

# L'ANALISI STATISTICA

## Introduzione

Contrariamente a quanto saremmo portati a pensare, la possibilità di attingere ad una massa sterminata di informazioni rischia di impedirci di fatto di utilizzarne anche solo una parte: non basta infatti avere solo l'accesso *teorico* ad una informazione, ma occorre che essa sia effettivamente fruibile.

E' forse questo il problema centrale della **statistica**: rendere davvero utilizzabili grandi quantità di informazioni, teoricamente disponibili, ma di fatto difficilmente gestibili, relative agli oggetti della propria indagine. Infatti tutte le informazioni - per contribuire effettivamente ad accrescere la conoscenza di un fenomeno - hanno bisogno di essere *trattate* da vari punti di vista: occorrono tecniche accurate di rilevazione, occorre procedere a accurate **selezioni**, occorre un lavoro di **organizzazione** e di **sintesi**. D'altra parte il lavoro statistico ha senso solo se si confronta con grandi quantità di informazioni.

Ricapitolando la statistica raccoglie e restituisce in forma organizzata grandi quantità di informazioni. Nel fare ciò "obbedisce" ad una duplice esigenza: quella *predittiva* e quella *descrittiva*.

Ogni comunità sente il bisogno - a fini di documentazione - di raccogliere una serie di dati sugli usi, sui costumi, sulle attività sociali e economiche dei suoi componenti; i censimenti costituiscono uno strumento fondamentale attraverso cui la statistica esplica questa funzione. L'immagine della complessità sociale che ne risulta è parziale e selettiva, ma proprio per questo ha una sua efficacia. Oltre al carattere descrittivo, un'esigenza, forse la principale, a cui risponde la statistica è quella *predittiva*: la raccolta e l'elaborazione dei dati, e quindi la "fotografia" del passato e del presente, serve per prevedere i comportamenti futuri, per operare scelte, per assumere decisioni. La statistica, mettendo i dati raccolti ed elaborati a disposizione delle attività di previsione, fornisce i presupposti conoscitivi per orientarsi secondo criteri ragionevoli (anche se non privi di un margine di indeterminazione) nelle situazioni in cui la quantità di informazioni effettivamente utilizzabili si rivela insufficiente a garantire sicurezze.

Si presti infine attenzione al fatto che, essendo l'elaborazione statistica frutto di varie operazioni (selezione, organizzazione e sintesi) può capitare che errori commessi casualmente (o volutamente) in una fase qualsiasi dell'interpretazione dei dati contribuiscano a fornirci un quadro, in parte o completamente, distorto del fenomeno analizzato.

# TRAPPOLE STATISTICHE

Con la dicitura *trappole statistiche* intendiamo tutte le fonti di errore o, nel caso in cui vengano sfruttate consciamente, di inganno nella interpretazione dei dati.

Infatti, poiché sappiamo che la statistica serve ad interpretare una realtà fatta di elevatissimi numeri di elementi, abbiamo deciso di fissare la nostra attenzione sulla difficoltà "percettiva" dei numeri, sulle variazioni dovute a trasformazioni numeriche o cambiamenti di scala e sugli effetti di diversi tipi di campionamento.

Consideriamo, ad esempio, le trasformazioni numeriche dei dati di partenza. Esse si riflettono molto spesso in un mascheramento se non talora in una distorsione della realtà: si pensi ad esempio a come cambia l'aspetto di un grafico a seconda che si usi una scala lineare o logaritmica oppure si grafichi una nuova variabile traslata rispetto all'origine. Inoltre lo zoom su una porzione della figura può amplificare o ridurre gli effetti interpretativi del fenomeno descritto.

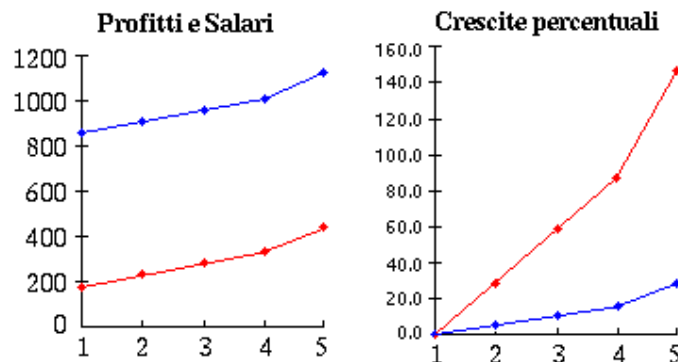
Un'altra trasformazione numerica che può creare problemi è la trasposizione dei dati iniziali in forma percentuale: vediamo a tal proposito un esempio.

