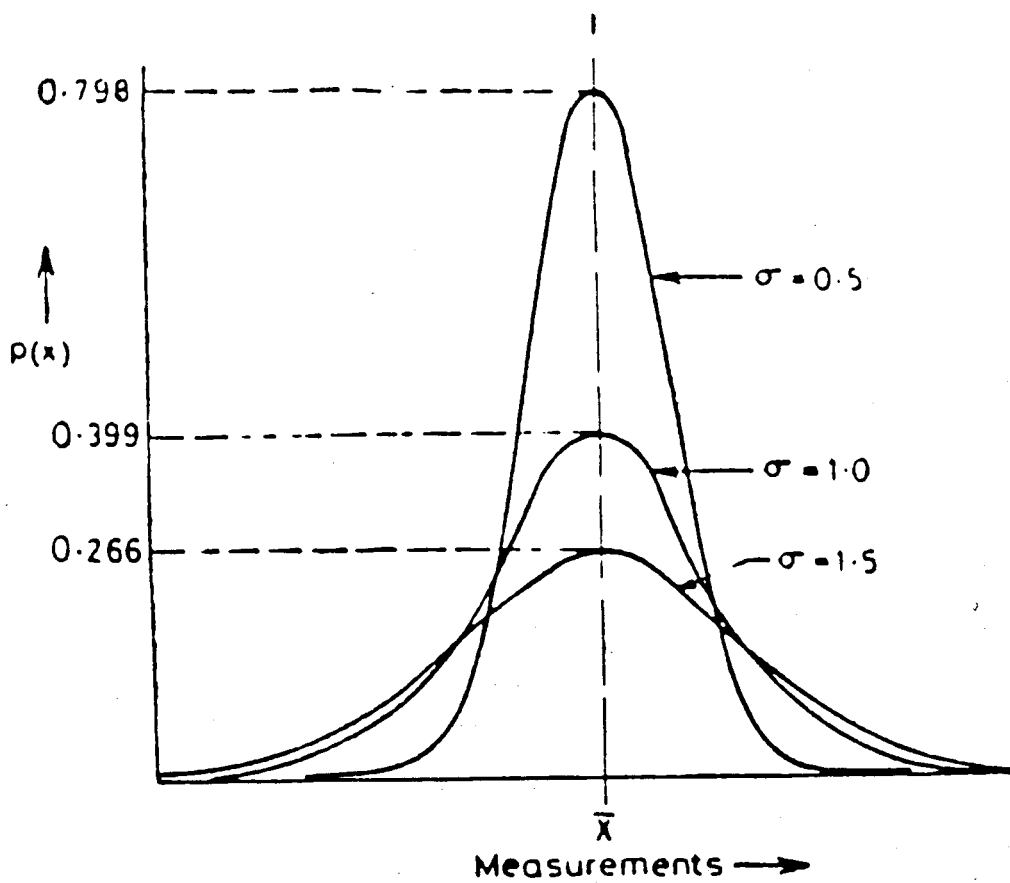
\bar{X} e' la media dei valori misurati.

Tale modello di distribuzione e' utilizzato nel processo decisionale per valutare le probabilita' che un valore misurato cada in un determinato intervallo. In alternativa, se il livello di probabilita' (o livello di confidenza) e' preassegnato, essa consente di valutare la dispersione permessa dei dati intorno al valore medio. Inoltre le proprietA di tale distribuzione sono utilizzate per confrontare diversi gruppi di misure attraverso criteri statistici noti come test di significativita'. E' possibile infine determinare la "bonta'" o l'idoneita' dei valori misurati attraverso il confronto con i valori attesi dalla distribuzione normale. Questo viene fatto utilizzando il test del χ^2 (chi quadro). Tale criterio e' anche applicabile per valutare se una distribuzione qualunque, non normale, si adatta o meno a qualche altra distribuzione teoricamente conosciuta.

