COMPLEMENTI DI STATISTICA

Parte III

Fulvio Ricci

Dipartimento di Fisica, Università di Roma *La Sapienza*

INDICE

INTRODUZIONE ALLA TEORIA DELLA STIMA

.

Consideriamo un esperimento teso allo studio della variabile X. L'esperimento consiste nell'effettuare n misure indipendenti della stessa grandezza. Avremo quindi n dati relativi ad X, che indicheremo usando il linguaggio statistico come le n *realizzazioni* della variabile aleatoria X. Utilizzando termini tipici della demografia, diremo che l'insieme di queste realizzazioni costituiscono un *campione* dell'intera *popolazione* delle X. Possiamo poi pensare di ripetere n volte ancora l'esperimento ottenendo un altro campione della stessa popolazione. Se ogni raccolta di dati è statisticamente indipendente una dall'altra, e se la funzione densità di probabilità è rimasta immutata durante tutto il processo di campionamento, allora possiamo pensare di dedurre dall'insieme dei dati delle stime dei parametri della distribuzione concernente l' intera popolazione. Precisiamo che solo se la condizione d'indipendenza è verificata, allora il campione viene considerato *rappresentativo* della popolazione. Formalizziamo quanto sin qui espresso:

$X_1 \longrightarrow$ realizzazioni $x_1^{'}, x_1^{''}, \cdots x_1^K$

$X_2 \longrightarrow$ realizzazioni $x_2^{'}, x_2^{''}, \cdots x_2^K$

$X_3 \longrightarrow$ realizzazioni $x_3^{'}, x_3^{''}, \cdots x_3^K$

A loro volta le quantità $X_1, X_2, \cdots X_n$ sono variabili

aleatorie aventi la stessa distribuzione di probabilità.

$$E[X_i] = \mu \qquad Var[X_i] = \sigma^2 \qquad \forall \quad i$$

Per la condizione di indipendenza delle variabili, potremo scrivere che la probabilità di osservare quell'insieme di n dati, nel caso di variabili discrete, è data da

$$P[X_1 = x_1, X_2 = x_2, \cdots X_n = x_n] = f(x_1)f(x_2)\cdots f(x_n)$$

Nel caso di variabili continue, ci riferiremo alla funzione densità di probabilità associata al campione e scriveremo:

$$f(x_1, x_2 \cdots x_n) = \prod_{i=1}^{n} f(x_i) = L(x_i)$$

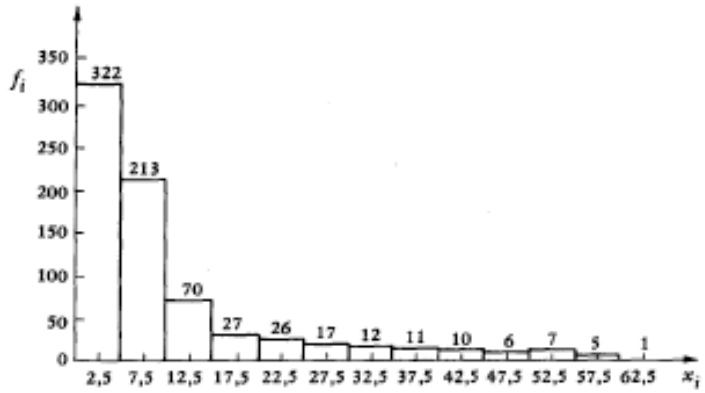
Tale probabilitá é indicata con $L(x_i)$ ed é denominata funzione di *verosimiglianza* del campione (*likelihood function.*

Ottenuto il campione di una certa variabile aleatoria X, classifichiamo le realizzazioni sulla base della frequenza con cui si presentano i differenti risultati della X.
A questo scopo, suddividiamo l'intervallo in cui si estende il campione in intervalli più piccoli di uguale ampiezza e che non si sovrappongono, le *classi.* In questo modo siamo nella condizione di valutare il numero di volte n f_i che nel campione di n dati, è stato ottenuto un risultato il cui valore cade all'interno della

i-esima classe. . Se le classi sono in numero pari ad r deve essere verificata la seguente condizione di normalizzazione

$$\sum_{i=1}^{r} f_i = n$$

f_i è detta frequenza dell'$i - esima$ classe; l'insieme dei valori delle f_i costituisce la *distribuzione empirica* dei risultati.



Esempio d'istogramma costruito sulla base di un campione di 727 persone di una scuola classificate secondo l'età.

Data l'analogia tra distribuzioni empiriche e teoriche (o di probabilità), per definire le proprietà della distribuzione empirica introdurremo concetti del tutto simili a quelli già incontrati nel calcolo delle probabilità. Definiamo

a) il Momento di ordine k del campione

$$\bar{x^k} = \sum_{i=1}^{n} x_i^k p_i = \frac{1}{n} \sum_{i=1}^{n} x_i^k$$

L'analogia con la definizione data in campo probabilistico è immediata se si rammenta che ciascun dato del campione è equiprobabile , quindi abbiamo che $p_i = \frac{1}{n}$.

b) la Media Empirica

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

c) il Momento Centrale di ordine k

$$S^k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^k$$

d) la Varianza Empirica

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

e) Covarianza Empirica (nel caso delle popolazioni bivariate)

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \bar{xy} - \bar{x} \cdot \bar{y}$$

f) Coefficiente di Correlazione Empirico (nel caso delle popolazioni bivariate)

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\Sigma(x_i - \bar{x})^2)(\Sigma_{i=1}^{n}(y_i - \bar{y})^2))}}$$

ovvero

$$r = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\sqrt{(\bar{x^2} - \bar{x}^2)(\bar{y^2} - \bar{y})}}$$

Vogliamo ora cercare di stabilire formalmente quale sia un criterio statisticamente significativo che ci consenta di stimare i parametri della distribuzione di probabilità, sulla base delle nostre conoscenze, limitate dalla natura finita dei campioni disponibili. Enunciamo allora in forma generale tale problema.

