Insiemi

Un branco di elefanti, un grappolo d'uva o un volo di storni sono tutti esempi di insiemi di oggetti. Il concetto matematico di *insieme* è il fondamento di tutta la matematica conosciuta fino ai giorni nostri. Non si dà la definizione di insieme, preferendo descrivere cosa si può fare con gli insiemi, analogamente alla trattazione usuale della geometria che non definisce il punto e la retta ma descrive cosa si può fare con questi oggetti. Si dice che insieme è sinonimo di *collezione, lista, classe* di oggetti ben definiti; analogamente si dice che oggetti è sinonimo di *elementi, membri.*

Esempi

1. I numeri 1, 2, 3, 4, 5;
2. le soluzioni dell'equazione $x^2 - 3x + 2 = 0$;
3. le vocali dell'alfabeto: a, e, i, o, u;
4. i paesi d'Europa.

Indichiamo gli insiemi con le lettere maiuscole dell'alfabeto:

$$A,\ B,\ C, \ldots, X,\ Y,\ Z\,.$$

Indichiamo gli elementi di un insieme con le lettere minuscole dell'alfabeto:

$$a,\ b,\ c, \ldots, x,\ y,\ z\,.$$

Per indicare che l'insieme A è costituito dagli elementi $a, b, c, \ldots$, questi si elencano, scrivendo:

$$A = \{a, b, c, \ldots\}\,.$$

Per descrivere un insieme A si può usare una proprietà $P(x)$ di cui godono gli elementi di A. Si scrive:

$$A = \{x |\, P(x)\}$$

che significa: A è l'insieme degli x tali che valga $P(x)$, per cui vale $P(x)$.

Esempi

1. $X = \{1, 2, 3\}$ l'insieme X ha come elementi i numeri 1,2,3.
2. $A = \{x \in \mathbf{Z} \,|\, 1 < x < 5\} = \{2, 3, 4\}$
3. $B = \{x \in \mathbf{R} \,|\, x^2 - 3x + 2 = 0\} = \{1, 2\}$;
4. Con $\mathbf{N}$ si indica l'insieme dei numeri naturali;
5. Con $\mathbf{Z}$ si indicano i numeri interi;
6. Con $\mathbf{Q}$ si indicano inumeri razionali;

7. Con $\mathbf{R}$ si indicano i numeri reali.

- $\mathbf{N}, \mathbf{Z}, \mathbf{Q}, \mathbf{R}$ sono insiemi formati da infiniti elementi.

Notazione

Se un oggetto x è un elemento dell'insieme A si scrive: $x \in A$, che si legge: x appartiene ad A oppure x è in A. D'altra parte, se x non appartiene ad A, non è un elemento di A si scriverà $x \notin A$.

Esempi

1. $V = \{a, e, i, o, u\}$, allora $a \in V,\ e \in V,\ b \notin V,\ c \notin V$.

2. $P = \{x \in \mathbf{N} \,|\, x \text{ pari}\}$, allora $2, 4, 6, \ldots \in A,\ 1, 3, 5, \ldots \notin A$.

3. L'insieme A privo di elementi, tale che $x \notin A$ per ogni x, si dice *insieme vuoto* e si segna con $\emptyset$.

Diagrammi di Venn

Un modo semplice per visualizzare i nessi tra insiemi è dato dai diagrammi di Venn. Ogni insieme è rappresentato da una regione del piano limitata da una curva chiusa.

Sottoinsiemi

Dati due insiemi $S,\ T$ si dice che S è *sottoinsieme* di T (T contiene S, S è incluso in T, $T \subseteq S$) se *ogni* elemento di S è anche elemento di T; se in T vi sono elementi che non sono in S allora S è un sottoinsieme *proprio* di T e si scrive $S \subset T$. Può essere che *ogni* elemento di T sia a sua volta elemento di S, e si possa scrivere $T \subset S$; allora dire che $S \subset T$ e $T \subset S$ equivale a dire $S = T$.

Esempi

1. $\mathbf{N} \subset \mathbf{Z} \subset \mathbf{Q} \subset \mathbf{R}$ i reali includono i razionali, i razionali includono gli interi, gli interi includono i naturali.

2. Tra i sottoinsiemi di un insieme $A \neq \emptyset$ vi sono A stesso e l'insieme vuoto $\emptyset$.

3. Tutti i sottoinsiemi di $A = \{a, b, c\}$ sono:
$\emptyset$, $\{a\}$, $\{b\}$, $\{c\}$, $\{a, b\}$, $\{a, c\}$, $\{b, c\}$, $\{a, b, c\}$.

Operazioni

Intersezione

Si chiama *intersezione* di due insiemi $A,\ B$ l'insieme formato da tutti gli elementi che appartengono *sia* ad A *che* a B. L'insieme intersezione si indica con $A \cap B$:

$$A \cap B = \{x \,|\, x \in A \text{ e } x \in B\}$$

Esempi

1. $A = \{a, b, c, d, e, f\},\ B = \{a, e, i, o, u\}$, allora $A \cap B = \{a, e\}$.

2. Siano $D = \{1, 2, 3, 4, 5, 6\}$ e $M = \{x \in \mathbf{Z} \,|\, x > 3\}$, allora è $D \cap M = \{4, 5, 6\}$.

3. Per le rette $r,\ s$ del piano considerate come insiemi di punti può essere:

- $r \cap s \to r = s$: le rette coincidono, tutti i punti in comune;
- $r \cap s \to \emptyset$: rette parallele, nessun punto in comune;
- $r \cap s \to \{P\}$: rette incidenti, il punto P in comune.

