

# Appunti di Teoria degli Errori

Marzo 2002

# Indice

<b>1</b>	<b>Errori e loro propagazione</b>	<b>2</b>
1.1	Introduzione al problema . . . . .	2
1.2	Origine e stima degli errori . . . . .	3
1.3	Come rappresentare gli errori . . . . .	4
1.3.1	Cifre significative . . . . .	4
1.3.2	Errori relativi . . . . .	4
1.4	Propagazione degli errori . . . . .	5
1.4.1	Somma e differenza . . . . .	5
1.4.2	Moltiplicazione e divisione . . . . .	6
1.4.3	Errori indipendenti e somma quadratica . . . . .	7
1.4.4	funzioni di una variabile . . . . .	7
1.4.5	Formula generale . . . . .	7
<b>2</b>	<b>Media, deviazione standard e distribuzione normale</b>	<b>9</b>
2.1	Errori casuali e sistematici . . . . .	9
2.2	La media, la deviazione standard, la deviazione standard della media . . . . .	9
2.3	Istogrammi e distribuzioni . . . . .	11
2.3.1	Distribuzione limite . . . . .	13
<b>3</b>	<b>Regressione lineare</b>	<b>16</b>
3.1	Relazioni lineari . . . . .	16
3.2	Metodo dei minimi quadrati . . . . .	17
3.2.1	Errori nella misura di $y$ . . . . .	18
3.2.2	Errore nella valutazione di $A$ e $B$ . . . . .	19
3.3	Relazioni non lineari . . . . .	19
3.4	Coefficiente di correlazione lineare . . . . .	20

# Capitolo 1

## Errori e loro propagazione

### 1.1 Introduzione al problema

Taylor,  
1.1-1.4

Quando effettuiamo delle misure è inevitabile incorrere in errori, in incertezze: da questi non possiamo liberarci, però possiamo controllarli.

Per delineare i problemi legati all'effettuazione di misure consideriamo un carpentiere che debba misurare l'altezza di uno stipite per poi costruire una porta: riflettendo su questa situazione impariamo che:

- l'accuratezza della nostra misura dipende dagli strumenti ma anche usando gli strumenti più accurati (metro, micrometro, interferometro...) ci sono dei limiti di **sensibilità** (interferometro  $10^{-9}$ m);
- l'azione di misura deve essere effettuata facendo un bilancio tra necessità ed economia: sarebbe inutile usare un interferometro per misurare lo stipite di una porta d'altra parte devo avere uno strumento sensibile al millimetro;
- è necessario avere sotto controllo l'incertezza della misura: se dico che lo stipite è alto 210 cm devo sapere se questo vuol dire che è  $210 \pm 5$  cm o  $210 \pm 0.3$  cm: la seconda misura sarebbe accettabile per i miei scopi, non la prima.

Altri esempi:

- problema di Archimede: per discriminare due leghe attraverso la misura delle loro densità, devo essere sicuro che l'incertezza con cui misuro le densità, minore della differenza delle densità delle leghe.

- verifica della validità della teoria della relatività generale: si doveva discriminare tra deflessione di luce di una stella di un angolo nullo (meccanica di Newton), di un angolo di  $0.9''$  (relatività ristretta), di un angolo di  $1.8''$  (relatività generale). Il risultato fu  $\alpha = 2'' \pm 0.3''$ .

## 1.2 Origine e stima degli errori

Molteplici sono le sorgenti d'errore: in alcuni casi è facile metterle in evidenza mentre in altri è difficile e fare una stima delle incertezze che provocano:

Taylor,  
1.5-  
1.6,3.1

1. errori legati a letture di scale graduate: incertezze dell'ordine della scala graduata; nella lettura di scale graduate possibile errori di **parallasse**;
2. errori dovuti ad una errata taratura dello strumento di misura rivelabili attraverso il confronto con altri strumenti;
3. errori dovuti a qualche processo non controllabile ma random, come il tempo di reazione nella misura di un intervallo temporale con l'ausilio di un cronometro;
4. incertezza nello stabilire il punto di cui fare la misura come ad esempio nella determinazione del fuoco di una lente devo decidere dove l'immagine a fuoco;
5. ...

Il terzo e quarto tipo di errore sono **casuali** nel senso che spingono il dato misurato con ugual probabilità in su ed in giù rispetto al valore vero e queste incertezze possono essere rivelate attraverso una ripetizione delle misure.

Il secondo tipo di errore è **sistematico** e spinge la misura sempre nella stessa direzione rispetto al valore vero quindi non posso rivelare questo errore ripetendo le misure.

Il primo tipo di errore è di norma casuale ma pu diventare sistematico se non faccio attenzione alla parallasse.

Gli **errori casuali**, evidenziabili attraverso una ripetizione delle misure, saranno oggetto delle nostre attenzioni e conosceremo gli strumenti teorici attraverso cui controllare questi errori.