Siano λ_j i parametri della distribuzione da stimare. Per far questo noi faremo uso di opportune funzioni Θ del campione $x_1 \cdots x_n$, gli *stimatori*.

Questi stimatori Θ doovrebbero soddisfare la condizione di **consistenza**. In altre parole se l'estensione del campione é infinita ($n \to \infty$) allora deve accadere che

$$\lim_{n\to\infty} P[\Theta_j(x_1 \cdots x_n) - \lambda_j| > \varepsilon] = 0 \qquad \forall \ \ \varepsilon$$

Inoltre i diversi Θ dovrebbero essere **esenti** da **distorsione** (*correttezza dello stimatore*). Questo si esprime dicendo che, qualunque sia l'estensione del campione, deve accadere che

$$E[\Theta_j(x_1 \cdots x_n)] = \lambda_j \qquad \forall\ n$$

Nel caso in cui la X è una variabile aleatoria a distribuzione normale , abbiamo già visto che $\bar{x}$ é una stima di $E[X]$ mentre lo scarto quadratico S^2

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

é una stima non distorta della varianza.
Se la X non è variabile aleatoria normale , si può far vedere che
- $\bar{x}$ è ancora stima di $E[X]$,
- ma non è sempre verificata la condizione di *assenza di distorsione* per lo stimatore S^2 della varianza.

IL METODO DELLA MASSIMA VEROSIMIGLIANZA.

Sia dato un campione di dimensione n della variabile aleatoria X di cui è nota la forma della distribuzione ma non i suoi parametri.
Il principio della Massima Verosimiglianza afferma che *la miglior* stima dei parametri è quella che rende massimo la probabilità di verificarsi dell'evento costituito dalla n-pla dei risultati ottenuti.
Per esprimere in termini formali tale affermazione, imponiamo la condizione di massimo alla funzione che esprime la probabilità P rispetto ai parametri $\lambda_1 \cdots \lambda_k$ della distribuzione associata all'evento $(x_1, \cdots x_n)$. Nel caso di variabili discrete imporremo tale condizione alla funzione

$$P(x_1, \cdots x_n) = \prod_{i=1}^{n} P(x_1, \lambda_1 \cdots \lambda_k)$$

Nel caso di variabili continue cercheremo per quali valori di $\lambda_1 \cdots \lambda_k$ si ha un massimo della funzione di verosimiglianza

$$L(x_1, \cdots x_n, \lambda_1 \cdots \lambda_k) = \prod_{i=1}^{n} f(x_1, \lambda_1 \cdots \lambda_k)$$

Limitiamoci qui a sviluppare il caso delle variabili continue. Per esse avremo

$$\frac{\partial L}{\partial \lambda_j} = 0 \quad ; \quad \frac{\partial^2 L}{\partial \lambda_j^2} \Big| < 0 \qquad j = 1, \cdots k$$

Indichiamo con λ_j^* le soluzioni del sistema di k equazioni sopra riportato; queste sono le *stime di massima verosimiglianza* dei parametri della distribuzione di probabilità.

ESEMPI DI APPLICAZIONE DEL METODO DELLA MASSIMA VEROSIMIGLIANZA.

Supponiamo di aver posto sotto osservazione una variabile aleatoria discreta che fluttua secondo la distribuzione binomiale. Per caratterizzare completamente il sistema dobbiamo stimare l'unico parametro caratteristico della binomiale: la probabilità di successo associata ad una singola prova p. Consideriamo allora la probabilità di avere osservato in n tentativi, k successi

$$P_{n_i,k}(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

In questo caso la funzione di verosimiglianza risulta proporzionale alla quantità $p^k(1-p)^{n-k}$. Osserviamo inoltre che il valore di p per cui si ha il massimo della funzione L, coincide con quello in corrispondenza del massimo della funzione log L .Ma in quest'ultimo caso la ricerca del massimo implica uno sviluppo algebrico molto più semplice. Infatti abbiamo

$$\frac{dLog(L(n,K,p))}{dp} = \frac{K}{p} - \frac{n-K}{1-p} = 0$$

Otteniamo quindi

$$p^* = \frac{K}{n}$$

Verifichiamo se questi stimatori rispettano le condizioni di consistenza ed assenza di distorsione. Per la

legge dei grandi numeri abbiamo

$$E\left[\frac{K}{n}\right] = p$$

e ciò ci garantisce la consistenza.
Inoltre, per la diseguaglianza di Tchebychev, abbiamo che

$$P\left[|\frac{k}{n} - p| \geq K\sigma\right] \leq \frac{1}{K^2}$$

e ciò garantisce l'assenza di distorsione.

Stimiamo ora con lo stesso metodo μ e σ di una distribuzione normale.
Supponiamo di avere un campione di dimensione n estratto da una popolazione a distribuzione normale. La funzione di verosimiglianza associata è

$$L = \left(\frac{1}{\sqrt{2\ \pi}\sigma}\right)^n \prod_i e^{\frac{(x_i-\mu)^2}{2\ \sigma^2}}$$

Al solito passiamo al $\log L$ ed imponiamo la condizione di massimo rispetto a μ e σ.

$$\log L = -n \log \sqrt{2\ \pi} - n \log \sigma - \sum_i \frac{(x_i - \mu)^2}{2\ \sigma^2}$$

Per la stima di μ avremo

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i - n\mu = 0$$

da cui

$$\widehat{\mu} = \frac{1}{n} \sum_i x_i$$

È facile verificare che

$$(\frac{\partial^2 \log L}{\partial \mu^2})_{\mu=\widehat{\mu}} = -\frac{n}{\sigma^2} < 0$$