NB: Se due insiemi A, B non hanno elementi comuni allora la loro intersezione è l'insieme vuoto, ovvero $A \cap B = \emptyset$.

Unione

Si chiama *unione* di due insiemi $A,\ B$ l'insieme formato da tutti gli elementi che appartengono ad A *oppure* a B, ciascuno contato una sola volta. Qui 'oppure' è usato in senso non esclusivo, non di alternativa tra due scelte. L'insieme unione si indica con $A \cup B$.

$$A \cap B = \{x \,|\, x \in A \text{ oppure } x \in B\}$$

Esempi

1. $A = \{a, b, c\},\ B = \{d, e, f\}$, allora $A \cup B = \{a, b, c, d, e, f\}$.

2. $P = \{2, 4, 6, \ldots, 2n, \ldots\},\ D = \{1, 3, 5, \ldots, 2n + 1, \ldots\}$, allora $P \cup D = N$.

3. $A = \{x \in \mathbf{R} \,|x > 3\}$, $B = \{x \in \mathbf{R} \,|x < 2\}$. $A \cup B$ ricopre $\mathbf{R}$, escluso l'insieme $C = \{x \in \mathbf{R} \,|2 \leq x \leq 3\}$.

Proprietà

1. $A \cup A = A$

2. $A \cup B = B \cup A$ commutativa;

3. $(A \cup B) \cup C = A \cup (B \cup C)$ associativa;

4. $A \cap A = A$

5. $A \cap B = B \cap A$ commutativa;

6. $(A \cap B) \cap C = A \cap (B \cap C)$ associativa.

Differenza

Si chiama *differenza* fra A e B e si scrive $A - B$ l'insieme formato dagli elementi di A che *non* appartengono a B, cioè

$$A - B = \{x \in A \,|\, x \notin B\}.$$

Esempi

1. $A = \{a, b, c, d, e, f\},\ B = \{a, e\}$, allora $A - B = \{b, c, d, f\}$.

2. $D = \{1, 2, 3, 4, 5, 6\},\ M = \{2, 4, 6\}$, allora $D - M = \{1, 3, 5\}$.

3. **R − Q** è l'insieme dei numeri irrazionali.

Differenza simmetrica

La differenza simmetrica di due insiemi A, B è l'insieme che contiene gli elementi di A che non sono in B e gli elementi di B che non sono in A. Per come è definita vale $(A \cup B) - (A \cap B)$ ed è un esempio di un operazione costituita da più operazioni elementari. Si può anche considerare come l'insieme degli elementi che stanno in A oppure in B, ma non in entrambo. Qui oppure è usato in senso esclusivo.

Esempi

1. $A = \{a, b, c, d, e, f\},\ B = \{a, e, f, g, h\}$,
   allora $(A \cup B) - (A \cap B) = \{b, c, d, g, h\}$.
2. $D = \{1, 2, 3, 4, 5, 6\},\ M = \{2, 4, 6, 7, 8\}$,
   allora $(D \cup M) - (D \cap M) = \{1, 3, 5, 7, 8\}$.

Prodotto cartesiano

Si dice *prodotto cartesiano* di due insiemi X e Y e si indica con $X \times Y$ l'insieme i cui elementi sono le coppie ordinate (x, y) con $x \in X,\ y \in Y$.

$$X \times Y = \{(x, y) \,|\, x \in X,\, y \in Y\}$$

Se $X = Y$, allora $X \times Y = X^2$. Due coppie ordinate $(x, y),\ (x', y')$ si dicono uguali se $x = x',\ y = y'$.

Esempio

Una rappresentazione del prodotto cartesiano degli insiemi $X\{1, 2, 3, 4, 5\}$ e $Y\{a, b, c, d\}$ si ottiene in forma di tabella.
$P = X \times Y$ si rapprenta come

d	•	•	•	(4,d)	•
c	•	(2,c)	•	•	•
b	•	•	•	•	(5,b)
a	(1,a)	•	(3,a)	•	•
	1	2	3	4	5

mentre $Q = Y \times X$ come

5	•	(b,5)	•	•
4	•	•	•	(c,4)
3	(a,3)	•	•	•
2	•	•	•	•
1	•	•	(c,1)	•
	a	b	c	d

- Se $X \neq Y$ è anche $X \times Y \neq Y \times X$. Inoltre
- $X \times \emptyset = \emptyset \times X = \emptyset$;
- $A \times B \subset X \times Y$ se $A \subset X$ e $B \subset Y$.

Esempi

1. Le coppie di numeri che si ottengono lanciando due dadi $D_{1,2} = \{1,2,3,4,5,6\}$ sono elementi del prodotto cartesiano $D_1 \times D_2 = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$ e sono tutte le coppie ordinate $(1,1),\ldots,(1,6),(2,1),\ldots,(5,6),(6,1),\ldots,(6,6)$.

2. Le coordinate (x,y) del piano cartesiano sono elementi di $\mathbf{R}^2$.

3. Le coordinate (x,y,z) nello spazio euclideo tridimensionale sono elementi di $\mathbf{R}^3$.

4. Le liste ordinate (vettori) di n numeri reali $(a_1,a_2,\ldots,a_n)$ sono elementi del prodotto cartesiano $\mathbf{R}\times\mathbf{R}\times\cdots\times\mathbf{R} = \mathbf{R}^n$.

5. I *pixel* $p(i,j), 0 \le i \le m,\ 0 \le j \le n$ di un'immagine sono gli elementi del prodotto cartesiano $(m+1)\times(n+1)$ del numero di righe per il numero di colonne.

6. La coppia di indici $(i,j), 1 \le i \le r,\ 1 \le j \le c$ che individuano gli elementi di una matrice sono gli elementi del prodotto cartesiano $r \times c$ del numero di righe per il numero di colonne.