Le seguenti tabelle mostrano l'andamento di ipotetici salari e profitti di un'azienda nel corso di cinque anni e i rispettivi incrementi in forma percentuale, riferiti al dato del primo anno.

Anni	Profitti	Salari
1	170	850
2	220	900
3	270	950
4	320	1000
5	420	1100

Anni	+% Profitti	+% Salari
1	0.0	0.0
2	29.4	5.9
3	58.8	11.8
4	88.2	17.6
5	147.1	29.4

La "diversità" tra i due andamenti (la crescita salari - profitti e la crescita delle relative percentuali) si evidenzia ancora meglio nel momento in cui andiamo a rappresentarli graficamente: in blu abbiamo disegnato i salari e la loro crescita percentuale, mentre in rosso abbiamo graficato i dati relativi ai profitti.



Si nota facilmente la differenza tra i due tipi di andamenti.....

Questo piccolo "trucchetto" viene molto spesso usato dai mezzi di informazione per dare una diversa connotazione alla medesima notizia.

# MEDIA, MODA, MEDIANA E MOMENTI

- Media
- Moda
- Mediana
- Momenti

## LA MEDIA

Quando della stessa grandezza si possiede un campione di valori frutto di  $N$  misurazioni si possono "riassumere" le informazioni inerenti alla grandezza derivanti dalle singole misure attraverso la media che, in questo caso, costituisce la miglior stima possibile per la grandezza in esame.

Se le singole misure si possono considerare equivalenti l'una all'altra senza che ve ne siano di alcune più importanti o privilegiate, allora definiamo la media aritmetica come segue:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{\sum x_i}{N}$$

Nel caso in cui i singoli dati abbiano pesi diversi, allora si ricorre a quella che viene definita media pesata. Esistono anche altre definizioni di media.

## LA MODA

Per quanto riguarda una variabile aleatoria, si definisce *moda* il valore più probabile che questa può assumere: quando trattiamo campioni di dati frutto ad esempio di diverse misure della stessa grandezza, allora definiamo la *moda* come il valore più "popolare" del campione. Per capire meglio questa definizione pensiamo ad un istogramma: la moda è costituita in questo caso dal valore corrispondente alla colonna più alta.

## LA MEDIANA

La mediana è, ad esempio in un istogramma, l'ascissa corrispondente al punto in cui l'area delimitata dall'istogramma si divide in due parti uguali: in pratica il numero di dati che sta alla destra della mediana (quelli maggiori) è uguale al numero di dati alla sinistra della mediana (quelli minori).

## I MOMENTI

Accanto alle cosiddette *caratteristiche di posizione* quali la media, la moda e la mediana, esistono altre caratteristiche, ciascuna delle quali descrive una proprietà della distribuzione in esame. I **momenti** forniscono indicazioni sulla dispersione rispetto al valore centrale, sull'asimmetria della distribuzione, ecc.: nelle applicazioni pratiche si ha a che fare con *momenti iniziali* e con *momenti centrali*.

# LA SCELTA DEI DATI

Quando si hanno a disposizione diverse misure della stessa grandezza, può accadere che una o più misure siano in disaccordo con le restanti: quando ciò si verifica entra in gioco in modo preponderante la figura dello sperimentatore.

E' lui infatti che dovrà decidere il dato è compatibile o meno con gli altri e di conseguenza se conservarlo o *rigettarlo*.

Il fatto che in ultima analisi sia lo sperimentatore a decidere delle sorti del dato fa sì che il "criterio" per il rigetto o meno dei dati sia, alla fine, soggettivo e quindi diventi quasi impossibile delinearne un quadro preciso.

In verità esistono dei criteri già formalizzati e accettati dalla maggior parte degli scienziati: l'influenza dello sperimentatore, al momento dell'utilizzo di questi metodi risiede nello stabilire quale sia la "soglia di accettabilità" del dato.