Laddove possibile è buona norma **ripetere le misure** per mettere in evidenza errori casuali.

## 1.3 Come rappresentare gli errori

Taylor,  
2.1-  
2.2,2.7

Abbiamo detto che da una misura dobbiamo ottenere il valore associato ad una grandezza ma dobbiamo anche stabilire l'incertezza con cui quel valore è stato ottenuto.

Una maniera conveniente di rappresentare quest'informazione è quello di indicare il valore più probabile ( $x_{best}$ , il “migliore” valore di  $x$ ) e due estremi entro i quali siamo *sufficientemente sicuri* che il valore reale si trovi. Allora diciamo:

$$x = x_{best} \pm \delta x \Rightarrow \\ x_{best} - \delta x \leq x \leq x_{best} + \delta x$$

Specificheremo meglio in seguito cosa si deve intendere per sufficientemente certi.

### 1.3.1 Cifre significative

L'errore viene dato con una cifra significativa. Fa eccezione il caso in cui la cifra significativa è 1 ed allora si dà l'errore con 2 cifre significative.

$x_{best}$  viene dato approssimato alle decina o decimale corrispondente alla cifra significative dell'errore.

Esempio:

- Scriveremo  $6.32 \pm 0.03$ , oppure  $6320 \pm 30$
- **Non** si scrive invece né  $6322.37 \pm 30$  né  $6320 \pm 32.45$

### 1.3.2 Errori relativi

L'errore non ci dice tutta la storia: è spesso utile dare l'errore frazionario od errore relativo definito come

$$\left( \begin{array}{c} \text{errore} \\ \text{frazionario} \end{array} \right) = \frac{\delta x}{|x_{best}|}$$

così se devo dichiarare la misura di una lunghezza posso dire

$$l = 50 \pm 1 \text{ cm}$$

oppure

$$l = 50 \pm 2\%$$

Da notare come l'errore relativo è un numero puro, non ha unità di misura.

## 1.4 Propagazione degli errori

Le misure dirette sono poco frequenti, ben più frequentemente le grandezze di interesse sono espresse come funzioni di una o più quantità note con errore.

Ad esempio, se voglio conoscere l'area di una superficie rettangolare, quello che farò sarà misurare la sua base  $b_{best} \pm \delta b$  e la sua altezza  $h_{best} \pm \delta h$ . Verosimilmente mi aspetto che il valore dell'area sia

$$\text{Area}_{best} = b_{best} * h_{best}$$

Ma cosa posso dire sull'incertezza di Area? Come trovare l'errore sulle quantità finali a partire dall'errore sulle quantità misurate?

### 1.4.1 Somma e differenza

Par. 2.5,  
3.2

Supponiamo di voler conoscere una quantità  $q$  non accessibile da una misura diretta ma che sappiamo essere  $q = x + y$  dove le quantità  $x$  e  $y$  siano misurate con errore, e sono dunque note come  $x_B \pm \delta x$  e  $y_B \pm \delta y$ ; l'errore nella somma  $q = x + y$  si può *stimare in modo intuitivo* nel seguente modo: poiché il massimo valore probabile per  $x$  è

$$x_B + \delta x$$

e il massimo per  $y$  è

$$y_B + \delta y$$

il massimo valore probabile per  $q$  è

$$x_B + \delta x + y_B + \delta y$$

Analogamente il minimo valore probabile per  $q$  è

$$x_B - \delta x + y_B - \delta y$$

quindi rappresenteremo  $q = x + y$  come

$$q_B \pm \delta q \simeq (x_B + y_B) \pm (\delta x + \delta y)$$

Seguendo lo stesso ragionamento, si trova che per la differenza di due quantità vale la stessa regola: se  $q = x - y$  allora

$$q_B \pm \delta q \simeq (x_B - y_B) \pm (\delta x + \delta y)$$

Lo stesso risultato vale per un **numero arbitrario di addendi**.

## 1.4.2 Moltiplicazione e divisione

Il ragionamento del punto precedente si può estendere al caso della moltiplicazione. Scriviamo  $x$  e  $y$  in termini degli errori relativi per semplicità:

$$x_B(1 \pm \delta x/|x_B|)$$

$$y_B(1 \pm \delta y/|y_B|)$$

Allora il valore massimo per il prodotto è quello

$$x_B y_B (1 + \delta x/|x_B|)(1 + \delta y/|y_B|)$$

poiché gli errori relativi sono in genere piccoli possiamo trascurare il prodotto e scrivere:

$$\left(1 + \frac{\delta x}{|x_B|}\right)\left(1 + \frac{\delta y}{|y_B|}\right) = 1 + \frac{\delta x}{|x_B|} + \frac{\delta y}{|y_B|} + \frac{\delta x}{|x_B|} \frac{\delta y}{|y_B|} \simeq 1 + \frac{\delta x}{|x_B|} + \frac{\delta y}{|y_B|}$$

analogamente, il valore minimo è

$$x_B y_B (1 - \delta x/|x_B|)(1 - \delta y/|y_B|) = \left(1 - \frac{\delta x}{|x_B|}\right)\left(1 - \frac{\delta y}{|y_B|}\right) \simeq 1 - \left(\frac{\delta x}{|x_B|} + \frac{\delta y}{|y_B|}\right)$$

cosicché

$$\frac{\delta q}{|q_B|} = \frac{\delta x}{|x_B|} + \frac{\delta y}{|y_B|}$$

Lo stesso risultato vale (dopo un'algebra un po' più intricata) per la divisione, e per un numero arbitrario di fattori.

