

Trattamento delle Osservazioni

Generalità

Scopo: descrizione di fenomeni;

Metodologia: elaborazione di modelli e dotazione di strumenti per verificare il grado di approssimazione di tale elaborazione.

MODELLI: relazioni matematiche fra grandezze, che descrivano e prevedano il fenomeno;

VERIFICA: mediante interrogazione della realtà fisica, cioè misurando grandezze.

Misura di grandezze

Non è immediato decidere se un valore attribuito ad una certa grandezza (lunghezza, qualità di un processo produttivo, intelligenza) si possa definire misura di quella grandezza. E' quindi ovvio dover affermare che il concetto di misura non può prescindere dalla considerazione delle caratteristiche dello strumento di misura, dalle sue interazioni con l'ambiente e dalla definizione di un modello per la grandezza stessa.

Incerteza di misura

Osservazione sperimentale che la ripetizione della misura di una medesima grandezza in talune condizioni porta a risultati diversi.

Esempio

Misura della lunghezza di una trave rettilinea, di diversi metri.

La variabilità dei risultati dipende almeno da due cause:

*a) il **riporto** dello zero dello strumento;*

*b) la **stima** della tacca della graduazione millimetrica a cui corrisponde l'estremo della trave.*

*In ogni ripetizione del processo i riporti e la stima sono soggetti a **fluttuazioni accidentali** che generano piccole variazioni nel valore finale stimato della lunghezza.*

In generale si nota che:

- *la discordanza fra le ripetizioni della misura cresce con il numero dei riporti;*
- *se la lunghezza della trave fosse inferiore al metro, la discordanza sarebbe al più di 1 millimetro.*

In conclusione, qualsiasi strumento/metodo di misura ha una propria incertezza, che si evidenzia quando le condizioni di misura (es. la necessità del riporto) introducono un rumore superiore alla sensibilità dello strumento e quando si usa uno strumento ai limiti della sua sensibilità (es. volendo stimare il decimo di millimetro con una riga millimetrata).

. **il risultato di un'operazione di misura è dato dall'associazione del valore numerico della grandezza misurata con la valutazione dell'incertezza con cui tale valore è stato ricavato.**

Oltre alle cause accidentali, esistono **cause sistematiche d'errore**, legate al modello impiegato per descrivere il fenomeno.

Approssimazione del modello

Descrivere matematicamente un fenomeno fisico comporta la definizione di un modello con un certo grado di semplificazione.

Il modello deve essere:

- a) **il più semplice possibile** (es.: non dipendente da troppi parametri);
- b) **complicato quanto necessario**, in relazione alla approssimazione che si richiede ai valori predetti dal modello stesso.

Nel modello sono presenti due componenti, quella funzionale e quella stocastica, strettamente connesse.

Componente funzionale: descrive analiticamente la relazione fra la grandezza osservabile e i parametri ad essa collegati.

Componente stocastica: è legata al complesso delle cause accidentali di variabilità del valore osservato non incluse esplicitamente nel modello funzionale, o perché sfuggono alla modellizzazione analitica, o perché troppo complesse per decidere di modellarle analiticamente.

Esempio

Misura della distanza piana L con una rotella metrica centimetrata lunga 50 m.

Sia $a = 5 \cdot 10^{-5} \text{ kg}^{-1}$ il coefficiente di allungamento del materiale della rotella a cui viene applicata una tensione di $F = 5 \text{ kg}$.

Sia $b = 10^{-5} \text{ }^\circ\text{C}^{-1}$ il coefficiente di dilatazione termica della rotella e la temperatura ambiente pari a $\Delta T = 20^\circ\text{C}$.