Per la stima di σ^2 avremo:

$$\frac{\partial \log L}{\partial \sigma^2} = \frac{1}{2}\frac{\sum_i (X_i - \mu)^2}{\sigma^4} - \frac{1}{2}\frac{n}{\sigma^2}$$

da cui deduciamo

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_i (x_i - \mu)^2$$

Si può inoltre verificare che

$$(\frac{\partial^2 \log L}{(\partial \sigma^2)^2})_{\sigma^2=\widehat{\sigma}^2} = -\frac{n}{2\ \sigma^4} < 0$$

Notiamo che lo stimatore della varianza che abbiamo ricavato, dipende dal valore vero μ. Se anch'esso è incognito, allora puó sembrare ovvio utilizzare al suo posto la stima $\widehat{\mu}$. Tuttavia in questo caso la quantità

$$\widehat{S}^2 = \frac{1}{n}\sum_i^n (x_i - \widehat{\mu})^2$$

risulta essere uno stimatore *consistente* ma *distorto*. Al contrario la stima

$$S^{2=}\frac{1}{n-1}\sum_i (x_i - \widehat{\mu})^2$$

verifica ambedue le condizioni. In particolare l'assenza di distorsione é facilmente dimostrabile ricordando che

$(n-1)S^2/\sigma^2$ è una variabile ${\chi^2}_{n-1}$ ad $(n-1)$ gradi di libertá e di pari valore aspettato, e quindi si avrá

$$E[S^2] = \sigma^2$$

Vediamo ora un caso ancora più generale. Siano x_i le realizzazioni di variabili aleatorie a distribuzione normale aventi lo stesso valore medio μ ma varianze differenti σ_i. Esplicitiamo la funzione di verosimiglianza:

$$L(x_i, \mu) = \prod_1^n \frac{1}{\sqrt{2\ \pi\sigma_i}} e^{\frac{(x_i,\mu)^2}{2\ {\sigma_i}^{.2}}}$$

$$\frac{d \log L(x_i, \mu)}{d\mu} = \frac{d}{d\mu} \sum_i \frac{(x_i - \mu)^2}{2\ {\sigma_i}^{.2}} = 0$$

ovvero

$$\sum_{i=1}^n \frac{x_i - \mu}{\sigma_i^2} = 0$$

Quindi otteniamo la *media pesata*

$$\widehat{\mu} = \sum_{i=1}^n \frac{1}{\sigma_i^2} x_i \Big/ \sum_{i=1}^n \frac{1}{\sigma_i^2}$$

dove $\frac{1}{\sigma_i^2}$ sono i pesi dei termini della media.

INTERVALLI FIDUCIARI

Abbiamo sinora discusso un metodo per ottenere una stima di uno o più parametri di una distribuzione di forma nota, non abbiamo peró alcuna valutazione statistica sulla bontá della stima. A questo scopo si introducono i cosidetti *intervalli fiduciari* che consentono di valutare quale sia l'intervallo associato alla quantità da stimare all'interno del quale cade il suo valor vero con una prefissata probabilità.
Il metodo è basato sulla scelta di una stimatore T di cui si conosce la funzione di distribuzione purché sia una funzione del parametro incognito Θ da stimare. Sia allora $T = T(\Theta)$ una funzione invertibile e P la probabilità associata a T. Indichiamo inoltre con t_1 e t_2 i valori per cui sia verificato

$$P[t_1 \leq T \leq t_2] = 1 - \alpha$$

α è un valore della probabilità stabilito *a priori*: esso è associato **all'errore che si compie affermando che**

$$t_1 \leq T \leq t_2$$

e quindi di conseguenza, essendo invertibile la funzione $T(\Theta)$, che

$$\theta_1 \leq \theta \leq \theta_2$$

$1 - \alpha$ è denominato il **livello di fiducia** dell'intervallo e $\theta_1 - \theta_2$ l'**intervallo fiduciario**.

In generale scriveremo, usando la funzione cumulativa di distribuzione $F(t) = \int_{-\infty}^{t} f(t^{'})\ dt^{'}$

$$P[t_1 \leq T \leq t_2] = F(t_2) - F(t_1)$$

Se la funzione $f(t)$ è simmetrica, allora avremo che $t_1 = -t$, $t_2 = t$ e

$$P[-t \leq T \leq t] = 2F(t) - 1 = 1 - \alpha$$

essendo quindi $\alpha = 2\ [1 - F(t)]$.

Vediamo ora un semplice esempio.
Sia X una variabile aleatoria normale avente varianza σ e valore aspettato incognito μ_x. Per fissare le idee supponiamo che X rappresenti i possibili risultati di una misura di lunghezza.
Supponiamo allora di aver effettuato $n = 200$ misure di X ed aver ottenuto una media $\bar{x} = 82.40$ cm.
A ciascuna di queste misure attribuiamo una deviazione standard $\sigma = 4.2$ cm.
Vogliamo ricavare l'intervallo fiduciario per μ_x, stabilendo un livello di fiducia del 95 %.