Caso particolare:  $q = Cx$  dove  $B$  non ha errore. dal caso generale deriva  $\delta q/q_B = \delta x/|x_B|$  ossia  $\delta q = q_B/|x_B|\delta x$ ; ma poiché  $q_B = x_B C_B = x_B C$  si ha:

$$\delta q = |C|\delta x$$

Par. 2.7  
e 3.2

### 1.4.3 Errori indipendenti e somma quadratica

Par. 3.3

Nel derivare le relazioni precedenti abbiamo assunto che il “valore massimo più probabile” sia quello ottenuto sommando (o moltiplicando) i valori massimi probabili degli addendi (o dei fattori). Se tuttavia  $x$  e  $y$  sono misurati indipendentemente, è poco probabile che ad una sovrastima di  $x$  corrisponda una sovrastima di  $y$ .

Senza addentrarci per ora nei dettagli, osserviamo che i risultati precedenti vanno così modificati:

$$\begin{aligned} \text{Somma/Sottrazione :} \quad \delta q &\simeq \sum_i \delta x_i && \rightarrow \delta q = (\sum_i \delta x_i^2)^{1/2} \\ \text{Prodotto/Divisione :} \quad \delta q/|q_B| &\simeq \sum_i \delta x_i/|x_{i,B}| && \rightarrow \delta q = (\sum_i (\delta x_i/x_{i,B})^2)^{1/2} \end{aligned}$$

Osserviamo come la **somma in quadratura** degli errori (la radice quadrata della somma dei quadrati) è certamente minore della semplice somma, dunque la stima dell'errore data da  $(\sum_i \delta x_i^2)^{1/2}$  è certamente più piccola di  $q \sum_i \delta x_i$ .

### 1.4.4 funzioni di una variabile

Par. 3.5

Se la grandezza di interesse è una funzione arbitraria (seno, coseno, radice quadrata, ...) seguendo il ragionamento dei paragrafi precedenti per la somma e il prodotto si trova

$$\delta q \simeq q(x_B + \delta x) - q(x_B) \simeq \frac{q(x_B + \delta x) - q(x_B)}{\delta x} d\delta x$$

poiché l'errore è auspicabilmente piccolo si ha:

$$q(x_B + \delta x) \simeq q(x_B) + \delta x \frac{dq}{dx}(x_B)$$

ne segue che

$$\delta q \simeq \delta x \left| \frac{dq}{dx}(x_B) \right|$$

### 1.4.5 Formula generale

Taylor,  
3.9

In generale, data una grandezza che funzione di  $n$  variabili  $q = q(x, y, z, \dots)$  si ha<sup>1</sup>

---

<sup>1</sup>Ricordiamo che effettuare una derivata parziale della funzione  $q(x, y, z, \dots)$  rispetto ad  $x$  ( $\partial q/\partial x$ ) significa effettuare la “solita” derivata della funzione  $q$  rispetto ad  $x$  considerando tutte le altre variabili  $y, z, \dots$  come delle quantità costanti.

$$q_{best} = q(x_{best}, y_{best}, z_{best}, \dots)$$

$$\delta q = \sqrt{\left(\frac{\partial q(x_B, y_B, z_B, \dots)}{\partial x} \delta x\right)^2 + \left(\frac{\partial q(x_B, y_B, z_B, \dots)}{\partial y} \delta y\right)^2 + \dots}$$

# Capitolo 2

## Media, deviazione standard e distribuzione normale

### 2.1 Errori casuali e sistematici

Taylor,  
4.1

Abbiamo detto che gli errori che intervengono in una misura possono essere distinti:

- **casuali**: errori che spingono il valore misurato con ugual probabilità in alto od in basso rispetto al valore vero (es. le misure di intervalli di tempo misurati con cronometro); questi errori **possono essere trattati statisticamente**
- **sistematici**: errori che spingono il valore misurato sempre nella stessa direzione rispetto al valore vero. L'origine di questi errori é solitamente in una sbagliata taratura dello strumento di misura od in una sbagliata procedura di misurazione (es. errori di parallasse). Questi errori **non possono essere trattati statisticamente**