Abbiamo detto che esistono dei criteri già formulati e, nella maggior parte dei casi, accettati tra cui citiamo:

- Il criterio di Chauvenet
- Il criterio "*a priori*"

E inoltre, ma merita un discorso a parte:

- Il test  $\chi^2$

Anche se attraverso metodi e procedimenti differenti, in linea di principio questi criteri forniscono a chi li usa una stima di quanto un dato si discosta dagli altri: è qui che interviene lo sperimentatore definendo la soglia oltre la quale un dato deve essere rigettato o meno. Tale soglia dipende da diversi fattori tra cui la metodologia e le tecniche di misura adottate.

La determinazione del valore oltre il quale scartare può rappresentare un'arma a doppio taglio per due motivi fondamentali.

Come prima cosa uno scienziato può essere tacciato di voler pilotare i dati del suo esperimento, scartando quelli che non sono in accordo con le sue previsioni teoriche e in secondo luogo esiste il rischio di scartare eventi considerandoli frutto di qualche errore, quando questi potrebbero essere il segnale di qualche importante fenomeno che lo scienziato aveva trascurato o del quale non era a conoscenza.

A questo proposito va sottolineato che molte importanti scoperte sono state frutto di misure all'apparenza anomale. Tali misure, prima di essere scartate, sono state analizzate in modo più approfondito: si è quindi visto che, quello che sembrava un evento scarsamente significativo, rappresentava invece un segnale di una realtà diversa da come la si era formalizzata.

# LA SCELTA DEI DATI

## Il criterio di Chauvenet

Per capire il criterio di Chauvenet per il rigetto dei dati consideriamo un esempio. Supponiamo di aver effettuato dieci misure di una certa grandezza  $X$  e di averle riassunte nella seguente tabella:

1	2	3	4	5	6	7	8
14.1	13.4	13.8	13.0	11.8	14.1	13.0	14.0

Se ora procediamo al calcolo della media ( $\bar{x}$ ) e della deviazione standard ( $\sigma$ ) troviamo i valori:

$$\bar{x} = 13.4$$

$$\sigma = 0.8$$

In questa serie di misure il quinto valore (11.8) è decisamente in disaccordo con tutti gli altri: vediamo come procedere nei confronti di tale valore.

Dobbiamo prima di tutto quantificare quanto la misura in questione sia anomala rispetto alle altre: per fare questo, notiamo che il valore 11.8 si discosta dal valor medio di due volte la deviazione standard.

Se assumiamo che le misure si conformino ad una distribuzione di Gauss avente media  $\bar{x}$  e deviazione standard  $\sigma$  allora siamo in grado di calcolare la probabilità di avere misure che differiscano dalla media di almeno due deviazioni standard.

La probabilità di avere tali misure si ottiene, secondo la proprietà degli eventi contrari, sottraendo da uno (il 100% rappresentante la globalità degli eventi) la probabilità di ottenere risultati *entro* due deviazioni standard, cioè:

$$\overline{P_{2\sigma}} = 1 - P_{2\sigma}$$

dove con  $\overline{P_{2\sigma}}$  si intende appunto la probabilità di ottenere valori al di fuori di  $2\sigma$  e con  $P_{2\sigma}$  la probabilità di ottenerne entro  $2\sigma$ .

Da quanto detto, andando a vedere il valore tabulato di  $P_{2\sigma}$ , otteniamo:

$$\overline{P_{2\sigma}} = 1 - 0.95 = 0.05$$

In pratica abbiamo il 5% di probabilità di ottenere una misura al di fuori di due deviazioni standard, cioè ci si deve aspettare che una misura su venti si discosti di più di 1.6 unità ( $2\sigma$ ) dal valor medio (che nel nostro caso era 13.4).

Avendo noi eseguito otto misure, per la proprietà delle probabilità di eventi indipendenti, il numero di misure oltre  $2\sigma$  è dato da:

$$n = 0.05 \times 8 = 0.4$$

Significa che mediamente ci si dovrebbe aspettare 2/5 di una misura anomala come il nostro 11.8: in questo modo abbiamo quantificato l'anomalia del valore in questione. Ora si tratta di stabilire quale sia la "soglia di accettabilità" per i dati dopodichè andiamo a vedere se il dato incriminato deve essere rigettato o meno.

Di solito viene stabilita tale soglia ad 1/2, per cui *se il numero atteso (n) di misure anomale è minore di 1/2, la misura sospetta deve essere rigettata*: da questo discende che il nostro valore 11.8 non è da considerarsi ragionevole e quindi deve essere rigettato.