Le variazioni di lunghezza corrispondenti saranno:

$$\Delta L_1 = a \times F \times L = 5 \cdot 10^{-5} \times 5 \times 50 = 1.25 \text{ cm}$$

$$\Delta L_2 = b \times \Delta T \times L = 10^{-5} \times 20 \times 50 = 1.00 \text{ cm}$$

E' evidente che se voglio misurare con un'incertezza dell'ordine del cm devo correggere i valori misurati Loss della quantità ΔL .

$$\Delta L = \Delta L_1 + \Delta L_2 \Rightarrow \text{Loss} = L (1 - a \times F - b \times \Delta T) \quad \text{Modello Funzionale}$$

Il modello funzionale contiene in questo caso due effetti sistematici, lineari nei parametri F e ΔT , attraverso i coefficienti a e b . Se opero in condizioni ambientali stabili, la caratteristica di questi errori è che posso prevederne l'entità, perché li ho descritti analiticamente. Se non lo faccio, la loro presenza denota una inadeguatezza del modello.

Se eseguo più misure in condizioni ambientali instabili (variazione "accidentale" della forza applicata e della temperatura) ottengo una maggiore dispersione dei risultati e le due cause d'errore assumono, se non corrette, un comportamento di tipo accidentale. **La distinzione tra errore sistematico e errore accidentale, quindi, pur netta concettualmente, non sempre è univoca in pratica.**

I fenomeni aleatori

Sono detti tali gli eventi il cui esito non è possibile prevedere a priori (lancio di un dado, estrazione di una carta, misura di una lunghezza). Per quanto incapaci di prevederne con esattezza il risultato, siamo però in grado di evidenziare delle regolarità, di descrivere un comportamento "in media", di assegnare delle probabilità agli eventi.

Ne deriva un approccio di tipo probabilistico, in cui le oscillazioni dei valori osservati sono rappresentabili come estrazioni di una variabile casuale.

La descrizione e l'interpretazione dei fenomeni aleatori sono oggetto di studio della **teoria della probabilità** e della **statistica**.

Teoria della probabilità

Essenzialmente deduttiva, insegna a costruire la probabilità di eventi complessi a partire da un modello stocastico noto.

Statistica

Di tipo induttivo, si occupa di ricostruire un modello stocastico a partire da eventi già realizzati. Si articola in: *Teoria della stima* (la ricerca della miglior strategia di interrogazione della realtà per estrarre informazioni sul fenomeno) e *Inferenza* (la verifica di ipotesi sul modello interpretativo sulla base di dati estratti dal fenomeno).

Variabili aleatorie

Quando si associa ad ogni punto dello spazio campione un valore numerico: lo spazio campione - diventa l'insieme dei numeri e prende il nome di variabile aleatoria.

La realizzazione di un evento corrisponde ora all'assegnazione di un valore (tra i possibili) alla variabile aleatoria; tale valore "prescelto" prende dunque il nome di realizzazione della v.a.

Distinguiamo inoltre tra **variabili aleatorie discrete** e **continue**, a seconda se la grandezza che descrivono abbia valori numerabili o continui.

La caratterizzazione della variabile aleatoria avviene attraverso le 2 funzioni di *densità di probabilità* e di *distribuzione di probabilità*.

I parametri statistici

Sono indici sintetici che riassumono fedelmente le informazioni contenute in una serie di dati raccolti su una popolazione, data l'inopportunità di mantenere tutte le misure acquisite (per ragioni di chiarezza e anche per difficoltà pratica).

Un parametro statistico è tanto più efficace quanto meglio riassume il contenuto informativo dei dati con la minor perdita di informazioni e quanto meglio si presta a calcoli e test.

Parametri statistici *efficaci*: la *media aritmetica* e la *varianza* o la *deviazione standard*.

La media

La media aritmetica è il valore centrale attorno a cui si distribuiscono i dati.

$$\mu = \sum_i X_i \cdot f_i$$

dove f_i sono le frequenze relative degli N valori argomentali X_i .

Contiene solo una parte dell'informazione sui dati, non affermando nulla sulla

distribuzione dei dati intorno ad essa.

Esempio

Si calcoli la media delle 3 serie di misure:

1a) 99; 100; 101; \emptyset media =

2a) 50; 100; 150; \emptyset media =

3a) 0.1; 100; 199.9; \emptyset media =

La media è un parametro significativo per confrontare le tre serie?