### 2.2 La media, la deviazione standard, la deviazione standard della media

Taylor,  
4.2-4.5

Supponiamo di dover misurare una grandezza  $x$  e, dopo aver ridotto gli errori sistematici a livello trascurabile, di ripetere la misura  $N$  volte e di trovare un set di valori simili ma non uguali:

$$x_1, x_2, x_3, \dots, x_N$$

Dati questi  $N$  valori, qual'è la miglior stima per  $x$ ? Si mostra che la miglior stima è:

$$x_{best} = \bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_N}{N} = \frac{\sum x_i}{N}$$

Dato la ricetta per determinare il miglior valore per  $x$ , come possiamo stimare l'incertezza sulle nostre misure? Come prima cosa possiamo considerare le deviazioni delle singole misure dalla media e cioè  $d_i = x_i - \bar{x}$ . Se le deviazioni sono tutte molto piccole, allora le nostre misure sono presumibilmente precise. Se vogliamo invece stimare l'affidabilità di  $x_{best}$ , abbiamo bisogno di una grandezza definita sulle deviazioni: potremmo pensare di fare la **media** delle deviazioni ma questa è **zero** perché  $d_i$  è a volte positivo ed a volte negativo. Il modo migliore di evitare questo inconveniente è elevare al quadrato ogni  $d_i$  e di questo nuovo insieme di valori fare la media. Definiamo allora **deviazione standard**:

$$\sigma_x = \sqrt{\frac{1}{N} \sum_{i=1}^N d_i^2} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (2.1)$$

Ci sono argomenti teorici per rimpiazzare  $N$  col fattore  $N - 1$  e quindi definire la deviazione standard

$$\sigma_x = \sqrt{\frac{1}{N-1} \sum_{i=1}^N d_i^2} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (2.2)$$

Per distinguere tra le due formule precedenti ci si riferisce alla 2.1 come **deviazione standard della popolazione** e alla 2.2 come **deviazione standard del campione**. La deviazione standard mi dá l'incertezza sulla **singola** misura: mostreremo che vi è la possibilità di affermare che "vi è un 70% di probabilità che una **singola** misura differisca meno di  $\sigma_x$  dal valore vero.

Abbiamo detto che il valore medio è la miglior stima del valore vero della grandezza che vogliamo misurare. Possiamo chiederci quale sia l'affidabilità con cui possiamo fare quella affermazione. Si mostra che l'incertezza nel risultato finale  $x_{best} = \bar{x}$  risulta essere la deviazione standard  $\sigma_x$  divisa per  $\sqrt{N}$ .

Questa grandezza é la **deviazione standard della media** ed é denotata con

$$\sigma_{\bar{x}} = \sigma_x / \sqrt{N}$$

Un punto importante della deviazione standard della media é il fattore  $\sqrt{N}$  nel denominatore. La deviazione  $\sigma_x$  rappresenta l'incertezza nelle singole misure e se dovessimo aumentare il numero di misure la deviazione standard non cambierebbe significativamente. D'altra parte, la deviazione standard della media dovrebbe diminuire all'aumentare di  $N$ . Sfortunatamente il fattore  $\sqrt{N}$  cresce lentamente all'aumentare di  $N$ . Se vogliamo ridurre l'incertezza sul valore medio di  $N$  di un fattore 100, dobbiamo effettuare 10000 misure.

## 2.3 Istogrammi e distribuzioni

Una seria analisi statistica ci richiede di fare molte misure: dobbiamo quindi sviluppare metodi adeguati per registrare e mettere in evidenza un gran numero di valori.

Per descrivere questi metodi supponiamo di effettuare un misura che restituisce dei numeri interi. Una maniera conveniente per registrare questi valori é organizzare una tabella in cui, in corrispondenza del valore misurato riportiamo il numero di volte in cui questo é stato registrato.

Es.

valore di $x_k$	valore di $n_k$
23	1
24	3
25	2
26	3
27	0
28	1

Se registriamo come detto il numero di misure, allora possiamo riscrivere la definizione della media come

$$\bar{x} = \frac{\sum_i x_i}{N} = \frac{\sum_k n_k x_k}{N} = \sum_k F_k x_k$$

Nello scrivere l'ultima uguaglianza abbiamo introdotto la grandezza  $F_k$  definita come:

Taylor,  
5.2

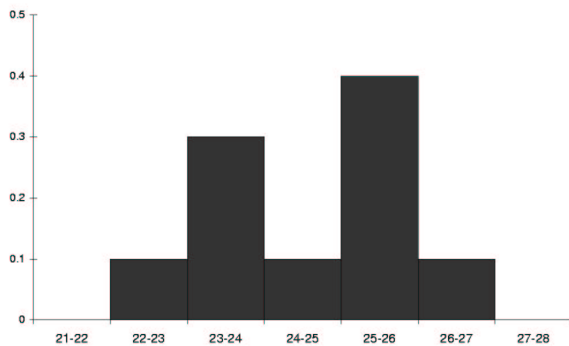


Figura 2.1: Esempio di istogramma a intervalli.