Una volta capito questo esempio, la generalizzazione del criterio ad un problema con più dati è immediata: si supponga di avere  $N$  misure ( $x_1, x_2, \dots, x_n$ ) della stessa grandezza  $X$ . Come prima cosa calcoliamo  $\bar{x}$  e  $\sigma$  dopodichè osserviamo i dati per vedere se esiste qualche valore sospetto. Nel caso ci sia un dato sospetto ( $x_i$ ) calcoliamo il numero di deviazioni standard ( $Z_i$ ) di cui  $x_i$  differisce da  $\bar{x}$  applicando la formula:

$$Z_i = \frac{x_i - \bar{x}}{\sigma}$$

Fatta questa operazione bisogna andare a vedere quale è la probabilità che una misura differisca da  $\bar{x}$  di  $Z_i$  volte la deviazione standard: per fare questo bisogna ricorrere ai valori della probabilità in funzione del numero di deviazioni standard che si trovano facilmente tabulati.

Alla fine, per ottenere il numero ( $n$ ) di misure anomale che ci si aspetta, moltiplichiamo la suddetta probabilità per il numero totale di misure ( $N$ ):

$$n = N \times P(\text{oltre } Z_i \sigma)$$

Se il numero  $n$  è minore di 1/2 allora non si attiene al criterio di Chauvenet e come tale deve essere rigettato.

A questo punto si presenta uno spinoso problema:

come agire con i dati rimasti?

C'è chi sostiene che si debba applicare nuovamente il criterio di Chauvenet ai dati rimasti (tenendo conto che dopo l'eliminazione del primo dato si hanno diversi valori di  $\bar{x}$  e  $\sigma$ ) fintanto che tutti i dati rimasti siano conformi al criterio di Chauvenet, mentre altri sostengono che tale metodo non vada applicato una seconda volta ricalcolando la media e la deviazione standard. Esiste però anche un terzo modo, forse il più equilibrato anche rispetto a coloro che ritengono che il rigetto di un dato non sia mai giustificato, di affrontare il problema: molti scienziati infatti utilizzano il criterio di Chauvenet non per scartare immediatamente il dato, bensì solamente per *individuare* il dato: una volta individuato il valore sospetto si procede alla verifica della sua attendibilità attraverso la riproduzione delle misure e una successiva rianalisi dei dati.

## IL CRITERIO "A PRIORI"

Al solito supponiamo di avere una serie di  $N$  dati di cui calcoliamo la media ( $\bar{x}$ ) e la deviazione standard ( $\sigma$ ): in base al criterio a priori il rigetto o meno dei dati viene eseguito rispetto ad una soglia di accettabilità stabilita precedentemente.

In questo modo si fissa un intervallo, eventualmente molto ampio, entro il quale i dati vengono accettati, mentre quelli che cadono al di fuori vengono scartati. Ad esempio, nel caso di una distribuzione gaussiana, un intervallo relativo alla probabilità dell'1% ha una semiampiezza di 3.29  $\sigma$ . La larghezza dell'intervallo va anche considerata in base al numero di misure che si effettuano: se si ha un numero di misure molto inferiori a 1000 si può ritenere che il fatto che un dato sia eventualmente esterno all'intervallo non sia dovuto alla sola fluttuazione statistica, mentre se si è in presenza di qualche migliaio di dati ci si deve aspettare per ragioni statistiche che un certo numero di valori capiti al di fuori dell'intervallo.

# LA SCELTA DEI DATI

## Il criterio del $\chi^2$

Il test del  $\chi^2$  (o, come talvolta viene chiamato, di Pearson) viene usato come una sorta di "verifica delle ipotesi".

Le ipotesi in questione possono riguardare semplicemente la presenza o meno di una correlazione tra diverse variabili (in questo caso si parla di *verifica dell'indipendenza*); oppure riferirsi alla distribuzione teorico-matematica che meglio riproduce i dati sperimentali (allora si parla di *verifica dell'aggiustamento*).

In entrambe i casi il problema è quello di paragonare i risultati sperimentali con le previsioni teoriche e di valutare la distanza globale tra i due insiemi sommando i contributi di ciascun elemento. Questi contributi sono valutati percentualmente, cioè come rapporto tra lo scarto (*differenza tra risultato e previsione*) e la previsione. Gli scarti sono successivamente elevati al quadrato per sopprimere il segno.