Intervalli

Chiamiamo *intervalli* gli insiemi di punti sulla retta. Li denotiamo con diverse scritture:

$x < 1$	$]-\infty,1[$	aperto
$x \le 1$	$]-\infty,1]$	ne' aperto, ne' chiuso
$x > 1$	$]1,+\infty[$	aperto
$x \ge 1$	$[1,+\infty[$	ne' aperto, ne' chiuso
$-2 < x < 1$	$]-2,1[$	aperto
$-2 \le x < 1$	$[-2,1[$	ne' aperto, ne' chiuso
$-2 < x \le 1$	$]-2,1]$	ne' aperto, ne' chiuso
$-2 \le x \le 1$	$[-2,1]$	chiuso

Applicazioni

Una *applicazione* di un insieme S in un insieme T è una "legge" o "regola" che associa ad ogni elemento $s \in S$ un *unico* elemento $t \in T$.

Esempio

Se $S = T = \mathbf{R}$ è l'insieme del reali, l'insieme della coppie $S\times T$ è il piano cartesiano dove la coppia ordinata (x,y) corrisponde a un punto le cui coordinate sono rispettivamente x e y. Se, tra le infinite coppie si considera il sottoinsieme formato dalle coppie (x,x^2), l'insieme di punti si dice *grafico* dell'applicazione di S in T che manda l'elemento $x \in S$ nell'elemento y di T tale che $y = x^2$.

La nozione di applicazione si ritrova in tutta la matematica ed è, senza dubbio, una delle idee piú generali e produttive, con cui tutti hanno familiarità: mediata dal linguaggio, magari impreciso, non rigorosa, ma certamente l'idea che esista una "regola" che

"collega" gli elementi (numeri, cose, persone, ...) di un insieme con gli elementi di un altro, è diffusa e consolidata.

Una applicazione è una regola che associa a ciascun $s \in S$ un elemento $t \in T$ purché la coppia (s,t) stia in $S \times T$.

Questa definizione è meno intuitiva, ma fissa una condizione perchè si possa parlare di applicazione: applicare s in t vuol dire che la coppia (s,t) sta nel prodotto cartesiano $S \times T$; l'applicazione agisce come una regola per selezionare dall'insieme delle coppie (s,t) un preciso sottoinsieme.

L'applicazione si scrive come $f\colon s \mapsto t$, oppure $t = f(s)$, dove t si dice *immagine* di s in f.

Esempi

1. Un'applicazione i di S in se' stesso ($S \mapsto S$) tale che $i\colon s \mapsto s$ si chiama *identità*.

2. Sia I l'insieme degli interi, $C = \{(m,n) \in I \times I \,|n \neq 0\}$ e Q l'insieme dei numeri razionali. L'applicazione $r : C \mapsto Q$ associa a ciascuna coppia m, n il numero razionale m/n.

3. Sia I l'insieme degli interi, $C = \{(m,n) \in I \times I\}$. L'applicazione $s : C \mapsto I$ associa a ciascuna coppia m, n il numero $m + n$.

4. L'applicazione p di $S \times T$ (S, T qualsiasi) tale che $s\colon (x,y) \mapsto x$ si chiama *proiezione* di $S \times T$ su S.

Una applicazione f di S in T si dice *suriettiva* se per ogni elemento $t \in T$ esiste almeno un $s \in S$ tale che $t = f(s)$.

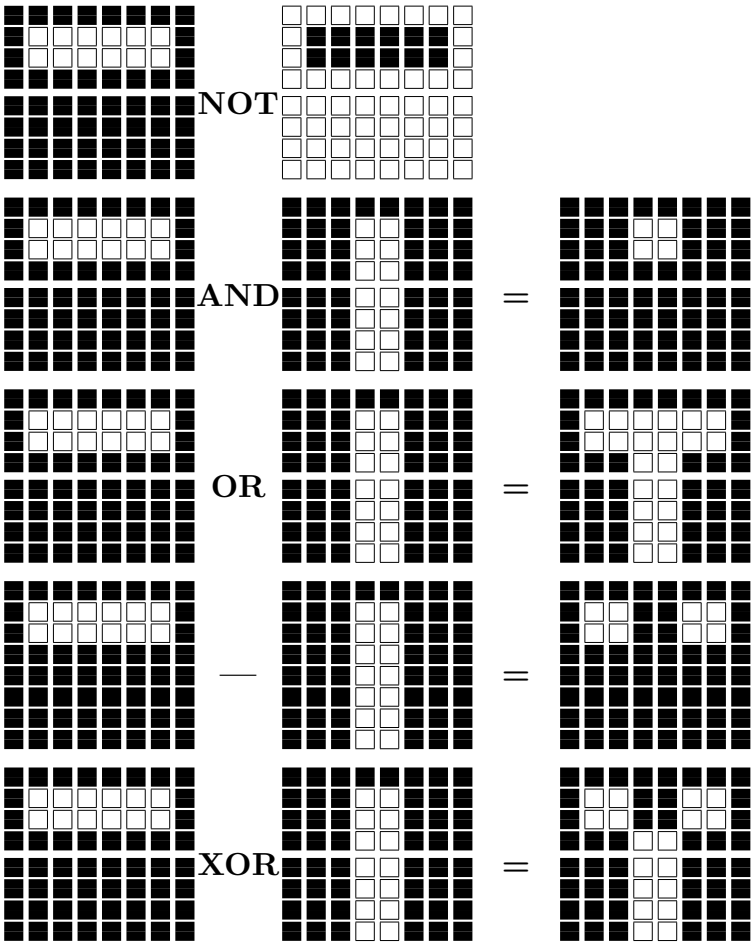
Una applicazione f di S in T si dice *iniettiva* se, per $s_1, s_2 \in S$, $s_1 \neq s_2$ è anche $f(s_1) \neq f(s_2)$

Operazioni logiche su insiemi

Se si considera un'immagine come una matrice (prodotto cartesiano) di pixel con assegnato a ciascuna posizione, per esempio, il valore 0 oppure il valore 1, è possibile pensare di applicare elemento per elemento le operazioni logiche NOT, AND, OR e le loro composizioni agli insiemi di pixel di due (o più immagini). In tal modo, ad esempio è possibile modificare l'immagine, estrarne parti, ecc.