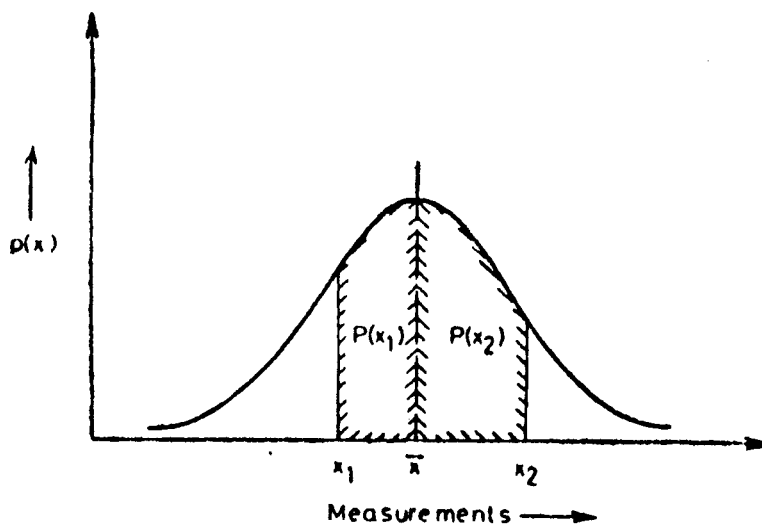
6.2 ProprietA della distribuzione normale

Ogni tipica curva gaussiana ha le seguenti caratteristiche (fig.6):

- 1) essa ha un massimo per $x=\bar{X}$ (cioe in corrispondenza del valore medio);



- FIG. 6 -



- FIG. 7 -

- 2) i punti di flesso sono a $x = \bar{x} \pm \sigma$;
- 3) a causa della sua simmetria, la mediana e' uguale alla media. Inoltre, giacche' la media capita sul valore di picco della curva, essa rappresenta anche la moda;
- 4) l'asse delle ascisse e' un asintoto per la curva;
- 5) l'area sottesa alla curva e' unitaria, cioe':

$$\int_{-\infty}^{+\infty} P(x) dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \exp\left\{-\frac{(x-\bar{x})^2}{2\sigma^2}\right\} dx = 1$$

- 6) Per uno stesso valore di media, la curva assume un picco appuntito per piccoli valori di σ ed un picco appiattito per elevati valori di σ . Se σ e' piccolo, cio' significa che c'e' una piccola dispersione tra i dati e, di conseguenza, si hanno piu' valori concentrati intorno al valor medio e la curva assume un massimo piu' elevato. D'altra parte il massimo e' inversamente proporzionale a σ :

$$\text{MAX}(P(x)) = \frac{1}{\sqrt{2\pi}\sigma} = \frac{0,399}{\sigma}$$

- 7) La probabilita' che la media cada tra x_1 e x_2 e' uguale all'area della curva gaussiana compresa tra le ascisse x_1 ed x_2 , come mostrato in fig. ; tale valore e' noto col nome di integrale gaussiano di probabilita' nell'intervallo (x_1, x_2) e si indica con $|P(x)|_{x_1}^{x_2}$. Svolgendo questo integrale si ottiene

$$|P(x)|_{x_1}^{x_2} = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\bar{x})^2}{2\sigma^2}\right\} dx = |P(x)|_{x_1}^{\bar{x}} + |P(x)|_{\bar{x}}^{x_2} = P(x_1) + P(x_2)$$

dove $P(x_1)$ e $P(x_2)$ sono gli integrali gaussiani tra x_1 e \bar{x} e tra \bar{x} ed x_2 rispettivamente.

6.3 Area sottesa dalla curva normale

L'area sottesa alla curva di distribuzione normale tra i limiti $-\infty, +\infty$ e' l'integrale gaussiano della probabilita' con cui i valori misurati cadono nel suddetto intervallo. Di conseguenza, l'integrale gaussiano dovrebbe essere unitario poiche' tutti i possibili valori misurati cadono tra $-\infty$ e $+\infty$. Per ottenere questo risultato, si integra l'equazione della distribuzione gaussiana, come segue:

l'area elementare dA della curva e' $dA = p(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\bar{x})^2}{2\sigma^2}\right\} dx$

integrando tra $(-\infty, +\infty)$ otteniamo l'area sottesa alla curva:

$$[A]_{\text{norm}} = \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\bar{x})^2}{2\sigma^2}\right\} dx$$

operando la sostituzione $z=(x-\bar{X})/\sigma$ (da cui $dz=dx/\sigma$) si ha

$$[A]_{\text{norm}} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{z^2}{2}\right) dz$$

Il termine $\int_{-\infty}^{+\infty} \exp(-z^2/2) dz$ e' conosciuto come funzione normale di errore ed il suo valore, calcolato attraverso integrazione grafica oppure in maniera differente, e' di $\sqrt{2\pi}$. Sostituendo il valore della funzione d'errore cosi' ottenuta nell'equazione precedente, abbiamo come risultato che l'area sottesa alla curva normale e' unitaria.