A questo scopo introduciamo lo stimatore T così definito:

$$T = \frac{(\bar{x} - \mu_x)}{\frac{\sigma}{\sqrt{n}}}$$

Essendo noto σ, lo stimatore è una funzione della sola μ_x ed è una variabile normale ridotta. La distribuzione di Gauss è simmetrica, quindi procediamo notando che, $P = 1 - \alpha = 0.95$ e quindi si ha $F(t^*) = 1 - \alpha/2 = 0.975$. Consultando la tabella in cui sono riportati i valori della funzione cumulativa gaussiana in funzione

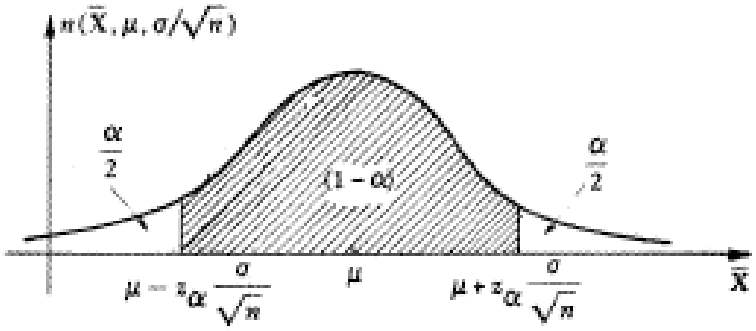
della variabile ridotta, é possibile verificare che ad $F = 0.975$ corrisponde $t^* = 1.96$.

Quindi l'intervallo fiduciario al 95 % di probabilità è calcolato invertendo la funzione $T(\mu_x)$, ottenendo

$$\Delta\mu_x = t^* \frac{\sigma}{\sqrt{n}} = 0.58 \quad cm$$

ovvero

$$\bar{x} - t^* \frac{\sigma}{\sqrt{n}} \leq \mu_x \leq \bar{x} + t^* \frac{\sigma}{\sqrt{n}}$$



Vediamo ora un secondo esempio. Supponiamo di aver effettuato stavolta $n = 20$ misure di lunghezza e di non conoscere la varianza della distribuzione associata. Sulla base dei risultati ottenuti possiamo comunque dedurre $\bar{x} = 1.832$ cm e lo scarto quadratico medio

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{20} (x_i - \bar{x})^2} = 0.497 \quad cm$$

Vogliamo di nuovo definire l'intervallo fiduciario di μ_x ad un livello del 90 %.

Introduciano allora un nuovo stimatore T, definito come

$$T = \sqrt{n}\frac{(\bar{x} - \mu_x)}{s}$$

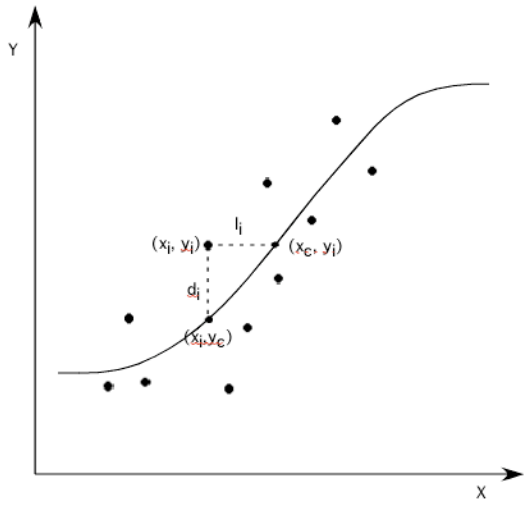
Quando abbiamo parlato delle proprietà delle variabili T di Student, abbiamo incontrato questa definizione. In quella occasione abbiamo dimostrato esplicitamente che una variabile cosí definita segue la distribuzione di Student a $(n-1)$ gradi di libertá ($(n-1) = 19$ per l'attuale esempio). Quindi stiamo utilizzando uno stimatore, funzione di μ_x, che segue una nota distribuzione di probabilità; notiamo inoltre che si tratta di una distribuzione simmetrica. Il livello di significatività, stabilito *a priori*, è $(1-\alpha) = 0.9$: quindi abbiamo $\alpha = 0.1$ e $F(t^*) = 0.95$.
Consultando le tabelle della F di un T di Student a 19 gradi di libertá, deduciamo che il valore di t^*, corrispondente ad $F(t^*) = 0.95$, è pari a $t^* = 1.729$. Concludiamo allora che

$$\Delta\mu_x = t^*\frac{s}{\sqrt{n-1}} = 0.206$$

.

STIMA DELLE CURVE DI REGRESSIONE DI UNA POPOLAZIONE BIVARIATA : METODO DEI MINIMI QUADRATI E METODO DELLA MASSIMA VEROSIMIGLIANZA

Il problema della stima dei parametri di una popolazione si estende al caso di popolazioni multivariate di probabilità. Noi ci limiteremo ad affrontare il caso delle popolazioni bivariate. In questo ambito il nostro interesse si focalizza sul problema di stimare i parametri caratterizzanti la relazione empirica che lega le due variabili aleatorie della particolare popolazione considerata. Evidentemente noi stiamo ipotizzando che tra le variabili aleatorie X e Y vi sia una correlazione non nulla e supponiamo che esista una dipendenza funzionale $Y(X)$ di forma nota (lineare, parabolica, esponenziale...) di cui occorra stimare i parametri.



Consideriamo allora la rappresentazione grafica del

campione della popolazione X, Y nel piano cartesiano (*diagramma di diffusione* o *scatter plot*). Per ogni punto (x_i, y_i si osserva una differenza tra l'ordinata del y_i e l'ordinata del corrispondente punto della curva (x_i, y_c) che approssima l'andamento dei dati

$$d_i = y_i - y_c(x_i)$$

Al fine di definire uno stimatore del livello di adattamento della curva considerata ai dati del campione, consideriamo una quantitá indipendente dal segno di questa differenza d_i:

$$S_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n} c_i(y_i - y_c(x_i))^2}$$

dove i coefficienti c_i che appaiono a moltiplicare i termini d_i^2 della sommatoria, hanno la funzione di pesare in modo differente i contributi dei vari termini d_i^2 nello stimatore. La curva che meglio approssima il campione sará quella i cui parametri **minimizzano** tale stimatore. Questo é il metodo dei *minimi quadrati*.