Questi procedimenti elementari sono esemplificativi dei metodi usati per trattare l'immagine digitale.

I disegni che seguono illustrano le operazioni logiche su insiemi di bit rappresentati da quadratini bianchi □, 1 logico, e neri ■ 0 logico. L'immagine è un reticolo di 8×8 pixel. L'operazione **NOT** è unaria, mentre **AND**, **OR**, — e **XOR** sono binarie, cioè operano su due insiemi.



Esercizi

1. In un paese di confine si parla francese e tedesco. Il 70% della popolazione parla tedesco, mentre il 60% della popolazione parla il francese. Quale percentuale della popolazione parla francese e tedesco?

2. Il 30% degli abitanti di una città sono stati vaccinati contro l'influenza. Statisticamente, in una epidemia di influenza il 80% delle persone non vaccinate si ammalano, contro solo il 10% di quelle vaccinate. Quanti si ammalano?

3. Descrivere a parole cosa fa l'applicazione c di S in T tale che per ogni $s \in S$ sia $c\colon s \mapsto t_0$, dove t_0 è un elemento di T.

4. Stabilire se le applicazioni di S in T sono suriettive o iniettive.
a) $S = \{s \in \mathbf{R}\}$, $T = \{t \in \mathbf{R} | t \geq 0\}$, $g : s \mapsto s^2$.
b) $S = \{s \in \mathbf{R} | s \geq 0\}$, $T = \{t \in \mathbf{R} | t \geq 0\}$, $g : s \mapsto s^4$.
c) $S = \{s \in \mathbf{R}\}$, $T = \{t \in \mathbf{R}\}$, $g : s \mapsto s^3$.

5. Applicare bit per bit (*bitwise*) le operazioni logiche AND, OR, XOR sulle due stringhe

```
01010001001001001111011010010111111
10000101010101011001010101001001
```

Nota storica

Lo sviluppo dei sistemi numerici e della numerazione scritta è una parte essenziale della storia della civiltà.

Saper svolgere calcoli elementari è una necessità della vita quotidiana. Quanto questo sia agevole dipende in primo luogo dal tipo di numerazione usata. I simboli usati dai Romani rappresentavano raggruppamenti dei segni usati per contare: V (5) è il raggruppamento di cinque segni I (1), C di 100, ecc. Eseguire un calcolo non era semplice e richiedeva l'impiego di personale addestrato; per eseguire il calcolo si inventò allora l'*abaco*[1], una tavoletta su cui si eseguiva il calcolo con gettoni; questo strumento di calcolo rimase in uso sino alla metà del 1500.

L'attuale *algoritmo*[2] usa l'informazione contenuta nella *posizione* della cifra; era quindi essenziale introdurre un segno apposito, lo *zero*, per indicare dove non vi era niente. Questa invenzione venne introdotta in Europa nel 1202 con il *Liber abbaci* di Leonardo Pisano, detto il Fibonacci.

Il sistema decimale

Il sistema di numerazione comunemente usato è quello *decimale*, ovvero in *base* 10. La grandezza della base è il numero dei segni (*cifre*) con cui si scrivono i numeri. Nel sistema decimale i segni sono dieci $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$. I numeri sono formati da sequenze di cifre di varia lunghezza come $c_n c_{n-1} \cdots c_2 c_1 c_0$, dove c_0 si dice la cifra meno *significativa*, c_n la più significativa. Il numero associato ad una certa sequenza viene dalla notazione posizionale. Il numero associato a ciascun segno (cifra) viene moltiplicato per una potenza di 10 (della base), associando alla prima cifra da sinistra (delle unità) 10^0, alla seconda cifra (delle decine) 10^1, e via di seguito. Quindi la stessa cifra rappresenta numeri diversi secondo la sua posizione nella successione di cifre. Il numero associato a questa scrittura è la somma di tutti questi n prodotti, secondo lo schema:

$$c_n c_{n-1} \cdots c_2 c_1 c_0 = c_n \cdot 10^n + c_{n-1} \cdot 10^{n-1} + \cdots + c_2 \cdot 10^2 + c_1 \cdot 10^1 + c_0 \cdot 10^0$$

$$\begin{array}{rrcll} & 4 & \times & 10^0 & + \\ & 3 & \times & 10^1 & + \\ 10234 := & 2 & \times & 10^2 & + \\ & 0 & \times & 10^3 & + \\ & 1 & \times & 10^4 & + \end{array}$$

---

[1] È il pallottoliere usato nell'istruzione elementare.

[2] Sinomimo di *metodo, procedimento di calcolo*; deriva dal nome, al - Khuwaritzmi, dell'autore del trattato in cui viene descritta la notazione posizionale e come con essa condurre i calcoli.

La notazione posizionale funziona perchè un segno opportuno è previsto per lo zero. Esiste quindi un moltiplicatore nullo, che azzera il contributo della corrispondente potenza di 10.