Si considerino ora le seguenti 3 serie:

1b) 107; 105; 103; \emptyset media =

2b) 51; 110; 154; \emptyset media =

3b) 0.1; 115; 199.9; \emptyset media =

La differenza delle medie tra le nuove serie e le precedenti ha sempre significato?

La varianza e la deviazione standard

La **varianza** σ^2 si definisce matematicamente come:

$$\sigma^2 = \sum_i (X_i - \mu)^2 \cdot f_i$$

La **deviazione standard** σ è la sua radice quadrata.

Esempio

Valori X_i [m]	Scarti $X_i - \mu = x_i$ [m]	Quadrati degli scarti:
10.122
10.120
10.119
10.124
10.121
10.129
$\mu = 10.1225$; Numerosità = $N = 6$; $\sigma^2 = 0.000013$; $\sigma = 0.00361939$		

In aggiunta si definiscono il **coefficiente di variazione** $C.V. = \frac{\sigma}{\mu}$ e l'**errore standard**

della media $\sigma_m = \frac{\sigma}{\sqrt{N}}$

Nell'esempio precedente

$$C.V. = \frac{\sigma}{\mu} = 0.000357559$$

$$\sigma_m = \frac{\sigma}{\sqrt{N}} = 0.00147761$$

Il coefficiente di variazione si esprime in %: 0.0358%

Perché il denominatore di σ^2 contiene (N-1) e non N?

La serie X del nostro esempio contiene 6 misure **indipendenti fra loro**.

Si dice che la serie di N valori ha N gradi di libertà. I gradi di libertà sono la differenza tra il numero di dati disponibili e il numero di relazioni che li vincolano.

Se consideriamo la serie x degli scarti, l'indipendenza si conserva per N-1 elementi; quello rimanente si può determinare a partire dai primi 5 valori perché la somma algebrica degli scarti deve essere uguale a zero, come dimostrato nella formula seguente (tenendo conto che la somma delle frequenze relative è pari all'unità)

$$\sum_i (X_i - \mu) \cdot f_i = \sum_i X_i \cdot f_i + \sum_i \mu \cdot f_i = \mu - \mu \cdot \sum_i f_i = \mu - \mu = 0$$

Quindi i gradi di libertà sono N-1 e non N.

Altri parametri statistici, comunemente impiegati sono i seguenti:

Indici di posizione:

- **Moda**
 - o Ascissa del punto di massimo della distribuzione

- **Mediana**
 - o La mediana M di un insieme di n dati ordinati in ordine di grandezza crescente è il valore centrale dei dati, se il numero di dati è dispari, o la media aritmetica dei due valori centrali, se il numero dei dati è pari. Questa definizione della mediana assicura che lo stesso numero di dati cade sia a sinistra che a destra della mediana stessa. L'uso della mediana come indice per descrivere le caratteristiche dei dati ha lo svantaggio di dover prima riordinare i dati in ordine crescente, il che non è richiesto per il calcolo della media.

Indici di dispersione:

- **Intervallo di variazione (o range)**
 - o definito come la differenza tra la più piccola e la più grande delle misure, pur semplice e intuitivo, è inadeguato. Poco robusto (sensibile agli errori grossolani) e molto instabile (perché legato ai valori estremi, maggiormente influenzati dalle oscillazioni accidentali).
- **Deviazione media o scarto medio**
 - o E' la differenza fra un generico valore della serie ed il valore medio. Lo scarto medio si ottiene facendo la media dei valori assoluti degli scarti.

