Classificazione degli errori

Escludendo gli errori grossolani, derivanti, ad esempio, da macroscopiche sviste nell'applicazione della procedura o da un problema strumentale improvviso, gli errori che caratterizzano una misura sperimentale possono essere distinti in:

Errori casuali (random)

- ✓ derivano dalla naturale variabilità nell'esito di una misura, legata all'operatore, alla strumentazione e/o alla metodologia impiegata;
- ✓ influenzano la precisione di una misura, ossia la dispersione dei dati intorno al loro valore medio;
- ✓ non sono eliminabili ma possono essere ridotti;
- ✓ la loro entità può essere espressa mediante la deviazione standard.

Errori sistematici

- ✓ derivano da scostamenti costanti delle misure dal valore vero, dovuti all'operatore, alla strumentazione e/o alla metodologia impiegata;
- ✓ influenzano l'accuratezza della misura, ossia la differenza fra il valore misurato ed il valore vero;
- ✓ sono potenzialmente eliminabili, anche completamente, se riconosciuti;
- ✓ sono quantificabili dalla differenza fra il valore misurato ed il valore vero (ad esempio analizzando un campione certificato).

Media

Date n misure sperimentali replicate (ad esempio i volumi equivalenti determinati da n titolazioni indipendenti), x_1, x_2, …x_n, la media dei loro valori si definisce come:

$$\bar{x} = \sum_i \frac{x_i}{n}$$

Mediana

Date n misure sperimentali replicate, con n dispari, la mediana è il valore che si trova a metà della serie compresa fra il valore più piccolo e quello più grande determinati:

10.10, 10.20, 10.40, 10.46, 10.50, 10.54, 10.60, 10.80, 10.90

Se n è un numero pari, la mediana è la media della coppia di valori centrali:

10.10, 10.20, 10.40, 10.46, 10.50, 10.54, 10.60, 10.80, 10.90, 11.02

10.52

Deviazione standard

Misura la dispersione dei valori misurati intorno al valore medio, dovuta alla presenza di errori random.
Date n misure sperimentali replicate, x_1, x_2, ...x_n, la deviazione standard si definisce come:

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

Si utilizzano talvolta anche grandezze correlate alla deviazione standard:

la deviazione standard relativa: $RSD = \frac{s}{\bar{x}}$

la deviazione standard relativa percentuale: $RSD\% = RSD \times 100$ e

la varianza campionaria: $V = s^2$

Ripetibilità e riproducibilità

Entrambi i termini indicano quanto i valori ottenuti per una misura siano dispersi intorno al loro valore medio, tuttavia, secondo le direttive ISO (International Standard Organization):

- ✓ la ripetibilità corrisponde alla dispersione dei dati relativi ad uno stesso campione ottenuti nelle stesse condizioni (operatore, apparato strumentale, laboratorio) e in un breve lasso di tempo (ad esempio nel corso della stessa giornata);

- ✓ la riproducibilità corrisponde alla dispersione dei dati relativi allo stesso campione ma ottenuti da diversi laboratori oppure nello stesso laboratorio ma da diversi operatori o con apparati diversi o in tempi diversi (ad esempio in giorni diversi).

In generale la ripetibilità è migliore (ossia inferiore) della riproducibilità perché quest'ultima è influenzata, almeno potenzialmente, da un numero maggiore di fonti di variabilità rispetto alla ripetibilità.

I contributi all'errore complessivo su una misura

Supponendo di conoscere il valore vero della grandezza da misurare, detto T (*True*), l'errore complessivo associato alla generica determinazione x_i di quella grandezza si può esprimere come:

$$E_i = x_i - T$$

Sommando e sottraendo il valor medio delle determinazioni l'errore si può esprimere anche come:

$$E_i = (x_i - \overline{X}) + (\overline{X} - T)$$

il primo termine, $(x_i - \overline{X})$, rappresenta il contributo dovuto all'errore casuale

il secondo termine, $(\overline{X} - T)$, rappresenta il contributo dovuto all'errore sistematico (bias o distorsione)

L'errore sistematico può avere a sua volta diversi contributi, che possono essere valutati mediante test intra-laboratorio e inter-laboratorio, analizzando un campione a concentrazione nota (il valore T).

Ripartizione degli errori sistematici

In un singolo laboratorio

L'errore sistematico associato alla misura di un campione certificato in un singolo laboratorio comprende:

- l'errore sistematico dovuto al metodo di analisi (*method bias*), ad esempio l'incompleta formazione di un complesso di cui si misura poi l'assorbanza, laddove si supponga, su base teorica, che la sua formazione sia completa

- l'errore sistematico dovuto al laboratorio (*laboratory bias*), ad esempio un problema strumentale che porta a sottostimare l'assorbanza

errore sistematico intra-laboratorio =
method bias + laboratory bias

Confronto fra laboratori diversi

Quando un campione di concentrazione nota (campione certificato) viene analizzato da diversi laboratori con la medesima procedura (inter-laboratory test) è possibile separare il *method bias* dal *laboratory bias*.

Purché il numero di laboratori che partecipano al test sia sufficientemente elevato, è altamente probabile che i *laboratory bias* siano distribuiti in modo casuale, ossia che i loro valori positivi o negativi siano più o meno equivalenti, dunque che il loro valore medio sia prossimo a 0.

In una situazione di questo tipo l'eventuale persistenza di un errore sistematico indica la presenza di un *method bias*.

errore sistematico inter-laboratorio = *method bias*

Numero di cifre significative in una misura

Un dato sperimentale ha significato solo se accompagnato da una stima dell'errore associato alla misura con cui è stato determinato.

L'errore casuale può essere determinato, nell'approccio più semplice, effettuando un numero n di replicati della stessa misura e calcolando la deviazione standard secondo la formula abituale.

Quando non viene esplicitamente riportata l'incertezza, si può dedurre SOLTANTO quale/i cifra/e sia/siano affetta/e da incertezza, purché sia stata rispettata la convenzione delle cifre significative. Una di esse si esprime nel modo seguente:

le cifre significative di una misura sono tutte quelle note con certezza più la prima affetta da incertezza (ed anche la seconda, se la prima cifra affetta da incertezza è affetta da un errore pari ad 1).

Esempio

Il dato 153.7, così riportato, presuppone che la cifra 7 sia la prima affetta da incertezza (e che dalla cifra 3 verso sinistra non ci sia incertezza). Ad esempio l'errore sulla misura potrebbe essere 0.4.
Le cifre significative del dato sono dunque, per definizione, quattro.