Questo approccio al problema è intuitivo nel caso in cui si assuma X come variabile indipendente e Y variabile dipendente ed é tanto piú naturale quanto meno é aleatorio il carattere della variabile X. Nel caso piú generale é possibile invertire i ruoli delle variabili X e Y, definire un analogo stimatore basato sui quadrati delle distanze $l_i = x_i - x_c(y_i)$ minimizzandolo rispetto ai parametri della curva. Le due procedure concorrenti porteranno a stimare due curve di regressione diverse. In genere peró, nell'applicare metodo dei minimi quadrati

i valori della X sono stati *prescelti* cosí che lo schema tipico di dipendenza funzionale sará del tipo

$$Y = f(x, \lambda_j) + Y_R$$

dove $f(x, \lambda_j)$ é la dipendenza funzionale ipotizzata che dipende dai parametri λ_j e la variabile aleatoria residua Y_R verifica le proprietá $E[Y_R] = 0$ e $Var[Y_R] = \sigma_R^2$ e $E[Y_{Ri}, Y_{Rj}] = 0$ *per* $\forall i \neq j$.

Se le Y_R **sono variabili a distribuzione normale** allora le stime dei parametri ottenute con il metodo dei **minimi quadrati** **coincidono** con quelle ottenute con il metodo della **massima verosimiglianza**.

Infatti, vediamo quali siano i valori dei parametri che massimizzano la funzione di verosimiglianza, che nel caso di dati a distribuzione normale prende la forma:

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[\frac{\left(y_i - f(x_i, \lambda_1, \lambda_2, ..., \lambda_m)\right)^2}{2\sigma_i^2}\right]$$

Poiché utilizziamo per la ricerca del massimo l'usuale artificio matematico di calcolare prima il $\log L$ e poi derivare rispetto ai parametri λ_j, ci limiteremo a cercare il minimo della quantitá

$$S_y = \sum_{i=1}^{n} \left[\frac{\left(y_i - f(x_i, \lambda_1, \lambda_2, ..., \lambda_m)\right)^2}{\sigma_i^2}\right]$$

che coincide con lo stimatore utilizzato nel caso del metodo dei minimi quadrati.

REGRESSIONE LINEARE

Specifichiamo ora quanto é stato esposto nel paragrafo precedente nel caso in cui la forma funzionale $f(x_i, \lambda_j)$ sia una retta di equazione

$$y = a + bx$$

Supponiamo quindi di avere un campione bivariato di dati (x_i, y_i), di ordine n. Assumiamo inoltre che sia x la variabile indipendente e che, per ciascun valore y_i sia associabile una deviazione standard σ_i. Allora l'applicazione del metodo dei minimi quadrati corrisponde a calcolare il minimo rispetto ai parametri a e b della quantitá

$$S_y = \sum_{i=1}^{n} [\frac{(y_i - a - bx_i)^2}{\sigma_i^2}]$$

Si tratta quindi di trovare le soluzion del sistema lineare definito dalle seguenti due equazioni nelle incognite a, b:

$$\frac{\partial}{\partial a}\{ \sum_{i=1}^{n} [\frac{(y_i - a - bx_i)^2}{\sigma_i^2}] = 0$$

$$\frac{\partial}{\partial b}\{ \sum_{i=1}^{n} [\frac{(y_i - a - bx_i)^2}{\sigma_i^2}] = 0$$

Sviluppando otteniamo

$$\sum_{i=1}^{n} \frac{y_i}{\sigma_i^2} - a \sum_{i=1}^{n} \frac{1}{\sigma_i^2} - b \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} = 0$$

$$\sum_{i=1}^{n} \frac{x_i y_i}{\sigma_i^2} - a \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} - b \sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2} = 0$$

Da tale sistema é facile dedurre le soluzioni a^* e b^* :

$$a^* = \frac{\sum_{i=1}^n \frac{x_i^2}{\sigma_i^2} \ \sum_{i=1}^n \frac{y_i}{\sigma_i^2} - \sum_{i=1}^n \frac{x_i}{\sigma_i^2} \ \sum_{i=1}^n \frac{x_i y_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{1}{\sigma_i^2} \ \sum_{i=1}^n \frac{x_i^2}{\sigma_i^2} - (\sum_{i=1}^n \frac{x_i}{\sigma_i^2})^2}$$

$$b^* = \frac{\sum_{i=1}^n \frac{1}{\sigma_i^2} \ \sum_{i=1}^n \frac{x_i y_i}{\sigma_i^2} - \sum_{i=1}^n \frac{x_i}{\sigma_i^2} \ \sum_{i=1}^n \frac{y_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{1}{\sigma_i^2} \ \sum_{i=1}^n \frac{x_i^2}{\sigma_i^2} - (\sum_{i=1}^n \frac{x_i}{\sigma_i^2})^2}$$

Tali relazioni assumono una forma piú semlice nel caso in cui a tutti i punti del campione é associata la stessa varianza

$$\sigma_i = \sigma \qquad \forall i$$

In tal caso le stime a^* e b^* non dipendono piú dalla deviazione standard σ ed assumono una forma piú compatta se si introducono i valori medi di $\bar{x}$ e $\bar{y}$, i valori quadratici medi $\bar{x^2}$ e $\bar{y^2}$ e la correlazione $\bar{xy}$:

$$a^* = \frac{\bar{y} \sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\bar{y} \cdot \bar{x^2} - \bar{x} \cdot \bar{xy}}{\bar{x^2} - \bar{x}^2}$$

$$b^* = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\bar{x^2} - \bar{x}^2}$$