$$F_k = \frac{n_k}{N}$$

Le frazioni  $F_k$  specificano la **distribuzione**  
Ovviamente deve valere

$$\sum_k n_k = N \tag{2.3}$$

e

$$\sum_k F_k = 1$$

Se nella mia misura non vado a registrare valori interi ma ad esempio

26.4, 23.9, 25.1, 24.6, 22.7, 23.8, 25.1, 23.9, 25.3, 25.4

allora meglio costruire un secondo tipo di tabella

---

inter.	22-23	23-24	24-25	25-26	26-27	27-28
eventi						
per inter.	1	3	1	4	1	0
fraz. di						
misure per int.	0.1	0.3	0.1	0.4	0.1	0

---

I risultati di questa tabella si possono mettere in un grafico "istogramma a intervalli" (cf. fig. 2.1). L'altezza  $f_k$  delle barre deve essere tale che  $f_k * \Delta_k =$  è la frazione misure che cadono nell'intervallo  $\Delta_k$

### 2.3.1 Distribuzione limite

Taylor,  
5.1,5.2

Si osserva dall'esperienza e gli statistici hanno anche dimostrato, che per misure soggette ad errori casuali l'istogramma a intervalli tende, all'aumentare del numero delle misure, ad una curva a campana chiamata **distribuzione normale** o **gaussiana**

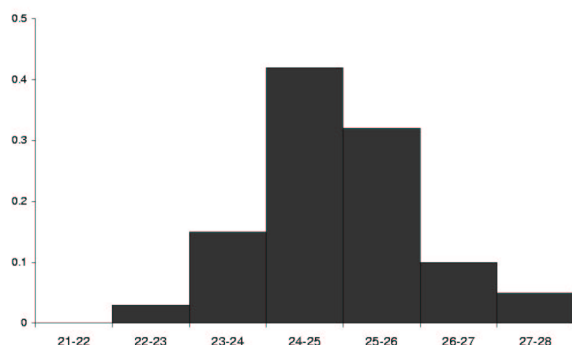


Figura 2.2: Comportamento dell'istogramma a barre all'aumentare del numero delle misure

La distribuzione normale è una curva che rappresenta come si **distribuiscono** le misure soggette ad errori casuali.

Come ogni funzione possiamo rappresentare la distribuzione gaussiana su un piano cartesiano e la sua forma analitica è

$$y_{X,\sigma} = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-X)^2/2\sigma^2}$$

La variabile  $x$  è un possibile valore in uscita dal nostro processo di misura mentre la variabile  $y$  rappresenta la frequenza con cui mi devo aspettare che la variabile  $x$  esca effettivamente dal mio processo di misura. I pedici  $X$  e  $\sigma$  indicano come la funzione gaussiana sia dipendente da questi due parametri:

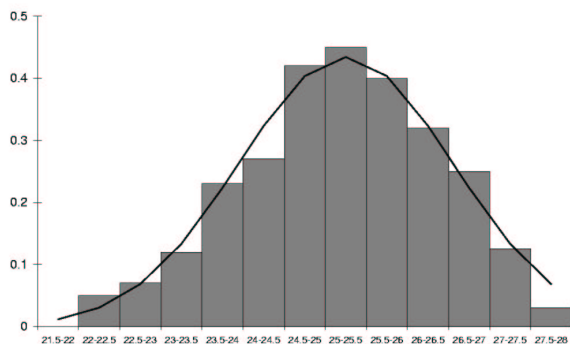


Figura 2.3: Esempio di possibile istogramma a barre e curva gaussiana

per ogni tipo di misura e per ogni grandezza che voglio fissare, per sapere come si distribuiranno i valori in uscita, devo fissare quei due valori.

La figura 2.4 mostra come la gaussiana dipenda dai parametri  $X$  e  $\sigma$ :  $X$  rappresenta l'ascissa attorno a cui è centrata la campana mentre  $\sigma$  governa la larghezza della campana. Un valore di  $\sigma$  grande allarga la campana mentre un valore di  $\sigma$  piccolo restituisce una gaussiana che ha un collo più sottile.

Il parametro  $X$  rappresenta la media delle misure, il parametro  $\sigma$  rappresenta la deviazione standard.