6.4 Distribuzione normale standardizzata

Allo scopo di ridurre differenti distribuzioni normali in una forma generale, e' usuale standardizzarle spostando l'origine delle coordinate della media e scegliere una scala sull'asse delle ascisse in funzione di σ .

Una conveniente variabile e' $z=(x-\bar{X})/\sigma$, detta variabile normale standardizzata.

La probabilita' di occorrenza nell'intervallo $[x_1, x_2]$ ricavabile dall'equazione della distribuzione normale e':

$$P(x) \{x_1 \leq x \leq x_2\} = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\bar{X})^2}{2\sigma^2}\right\} dx \quad (12)$$

Questa equazione puo' essere trasformata secondo il cambiamento di variabili proposto ottenendo:

$$z \triangleq \frac{x-\bar{X}}{\sigma}; \quad z_1 = \frac{x_1-\bar{X}}{\sigma}; \quad z_2 = \frac{x_2-\bar{X}}{\sigma}; \quad ; \quad dz = \frac{dx}{\sigma}$$

da cui

$$P(z) \{z_1 \leq z \leq z_2\} = \int_{z_1}^{z_2} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz \quad (13)$$

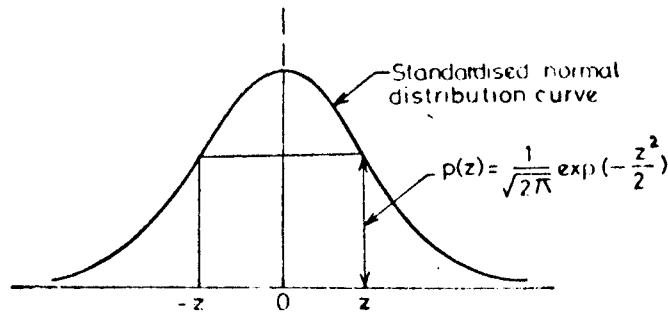
Tale formula e' denominata equazione normale standardizzata. Confrontando le (12) e (13) e' ovvio che ogni qualvolta e' richiesta la probabilita' nell'intervallo $[x_1, x_2]$, e' richiesto il calcolo dell'integrale in (12). Quest'ultimo e' piuttosto complesso comparato all'espressione (13) in quanto nel secondo caso, l'uso di valori tabulati della funzione di errore, cioe' $\int \exp(-z^2/2) dz$ semplifica notevolmente i calcoli. In secondo luogo, la forma della curva (12) dipende dai valori di \bar{X} e σ ed e' differente da caso a caso. Al contrario l'equazione (13) mostra che tutte le distribuzioni normali in x , con qualsiasi valore di \bar{X} e σ , si riducono alla stessa forma di distribuzione normale standardizzata in termini di variabile normale standardizzata. Essa e' caratterizzata dall'aver valore medio uguale a zero e deviazione standard uguale ad 1.

L'integrale gaussiano di probabilita' $P(z) = \int p(z) dz$ e' dunque tabulato rispetto ai valori di Z (vedi tab. I e tab II).

Table 18.1 Normal Probability Density Function $p(z)$

$$p(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$$

Each entry in the table indicates normal probability density function $p(z)$ corresponding to $\pm z$



To illustrate: the ordinate $p(z)$ of the standardised normal distribution curve corresponding to $z = \pm 1.0$ is 0.2420

$\pm z$	0.00	0.02	0.04	0.06	0.08
0.0	0.3989	0.3989	0.3986	0.3982	0.3977
0.1	0.3970	0.3961	0.3951	0.3939	0.3925
0.2	0.3910	0.3894	0.3876	0.3857	0.3836
0.3	0.3814	0.3790	0.3765	0.3739	0.3712
0.4	0.3683	0.3658	0.3621	0.3589	0.3555
0.5	0.3521	0.3485	0.3448	0.3410	0.3372
0.6	0.3332	0.3292	0.3251	0.3209	0.3166
0.7	0.3123	0.3079	0.3034	0.2989	0.2943
0.8	0.2897	0.2850	0.2803	0.2756	0.2709
0.9	0.2661	0.2613	0.2565	0.2516	0.2468
1.0	0.2420	0.2371	0.2323	0.2275	0.2227
1.1	0.2179	0.2131	0.2083	0.2036	0.1989
1.2	0.1942	0.1895	0.1849	0.1804	0.1758
1.3	0.1714	0.1669	0.1626	0.1582	0.1539
1.4	0.1497	0.1456	0.1415	0.1374	0.1334
1.5	0.1295	0.1257	0.1219	0.1182	0.1145
1.6	0.1109	0.1074	0.1040	0.1006	0.0973
1.7	0.0940	0.0909	0.0878	0.0848	0.0818
1.8	0.0790	0.0761	0.0734	0.0707	0.0681
1.9	0.0656	0.0632	0.0608	0.0584	0.0562
2.0	0.0540	0.0519	0.0498	0.0478	0.0459
2.1	0.0440	0.0422	0.0404	0.0387	0.0371
2.2	0.0355	0.0339	0.0325	0.0310	0.0297
2.3	0.0283	0.0270	0.0258	0.0246	0.0235
2.4	0.0224	0.0213	0.0203	0.0194	0.0184
2.5	0.0175	0.0167	0.0158	0.0151	0.0143
2.6	0.0136	0.0129	0.0122	0.0116	0.0110
2.7	0.0104	0.0099	0.0093	0.0088	0.0084
2.8	0.0079	0.0075	0.0071	0.0067	0.0063
2.9	0.0060	0.0056	0.0053	0.0050	0.0047
3.0	0.0044	0.0042	0.0039	0.0037	0.0035