La distanza globale che cerchiamo, detta appunto  $\chi^2$ , è data da:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - T_i)^2}{T_i}$$

dove  $O_i$  sono i risultati osservati e  $T_i$  le previsioni. Il  $\chi^2$  è tanto più grande quanto meno l'osservazione è compatibile con le ipotesi. Siccome il  $\chi^2$  è calcolato utilizzando dati sperimentali è esso stesso una variabile aleatoria che segue, appunto, la distribuzione  $\chi^2$  nel caso in cui i dati seguano la distribuzione normale.

Esistono tabelle e grafici che forniscono, in funzione del numero di gradi di libertà dell'insieme, cioè del numero di contributi che si sommano, i livelli di confidenza che corrispondono a diversi valori di  $\chi^2$ , dicono cioè quale è la probabilità di ottenere un  $\chi^2$  maggiore o uguale a quello osservato se l'ipotesi è vera.

L'uso del  $\chi^2$  come verifica dell'aggiustamento è concettualmente simile al procedimento che abbiamo descritto. Si costruisce l'istogramma dei dati e si ipotizza che essi seguano una certa distribuzione teorica. Per verificare l'ipotesi si calcola il  $\chi^2$  secondo la stessa definizione data sopra, utilizzando come  $O_i$  le altezze delle colonne dell'istogramma e come  $T_i$  i corrispondenti valori teorici. Anche in questo caso il  $\chi^2$  ottenuto è una variabile aleatoria e bisogna consultare le tabelle per poterne derivare una conclusione quantitativa in termini di probabilità. Per la verifica dell'aggiustamento il numero di gradi di libertà è dato da

$$n = k - n_p - 1$$

dove  $k$  è il numero di colonne dell'istogramma e  $n_p$  il numero di parametri della distribuzione teorica (ad esempio 3 per la distribuzione di Gauss, 2 per la distribuzione di Poisson, ...). Se il  $\chi^2$  ottenuto è inferiore al valore che secondo le tabelle corrisponde ad un livello di confidenza maggiore o uguale al 95%, di solito l'ipotesi viene accettata: questo significa infatti che, *se i dati seguono effettivamente la distribuzione teorica che stiamo verificando*, la probabilità di avere per fluttuazione statistica un  $\chi^2$  minore di quello osservato è soltanto del 5%.

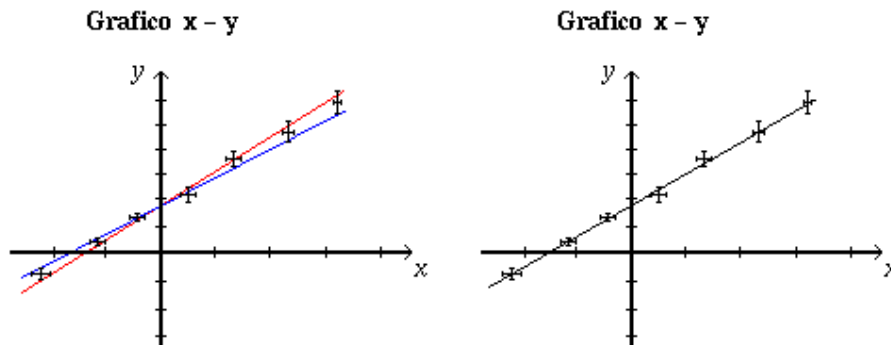
# RAPPRESENTAZIONE GRAFICA DEI DATI

## L'interpolazione

Supponiamo di essere abbastanza soddisfatti della linearità mostrata da un ipotetico insieme di dati: il passo successivo è di chiedersi quale sia la retta che meglio modella i punti sul grafico e successivamente tentare di scoprire quale sia l'equazione che rappresenta la relazione tra le due grandezze.

A causa delle incertezze sperimentali, difficilmente tutti i punti di un grafico giacciono su una retta per cui spetta all'abilità dello sperimentatore trovare quale sia la retta che meglio si adatta alla distribuzione dei punti, ossia quella retta che meglio **interpola** i punti del grafico.

Tale retta è detta **retta di best fit**. Nel primo grafico sottostante sono state disegnate le *rette limite*, le rette, cioè, aventi la massima e la minima pendenza tra tutte quelle che si adattano ai punti del grafico, mentre il secondo rappresenta proprio la retta di best fit.