E' spesso utile sapere quale è la probabilità che una certa misura cada in un intervallo centrato attorno al valore medio  $X$ . Per rispondere a questa domanda si usano tabelle del tipo

t	P
0.1	0.08
...	...
1	0.68
3	0.99

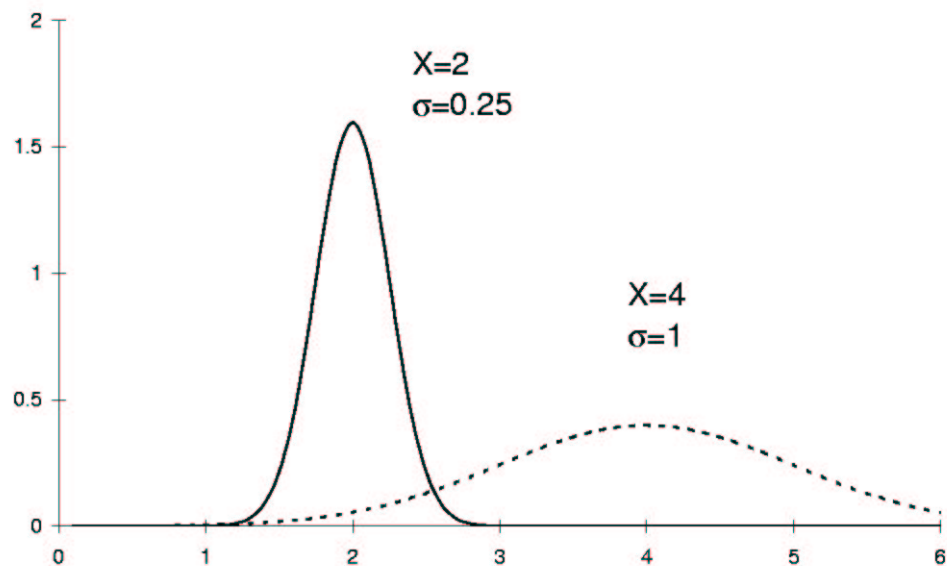


Figura 2.4: Due esempi di curve gaussiane

dove  $t$  è la frazione di  $\sigma$  che dà la larghezza dell'intervallo mentre  $P$  moltiplicato per 100 dà la percentuale che la misura cada nell'intervallo

$$X - t\sigma_x < x < X + t\sigma_x$$

$$P(\text{entro } \sigma) = \int_{\bar{x}-\sigma_x}^{\bar{x}+\sigma_x} \left( \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-X)^2/2\sigma^2} \right) dx = 0.68$$

In riferimento alla tabella precedente possiamo dire che

**La deviazione standard mi definisce quegli intervalli per cui ho il 68% di probabilit che se effettuo una misura questa cade in quell'intervallo**

# Capitolo 3

## Regressione lineare

### 3.1 Relazioni lineari

8.1

Uno dei più importanti compiti degli esperimenti è quello di investigare la relazione tra due variabili. Il caso più importante (e a cui spesso ci si riconduce, come vedremo) è quello in cui la relazione che si intende studiare è lineare:

$$y = A + Bx$$

Ad esempio, se un corpo si muove a velocità costante, con  $x$  denotiamo il tempo trascorso dalla partenza e con  $y$  la distanza percorsa,  $x$  e  $y$  sono legate da una relazione lineare.

Si osservi che qualora le due grandezze in questione *fossero* misurate senza incertezze o errori, una serie di misure di  $x$  e  $y$  porterebbe immediatamente a verificare o rigettare l'ipotesi della relazione lineare. Infatti in tal caso i punti  $(x, y)$  si disporrebbero lungo una retta. Inoltre le costanti  $A$  e  $B$  sarebbero immediatamente determinabili dall'analisi della suddetta retta.

Tuttavia, poichè le grandezze in questione spesso *non sono* note senza incertezze o errori, e dunque le coppie  $(x, y)$  non si disporrebbero su di una retta neppure se *fossero* legati da una relazione lineare; e quindi neppure la linearità della relazione è determinabile; Ci si possono porre le seguenti domande:

- supponendo che la relazione tra le variabili misurate sia lineare, come si possono determinare  $A$  e  $B$ ?

8.2

- come valutare dalla misurazione se l'ipotesi della linearità è verificata?

## 3.2 Metodo dei minimi quadrati

Quanto un dato esperimento si discosta dalla situazione ideale? In assenza di errore per una data  $x_i$  si avrebbe per la  $y$  il valore

$$\hat{y}_i = A + Bx_i.$$

Tuttavia, a causa delle incertezze su  $y$ , il risultato della misura  $y_i$  sarà, in generale diverso da  $\hat{y}_i$ . La misura

$$\sum (\hat{y}_i - y_i)^2$$

(la *somma degli scarti quadratici*) è una misura di quanto i punti misurati si discostano dalla retta  $A + Bx$ , e dunque la retta stessa è tanto “migliore” quanto piccola è la somma suddetta. Qualora le misure non fossero soggette ad errore ( $\hat{y}_i = y_i$ ) si avrebbe  $\sum (\hat{y}_i - y_i)^2 = 0$ .