E' da notare che mentre si usano le tavole per valutare le probabilita' riferite alle varie distribuzioni normali, e' consigliabile disegnare l'area cui corrisponde la probabilita' richiesta. Inoltre, bisogna stare attenti nell'uso delle tabelle, in quanto non tutte danno gli stessi valori di area. Alcune tabelle danno l'area tra 0 e z, altre la danno tra -z e z, altre ancora da z a ∞ o da $-\infty$ a z.

6.5 Livelli di confidenza

Si e' gia' detto che l'area sottesa alla distribuzione normale tra $(-\infty, +\infty)$ e' unitaria. In realta' dovrebbe essere cosi' in quanto la probabilita' che tutti i possibili valori misurati cadano tra $(-\infty, +\infty)$ deve essere unitaria. Nella pratica, si specifica generalmente un certo range di valori ammessi intorno al valore medio e si determina la probabilita' con cui il valore misurato cade in tale range. Tale probabilita' puo' essere valutata calcolando l'area sottesa dalla curva normale di distribuzione in quell'intervallo prefissato. Quando questa probabilita' viene espressa in percentuale essa prende il nome di livello di confidenza.

Ad esempio, si vuole valutare qual'e' la probabilita' che un valore misurato cada nell'intervallo $[X-\sigma, X+\sigma]$; usando la tabella si trova che questa probabilita' e' uguale a 0,6826. Dunque si puo' dire che le probabilita' che un valore cada nell'intervallo prefissato siano di 2:1, (in realta' 68,26: (100-68,26)). Si puo' dire che il livello di confidenza in tal caso e' del 68,26%.

Generalmente sia le tabelle sia i livelli di confidenza per casi diversi possono essere determinati. Alcuni valori tipici di livelli di confidenza sono indicati in figura 18.6 (pag. 451). Usualmente, in ogni esperimento individuuiamo un insieme di livelli di confidenza all'interno dei quali ci aspettiamo che cadano i valori misurati con una certa probabilita' (pag. 451); per fare questo dobbiamo decidere le probabilita' di errore che siamo disposti ad accettare. La percentuale di probabilita' di errore e' definita come uguale a 100-livello di confidenza.

Di solito, accettiamo errori del 5%, cioe' 95% e' il livello di confidenza per il quale $z=1.96$ (dalla tabella di pag. 449) considerando entrambi i lati del diagramma e $z=1,645$ considerando solo un lato. Comunque, in ogni luogo in cui la vita umana si e' sviluppata, si considera come bassa probabilita' di errore, quella dell'ordine del 1%; questo assicura un livello di confidenza del 99% per il quale $z=2,326$ per un solo lato e $z=2,575$ per entrambi i lati.

Nella pratica, possiamo incontrare due tipi di problemi: problema diretto nel quale i limiti di X sono assegnati ed e' richiesta la determinazione della probabilita' di occorrenza in quel dato campo; l'altro e' un problema di tipo inverso nel quale e' assegnata la probabilita' (il livello di confidenza assegnato e' usualmente il 95%) e sono richiesti i limiti di X.