La migliore retta di regressione ha quindi equazione $y = a^* + b^8 x$. É facile, ma noioso, rielaborare le formule ottenute per a^* e b^*, ottenendo una espressione ancora piú compatta per le due stime:

$$a^* = \bar{y} - b^* \bar{x}$$

$$b^* = r \frac{s_y}{s_x}$$

dove r, s_x^2 ed s_y^2 sono rispettivamente il coefficiente di correlazione empirico, gli scarti quadratici della x e della y del campione , ovvero i momenti campionari del secondo ordine.
Quindi la retta di regressione ha la forma

$$y - \bar{y} = r\frac{s_y}{s_x}(x - \bar{x})$$

Si noti che il baricentro del campione $(\bar{x}, \bar{y})$ é un punto della retta di regressione. Il coefficiente angolare b^* é detto *coefficiente di regressione.*

Poiché il coefficiente di regressione lineare b^* é proporzionale ad r, esaminiamo in maggior dettaglio il legame logico esistente tra queste due quantitá. A questo scopo consideriamo lo scarto quadratico della y, s_y, e riscriviamolo facendo apparire esplicitamente le quantitá t $y_c^i = a^* + b^* x_i$

$$s_y = \frac{1}{n}\sum_{i=1}^{n}\left[(y_i - y_c^i) + (y_c^i - \bar{y})\right]^2 = \frac{1}{n}\left[\sum_{i=1}^{n}(y_i - y_c^i)^2 + \sum_{i=1}^{n}(y_c^i - \bar{y})^2\right]$$

Questo é vero perché il termine derivante dal doppio prodotto

$$\sum_{i=1}^{n}(y_i - y_c^i)(y_c^i - \bar{y}) = \sum_{i=1}^{n}(y_i - a^* - b^* x_i)(a^* + b^* x_i - \bar{y})$$

risulta essere nullo, una volta esplicitate le definizioni di a^* e b^*.
Abbiamo cosí messo in evidenza che la varianza empirica della y é costituita da due termini;

$$S_c^2 = \frac{1}{n}\sum_{i=1}^{n}(y_c^i - \bar{y})^2$$

$$S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_c^i)^2$$

Il primo termine S_c^2 è la somma delle deviazioni quadratiche dei corrispondenti punti sulla retta y_c^i rispetto al valor medio $\bar{y}$. Esso é indicato con il termine *media dei quadrati dovuti alla regressione*. Questo termine sarebbe identico a s_y^2 se tutti i punti giacessero sulla retta calcolata. Esso valuta la dispersione dei valori della regressione rispetto alla media. Questa parte della dispersione dipende strettamente da come é distribuito il campione delle x_i, ed é per questo che si parla di parte **spiegata** della varianza.

Il secondo termine S_y^2 é basato sulla differenza tra i valori di y_i ed i corrispondenti valori sulla retta di regressione y_c^i. Questo termine viene indicato con tre sinonimi: (*media dei quadrati attorno alla linea di regressione, media dei quadrati dei* **residui** o ancora *media dei quadrati degli* **errori**. Questa dispersione residua non é connessa all'esistenza di una relazione di tipo lineare tra X e Y e non é quindi legata alla dispersione di X. Allora essa é attribuibile al carattere intrinsecamente aleatorio della Y e viene chiamata la varianza **non spiegata**.

Consideriamo ora il rapporto tra il primo termine *somma dei quadrati dovuti alla regressione* e la varianza empirica totale della Y:

$$\frac{S_c^2}{s_y^2} = \frac{\sum_{i=1}^{n}(y_c^i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Con un pó di algebra semplice ma laboriosa, sostituendo nella formula precedente le relazioni esplicite di r , di a^*

e di b^*, si dimostra che tale rapporto é pari al quadrato del coefficiente di regressione empirico r:

$$r^2 = \frac{S_c^2}{s_y^2}$$

Questa relazione si fornisce una interessante interpretazione del ruolo di r^2. *Esso misura la frazione di varianza rispetto al totale attribuibile alla bontá della retta di regressione*, ovvero a quanto essa é capace di spiegare la dispersione dei dati in y.
Per $r = \pm 1$ tutti i punti giacciono sulla retta e la dispersione totale delle y é rappresentata dal solo termine S_c^2. Per $r = 0$ la retta di regressione *non da alcuna informazione* in merito alla dispersione dei dati di y. Ad esempio, ottenere una retta di regressione a cui é associato un valore $r^2 = 0.912$ significa che il 91.2 % della dispersione dei dati é dovuto all'esistenza di una relazione lineare tra X e Y, mentre il restante 8.8 % , detto anche *errore sperimentale*, é dovuto alle fluttuazioni casuali dei dati, cioé a fattori non riconducibili all'esistenza di una relazione lineare tra le variabili.

Una relazione, utile come la precedente e ricavabile a partire da essa, é:

$$S_y^2 = s_y^2 - S_c^2 = s_y^2(1 - r^2)$$

Questa relazione ci fa porre l'attenzione sulla somma dei quadrati dei residui, l'*errore* associato alla retta di regressione. Per $r = \pm 1$ tutti i punti sono sulla retta e l'errore della stima é nullo.

Concludiamo questo paragrafo, facendo notare che, assumendo Y come variabile indipendente e X vari-

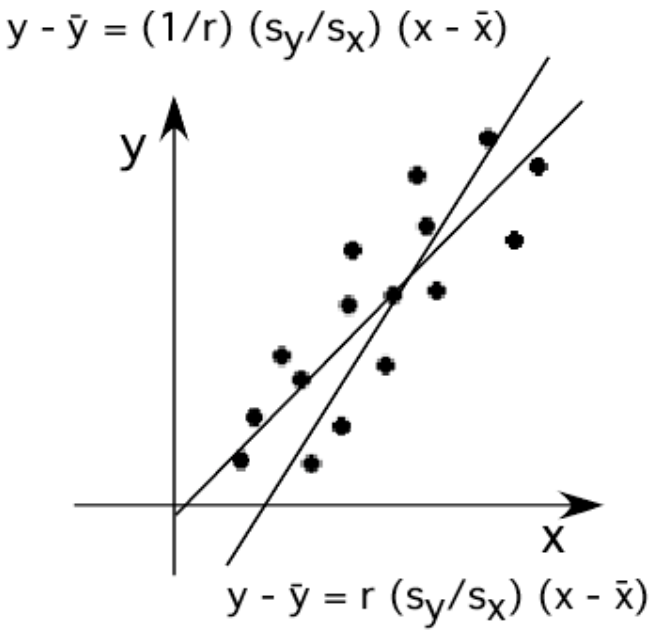
abile dipendente, e procediamo stavolta minimizzando i quadrati delle differenze tra l'ascissa x_i del campione e l'ascissa del corrispondente punto sulla retta di regressione $x_c^i = c + dy_i$, giungiamo a definire la retta di regressione di equazione