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

Indici di forma:

- **Indice di asimmetria (Skewness)**
 - o L'asimmetria misura quanto i dati sono distribuiti da un lato della distribuzione rispetto alla media aritmetica, cioè se da un lato sono tutti molto vicini e dall'altro molto distesi verso valori lontani dalla media. La skewness assume valore 0 se c'è simmetria, presenta valori < 0 con asimmetria negativa, cioè quando la moda è spostata verso i valori massimi della distribuzione ed è > 0 se la moda è spostata verso l'estremo inferiore della distribuzione (asimmetria positiva).

$$\frac{\sum_i (X_i - \mu)^3 \cdot f_i}{\sigma^3}$$

- **Indice di Curtosi**
 - o Misura il grado di appiattimento, cioè misura la concentrazione o dispersione dei dati attorno al valore centrale. La Curtosi assume valore 0 se la distribuzione è *mesocurtica* (come la distribuzione Normale tratta nel seguito).
 - o Con valori < -3 la distribuzione è detta *platicurtica* e presenta una forma appiattita con valori maggiormente concentrati nelle code, per Curtosi > 3 la distribuzione è *leptocurtica* con picco accentuato dato dalla

concentrazione dei dati intorno al valore massimo.

$$\frac{\sum_i (X_i - \mu)^4 \cdot f_i}{\sigma^4}$$

Campione di una popolazione

Il metodo della statistica è di tipo **induttivo**: si traggono conclusioni generali da dati particolari.

Problema: fino a che punto il campione esprime le caratteristiche della popolazione originaria? L'induzione è garantita? L'esperienza ha dimostrato che la maggior parte delle misurazioni può considerarsi estratta da popolazioni **distribuite normalmente**.

Una distribuzione normale, o di Gauss ha una espressione matematica definita (Karl Friedrich Gauss descrisse la distribuzione Normale studiando il moto dei corpi celesti [http://it.wikipedia.org/wiki/Variabile_casuale_normale]), funzione di soli due parametri: l'ascissa della sua sommità (valor medio μ) e l'ascissa dei punti di flesso (deviazione standard, $\pm\sigma$).

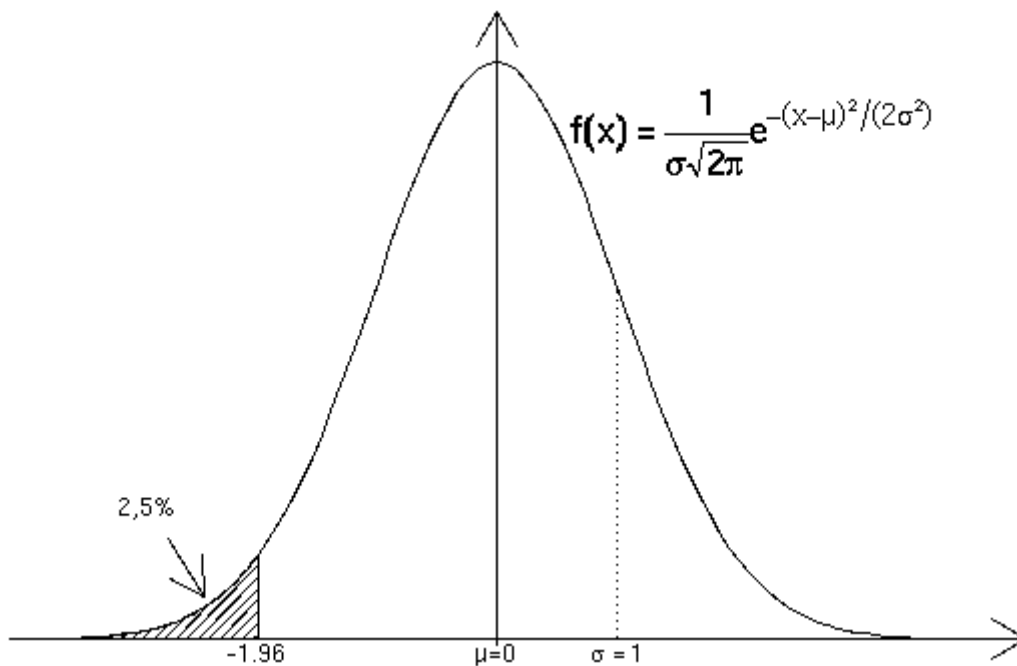


Figura 1 - Funzione di densità di probabilità della variabile casuale normale (fonte Wikipedia).