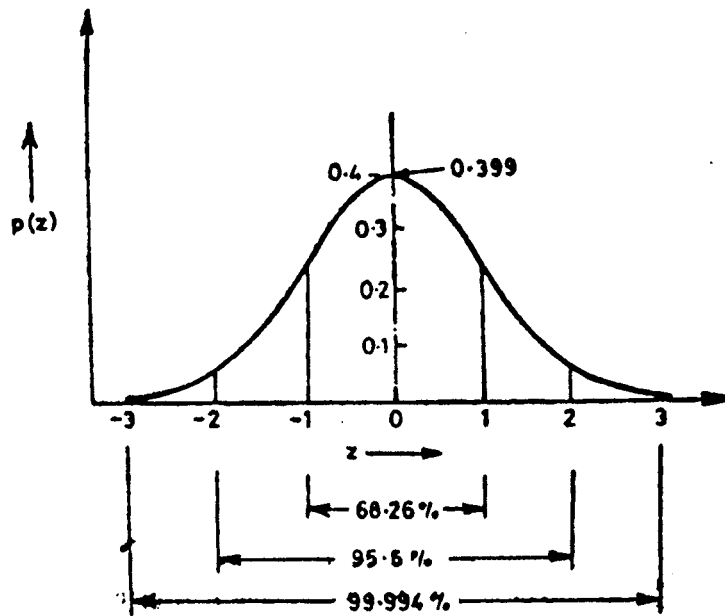


Fig. 18.6 Confidence levels for different ranges of measured values

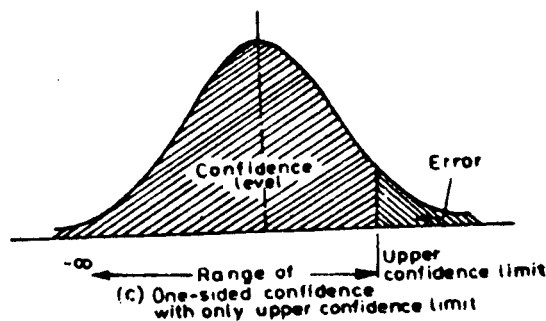
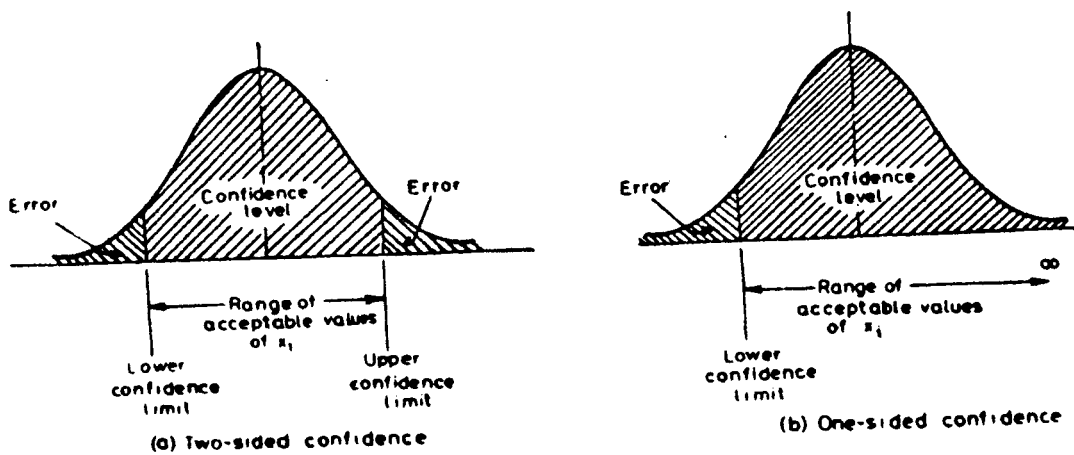


Fig. 18.7 Pictorial representation of Gaussian distribution with given confidence limits

TEOREMA DEL LIMITE CENTRALE.

Il teorema del limite centrale e' molto importante nel campo statistico. Esso asserisce che il campione di un popolazione segue l'andamento gaussiano, qualunque sia la distribuzione delle misure individuali. La variabile normale standardizzata e' definita come

$$Z = \frac{\bar{X} - \bar{X}}{\sigma_m}$$

dove \bar{X} e' il valore medio di ogni campione, \bar{X} e' la media della popolazione e σ_m e' deviazione media standard della media che e' circa uguale alla stima interna di incertezza (oppure errore standard interno) σ , quando il numero di osservazioni e' abbastanza alto.