Il ruolo della cifra 0 nell'espressione di una misura

- Lo zero che precede il punto decimale non è MAI una cifra significativa (ad esempio nel dato 0.715 le cifre significative sono 3)

- Lo zero compreso fra cifre significative è SEMPRE una cifra significativa (ad esempio nel dato 7.015 le cifre significative sono 4)

- Lo zero che segue delle cifre significative PUO' essere una cifra significativa. In particolare lo è sempre se segue il punto decimale (ad esempio nel dato espresso come 2.0)

Per evitare equivoci, specialmente nel caso di valori molto elevati, può essere utile impiegare la notazione scientifica, nella quale le cifre significative sono SOLTANTO quelle che precedono la potenza di 10.

Ad esempio il dato 1357000, così espresso, implica che l'ultima cifra significativa sia lo zero nella posizione delle unità. In notazione scientifica il numero andrebbe dunque scritto come 1.357000×10^6.

Se però i tre zeri finali non fossero significativi, insistendo l'errore già sulla cifra delle migliaia (7), il numero andrebbe scritto come 1.357×10^6.

Arrotondamento delle cifre associate ad una misura a partire dalle cifre significative associate all'errore che è stato determinato per essa

La procedura di arrotondamento di una misura dovrebbe prevedere i seguenti passaggi, in rigoroso ordine cronologico:

1) Calcolare l'errore random (come semplice deviazione standard o con calcoli più complessi, che verranno illustrati in seguito)

2') Conservare, del dato numerico relativo all'errore, soltanto una cifra significativa, se essa è maggiore di 1, e procedere all'arrotondamento secondo la regola abituale (arrotondamento verso l'alto se la cifra che la segue è ≥ 5 e verso il basso se essa è ≤ 4):

Ad esempio:

se l'errore è risultato pari a 0.7862348 si conserverà soltanto il 7 che segue lo 0 (lo 0 posto davanti agli altri numeri non è significativo) ma arrotondato verso l'alto perché il 7 è seguito da un 8: 0.8;

se l'errore è risultato pari a 0.7342348 si conserverà soltanto il 7 ma non modificato, perché il 7 è seguito da un 3: 0.7.

se l'errore è risultato pari a 7.862348 si conserverà soltanto 8

se l'errore è risultato pari a 7.342348 si conserverà soltanto 7

2'') Conservare, del dato numerico relativo all'errore, la prima cifra significativa e quella subito successiva, se la prima cifra risulta uguale ad 1. Per l'arrotondamento della seconda cifra significativa si procede come illustrato al punto 2'

Ad esempio:

se l'errore è risultato pari a 0.1862348 si conserverà 0.19

se l'errore è risultato pari a 0.1342348 si conserverà 0.13

se l'errore è risultato pari a 1.862348 si conserverà 1.9

se l'errore è risultato pari a 1.342348 si conserverà 1.3

3) Arrotondare il valore della misura (media) in modo che sia COERENTE, in termini di cifre, con le cifre significative presenti nel suo errore

Ad esempio, per una misura pari a 9.21567:

se l'errore è risultato pari a 2.67821, ossia 3, arrotondato alla prima cifra significativa (l'unica considerabile in questo caso, secondo la convenzione adottata), la misura dovrà essere espressa come 9 ± 3;

se l'errore è risultato pari a 2.17821, ossia 2, arrotondato, la misura dovrà essere espressa come 9 ± 2;

se l'errore è risultato pari a 0.267821, ossia 0.3, la misura dovrà essere espressa come 9.2 ± 0.3;

se l'errore è risultato pari a 0.167821, ossia 0.17, arrotondato, la misura dovrà essere espressa come 9.22 ± 0.17

se l'errore è risultato pari a 0.163821, ossia 0.16, la misura dovrà essere espressa come 9.22 ± 0.16

La coerenza fra misura ed errore implica che la misura sia arrotondata su una posizione identica a quella su cui è stato in precedenza arrotondato l'errore.

Arrotondamento delle cifre associate ad una misura quando l'errore corrispondente ha la sua prima cifra significativa su una posizione molto lontana dall'unità

Se l'errore risulta avere la sua prima cifra significativa su una posizione distante dall'unità (ad esempio sulle migliaia, sui millesimi, ecc.) occorre fare molta attenzione negli arrotondamenti, soprattutto quando l'entità della misura è molto diversa da quella dell'errore.

Ad esempio:

se l'errore è risultato pari a 0.0034782, il suo arrotondamento sarà 0.003, quindi si avranno i seguenti arrotondamenti:

1.90843219 ⇨ 1.908 ± 0.003

0.01908432 ⇨ 0.019 ± 0.003

290838.983265 ⇨ 290838.983 ± 0.003

0.00067356 ⇨ 0.001 ± 0.003, più facilmente esprimibile come $(1 \pm 3) \times 10^{-3}$

0.000067356 ⇨ 0.000 ± 0.003, più facilmente esprimibile come $(0 \pm 3) \times 10^{-3}$

La notazione scientifica diventa fondamentale quando errore e misura hanno entrambi valori molto superiori all'unità.

Ad esempio:

190843219 ± 238981 va arrotondato come: $(1908 \pm 2) \times 10^5$

190843219 ± 138981 va arrotondato come: $(1908.4 \pm 1.4) \times 10^5$

Particolare attenzione va prestata alla situazione, non rara in chimica analitica, in cui errore e misura siano sì entrambi superiori all'unità ma, in aggiunta, l'errore è superiore, anche di molto, rispetto alla misura:

1908341 ± 2389819 ⇨ $(2 \pm 2) \times 10^6$

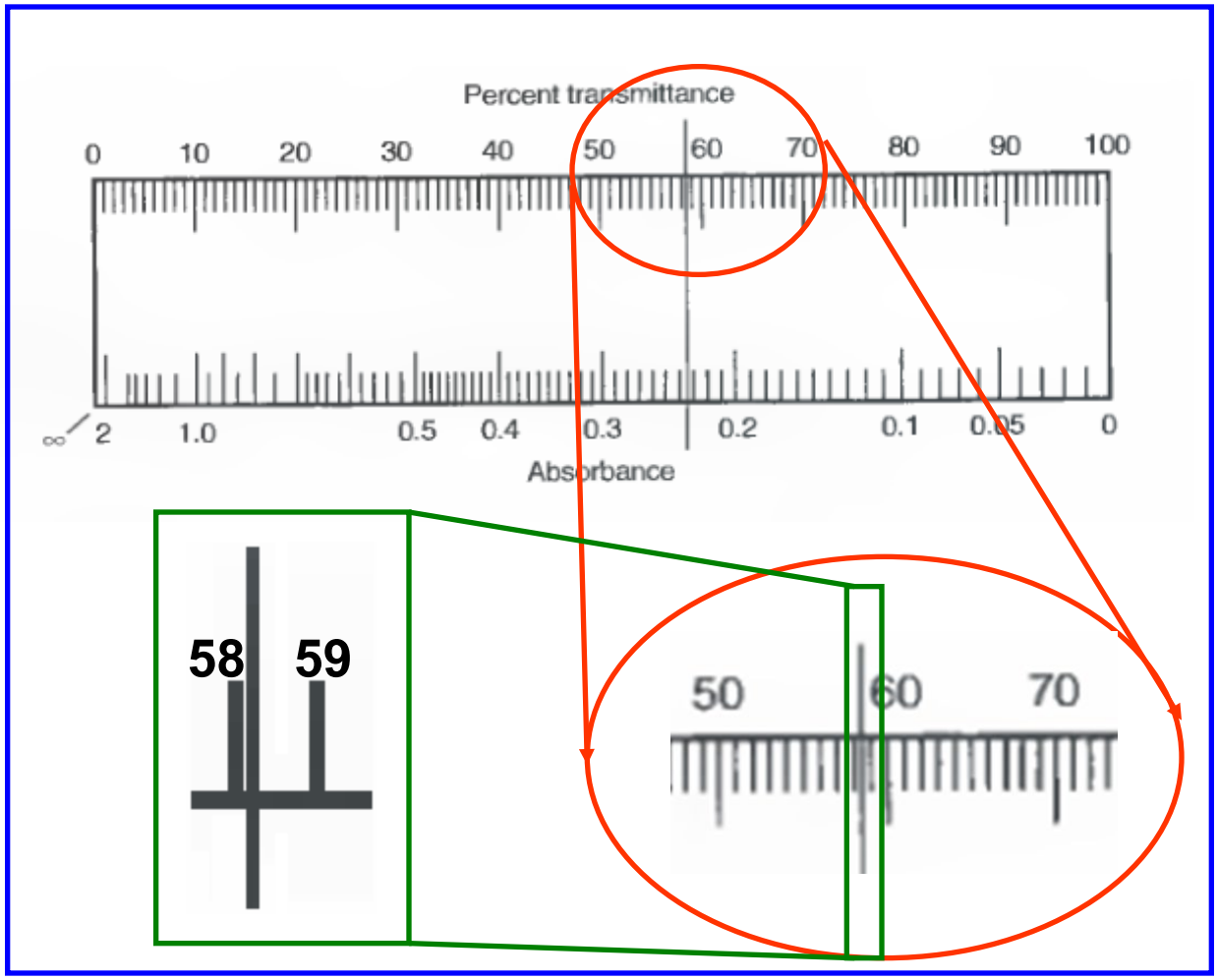
1908341 ± 23898191 ⇨ $(0.\cancel{19} \pm 2.\cancel{39}) \times 10^7$ ⇨ $(0 \pm 2) \times 10^7$

7908341 ± 13898191 ⇨ $(0.7\cancel{9} \pm 1.3\cancel{9}) \times 10^7$ ⇨ $(0.8 \pm 1.4) \times 10^7$

7908341 ± 138981911 ⇨ $(0.0\cancel{79} \pm 1.3\cancel{90}) \times 10^8$ ⇨ $(0.1 \pm 1.4) \times 10^8$

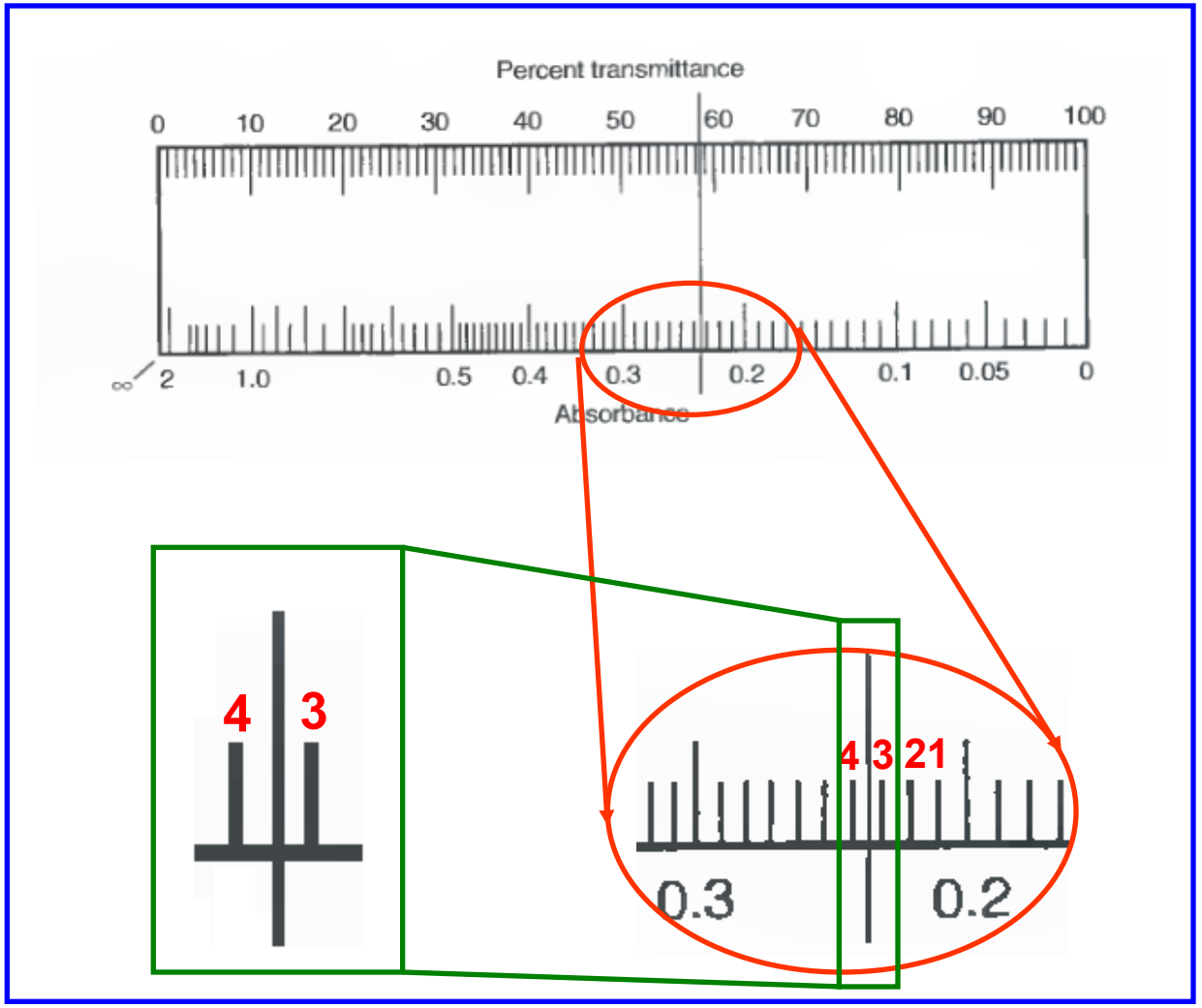
Arrotondamento nel caso della lettura derivante dalla posizione di una lancetta rispetto ad una scala graduata

Nel caso della lettura mostrata in figura, basata su una normale scala lineare (relativa alla trasmittanza percentuale, in questo caso), l'operatore è in grado di apprezzare certamente il valore 58 ma, di fatto, anche di stimare la corrispondente prima cifra decimale, che, a seconda di chi legge, potrà essere apprezzata come 1, 2 o 3, ragionevolmente.