Occorre far notare come le cifre che compongono il numero siano i resti delle successive divisioni per 10:

$$\begin{array}{rcrcll} 1234 & \div & 10 & = & 123 & R\ 4 \\ 123 & \div & 10 & = & 12 & R\ 3 \\ 12 & \div & 10 & = & 1 & R\ 2 \\ 1 & \div & 10 & = & 0 & R\ 1 \end{array}$$

In effetti allora si può anche scrivere $c_n c_{n-1} \cdots c_2 c_1 c_0 := c_0 + 10(c_1 + 10(c_2 + 10(\cdots + 10(c_{n-1} + 10c_n)\cdots)))$

Il sistema decimale si estende ai numeri minori di uno, o alla parte detta *decimale* di un numero. L'elemento di separazione è la virgola [3]. Alla prima posizione a destra della separazione (prima cifra decimale) si associa 10^{-1}, alla seconda 10^{-2}, alla n - esima 10^{-n}.

$$\begin{array}{rrcll} & 5 & \times & 10^{-1} & + \\ 0.567 = & 6 & \times & 10^{-2} & + \\ & 7 & \times & 10^{-3} & + \end{array}$$

Occorre far notare come le cifre si ottengano ora moltiplicando per 10 il numero decimale sinché questo non *svanisce*:

$$\begin{array}{rcrcll} 0.567 & \times & 10 & = & 5.67 & E\ 5 \\ 0.67 & \times & 10 & = & 6.7 & E\ 6 \\ 0.7 & \times & 10 & = & 7.0 & E\ 7 \end{array}$$

Il risultato delle operazioni $+$, $\times$ sulle 10 cifre vengono dati con due tavole *pitagoriche* che associano a ciascuna coppia di cifre il risultato. Queste tavole vengono memorizzate durante la prima infanzia e costituiscono la base indispensabile per il calcolo numerico mentale e scritto. Le operazioni su numeri di molte cifre si svolgono cifra per cifra, e in tal modo vengono ricondotte alle operazioni sui numeri di una cifra contenute nelle tavole.

Ad esempio, per sommare numeri di molte cifre si procede ripetendo due passi uguali: allineati i numeri, si sommano due cifre corrispondenti e al risultato intermedio si somma ancora il riporto della somma delle cifre precedenti.

L'algoritmo, semplice e affidabile, può essere svolto automaticamente (macchine calcolatrici).

Le stesse tavole vengono usate (più faticosamente) nelle (operazioni) sottrazione e divisione. Per la sottrazione si dice che $8 - 3 = 5$ perchè $5 + 3 = 8$, analogamente $6 \div 3 = 2$ perchè $2 \times 3 = 6$. Le difficoltà dipendono dal fatto che l'associazione è indiretta, mediata da una verifica con addizione o moltiplicazione (dirette). Questo è particolarmente vero per la divisione dove si

[3] Nei paesi anglosassoni, e quindi in molta letteratura tecnica, si usa il punto (*dot*) decimale, mentre la virgola separa le migliaia.

tratta sempre e comunque di ottenere i resti della divisione per il divisore.

Altri sistemi di numerazione

La necessità che una cifra significhi zero, fa sí che la base minima sia 2, con le cifre 0, 1, e le associate tavole di addizione e moltiplicazione. Leibniz per primo concepì il sistema binario. Gli sviluppi vennero molto più tardi. Di seguito, il sistema in base 3 ha le cifre 0, 1, 2, ecc. Una base minore/maggiore di 10 richiede più/meno cifre per rappresentare la stessa quantità. Ad esempio, un numero di dieci cifre binarie (*bit*, binary digit) si scrive in base 10 con solo tre cifre (è $2^{10} = 1024 \approx 10^3$). La base in effetti può essere anche maggiore di 10: ad esempio, gli assiro - babilonesi usavano un sistema in base 60[(1)], le cui tracce permangono nella consueta misura del tempo in minuti e secondi.

In informatica si usa la base 16, *esadecimale*, ottenuta aggiungendo alle cifre $0 \ldots 9$ i segni A, B, C, D, E, F, dove $\mathrm{A}_{16} = 10_{10}$, $\mathrm{B}_{16} = 11_{10}$, $\ldots$, $\mathrm{F}_{16} = 15_{10}$.

Si tratta di una contrazione del binario allo scopo di risparmiare spazio e tempo. Inoltre quattro bit sono mezzo byte (otto bit), quindi due cifre esadecimali rappresentanon un byte, un carattere, l'unità di elaborazione.

Ad esempio $48_{10} = 30_{16} = 0011\,0000_2$, mentre $\mathrm{FF}_{16} = 16 \times 15 + 15 = 255_{10}$

Nella tabella si riporta un confronto tra varie basi.

sistema	base	cifre	esempio
decimale	10	0123456789	2 009
binario	2	01	0111 1101 1001
ottale	8	01234567	3 713
esadecimale	16	0123456789ABCDEF	7D9

Un algoritmo converte un numero naturale da base 10 in binario: si tengono come cifre (binarie) i resti successivi delle ripetute divisioni per due, tenendo conto che il primo resto è il bit meno significativo, finchè non resta nulla da dividere.

$$\begin{array}{rcrcrcl} 29 & \div & 2 & = & 14 & R & 1 \\ 14 & \div & 2 & = & 7 & R & 0 \\ 7 & \div & 2 & = & 3 & R & 1 \\ 3 & \div & 2 & = & 1 & R & 1 \\ 1 & \div & 2 & = & 0 & R & 1 \end{array}$$

Lo stesso algoritmo funziona per convertire un numero binario in decimale: ora occorre dividere per 1010_2 (10_{10}) e tenere i resti successivi:

(1) È $60 = 2^2 \times 3 \times 5$, quindi i divisori di 60 sono ben $3 \times 2 \times 2 = 12$; le divisioni per numeri che hanno come fattori 2, 3 e 5 danno luogo a numeri (sessagesimali) con sviluppo finito.