Puo' essere notato che la distribuzione normale delle misure di un campione ha minor precisione se comparata alla corrispondente distribuzione normale dal campione medio della popolazione. In altre parole, l'ultima distribuzione ha un picco piu' appuntito rispetto alla prima considerata. E cio' avviene a causa della relazione tra $\sigma_m = \frac{\sigma}{\sqrt{n}}$

Nella maggior parte dei casi siamo interessati all'analisi dell'errore standard interno dei dati e nell'impiego della variabile standardizzata Z invece che alla media della popolazione ed alla deviazione standard delle medie. Comunque, in alcuni problemi, siamo interessati alla deviazione standard del solo campione J ; in casi del genere il valore di Z e' calcolato usando la media e la deviazione standard dal campione.

TEST DI SIGNIFICATIVITA'

Se la distribuzione normale di una particolare grandezza e' nota e tale grandezza e' misurata di nuovo sotto condizione diverse da quella precedente, il valor medio sar  probabilmente diverso dalla distribuzione originale. Se la differenza delle medie e' piccola, sarebbe ragionevole assumere che la distribuzione e' proveniente dalla stessa popolazione.

Invece, se la differenza e' considerevole, allora sarebbe ragionevole assumere che le circostanze diverse in cui e' stato effettuato il secondo rilevamento hanno alterato i valori ed il risultato e' da prendere in considerazione. In altre parole, possiamo dire che i dati originali ed i dati susseguenti non sono derivati dalla stessa popolazione.

Per testare se i due rilevamenti hanno riportato una differenza significativa, noi usiamo un test di significativita' basato sulla differenza delle medie. Il criterio statistico di tali medie e' quello secondo cui se la differenza dei valori medi dei due campioni devia di 1,96 volte l'errore standard interno della differenza delle medie, allora il cambio dei dati e' significativo.

Diciamo che un particolare campione e' individuato da n_1 valori ed ha \bar{X}_1 come valor medio e σ_1 come deviazione standard

..; analogamente, un altro campione rilevato sotto condizioni diverse ha n_2 valori, ed \bar{x}_2 come media e deviazione standard σ_2 . Si ricorda che l'errore standard interno (o stima interna di incertezza) di un campione di n valori con deviazione standard σ , e' dato da

$$U = \frac{\sigma}{\sqrt{n-1}}$$

Tenendo questo a mente si procede per determinare se il cambio di dati e' significativo o meno:

- 1) facciamo la differenza tra i valori medi che da' il campo di variazione di essi, cioè'

$$R(\bar{x}) = \bar{x}_1 - \bar{x}_2$$

- 2) determiniamo l'errore standard medio di $R(\bar{x})$ cioè':

$$\{(U^2)\}_{R(\bar{x})} = \left[\frac{\partial R(\bar{x})}{\partial \bar{x}_1}\right]^2 \cdot U_1^2 + \left[\frac{\partial R(\bar{x})}{\partial \bar{x}_2}\right]^2 \cdot U_2^2 = U_1^2 + U_2^2$$

Sostituendo i valori di U_1 ed U_2 in termini delle corrispondenti deviazioni standard dei campioni, abbiamo:

$$(U)_{R(\bar{x})} = \sqrt{U_1^2 + U_2^2} = \left[\left\{ \frac{\sigma_1}{\sqrt{n_1-1}} \right\}^2 + \left\{ \frac{\sigma_2}{\sqrt{n_2-2}} \right\}^2 \right]^{1/2}$$

- 3) Noti i valori di $R(\bar{x})$ e $(U)_{R(\bar{x})}$ dalle equazioni precedenti possiamo adesso impiegare il test di significatività che e' il seguente:

$$R(\bar{x}) > |1.96 \cdot (U)_{R(\bar{x})}|$$

allora la differenza delle medie e' significativa, altrimenti possiamo assumere che i dati originari e quelli seguenti sono appartenenti alla stessa popolazione.

In altre parole possiamo dire che sopra il 95% di livello di confidenza, le medie dei due campioni non sono della stessa popolazione se il test di significatività e' positivo.

Al contrario se la differenza delle medie $R(\bar{x})$ e' minore di 1,96 volte l'errore standard medio allora il risultato non e' significativo e possiamo dire che la probabilità di errore e' minore del 5% considerando entrambi i campioni appartenenti alla stessa popolazione. Invece se $R(\bar{x})$ si allontana rispetto a 2,58 volte l'errore combinato interno standard $(U)_{R(\bar{x})}$ allora il risultato e' altamente significativo ed in questo caso il livello di confidenza considerato e' del 99%.