La misura potrà essere dunque fornita come 58.1, 58.2 o 58.3 % e ciò implicherà che l'ultima cifra a destra, la prima decimale, sia affetta da un errore dell'ordine di 1, massimo 2 unità, realisticamente, quindi sarà espressa, ad es., come 58.2 ± 0.2 %.

Nel caso della lettura di assorbanza mostrata in figura, basata su una scala logaritmica, l'operatore è in grado di apprezzare certamente non solo il valore 0.23 ma, di fatto, anche la successiva cifra decimale, che sarà assegnata come 2, 3 o 4, a seconda di come viene stimata da chi effettua la lettura (si noti che fra 0.2 e 0.3 la scala è quasi lineare).

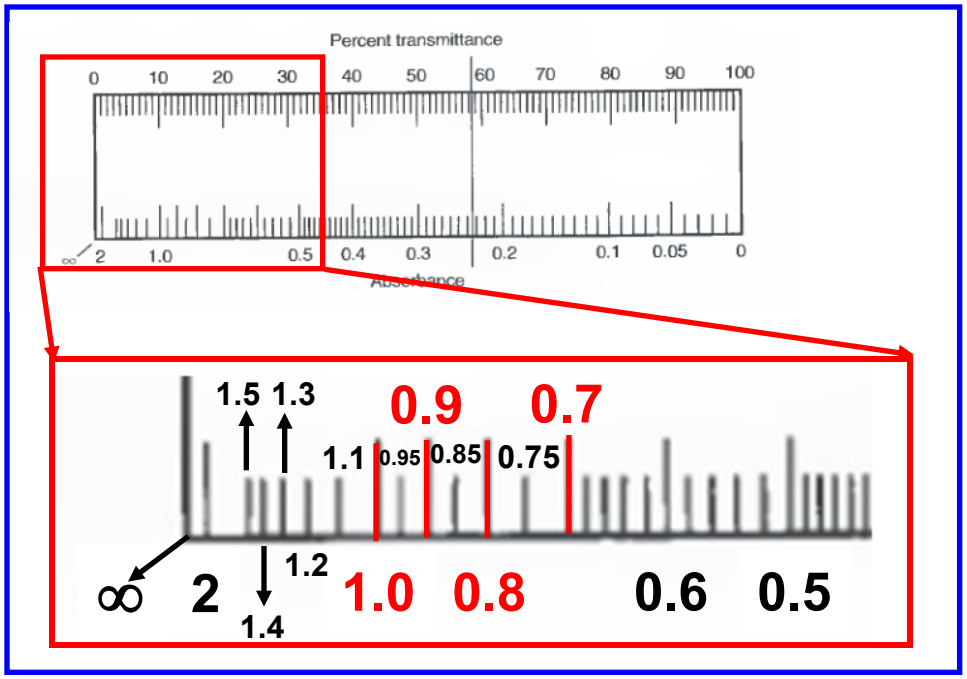


La misura potrà essere dunque fornita come 0.232, 0.233 o 0.234 e ciò implicherà che l'ultima cifra a destra, la terza decimale, sia affetta da un errore ragionevolmente pari a 1-2 unità, quindi, ad esempio, 0.233 ± 0.002.

In generale, quindi, si accetta che per un dato derivante dalla lettura di una scala graduata l'operatore possa attribuire come ultima cifra significativa un multiplo della quantità pari ad un decimo della più piccola spaziatura della scala. Nella valutazione più pessimistica il multiplo in questione sarà 5.

L'incertezza legata alla lettura contribuisce alla variabilità osservata sulla quantità misurata quando si replica più volte la misura.

Nel caso specifico va considerato con attenzione il fatto che la suddivisione della scala logaritmica diventi sempre più irregolare via via che il valore di assorbanza aumenta:



Fra le tacche 0.7 e 1.0 le spaziature interne corrispondono a 0.05 unità di assorbanza, invece che 0.01, e questo, associato alla non linearità delle spaziature, può portare a maggiore imprecisione, pertanto è consigliabile leggere la trasmittanza percentuale, per poi trasformarla in trasmittanza e, infine, in assorbanza.

Propagazione dell'errore

Quando una grandezza viene determinata combinando i valori di altre grandezze determinate sperimentalmente l'errore ad essa associato si può calcolare a partire da quelli che caratterizzano le varie grandezze misurate

Errori casuali

Se la grandezza da determinare y è una funzione di n grandezze, x_1, x_2, …, x_n, sperimentalmente determinate ed indipendenti, l'errore casuale (deviazione standard) ad essa associato $\mathbf{s_y}$ è dato dalla relazione:

$$s_y^2 = \sum_{i=1}^{n} \left(\frac{\partial y}{\partial x_i}\right)^2 s_{x_i}^2$$

dove s_{xi} è la deviazione standard associata alla grandezza x_i.

Se nel calcolo delle derivate parziali occorre introdurre uno o più valori per le grandezze x_i, si utilizzano i valori medi ottenuti per queste.

Errori sistematici

L'errore sistematico su una grandezza y determinata a partire dalle grandezze misurate x_1, x_2, …, x_n, fra loro independenti, è dato dall'equazione:

$$\Delta y = \sum_{i=1}^{n} \frac{\partial y}{\partial x_i} \Delta x_i$$

dove Δx_i sono i vari errori sistematici, considerati con il proprio segno.

A differenza dell'errore casuale l'errore sistematico finale può essere anche nullo, se i termini di segno positivo e negativo dell'espressione sopra indicata si compensano vicendevolmente.

Variabili random discrete e continue

Una variabile si definisce random se ad ogni suo valore può essere associata una certa probabilità.

In particolare una variabile random è:

discreta, se può assumere soltanto i valori di un sottoinsieme del campo reale, ad esempio soltanto numeri interi (come il numero delle particelle emesse da una sorgente radioattiva);

continua, se può assumere, potenzialmente, tutti i valori del campo reale (o quantomeno, quelli positivi, nel caso di molte grandezze fisiche), ad esempio la concentrazione, la temperatura, il volume, ecc.