$100101001_2 = 153_{10}$

$$\begin{array}{rcrcr} 100101001 & \div & 1010 = 1111 & R & 11 = 3_{10} \\ 1111 & \div & 1010 = 1 & R & 101 = 5_{10} \\ 1 & \div & 1010 = 0 & R & 1 = 1_{10} \end{array}$$

Per convertire il numero n_a (in base a) nel numero m_b (base b) si adotta allora la regola: le cifre del numero m_b sono i resti delle successive divisioni di n_a per b_a, cioè b espresso in base a.

Nelle macchine il trasferimento delle informazioni avviene con stringhe di caratteri e i dati numerici non sfuggono a questa regola. La conversione dipende dalla rappresentazione interna (codice) del carattere. Un codice a caratteri di 8 bit è l'ASCII esteso, mentre il Braille è un codice a 6 bit (punti in rilievo su cartoncino).

Un algoritmo converte successioni di caratteri (stringhe) in numeri in base due per la successiva elaborazione. La conversione di un carattere nel corrispondente numero binario usa il fatto che in tutti i codici i caratteri 0..9 occupano posizioni consecutive. Allora $c_2 = \text{ord}(C) - \text{ord}(0)$ dove la funzione $\text{ord}(C)$ restituisce il numero d'ordine di ciascun carattere. Questo tipo di realizzazione dell'algoritmo lo rende indipendente dal particolare codice usato. Ad esempio, in ASCII è $\text{ord}(0) = 48_{10}$.

Il sistema in base due si presta particolarmente a implementare il calcolo 'meccanico' (elettrico, elettronico). Alle cifre binarie 0, 1 si associano due stati ben distinguibili di un circuito elettrico. Le operazioni di somma e di confronto sono equivalenti ad una definita sequenza di operazioni logiche fisicamente implementate da particolari circuiti elettronici in grado di svolgere le operazioni ad alta velocità.

Basterà riprodurre la tabellina dell'addizione e quella della moltiplicazione per un bit

$+$	0	1
0	0	1
1	1	0

$\times$	0	1
0	0	0
1	1	1

Nell'addizione la cifra del risultato è 0 perché $1_2 + 1_2 = 10_2$, cioè vi è un riporto in avanti, analogamente a quanto accade nel sistema decimale quando la somma di due cifre supera 9. Bene, le tabelline per il sistema binario sono identiche a quelle delle operazioni logiche XOR, (una operazione composta dagli operatori elementari NOT,OR, AND) e AND.

Questi nomi non sono altro che la "traduzione" degli operatori elementari della logica formale, $\neg$, 'non', $\vee$, 'oppure', $\wedge$, 'e'. È notevole come il calcolo di espressioni formate con gli operatori della logica formale sia alla base del calcolo automatico.

Una rappresentazione dei numeri negativi detta *complemento a 2* permette di svolgere la sottrazione con lo stesso hardware della somma. Altre e più sofisticate disposizioni di elementi logici

eseguono prodotti e divisioni, mentre altre operazioni, ad esempio l'estrazione di radice come $\sqrt{2}$, vengono svolte con algoritmi complessi (software a virgola mobile).

Le operazioni logiche elementari $\neg$, 'non', $\vee$, 'oppure', $\wedge$, 'e' vengono chiamate rispettivamente NOT, OR, AND quando vengono usate, ad esempio, per generare la somma di due bit.

La notazione standard

• Circa 12 000 000 000 000 000 secondi fa i primi animali sono comparsi sulla Terra.

• La luce impiega circa 0.000 000 000 015 s per attraversare il vetro di una finestra.

• In un grammo di idrogeno si conta un numero enorme di atomi, circa 600 000 000 000 000 000 000 000.

Numeri come questi sono scomodi da usare, e contengono un numero di zeri tale che a prima vista è difficile comprenderne il valore e il significato. Tuttavia nelle scienze e nella tecnologia si incontrano spesso quantità espresse da numeri molto grandi o molto piccoli e altrettanto spesso si fanno calcoli con esse.

La *notazione standard*, uniformando la scrittura di numeri di questo tipo, li rende facilmente comprensibili e trattabili.

Poiché $10^0 = 1$, possiamo riscrivere i numeri $123,4$ e $0,0078$ come $123,4 \times 10^0$ e $0,0078 \times 10^0$. Ora possiamo spostare la virgola di un posto a sinistra aumentando di 1 l'esponente di 10:

$$123,4 = 123,4 \times 10^0 = 12,34 \times 10^1 = 1,234 \times 10^2$$

sino a che resta una sola cifra $\neq 0$ davanti al punto decimale.

Analogamente si può spostare la virgola di un posto a destra diminuendo di 1 l'esponente di 10:

$$0,0078 = 0,0078 \times 10^0 = 0,078 \times 10^{-1} = 0,78 \times 10^{-2} = 7,8 \times 10^{-3}$$

anche qui sino a che resta una sola cifra $\neq 0$ davanti al punto decimale.

Ora è semplice scrivere numeri molto grandi o molto piccoli:

$$1\,200\,000\,000\,000\,000 \rightarrow 1,2 \times 1\,000\,000\,000\,000\,000 \rightarrow 1,2 \times 10^{+15}$$

scrittura che si legge: '1,2 per 10 alla 15'. Analogamente per i numeri minori di uno:

$$0.000\,000\,000\,015 \rightarrow 1,5 \times 0.000\,000\,000\,01 \rightarrow 1,5 \times 10^{-11}$$

che si legge: '1,5 per dieci alla meno undici".