$$x - \bar{x} = r\frac{s_x}{s_y}(y - \bar{y})$$

Tale affermazione é facilmente verificabile; infatti basta scambiare i ruoli di x con y nelle formule precedentemente ricavate, notando che r rimane immutato. Tale equazione peró definisce una retta diversa dalla precedente. Infatti con un semplice passaggio ci rendiamo conto che in tal caso si ha:

$$y - \bar{y} = \frac{1}{r}\frac{s_y}{s_x}(x - \bar{x})$$

Questa retta si incrocia con l'altra in $(\bar{x}, \bar{y})$ ma é inclinata diversamente, pur conservando il segno del coefficiente angolare. *Le due rette coincidono solo nel caso in cui* $r = \pm 1$.


$y - \bar{y} = (1/r)\ (s_y/s_x)\ (x - \bar{x})$
y
x
$y - \bar{y} = r\ (s_y/s_x)\ (x - \bar{x})$

SIGNIFICATIVITÀ STATISTICA DELLA RETTA DI REGRESSIONE: INTERVALLI FIDUCIARI DI INTERCETTA E PENDENZA

.

Nel paragrafo precedente abbiamo discusso la natura della varianza empirica totale della variabile y decomponendola in due contributi distinti

$$s_y^2 = S_c^2 + S_y^2$$

L' analisi approfondita di questi due termini ci ha portato a concludere che stimare la retta di regressione aiuta a ridurre la nostra ignoranza sulla natura della variabilitá della Y. Infatti noi riduciamo al minimo la parte non spiegata della varianza,

$$S_y^2 = \frac{1}{n}\sum_i (y_i - y_c^i)^2$$

quantitá che abbiamo chiamato somma quadratica degli *errori*. Abbiamo cosí la possibilitá, tramite la retta di regressione stimata, di predire al meglio i valori di y per un assegnato valore di x.

Dobbiamo ora cercare di quantificare la bontá ed i limiti di tale stima. Per far questo noi ci limiteremo a considerare il caso piú semplice in cui la variabilitá del campione bivariata sia legata principalmente alla variabile Y. Inoltre supponiamo per ragioni di semplicitá che tale variabilitá sia descritta dalla distribuzione di Gauss.

In tal caso possiamo renderci conto agevolmente che le stime a^* b^*, ottenute con il metodo dei minimi quadrati , sono ancora interpretabili come realizzazioni di variabili aleatorie distribuite normalmente, essendo esse stesse combinazioni lineari di variabili normali indipendenti. Ció é dimostrabile in generale, ma per ragioni di semplicitá noi ci riferiremo al caso in cui tutti i punti del campione abbiano associata la stessa varianza. In tal caso abbiamo che

$$b^* = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\bar{y}\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

e poiché

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

allora b^* risulta essere una combinazione lineare delle y_i

$$b^* = \sum_{i=1}^{n} c_i \; y_i \qquad con \qquad c_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Segue immediatamente che anche a^* é una variabile a distribuzione normale. Infatti, poiché il punto $(\bar{x}, \bar{y})$ giace sulla retta di regressione, potremo scrivere

$$a^* = \bar{y} - b^* \bar{x}$$

ovvero

$$a^* = \sum_{i=1}^{n} d_i \; y_i \qquad con \qquad d_i = \sum_{i=1}^{n} (\frac{1}{n} - c_i \bar{x}) y_i$$

Ricaviamo allora le varianze associate ai parametri a^* e b^*. Ribadiamo qui che le y_i sono tra loro indipendenti, e per ragioni di semplicitá é stata fatta l'ipotesi

che abbiano tutte la stessa varianza σ^2. Avremo allora

$$Var[b^*] = Var[\sum_{i=1}^{n} c_i \; y_i] = \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Procedendo in modo analogo per a^* troviamo

$$Var[a^*] = \sigma^2 \sum_{i=1}^{n} d_i^2 = \sigma^2[\frac{1}{n} + \frac{\bar{x^2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}]$$

Notiamo come le varianze di a^* e b^* dipendano dalla dispersione dei valori di ascissa dei dati, $(x_i - \bar{x})^2$; quindi la stima della pendenza e dell' intercetta della retta di regressione sará tanto *migliore* quanto piú saranno *dispersi* i dati in x.

Stabilite le proprietá statistiche dei due parametri della retta di regressione, proviamo ora a definire la significativitá complessiva della retta stimata. Iniziamo con il riscrivere il termine della varianza $S_c^2 = \frac{1}{n}\sum_i (y_c^i - \bar{y})^2$, in funzione dei valori assunti dalla x. Troveremo che

$$S_c^2 = \frac{1}{n} b^{*2} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

dato che le y_c^i verificano la condizione $y_c^i - \bar{y} = b^*(x_i - \bar{x})$. Per fissati valori di x_i la varianza S_c^2 risulta dipendere dalla sola variabile aleatoria normale b^{*2}, quindi ha un sol grado di libertá. Poiché abbiamo giá dimostrato a suo tempo che la varianza empirica complessiva s_y^2 ha $n-1$ gradi di libertá, allora per sottrazione concludiamo che S_y^2 ne avrá $n-2$.