# RAPPRESENTAZIONE GRAFICA DEI DATI

## La regressione lineare

Occupiamoci ora del metodo analitico per ricavare la miglior linea retta che interpola una serie di punti sperimentali, metodo chiamato *regressione lineare*.

Per semplificare la trattazione, assumeremo d'ora in poi che le incertezze si abbiano solo sulle grandezze in ordinata e che tutte le incertezze siano uguali. Altro assunto (peraltro ragionevole) che faremo è che ogni misura in  $y$  sia governata dalla distribuzione di Gauss, con lo stesso parametro  $\sigma_y$  per tutte le misure.

Quello che in pratica vogliamo determinare sono le due costanti **A** e **B** che determinano la retta migliore avente la forma

$$y = Ax + B$$

Se conoscessimo le costanti **A** e **B** allora per ogni singolo  $x_i$  potremmo calcolare il corrispondente valore vero di  $y_i$  come

$$y_i = Ax_i + B$$

poichè la misura  $y_i$  è governata da una distribuzione normale centrata su questo valore vero con parametro  $\sigma_y$ , la probabilità di ottenere il valore osservato  $y_i$  è

$$P_{A,B}(y_i) \propto \frac{1}{\sigma_y} e^{-\frac{(y_i - Ax_i - B)^2}{2\sigma_y^2}}$$

dove i pedici **A** e **B** indicano che questa probabilità dipende dai loro valori che sono incogniti. La probabilità di ottenere il nostro insieme completo di valori osservati  $y_1 \dots y_N$  è

$$P_{A,B}(y_1, \dots, y_N) = P_{A,B}(y_1) \cdot \dots \cdot P_{A,B}(y_N) \propto \frac{1}{\sigma_y^N} e^{-\frac{\chi^2}{2}}$$

dove l'esponente è dato da

$$\chi^2 = \sum_{i=1}^N \frac{(y_i - Ax_i - B)^2}{\sigma_y^2}$$

Le migliori stime per le costanti si ottengono imponendo che la probabilità sia massima, in quanto gli  $y_i$  sono i valori effettivamente osservati: questo equivale ad imporre che la somma dei quadrati nell'esponente sia minima. Per trovare tali valori differenziamo rispetto ad **A** e **B** e poniamo le derivate uguali a zero:

$$\frac{\partial \chi^2}{\partial A} = \left(-\frac{2}{\sigma_y^2}\right) \sum_{i=1}^N x_i (y_i - Ax_i - B) = 0$$

e

$$\frac{\partial \chi^2}{\partial B} = \left(-\frac{2}{\sigma_y^2}\right) \sum_{i=1}^N (y_i - Ax_i - B) = 0$$

Queste due equazioni possono essere riscritte come:

$$A \sum x_i + B N = \sum y_i$$

e

$$A \sum x_i^2 + B \sum x_i = \sum x_i y_i$$

Tali equazioni, note anche come *equazioni normali*, una volta risolte, danno la stima delle costanti **A** e **B**

$$A = \frac{N(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\Delta}$$

e

$$B = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{\Delta}$$

dove abbiamo introdotto l'abbreviazione

$$\Delta = N(\sum x_i^2) - (\sum x_i)^2$$

Calcolate le miglior stime delle costanti **A** e **B** dai valori misurati  $(x_i, y_i)$  è spontaneo chiedersi quali siano le incertezze nelle nostre stime.

# RAPPRESENTAZIONE GRAFICA DEI DATI

## Adattamento ad altre curve col metodo dei minimi quadrati

Il caso di due variabili che soddisfano una relazione lineare del tipo  $y = Ax + B$  non è altro che un caso particolare di una vasta classe di problemi che riguardano le curve di adattamento, molti dei quali possono essere risolti in un modo simile. Vediamo ora alcuni di questi problemi.

### Adattamento ad una polinomiale

Spesso accade che ci si aspetti che una variabile  $y$  sia esprimibile come una funzione polinomiale di una seconda variabile  $x$ , ossia

$$y = A + Bx + Cx^2 + \dots + Wx^n$$

Dato un insieme di osservazioni delle due variabili, si può trovare la migliore stima delle  $A, B, \dots, H$  con un procedimento analogo a quello con cui abbiamo stimato le costanti  $A$  e  $B$  nel caso della relazione lineare. Per semplificare la trattazione supponiamo che la polinomiale sia di fatto una quadratica del tipo