TEST CHI-QUADRATO PER LA BONTA' E L'ADATTABILITA'

Quando e' stato fatto un certo insieme di misure, si ritiene che esse sono un campione di una distribuzione teorica nota. Per confrontare le differenti parti della distribuzione

osservata, noi suddividiamo i dati in n classi e determiniamo le frequenze osservate in ogni classe; quindi stimiamo le frequenze ottenute di ogni classe per assumere la distribuzione adottata alla particolare distribuzione teorica. Ad esempio, se la distribuzione assunta e' quella gaussiana allora la seguente procedura puo' essere adottata per calcolare i valori attesi di frequenza per un certo insieme di dati:

- 1) calcolare media e dati standard dei dati;
- 2) per ogni intervallo calcolare z_u e z_l cioe' il piu' basso ed il piu' alto valor limite rispettivamente;
- 3) dalla tabella dell'integrale gaussiano determinare la probabilita' tra Q e z_u e tra Q e z_l ;
- 4) la differenza dei suddetti valori di probabilita' da' l'integrale gaussiano nel dato intervallo se entrambi i limiti cadono tra Q ed ∞ oppure tra $-\infty$ e zero; La somma di questi valori da' l'integrale gaussiano di probabilita' se il piu' alto limite cade tra Q ed ∞ ed il piu' basso cade tra $-\infty$ e zero, e viceversa;
- 5) Moltiplicando l'integrale gaussiano di probabilita' in un dato intervallo per il numero totale di osservazione, si ottiene la frequenza attesa il quel dato intervallo;
- 6) la somma delle frequenze attese di tutte le classi non e' talvolta uguale al numero totale di osservazioni; cio' e' causato dall'arrotondamento nelle interpolazioni eseguite sulla tabella.

Quindi, le frequenze attese negli intervalli sono moltiplicate per un conveniente fattore di correzione in modo tale che la somma della frequenza eguaglia il numero di osservazioni. Dopo la determinazione delle frequenze attese nelle varie classi, determiniamo il χ^2 (si pronuncia chi-quadr) parametro come segue: diciamo che vi sono n classi e le frequenze attese ed osservate sono rispettivamente:

$$f_{e1}, f_{e2}, \dots, f_{en} \quad ; \quad f_{o1}, f_{o2}, \dots, f_{on}$$

Adesso, il nostro scopo sara' quello di determinare se le frequenze osservate e quelle attese sono tali per concludere se esse provengano o meno dalla stessa distribuzione di probabilita'.

→ vedi pag. 462-463

CRITERIO PER LA BONTA' E L'ADATTABILITA'

Il criterio statistico per la bonta' e l'adattabilita', cioe' come un insieme di dati osservatisi adatti ad una certa distribuzione teorica, sono i seguenti:

- 1) se il valore di probabilita' nel test χ^2 cade tra 0.1 e 0.9 allora la distribuzione osservata e' considerata come proveniente dalla distribuzione teorica assunta. In alcuni casi, il piu' basso valore della probabilita' χ^2 (definito come livello di significativita' o semplicemente livello) puo' essere ridotta a 0,05.
- 2) Se il valore di probabilita' nel test χ^2 e' al di sotto del piu' basso limite prescritto, allora, il risultato e' significativo ed il campione di dati e' considerato completamente differente dalla distribuzione assunta.
- 3) Se il valore del parametro χ^2 e' vicino allo zero o molto piu' piccolo, allora la probabilita' puo' oltrepassare il limite superiore di 0.9. Simili casi sono affrontati nella pratica; e se e' cosi', allora normalmente consideriamo i dati "buoni" ma con una certa diffidenza.

18.9 CHI-SQUARE TEST FOR GOODNESS OF FIT

When a set of measurements is obtained, it is believed that the measurements are a sample of some known theoretical distribution; say normal frequency distribution which is generally hypothesised in cases involving experimental statistics. For comparing the different parts of the observed distribution, we subdivide the data into a number of classes say n and determine the observed frequency in each class. Then we estimate the expected frequency of each class by assuming that the distribution conforms to a particular theoretical distribution.