Distribuzioni di frequenze

La probabilità che una variabile random assuma un certo valore può essere visualizzata attraverso la sua distribuzione di frequenza, che va costruita a partire da valori determinati sperimentalmente per quella variabile random. Se la variabile rappresenta il valore misurato in seguito ad un esperimento si procederà come segue:

- ✓ i dati ottenuti da n repliche dell'esperimento (*dati grezzi/raw data*) vengono prima ordinati in modo crescente o decrescente, costruendo una serie, caratterizzata da un campo di variazione, ossia la differenza fra il valore massimo e quello minimo;

- ✓ si individuano delle classi in cui raggruppare i dati, ossia intervalli di ampiezza costante che coprono l'intero campo di variazione;

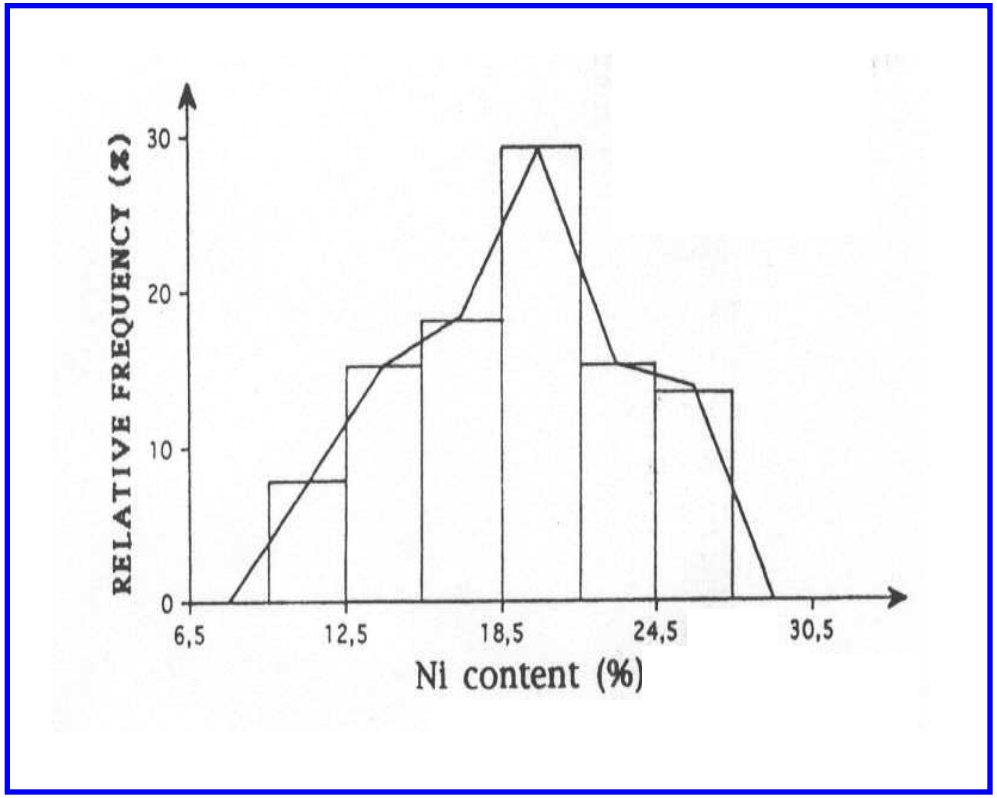
- ✓ per ogni classe si individua una frequenza assoluta, corrispondente al numero di dati che ricadono al suo interno;

- ✓ la distribuzione di frequenza può essere visualizzata con un'istogramma, un grafico costituito da rettangoli che hanno la base centrata sul valore centrale della classe e di lunghezza pari all'ampiezza della classe, mentre l'altezza corrisponde alla frequenza della classe (assoluta o anche relativa).

Esempio: istogramma delle frequenze per una serie di 65 dati corrispondenti al contenuto percentuale di nichel in una lega, misurato con arrotondamento alla prima cifra decimale.

Le frequenze di ogni classe possono essere espresse anche in termini relativi (ossia dividendo il numero di dati facenti parte della classe per quello totale), come mostrato in figura.

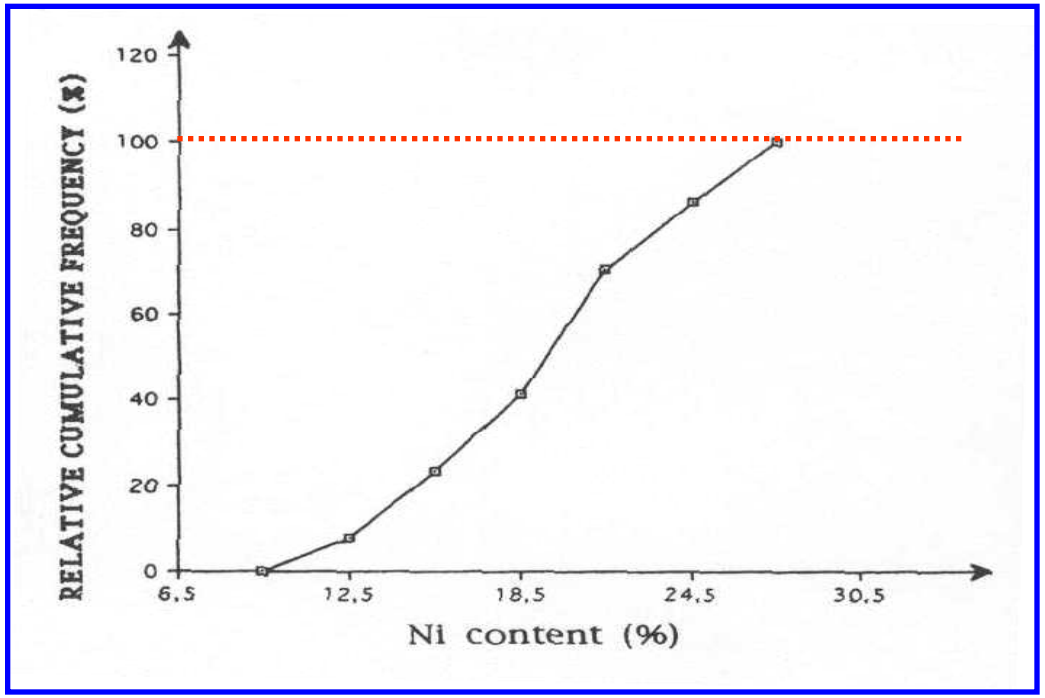
Il poligono di frequenza è la linea spezzata che collega i punti medi delle basi superiori di ogni rettangolo dell'istogramma.



Distribuzioni di frequenze cumulative: ogive

Data una certa classe della serie di dati, la somma delle frequenze di tutte le classi che la precedono e della frequenza della classe stessa si definisce frequenza cumulativa della classe.