Allora un numero è scritto in notazione standard quando è scritto come:

$$\text{m} \times 10^{\text{e}} = \text{mantissa} \times 10^{\text{esponente}}$$

La *mantissa* è un numero decimale e l'unica cifra davanti alla virgola (punto decimale) è compresa tra 1 e 9; l'*esponente* è un intero. Di solito, cioè a meno di non volerlo far risaltare in modo particolare, si omette il segno + davanti a mantissa ed esponente quando questi sono positivi.

Esempio

$a = -1.234 \times 10^{-2} \quad b = -0.1234 \times 10^{-1} \quad c = -12.34 \times 10^{-3}$

Qui a, b, c rappresentano la stessa quantità, ma solo a è scritto correttamente: in b la prima cifra è 0, mentre in c davanti alla virgola vi sono due cifre invece di una sola. Ora la cifra '1' davanti al punto decimale è la più significativa e '4' la meno significativa.

Calcoli in notazione standard

Moltiplicare e dividere è facile: si tratta di moltiplicare o dividere le due mantisse tra loro come solitamente si fa per i numeri decimali, e sommare (moltiplicazione) o sottrarre (divisione) gli esponenti, secondo le regole per le operazioni tra potenze di egual base: $a^x \times a^y = a^{x+y} \quad a^x \div a^y = a^{x-y} \quad (a^x)^y = a^{x \times y}$

Esempi

1. Calcolare $p = 1.5 \times 10^{-11} \times 1.2 \times 10^{16} = ?$
mantisse $1.5 \times 1.2 = 1.8$
esponenti $(-11) + (+16) = +5$
$p = 1.8 \times 10^5$

2. Calcolare $q = 1.5 \times 10^{-11} \div 1.2 \times 10^{16} = ?$
mantisse $1.5 \div 1.2 = 1.25 \approx 1.3$
esponenti $(-11) - (+16) = -27$
$q = 1.3 \times 10^{-27}$

3. Calcolare $s = \pi r^2 = ?$ con $r = 2.1 \times 10^3$ m
mantisse $3.14 \times 2.1 \times 2.1 = 13.8 \approx 14 = 1.4 \times 10^1$
esponenti $(+3) \times 2 + 1 = +7$
$s = 1.4 \times 10^7$

• Quando il prodotto delle mantisse supera 10 o il quoziente è minore di 1 si fa scorrere la mantissa finché rimane una sola cifra (non '0') davanti al punto decimale: ad ogni cifra spostata a destra si incrementa l'esponente di una unità, per ogni cifra spostata a sinistra lo si decrementa di una unità.

4. Calcolare $p = 5.5 \times 10^{-11} \times 6.2 \times 10^{16}$
mantisse $5.5 \times 6.2 = 34.1 \approx 34 = 3.4 \times 10^1$
esponenti $-11 + 16 + 1 = +6$
$p = 3.4 \times 10^6$

5. Calcolare $q = 5.5 \times 10^{-}11/6.2 \times 10^{16}$
mantisse $5.5/6.2 = 0.89 = 8.9 \times 10^{-1}$
esponenti $(-11) - (+16) - 1 = -28$
$q = 8.9 \times 10^{-28}$

- Si possono fare calcoli a catena (serie di moltiplicazioni e/o divisioni) separando le operazioni su mantisse ed esponenti come è stato appena fatto.

6. $g = 6.67 \times 10^{-11} \times 5.98 \times 10^{+24} \div (6.37 \times 10^{+6})^2$
mantisse $6.67 \times 5.98 \div (6.37)^2 = 0.983$
esponenti $-11 + 24 - 6 \times 2 = +1$
$g = 0.983 \times 10^1 = 9.83$

Addizione e sottrazione in notazione standard

Le cose si complicano per le addizioni e sottrazioni: queste operazioni sono possibili solo tra numeri che hanno lo stesso esponente. Quindi prima di operare è necessario riportare i numeri ad uno stesso esponente.

Esempi

1.
$$\begin{aligned} 1.234 \times 10^5 + 5.2 \times 10^3 &= 123.4 \times 10^3 + 5.2 \times 10^3 \\ &= 128.6 \times 10^3 = 1.286 \times 10^5 \end{aligned}$$

2.
$$\begin{aligned} 1.2 \times 10^{-4} - 2.34 \times 10^{-3} &= 1.2 \times 10^{-4} - 23.4 \times 10^{-4} \\ &= -22.2 \times 10^{-4} = -2.22 \times 10^{-3} \end{aligned}$$

Ordini di grandezza

Un vantaggio insito nella scrittura in notazione standard è la possibilità di paragonare due quantità semplicemente confrontando gli esponenti: se dopo aver arrotondato a una sola cifra la mantissa gli esponenti sono uguali le quantità sono dello stesso *ordine di grandezza*, altrimenti la differenza degli esponenti da' l'ordine di grandezza relativo di una quantità rispetto all'altra.

Spesso risulta assai utile verificare che l'ordine di grandezza del risultato di un calcolo è corretto, e ciò si può decidere esaminando il solo esponente.

Esempi

1. $a = 2.99 \times 10^8$
Qui a è dell'ordine di 10^8 (delle centinaia di milioni).

2. $b = 7.54 \times 10^3$
Qui b è dell'ordine di 10^4 (delle decine di migliaia).

3. $c = 9.81 \times 10^{-1}$ Qui c è dell'ordine di 10^0 (dell'unità).

4. Confrontare $c = 3.1 \times 10^3$ e $d = 1.8 \times 10^5$
Gli esponenti di c e d differiscono di 2: si può dire che c è di due ordini di grandezza minore di d, oppure che d è di due ordini di grandezza maggiore di c.