Abbiamo visto che le varianze di a^* e b^* dipendono da σ^2. Tuttavia, non é affatto detto che si abbia

l'informazione relativa al valore di σ. Per definire allora un intervallo fiduciario realtivo ai due parametri procederemo definendo degli opportuni stimatori.
A questo scopo indichiamo con a ed b i valori aspettati di a^* e b^*. Per la pendenza della retta di regressione b introduciamo lo stimatore T che, per mettere in evidenza le sue propretá statistiche, scriveremo nella seguente forma:

$$T = \sqrt{n-2}\frac{\frac{b^*-b}{\frac{\sigma}{\sqrt{\sum_i (x_i-\bar{x})^2}}}}{\sqrt{\frac{(n-2)s_y^2}{\sigma^2}}} = \frac{b^*-b}{\frac{s_y}{\sqrt{\sum_i (x_i-\bar{x})^2}}}$$

Avendo scritto lo stimatore in questa forma, dovrebbe risultare evidente al lettore che il numeratore di T é una variabile aleatoria normale ridotta, il denominatore é la radice quadrata di una variabile del tipo χ^2_{n-2} ad $n-2$ gradi di libertá. Ne consegue che T é una variabile di Student ad $n-2$ gradi di libertá.

Possiamo allora costruire l'intervallo fiduciario per b avendo scelto a priori il livello di significativitá dell'intervallo $1-\alpha$. Quindi partendo dalla condizione

$$P[-t_{\alpha/2,(n-2)} \leq T \leq t_{\alpha/2,(n-2)}] = 1-\alpha$$

deduciamo, utilizzando la tabella della funzione cumulativa del T di Student a $n-2$ gradi di libertá, il valore corrispondendente $t_{\alpha/2,n-2}$, e quindi l'intervallo fiduciario Δb^* associato alla stima b^*

$$\Delta b^* = t_{\alpha/2,(n-2)} \ \frac{s_y}{\sqrt{\sum_i (x_i-\bar{x})^2}}$$

ovvero

$$b^* - t_{\alpha/2,(n-2)} \frac{s_y}{\sqrt{\sum_i (x_i - \bar{x})^2}} \leq b \leq b^* + t_{\alpha/2,(n-2)} \frac{s_y}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

Per l'intercetta a^* possiamo seguire lo stesso sviluppo logico, introducendo uno stimatore di Student ad $n - 2$ gradi di libertá e concludendo che il corrispondente intervallo fiduciario é

$$\Delta a^* = t_{\alpha/2,(n-2)} \; s_y \frac{\sqrt{\sum_i x_i^2}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

ovvero

$$a^* - t_{\alpha/2,(n-2)} \; s_y \frac{\sqrt{\sum_i x_i^2}}{\sqrt{\sum_i (x_i - \bar{x})^2}} \leq a \leq a^* + t_{\alpha/2,(n-2)} \; s_y \frac{\sqrt{\sum_i x_i^2}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

INTERVALLO FIDUCIARIO DEL VALORE MEDIO DI Y PER UN ASSEGNATO VALORE DI x .

Molto spesso si presenta il problema di dover prevedere qual sia il valore che assumerá la variabile Y per un assegnato valore x od occorre affrontare l'analoga questione di quale sia l'intervallo fiduciario associabile alla media di Y quando x assume un determinato valore. Per fare un esempio, potremmo chiederci, osservando l'andamento nel tempo della tensione ai capi di un condensatore in fase di scarica, quale possa essere il valore della differenza di potenziale che raggiungerá dopo 35 secondi, sapendo che la legge di scarica é del tipo $V_o e^{-\frac{t}{\tau}}$, ed avendo stimato i parametri incogniti quali V_o e τ sulla base di una regressione lineare del tipo $y = a^* + b^* x$ dove $y = \log(V)$ e $b\ x = -\frac{t}{\tau}$.

Supponiamo che le y_i considerate siano di nuovo trattabili come variabili a distribuzione normale con la stessa varianza σ.
Uno stimatore, funzione del valore di tensione da prevedere in corrispondenza del determinato valore x^*, é

$$\bar{Y}^* = a^* + b^* x_o$$

Questo stimatore é una combinazione lineare di variabili aleatorie normali. Infatti si ha

$$\bar{Y}^* = \sum_{i=1}^{n} d_i y_i + x^* \sum_{i=1}^{n} c_i y_i = \sum_{i=1}^{n} (d_i y_i + c_i x^*) y_i$$

che ha come valore aspettato

$$E[\bar{Y}^*] = a + bx_o$$

dove a e b sono i valori aspettati di a^* e b^*; la sua varianza é:

$$\sigma^2_{\bar{Y}^*} = \sigma^2 \sum_{i=1}^{n} (d_i + c_i x_o)^2 =$$

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

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

Una volta definita la normalitá della variabile $\bar{Y}^*$, la sua varianza ed il suo valore aspettato possiamo costruire la variabile T di Student definita come il rapporto tra la variabile normale *ridotta*

$$\frac{\bar{Y}^* - a - bx_o}{\sigma^2_{\bar{Y}^*}}$$

e la variabile χ^2_{n-2}

$$(n-2)\frac{S_y^2}{\sigma^2}$$

Esplicitando tale stimatore, otteniamo

$$T = \frac{\bar{Y}^* - a - bx_o}{S_y\sqrt{\frac{1}{n} + \frac{x_o - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}}}$$

e quindi possiamo ricavare l' intervallo fiduciario connesso al livello di probabilitá $1 - \alpha$ nella forma

$$\bar{Y}^* \pm t_{\alpha/2,n-2}\ S_y \sqrt{\frac{1}{n} + \frac{x_o - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$