Il poligono delle frequenze cumulative viene anche definito ogiva:



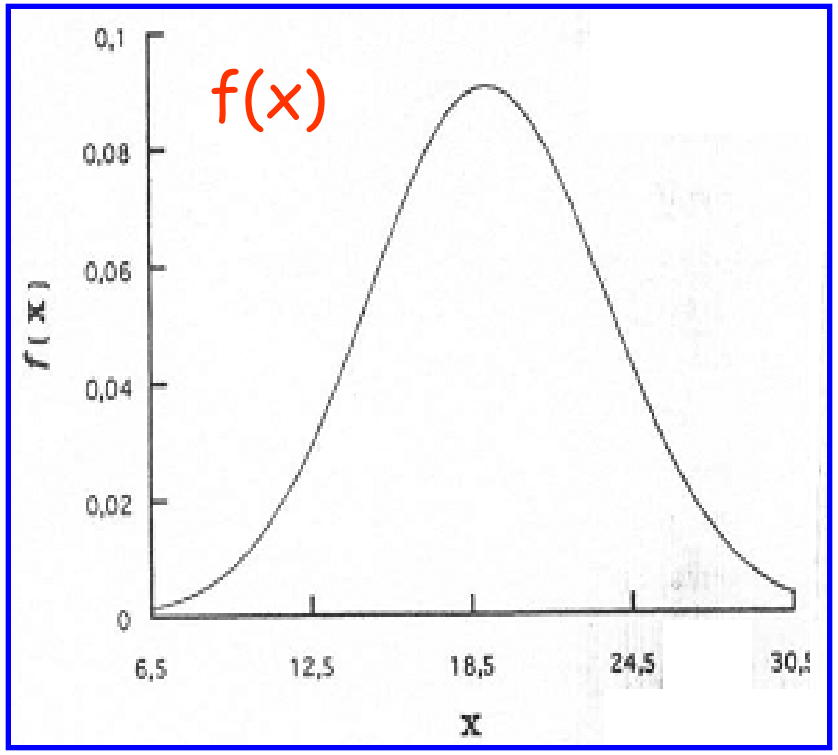
Densita' e distribuzione di probabilita'

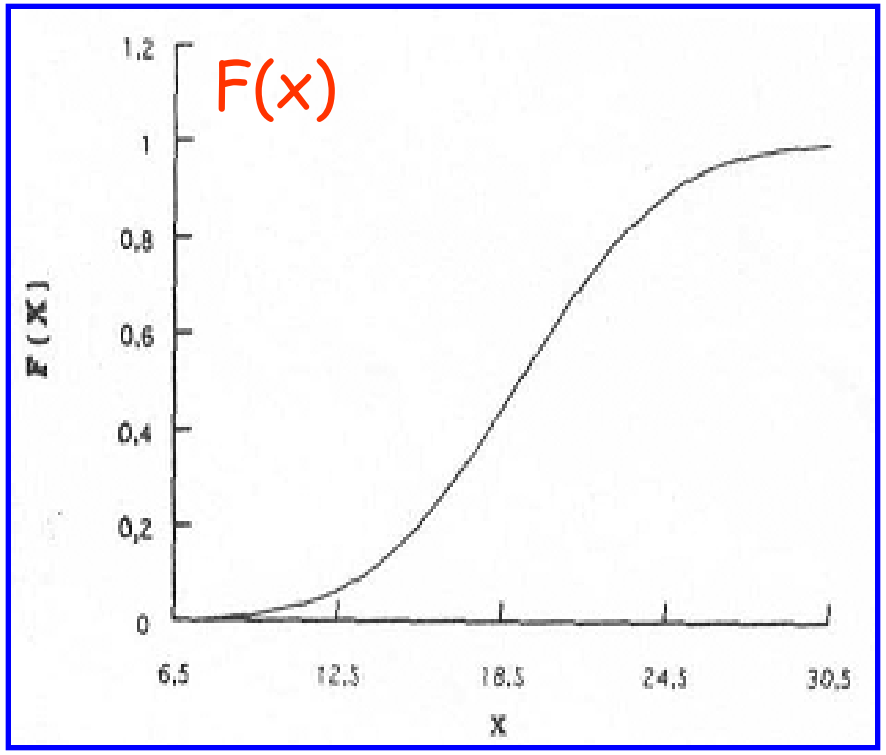
Se si immaginasse di aumentare notevolmente il numero dei dati raccolti e visualizzati in un istogramma di frequenze sarebbe possibile scegliere classi di ampiezza sempre più piccola e trovare comunque un certo numero di osservazioni comprese in ciascuna di esse.

L'istogramma delle frequenze diventerebbe costituito da rettangoli di base sempre più stretta ed il relativo poligono di frequenze una spezzata costituita da segmenti sempre più piccoli.

Per un numero infinito di replicati il poligono delle frequenze relative diventerebbe una curva; essa rappresenta la **funzione densità di probabilita' f(x)**.

La funzione derivante dal poligono delle frequenze cumulative, in presenza di un numero infinito di replicati, rappresenta invece la **distribuzione di probabilità F(x)**.