Intuitivamente ci si aspetta quindi che i migliori parametri  $A$  e  $B$  siano quelli che minimizzano la quantità:

$$\sum (\hat{y}_i - y_i)^2 = \sum (A + Bx_i - y_i)^2$$

e dunque<sup>1</sup>

---

<sup>1</sup>Solo per i più esigenti e senza pretese di rigore un accenno alla dimostrazione è il seguente. La funzione da minimizzare è

$$F(A, B) = \sum_i (A + Bx_i - y_i)^2$$

(Si tenga presente che qui  $x_i, y_i$  sono valori numerici *noti*, mentre le variabili sono  $A$  e  $B$ : si tratta di una funzione di due variabili.) in tal caso il noto teorema di Fermat (condizione solo necessaria!) che nel minimo la derivata sia nulla, diviene che le due derivate parziali si annullino:

$$\frac{\partial F}{\partial A} = \sum_i 2(A + Bx_i - y_i) = 0$$

$$\frac{\partial F}{\partial B} = \sum_i 2x_i(A + Bx_i - y_i) = 0$$

dunque, riscrivendo le somme e risolvendo per  $A$  e  $B$ , si trova il risultato riportato nel testo.

$$A = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{N(\sum x_i^2) - (\sum x_i)^2}$$

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

Tale metodo è (per ovvie ragioni) noto col nome di metodo dei minimi quadrati (*least squares method*). Il risultato intuitivo appena descritto può essere formalizzato (per la cronaca: un risultato chiamato dai matematici col nome di teorema di Gauss-Markov) una volta introdotte le seguenti importanti ipotesi:

- sebbene le misure di  $y$  siano soggette ad incertezza, le misure di  $x$  non lo sono, e sono dunque note con sicurezza.
- l'incertezza sulle misure di  $y$  sia la stessa per tutte le  $y$ ; <sup>2</sup>

La retta così ottenuta è detta retta di regressione lineare.

### 3.2.1 Errori nella misura di $y$

Nel valutare la linearità di certe relazioni, anziché effettuare molte volte la misura di un certo  $y_i$  si preferisce effettuare una misura di molti  $y_i$ . Questo modo di agire rende impraticabile la valutazione dell'errore usando la solita formula della deviazione standard che prevede diverse misure di una sola grandezza.

Taylor,  
8.3

È tuttavia importante avere una valutazione dell'errore che commettiamo nel misurare gli  $y_i$ : una formula ci viene data usando le seguenti ipotesi:

1. la deviazione standard è comune per tutti gli  $y_i$ , ossia  $\sigma_{y_i} = \sigma_y$ ;
2. il valore vero per  $y_i = A + Bx_i$

---

<sup>2</sup>occorre infatti fare la seguente osservazione importante: le misure  $y_i$  non sono misure della stessa grandezza (ad esempio potrebbero essere misure di distanze percorse  $y_i$  a tempi diversi  $x_i$ ). Ha pertanto senso immaginare situazioni in cui l'errore sulle diverse misure delle distanze sia diverso (ad esempio le distanze brevi sono misurate con uno strumento e quelle lunghe con un'altro meno preciso).

Con queste due ipotesi capiamo che anziché effettuare la media della deviazione standard su tante “deviazioni” di una singola misura, possiamo effettuare la media su una deviazione di molte misure cos che:

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

### 3.2.2 Errore nella valutazione di $A$ e $B$

8.4

Qual é l'errore che commettiamo nella valutazione di  $A$  e  $B$ ?

Ricordiamo la situazione:

1.  $A$  e  $B$  sono funzioni di tanti  $x_i$  e tanti  $y_i$

$$A = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{N(\sum x_i^2) - (\sum x_i)^2}$$

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

2.  $x_i$  è noto senza errore ( $\delta x_i = 0$ );
3. tutti gli  $y_i$  sono noti con lo stesso errore ( $\delta y_i = \sigma_y$ ).

Allora per calcolare  $\delta A$  e  $\delta B$  si può usare la **formula di propagazione degli errori**

Risultato:

$$\sigma_A^2 = \sigma_y^2 \sum x_i^2 / \left[ N(\sum x_i^2) - (\sum x_i)^2 \right]$$

$$\sigma_B^2 = N\sigma_y^2 / \left[ N(\sum x_i^2) - (\sum x_i)^2 \right]$$

### 3.3 Relazioni non lineari

Le relazioni lineari, studiate nel paragrafo precedente, non sono che una (importante) parte delle possibili relazioni. Il metodo intuitivo che abbiamo descritto (minimi quadrati) può tuttavia essere esteso con poco sforzo (Taylor, paragrafo 8.6) a casi non lineari.