$$y = Ax^2 + Bx + C$$

Al solito supponiamo di avere una serie di misure con incertezze solo sulla variabile dipendente ( $y$ ) e che tali incertezze siano tutte uguali. Allora per ogni  $x_i$ , il corrispondente valore vero di  $y_i$  è dato dalla funzione quadratica con  $A, B$  e  $C$  incogniti. Assumendo che le misure degli  $y_i$  siano governate da distribuzioni normali, ciascuna centrata sull'appropriato valore vero e tutte con lo stesso  $\sigma_y$ , possiamo calcolare la probabilità di ottenere proprio i valori osservati  $y_1 \dots y_N$  come

$$P(y_1, \dots, y_N) \propto e^{-\frac{\chi^2}{2}}$$

dove ora

$$\chi^2 = \sum_{i=1}^N \frac{(y_i - Ax_i^2 - Bx_i - C)^2}{\sigma_y^2}$$

La miglior stima per  $A, B$  e  $C$  è data da quei valori per cui la probabilità è massima, ossia quando l'esponente  $\chi^2$  è minimo.

Differenziando  $\chi^2$  rispetto a  $A, B$  e  $C$  e ponendo queste derivate uguali a 0, otteniamo le tre "equazioni normali"

$$A \sum x_i^2 + B \sum x_i + CN = \sum y_i$$

$$A \sum x_i^3 + B \sum x_i^2 + C \sum x_i = \sum x_i y_i$$

$$A \sum x_i^4 + B \sum x_i^3 + C \sum x_i^2 = \sum x_i^2 y_i$$

Per un dato insieme di misure queste equazioni simultanee per **A**, **B** e **C** possono essere risolte per trovare le miglior stime di **A**, **B** e **C**. Con **A**, **B** e **C** calcolati in questo modo l'equazione

$$y = Ax^2 + Bx + C$$

è chiamata *adattamento polinomiale dei minimi quadrati* per le misure date. Sfortunatamente, soprattutto quando il grado delle polinomiali aumenta, non sempre le equazioni normali sono risolvibili oppure sono estremamente difficili, nonostante questo esiste una grande classe di problemi che possono essere risolti.

## Funzioni esponenziali

Data l'importanza della funzione esponenziale in fisica, vediamo anche questo caso. Consideriamo la funzione

$$y = Ae^{Bx}$$

Per ricondurci ad una forma a noi più familiare, applichiamo una "linearizzazione" della funzione applicando la funzione logaritmo: in questo modo otteniamo

$$\ln y = \ln A + Bx$$

Questa operazione, anche se non rende lineare  $y$  nei confronti di  $x$ , fa sì che il logaritmo di  $y$  lo sia. L'utilità di questa nuova relazione è facilmente verificabile. Se riteniamo che  $x$  e  $y$  debbano soddisfare l'equazione esponenziale, allora le variabili  $x$  e  $z = \ln(y)$  dovrebbero soddisfare

$$z = \ln A + Bx$$

Se abbiamo una serie di misure  $(x,y)$ , allora per ogni  $y_i$  calcoliamo  $z = \ln(y)$ . Le coppie  $(z_i, y_i)$  dovrebbero giacere su una linea retta: tale retta può essere adattata con il metodo dei minimi quadrati, come abbiamo già visto, per ricavare le stime delle costanti **A** e **B**.

## Regressione multipla

Dopo aver parlato di problemi a due variabili, vediamo un esempio con tre variabili (in molti problemi reali più di due variabili che vanno considerate: si pensi ad esempio alla relazione tra pressione volume e temperatura nei gas).

L'esempio più semplice (noi ci limiteremo a questo) è quello in cui una variabile dipende linearmente dalle altre due:

$$z = A + Bx + Cy$$

Se abbiamo una serie di misure  $(x_i, y_i, z_i)$  con tutti gli  $z_i$  ugualmente incerti e gli  $x_i$  e  $y_i$  privi di incertezza, possiamo procedere in modo analogo al caso dell'adattamento ad una retta: in questo caso la miglior stima per le costanti è data dalle equazioni normali

$$AN + B \sum x_i + C \sum y_i = \sum z_i$$

$$A \sum x_i + B \sum x_i^2 + C \sum x_i y_i = \sum x_i z_i$$

$$A \sum y_i + B \sum x_i y_i + C \sum y_i^2 = \sum y_i z_i$$

Questo metodo è chiamato *regressione multipla*