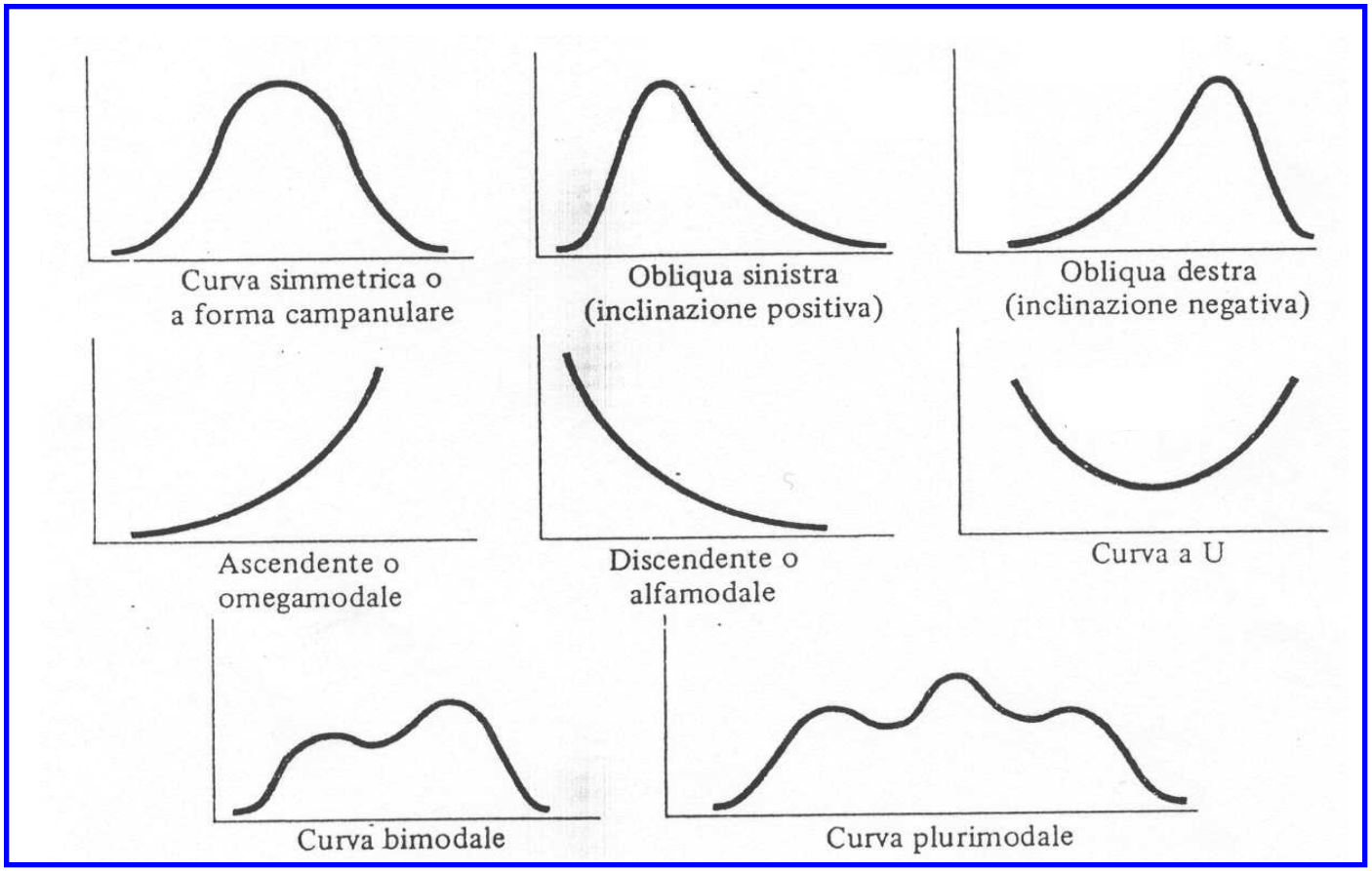




E' interessante notare che **f(x) = dF(x)/dx e**

$$\int_0^{\infty} f(x)dx = 1$$

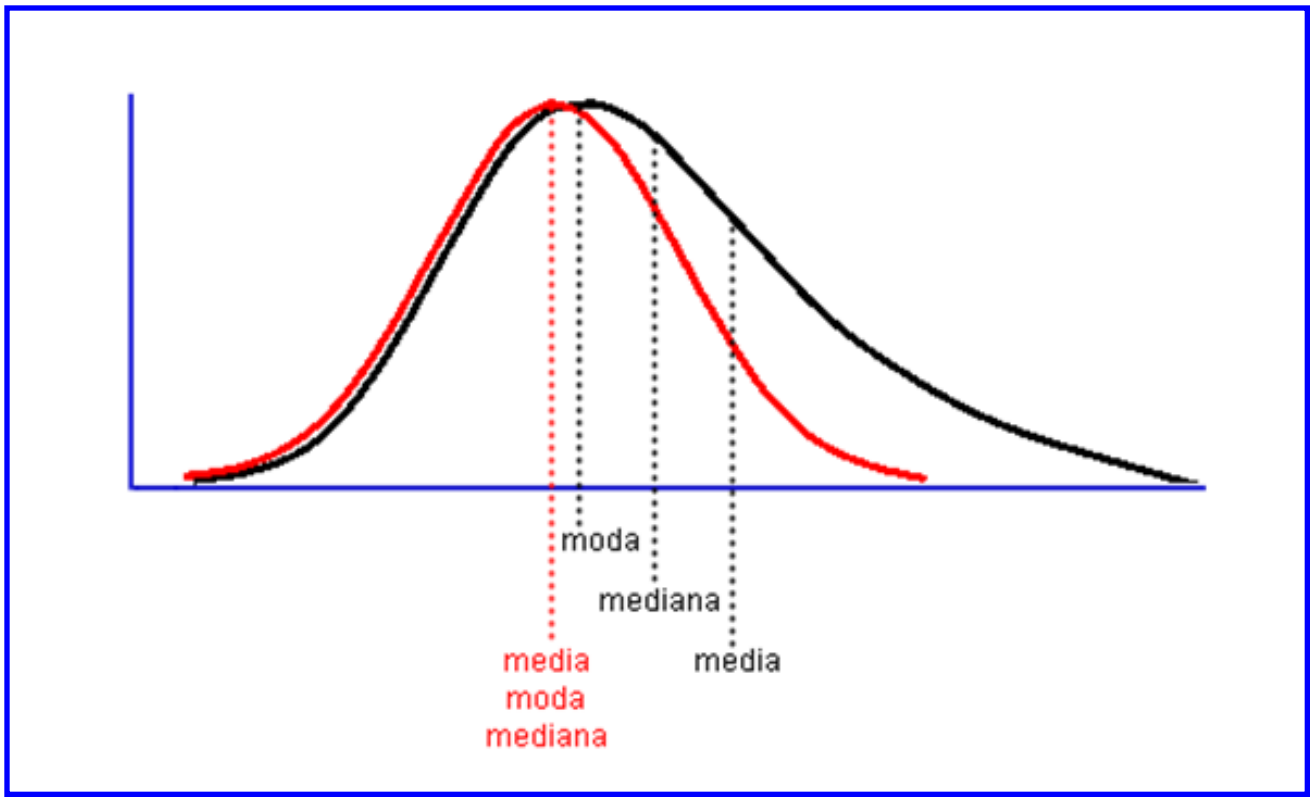
Funzioni di densità di probabilità: tipi più comuni



L'aggettivo **modale** deriva dal termine **moda**, che rappresenta il valore che si presenta con la più alta frequenza in una serie di dati, ossia il valore piu' comune.

Un esempio molto comune di funzione di densità di probabilità per dati sperimentali è la **curva gaussiana o normale**, che è una funzione simmetrica.

Nel caso di una funzione simmetrica media, mediana e moda coincidono, mentre in presenza di asimmetria la media è l'indice che risente maggiormente della distorsione, allontanandosi da moda e mediana:



La funzione *Speranza matematica (Expectation)*

Data una variabile random x, caratterizzata da una funzione di distribuzione di probabilita' f(x), la speranza matematica (o valore atteso) E di una generica funzione g della variabile random x è espressa dalla relazione:

$$E\{g(x)\} = \int_{-\infty}^{+\infty} g(x)f(x)dx$$

- Quando la funzione g(x) è del tipo x^r, la funzione E si definisce momento non centrale di ordine r della variabile x.

Per r = 1 si ottiene:

$$E\{x\} = \int_{-\infty}^{+\infty} xf(x)dx = \mu$$

μ si definisce media di popolazione della variabile x

❖ Quando la funzione $g(x)$ è del tipo $(x-\mu)^r$, la funzione E si definisce momento centrale di ordine r della variabile x.

Per r = 2 si ottiene:

$$E\{(x-\mu)^2\} = V(x) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)dx = \sigma^2$$

σ^2 e' la varianza di popolazione della variabile x

Differenza fra campione e popolazione in statistica

In statistica:

un campione di dimensioni n è un numero finito n di osservazioni ottenute per una variabile random;

la popolazione relativa alla stessa variabile random è rappresentata dal numero infinito di osservazioni che in teoria potrebbero essere effettuate su quella variabile.