Ci limiteremo qui a considerare il caso esponenziale che viene trattato con maggiore semplicità e che è molto importante per le applicazioni. Se  $x$  e  $y$  sono legate da una relazione esponenziale

$$y = Ae^{Bx},$$

prendendo il logaritmo di entrambi i membri si ha:

$$\log y = \log (Ae^{Bx})$$

ossia

$$\log y = \log A + Bx$$

rinominando  $Y = \log y$  e  $\alpha = \log A$  si ha dunque

$$Y = \alpha + Bx$$

ossia ci siamo riportati al modello lineare che sappiamo risolvere. in pratica sarà sufficiente, dalle misure di  $x$  e  $y$  prendere il logaritmo di  $y$  e calcolare i coefficienti  $\alpha$  e  $B$  secondo le relazioni viste sopra.  $A$  si può poi ricavare da  $A = \exp(\alpha)$ .

Un altro caso facile è:

$$y = Ax^B$$

da cui si ha

$$\log Y = \log A + B \log x$$

da cui chiamando rinominando  $Y = \log y$ ,  $X = \log x$  e  $\alpha = \log A$  si ha dunque

$$Y = \alpha + BX$$

### 3.4 Coefficiente di correlazione lineare

Avendo ora risposto alla domanda

- supponendo che la relazione tra le variabili misurate sia lineare, come si possono determinare  $A$  e  $B$ ?

ci rimane da considerare la seconda:

- come valutare dalla misurazione se l'ipotesi della linearità è verificata?

Taylor,  
Cap. 9

Per rispondere a questo quesito occorre introdurre una quantità che misura la correlazione di due diverse variabili.

Supponiamo di aver effettuato una serie di misure  $N$  accoppiate di  $x$  e  $y$ ; abbiamo già definito la media (campionaria) della variabile  $x$

$$\bar{x} = \frac{1}{N} \sum_i x_i,$$

e la sua analoga per  $y$ . Inoltre abbiamo definito gli scarti dalla media delle due serie di misure:

$$d_{x_i} = x_i - \bar{x}$$

e la sua analoga per  $y$ . La covarianza tra  $x$  e  $y$  è definita come:

$$\sigma_{xy} = \frac{1}{N} \sum_i d_{x_i} d_{y_i} = \frac{1}{N} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

Ricordiamo inoltre che dagli scarti  $d_{x_i}$  si può ottenere la varianza di  $x$ , definita da

$$\sigma_x^2 = \frac{1}{N} \sum_i d_{x_i}^2,$$

la cui radice quadrata è la deviazione standard  $\sigma_x$ .

Con queste quantità si definisce il *coefficiente di correlazione*, denotato con

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Il significato quantitativo di  $r$  può essere studiato con la teoria delle probabilità (vedi Taylor 9.4). Noi ci limiteremo a mostrarne le proprietà qualitative seguenti:

- $-1 \leq r \leq +1$ : coefficiente di correlazione è un numero con modulo minore uguale a 1. Infatti la disuguaglianza di Schwartz, afferma proprio che vale, per  $a_i$  e  $b_i$  qualsiasi,  $(\sum a_i b_i)^2 \leq (\sum a_i^2)(\sum b_i^2)$ , che usando  $a_i = d_{x_i}$  e  $b_i = d_{y_i}$  diventa proprio  $|\sigma_{xy}| \leq \sigma_x \sigma_y$ .
- se le due variabili sono perfettamente correlate (ossia giacciono su una retta) allora  $|r| = 1$ , ossia  $r = \pm 1$ . Il segno è positivo se la retta ha coefficiente angolare positivo (ossia se ad un aumento di una delle due variabili, la seconda cresce in proporzione), negativo se il coefficiente angolare è negativo (ossia ad un aumento di una delle due variabili la seconda *decrece* in proporzione.)

Infatti se  $y_i = A + Bx_i$  allora anzitutto  $\bar{y} = A + B\bar{x}$  e dunque anche gli scarti sono legati da  $d_{y_i} = Bd_{x_i}$  quindi:

$$r = \frac{B \sum d_{x_i}^2}{\sqrt{\sum d_{x_i}^2 B^2 \sum d_{x_i}^2}} = \frac{B}{|B|} = \pm 1$$

Se  $B > 0$  allora  $|B| = B$  e  $r = 1$ ; altrimenti  $|B| = -B$  e  $r = -1$ .

- se  $r = 0$  (o comunque molto piccolo) le due variabili non sono correlate e non hanno la tendenza a giacere su di una retta. Tale proprietà deriva dal fatto che se non vi fosse alcuna relazione tra  $x$  e  $y$ , qualunque sia il valore di  $x_i$  (e dunque di  $d_{x_i}$ ),  $y$  avrebbe la stessa probabilità di trovarsi sopra o sotto  $\bar{y}$ : quindi la somma che definisce  $\sigma_{xy}$  è composta di termini che hanno la stessa probabilità di essere positivi quanto negativi. Tale somma risulterà piccola e con lei  $r$ .

In definitiva, una misura di quanto bene i dati si dispongono su di una retta è data dal coefficiente di correlazione  $r$ : quanto più il suo modulo si avvicina a 1, tanto meglio la retta descrive la relazione tra i dati. Anziché  $r$  a volta si trova riportato  $r^2$ .