In essa, uno dei valori compare con frequenza massima e i valori inferiori e superiori compaiono con una frequenza tanto minore quanto più lontani dal valore più frequente. La curva ha forma a campana ed è simmetrica rispetto al valore di massima frequenza. La curva è asintotica. Tutti gli individui della popolazione stanno sotto la curva tra $-\infty$ e $+\infty$; la probabilità che un individuo preso a caso fra la popolazione presenti un valore compreso entro un intervallo assegnato è data dal calcolo dell'area sottesa dalla curva in quell'intervallo.

Il 68.26% della popolazione si trova nell'intervallo $\mu \pm \sigma$, il 94.44% nell'intervallo $\mu \pm 2\sigma$, il 99.73% nell'intervallo $\mu \pm 3\sigma$, il 100% fra $-\infty$ e $+\infty$.

La media aritmetica \bar{X} del campione è la migliore stima della media μ della popolazione.

La deviazione standard s del campione è la migliore stima della deviazione standard σ della popolazione.

Immaginiamo ora di estrarre dalla medesima popolazione (distribuita normalmente con media μ e deviazione standard σ) più campioni analoghi a quello del nostro esempio, con la stessa numerosità ($N=6$). Per ogni k -esimo campione otterremmo diverse medie \bar{X}_k e deviazioni standard σ_k . Potendo estrarre un numero infinito di campioni, le medie \bar{X}_k (una popolazione di medie) si distribuirebbero secondo una gaussiana di media μ (quella della popolazione di partenza) e deviazione standard $\sigma_m = \frac{\sigma}{\sqrt{N}}$ (errore standard della media) [TEOREMA FONDAMENTALE].

Le medie calcolate a partire da un campione oscillano meno attorno alla media di quanto non facciano gli individui del campione. E questo è vero tanto più quanto più numeroso è il campione.

Al limite, per $N \rightarrow \infty$, la deviazione standard tende a zero e quindi la media stimata tende alla media vera.

La propagazione dell'errore

Media e varianza di una variabile casuale monodimensionale ne rappresentano

rispettivamente il baricentro e la dispersione.

In topografia capita però raramente di misurare direttamente la quantità che si vuole determinare: si misurano angoli azimutali, angoli zenitali, distanze, dislivelli per determinare coordinate.

Si deve allora essere in grado di determinare media e varianza di variabili casuali che siano funzione di altre variabili casuali. Le caratteristiche di aleatorietà delle quantità misurate indirettamente dipendono dalla statistica delle quantità misurate direttamente, di cui le prime sono funzione.

1° Caso: funzioni lineari o non lineari di grandezze osservate indipendenti

Sia f una funzione lineare nelle grandezze X, Y, Z, \dots indipendenti e direttamente misurabili:

$$f(X, Y, Z, \dots) = aX + bY + cZ + \dots$$

Estraendo n serie $\{X_k, Y_k, Z_k, \dots\}$ (per $k=1, \dots, n$) il corrispondente valore di f sarà:

$$f_k = f(X_k, Y_k, Z_k, \dots) = aX_k + bY_k + cZ_k + \dots \quad k = 1, \dots, n$$

Sottraendo membro a membro ogni f_k alla f , si ottiene:

$$f - f_k = a(X - X_k) + b(Y - Y_k) + c(Z - Z_k) + \dots \quad k = 1, \dots, n$$

cioè l'errore indotto in f dalla serie k -esima, dovuto agli errori di misura:

$$\epsilon_k = ax_k + by_k + cz_k + \dots \quad k = 1, \dots, n$$

avendo posto:

$$\epsilon_k = (f - f_k), \quad x_k = (X - X_k), \quad y_k = (Y - Y_k), \quad z_k = (Z - Z_k)$$

Quadrando l'espressione ϵ_k e sommando membro a membro i quadrati $\forall k$:

$$\sum_{k=1}^n \epsilon_k^2 = a^2 \sum_{k=1}^n x_k^2 + b^2 \sum_{k=1}^n y_k^2 + c^2 \sum_{k=1}^n z_k^2 + \dots + 2ab \sum_{k=1}^n x_k y_k + \dots$$