L'introduzione della funzione E consente di stabilire un confronto fra media e varianza campionarie e i relativi parametri di popolazione:

	media	varianza
campione (dim. n)	$\overline{x} = \sum_{i=1}^{n} x_i / n$	$s^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 / (n-1)$
popolazione	$\mu = \int_{-\infty}^{+\infty} x f(x) dx$	$\sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$

E' possibile applicare alla media e alla varianza di un campione di dimensioni n, dette appunto media e varianza campionarie, la funzione E, tenendo conto di alcune sue proprietà generali, ossia:

$E(a) = a$ se a è una costante;

$E(ax + b) = aE(x) + b$, se a e b sono costanti e x è una variabile random

$E(x_1 + x_2 + \ldots + x_n) = E(x_1) + E(x_2) + \ldots E(x_n)$,

dove $x_1, x_2, \ldots, x_n$ sono valori della variabile random ottenuti dalla stessa popolazione.

Il valore atteso per la media campionaria è dunque dato da:

$$E(\bar{x}) = E\left(\sum_{i=1}^{n} x_i/n\right) = \frac{E(x_1) + E(x_2) + ... + E(x_n)}{n} = \frac{n\mu}{n} = \mu$$

Nel caso della varianza campionaria si ottiene:

$$E(s^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = E\left[\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^{\,2} - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2\right)\right]$$

$$= E\left[\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^{\,2} - 2n\bar{x}^2 + n\bar{x}^2\right)\right] = \frac{1}{n-1}\left(\sum_{i=1}^{n} E\left(x_i^{\,2}\right) - nE\left(\bar{x}^2\right)\right)$$

$$= \frac{n}{n-1}\left[E\left(x^2\right) - E\left(\bar{x}^2\right)\right]$$

Va sottolineato che poiché le varie x_i rappresentano, in questo caso, valori derivanti dalla stessa popolazione, quella della variabile x, si può scrivere, per ciascuna x_i, che $E(x_i^2) = E(x^2)$, quindi la sommatoria delle $E(x_i^2)$ corrisponde a $n\,E(x^2)$.

Le due speranze matematiche indicate in parentesi quadra sono correlate alle varianze associate alla variabile x stessa e alla variabile rappresentata dalla sua media campionaria, infatti risulta:

$$V(x) = E(x-\mu)^2 = E(x^2 - 2x\mu + \mu^2) = E(x^2) - 2\,\mu E(x) + \mu^2 = E(x^2) - 2\mu^2 + \mu^2$$

$$V(\bar{x}) = E(\bar{x}-\mu)^2 = E(\bar{x}^2 - 2\bar{x}\mu + \mu^2) = E(\bar{x}^2) - 2\,\mu E(\bar{x}) + \mu^2 = E(\bar{x}^2) - \mu^2$$

Sostituendo nell'ultima espressione ottenuta per la $E(s^2)$ si ottiene quindi:

$$\frac{n}{n-1}\left[E\left(x^2\right) - E\left(\bar{x}^2\right)\right] = \frac{n}{n-1}\left[V(x) - V(\bar{x})\right]$$

D'altro canto, considerando la definizione di media campionaria e le seguenti proprietà generali della varianza:

$V(ax + b) = a^2V(x)$ e $V(x + y) = V(x) + V(y)$

la seconda delle quali valida per x e y variabili random indipendenti,

si può scrivere:

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

e, in definitiva:

$$E(s^2) = \frac{n}{n-1}\left[V(x) - V(\bar{x})\right] = \frac{n}{n-1}\left[V(x) - \frac{1}{n}V(x)\right] = \frac{n}{n-1}\left[\frac{(n-1)\,V(x)}{n}\right] = \sigma^2$$

Si noti che l'uguaglianza finale vale soltanto in virtù della presenza del termine n-1 al denominatore dell'espressione della varianza campionaria.

Tale termine viene definito correzione di Bessel.

In termini statistici si dice che:

- ✓ la variabile $\overline{X}$ (media campionaria) rappresenta uno stimatore corretto (*unbiased*) della media di popolazione μ;

- ✓ la variabile s^2 (varianza campionaria) rappresenta uno stimatore corretto (*unbiased*) della varianza di popolazione σ^2.

Distribuzione campionaria della media

E' possibile applicare la funzione V definita in precedenza, ossia il momento centrale di ordine 2, alla media campionaria.

Ricordando le seguenti proprieta' della funzione V:

$V(ax + b) = a^2V(x)$ e

$V(x + y) = V(x) + V(y)$ se x e y sono variabili random **indipendenti**

si ottiene:

$$V(\overline{X}) = V\left(\sum_{i=1}^{n} x_i/n\right) = \frac{V(x_1) + V(x_2) + \ldots + V(x_n)}{n^2} = n\sigma^2/n^2 = \sigma^2/n$$

In definitiva:

se più serie di misure, ciascuna costituita da n replicati, vengono effettuate sulla stessa popolazione e si calcolano le corrispondenti medie, i valori di queste ultime hanno essi stessi una distribuzione che è centrata sul valore della media di popolazione μ ed ha una varianza σ^2/n, ossia una deviazione standard $\sigma/n^{1/2}$.

Tale distribuzione prende il nome di **distribuzione della media campionaria**.

Covarianza e coefficiente di correlazione

Covarianza

Quando si considerano due variabili random, x e y, le cui popolazioni siano caratterizzate dalle medie μ_x e μ_y, è possibile definire la funzione covarianza usando la speranza matematica:

$$C(x, y) = E\{(x - \mu_x)(y - \mu_y)\}$$

Dalla definizione si deduce che la varianza è la covarianza di una variabile random con se stessa.

La covarianza fra due variabili X e Y può essere calcolata a partire da N determinazioni delle due variabili, impiegando la formula:

$$C(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}[(x_i - \bar{x})(y_i - \bar{y})]$$

La covarianza può essere utilizzata per calcolare in forma più generale la varianza di una grandezza f che sia funzione di m variabili random:

$$V(f) = \sum_{i,\,j\,=1}^{m} \left[\frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} C(x_i, x_j) \right]$$

Si noti che nel caso in cui la covarianza fra tutte le possibili coppie di variabili sia nulla la formula diventa analoga a quella già considerata per la propagazione dell'errore random.

Coefficiente di correlazione

Fornisce una misura di quanto due variabili siano correlate fra di loro e si definisce come:

$$r(x,y) = \frac{C(x,y)}{\sqrt{\{V(x)\,V(y)\}}}$$

Il valore massimo che r puo' assumere è 1 (perfetta correlazione), mentre il valore minimo è –1 (perfetta anti-correlazione).