For example, if the assumed distribution considered is Gaussian, then the following procedure may be adopted to calculate the expected values of frequencies for a given set of data:

1. Calculate the mean value and standard deviation of the data.
2. For each class interval, calculate the standard normal variates z_u and z_l for the upper and lower boundary values respectively.
3. From the integral Gaussian table, determine the integral Gaussian probability between 0 and z_u and 0 and z_l .
4. The difference in the above values gives the integral Gaussian probability in the given interval if both the upper and lower boundaries lie either between 0 and ∞ or 0 and $-\infty$. The sum of these values gives the integral Gaussian probability if the upper boundary lies between 0 and ∞ and the lower boundary lies between 0 and $-\infty$ and vice versa.
5. Multiplying the integral Gaussian probability in a given class interval by the total number of observations gives the expected frequency of occurrence of the variable in that interval.
6. The summation of expected frequencies in all classes sometimes does not equal the total number of observations. The slight difference is caused by small rounding-off errors due to interpolations in the integral Gaussian table. Therefore, the expected frequencies in step (5) are multiplied by a suitable correction factor so as to make the sum of expected frequencies equal to the number of observations.

After determining the expected frequencies in the various classes, we determine the χ^2 (pronounced chi-square) parameter as follows:

Let us say that there are n classes ($n > 1$) and the expected and observed frequencies in the various classes are denoted by

$$f_{e_1}, f_{e_2}, f_{e_3}, \dots, f_{e_n} \quad \text{and} \quad f_{o_1}, f_{o_2}, f_{o_3}, \dots, f_{o_n}$$

Now, our aim is to determine whether the observed frequencies and the expected frequencies are close enough for us to conclude that they come from the same probability distribution. To do so, we define the χ^2 parameter as:

$$\chi^2_{(n-m)df} = \sum_{i=1}^n \left\{ \frac{(f_{o_i} - f_{e_i})^2}{f_{e_i}} \right\} \quad (18.29)$$

where n is the number of values that are summed up to produce the values of χ^2
 m is the number of constants used in the calculation of expected frequencies
 $n - m$ is the degrees of freedom and
 subscript df stands for the degrees of freedom

The values of the numerator in the χ^2 (chi square) expression represent the squares of deviations between the expected and observed frequencies in various classes which is always positive. These values are normalised in each class by dividing them by the respective expected frequency of each class. It may be noted that the same order of deviation in the expected and observed frequencies causes relatively larger contribution in the χ^2 -parameter at the tail portions of the normally distributed data, as compared to the values close to the mean value of the data. This is because of relatively large values of the expected frequencies near the mean value of the data which are in the denominator of the χ^2 -parameter. In order to restrict the unusually large contributions in χ^2 parameter when the expected frequencies are small, the empirical criterion commonly used in practice is to regroup the various classes in such a way so that expected frequency in each class is not less than 5.0.

Further, a correction is sometimes applied to the chi-square values when the degree of freedom F , i.e. $(n - m)$ is of the order of 1.0. This is termed *Yate's* correction and accounts for the inaccuracies involved when the results of continuous distributions are applied to discrete data. The correction consists of writing Eq. (18.29) in the following form:

$$\chi_F^2 \approx 1.0 = \sum_{i=1}^n \left[\frac{\{ | (f_{o_i} - f_{e_i}) | - 0.5 \}^2}{f_{e_i}} \right] \quad (18.30)$$

If the sample distribution agrees with the assumed theoretical distribution then $\chi^2 = 0$. This is of course very unlikely because even if the sample is taken from the parent distribution, one would not expect *exact* agreement in every interval. But, the larger the value of χ^2 , the more the disagreement between the assumed distribution and the observed values. In other words, in such a case, the smaller is the probability that the observed distribution matches the expected distribution. Thus, the chi-square parameter is quite useful in statistical analysis of data as it helps to test a particular hypothesis in the given data.

In applying the chi-square test, we first determine the value of χ^2 for the given data. Then, we determine the values of degrees of freedom F which is equal to $(n - m)$. Knowing the values of χ^2 and F , we determine the probability that the actual measurements match the expected distribution from either the chi-square tables (Table 18.3) or from the χ^2 - F diagram (Fig. 18.14) which gives cross-plots of chi-square probability, $P(\chi^2)$, for various values of χ^2 and F .

18.10 CRITERIA FOR GOODNESS OF FIT

The statistical criteria for the goodness of fit, i.e. how well a set of observed data fit the assumed theoretical distributions are as follows:

1. If the value of probability in the χ^2 -test lies between 0.1 and 0.9, then the observed distribution is considered to follow the assumed distribution. In other words, there is no reason to suspect the hypothesis. In certain cases, the lower limit of chi-square probability (also termed *significance level* or simply *level*) may be reduced to 0.05.
2. If the value of the probability in the χ^2 -test is below the lower prescribed limit, then the result is significant and the sample data is considered to be entirely different from the assumed distribution. In such cases, the value of the χ^2 -parameter is usually quite large.