Le somme miste nelle variabili, dovute ai doppi prodotti, al crescere di n, tendono a zero, perché – in presenza di soli errori accidentali - la frequenza di termini positivi e negativi tende a livellarsi; quindi, per n grande, si ha:

$$\sum_{k=1}^n \epsilon_k^2 = a^2 \sum_{k=1}^n x_k^2 + b^2 \sum_{k=1}^n y_k^2 + c^2 \sum_{k=1}^n z_k^2 + \dots$$

e, dividendo per n ogni termine, si ottengono gli errori medi:

$$\frac{\sum_{k=1}^n \epsilon_k^2}{n} = a^2 \frac{\sum_{k=1}^n x_k^2}{n} + b^2 \frac{\sum_{k=1}^n y_k^2}{n} + c^2 \frac{\sum_{k=1}^n z_k^2}{n} + \dots$$

E quindi la

$$\sigma_f^2 = a^2 \sigma_x^2 + b^2 \sigma_y^2 + c^2 \sigma_z^2 + \dots$$

E per una funzione f non lineare in X, Y, Z,...?

Si ottiene un risultato analogo linearizzando la funzione f mediante sviluppo in serie di Taylor arrestato al primo ordine attorno ai valori osservati { Xk , Yk , Zk , ... }.

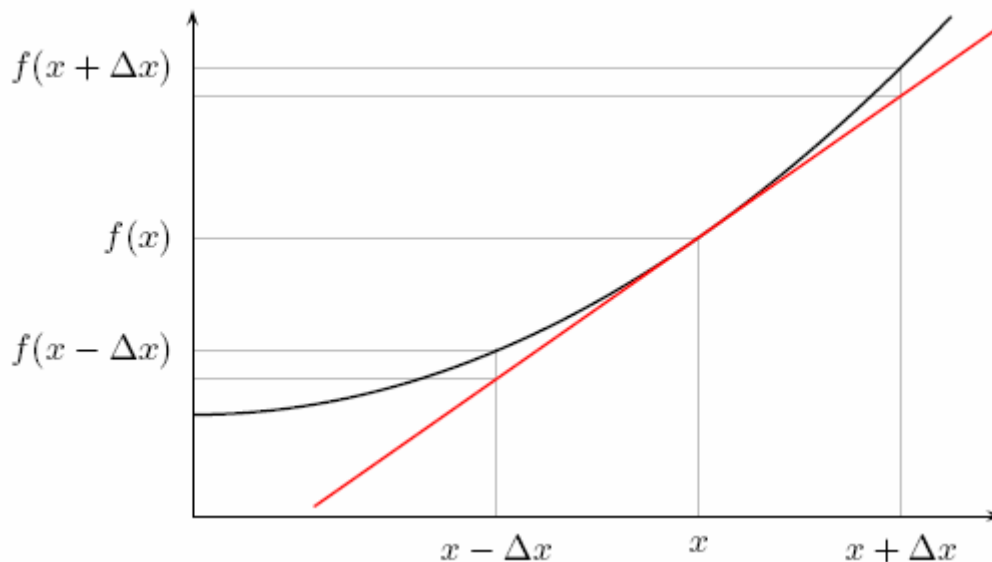


Figura 2 - Approssimazione in linearizzazione di una funzione (fonte: Lorenzo Roi – Elementi di teoria degli errori)

I coefficienti a_2, b_2, c_2, \dots saranno i quadrati delle derivate parziali di f rispetto alle grandezze osservabili X, Y, Z,... calcolate per un valore approssimato.

Infatti:

$$f(X, Y, \dots) = f(X_0, Y_0, \dots) + \left(\frac{\partial f}{\partial X}\right)_0 \cdot \Delta X + \left(\frac{\partial f}{\partial Y}\right)_0 \cdot \Delta Y + \dots$$

$$\text{con } \Delta X = X - X_0 = x_0$$

$$f(X, Y, \dots) = f_0 + a \cdot x_0 + b \cdot y_0 + \dots$$

2° Caso: funzioni lineari o non lineari di grandezze osservate dipendenti

Se le grandezze X, Y, Z,... non sono tra loro indipendenti, non è più possibile affermare che i termini misti tendano a zero al crescere di n.

Osservando che tali sommatorie corrispondono alla somma del prodotto degli scarti

$\sum_{k=1}^n x_k y_k$ e, dividendo tutto per n, alla media del prodotto degli scarti M(xy), si ha

rispettivamente, per il caso lineare:

$$\sigma_f^2 = a^2 \sigma_x^2 + b^2 \sigma_y^2 + \dots + 2 ab M(xy)$$

e per il caso non lineare:

$$\sigma_f^2 = \left(\frac{\partial f}{\partial X}\right)_0^2 \sigma_x^2 + \left(\frac{\partial f}{\partial Y}\right)_0^2 \sigma_y^2 + \dots + 2 \left(\frac{\partial f}{\partial X}\right)_0 \left(\frac{\partial f}{\partial Y}\right)_0 M(xy)$$

Si può dimostrare che l'espressione M(xy) è equivalente al prodotto delle deviazioni standard $\sigma_x \cdot \sigma_y$ per un coefficiente r_{xy} , detto **coefficiente di correlazione lineare**.

Il COEFFICIENTE DI CORRELAZIONE LINEARE r_{xy} misura il grado di dipendenza lineare fra le variabili; varia tra -1 e +1.

$$M(xy) = \sigma_{xy} = r_{xy} \cdot \sigma_x \sigma_y$$

La matrice di varianza-covarianza

Immaginiamo di avere una variabile casuale Y funzione di un vettore X di variabili casuali.

1) Se la funzione Y è lineare si può dire: $Y=AX+b$.

Data la linearità della media si ha $MY=A \cdot MX+b$ e quindi, sottraendo:

$$Y-MY=A(X-MX)$$

che è lo scarto della variabile casuale Y.

La covarianza di Y è il momento del 2° ordine, il cui termine generico vale

$$c_{ik} = M[(x_i - Mx_i)(x_k - Mx_k)]$$

dove:

per $i=k$ varianza della componente k-esima

per $i \neq k$ covarianza delle componenti i, k;

e in notazione matriciale diviene:

$$C_{YY} = M[(Y - MY)(Y - MY)^T] = M\{[A \cdot (X - MX)] \cdot [(X - MX)^T \cdot A^T]\}$$

e, grazie alla linearità della media,

$$C_{YY} = A \cdot M[(X - MX) \cdot (X - MX)^T] \cdot A^T = A \cdot C_{XX} \cdot A^T$$

che è la **legge di propagazione della varianza per funzioni lineari**.

Si noti che, per $Y = \text{scalare}$ ($A = \text{vettore}$), si avrà:

$$C_{YY} = \sigma^2_Y = A \cdot C_{XX} \cdot A = \sum_i a_i a_k c_{ik}$$

dove, se $c_{ik} = 0$ per $i \neq k$, le componenti X_i sono indipendenti e si può dire

$$C_{YY} = \sigma^2_Y = A \cdot C_{XX} \cdot A^T = \sum_i a_i^2 c_{ii} = \sum_i a_i^2 \sigma_X^2$$

2) Se la funzione Y non è lineare nelle componenti X si può comunque dire:

$$Y = g(X) \approx g(MX) + [\partial_X g(X)]_{MX} \cdot (X - MX)$$

e, avendo linearizzato, si può affermare che:

$$C_{YY} \approx A \cdot C_{XX} \cdot A^T = [\partial_X g(X)]_{MX} \cdot C_{XX} \cdot [\partial_X g(X)]_{MX}^T = J_X \cdot C_{XX} \cdot J_X^T$$

dove \mathbf{J}_X è **Jacobiano della variabile X**.

Se Y è uno scalare:

$$\sigma^2_Y = \sum_i [\partial_X g(X)]_i [\partial_X g(X)]_k \cdot c_{ik}$$