5. $3.1 \times 10^3 \times 1.8 \times 10^5 \div 5 \times 10^4$.

Poiché $3 \times 2 \div 5 \approx 1$, l'ordine di grandezza del risultato si ottiene come $10^{3+5-4} = 10^4$.

6. $3.1 \times 10^2 \times 1.8 \times 10^5 \times 5 \times 10^{-4}$.

Poiché $3 \times 2 \times 5 = 30$, l'ordine di grandezza è $10^{2+5-4+1} = 10^3$.

Le dimensioni delle cose

Ordini di grandezza di lunghezze determinate con mezzi meccanici e strumenti ottici. Valori in metri. (Da PSSC)

10^9	Raggio del Sole
10^8	Distanza Terra - Luna
10^7	Raggio della Terra, distanza Roma - San Francisco
10^6	Raggio della Luna, distanza Milano - Palermo
10^5	Distanza Milano - Torino
10^4	Distanza Venezia - Mestre, quota di crociera di un jet
10^3	Un km, un miglio
10^2	Un campo da calcio, la corsa più breve alle Olimpiadi
10^1	Un albero di alto fusto, una casa di 3 piani
10^0	Un passo umano
10^{-1}	La larghezza della mano e del mouse
10^{-2}	Il diametro di una matita
10^{-3}	Lo spessore del vetro di una finestra
10^{-4}	Lo spessore di un foglio di carta
10^{-5}	Un globulo rosso del sangue

Numeri reali

Toccò ai pitagorici, che matematizzavano il mondo in termini di rapporti e proporzioni, scoprire numeri che non possono venire scritti in forma di una frazione, cioè i numeri che si dicono non-razionali, *irrazionali*.

Un esempio di tali numeri è la misura della diagonale del quadrato di lato 1. Sappiamo che essa vale $\sqrt{2}$ ed è semplice dimostrarne l'irrazionalità, ovvero dimostrare che non esiste un razionale $r = \frac{p}{q}$ tale che $r = \sqrt{2}$.

Altri numeri irrazionali sono, ad esempio, $\sqrt[3]{2}$, $\sqrt{3}$, ..., oppure π, il rapporto tra la circonferenza e il diametro, e anche e, la base dei logaritmi naturali è irrazionale. Pur conoscendone relativamente pochi, i numeri irrazionali sono più 'numerosi' dei razionali.

I numeri razionali e i numeri irrazionali formano l'insieme dei numeri *reali*, che si segna con $\mathbf{R}$. I numeri reali possono vengono posti in corrispondenza con i punti della retta, che perciò viene chiamata *retta reale*.

Per i numeri reali valgono tutte le proprietà delle operazioni già viste per gli interi e i razionali. Nei reali tutte le operazioni godono della proprietà di chiusura, ovvero il risultato delle operazioni tra numeri reali è ancora un numero reale.
NB: non è definita la radice quadrata di un numero negativo, che richiederebbe l'introduzione di un altro tipo, completamente nuovo, di numeri.

Alcuni procedimenti di calcolo, essenzialmente quelli che implicano successive approssimazioni al valore cercato (ad esempio calcolare $\sqrt{2}$) hanno senso solo in $\mathbf{R}$.

Caratteristica fondamentale dei reali è la *completezza* che si può definire in modi equivalenti come:

- ogni successione di razionali ha limite nei reali;
- nei reali una successione di intervalli incapsulati isola un punto che appartiene a tutti gli intervalli;
- l'esistenza dell'estremo superiore (o inferiore) di una successione di numeri.

Ogni numero irrazionale si può approssimare quanto si vuole con un razionale, questa proprietà è la *densità* dei razionali sui reali. Anzi si può dimostrare che *solo* i numeri irrazionali possono venir approssimati a volontà con i razionali, tant'è che questa proprietà serve proprio a dimostrare che alcuni numeri, come e e π sono irrazionali.

Esempi

Gli esempi illustrano alcuni algoritmi per calcolare $\sqrt{2}$: i procedimenti generano effettivamente una successione di intervalli a cui si applica la definizione di completezza.

1. Un primo procedimento fa uso di una tabella di quadrati.

Si confronta 2 con $1^2 = 1$ e $2^2 = 4$; ne segue che $1 < \sqrt{2} < 2$. Ora si suddivide l'intervallo $(1, 2)$ in dieci parti ciascuna ampia 0.1 e si riempe la tabella:

x	1.1	1.2	1.3	1.4	1.5	1.6	...
x^2	1.21	1.44	1.69	1.96	2.25	2.56	...

Dall'esame della tabella si vede che $\sqrt{2}$ è compresa tra 1.4 e 1.5. Si ripete il procedimento, suddividendo l'intervallo $(1.4, 1.5)$ ancora in 10 parti ciascuna ampia 0.01, e riempendo la tabella:

x	1.41	1.42	1.43	...
x^2	1.9881	2.0164	2.0449	...

Allora $1.41 < \sqrt{2} < 1.42$. Possiamo immaginare di ripetere questo procedimento più volte, ogni volta aggiungendo una cifra (un decimale) a destra del precedente.

È da notare che:

• Gli intervalli $(1, 2)$, $(1.4, 1.5)$, $(1.41, 1.42)$, ... hanno per estremi i numeri razionali ottenuti ad ogni applicazione del procedimento. Gli estremi superiori sono una successione decrescente, quelli inferiori una crescente e gli intervalli sono ognuno interno al precedente (intervalli incapsulati).

$$1 < 1.4 < 1.41 < \ldots < \sqrt{2} < \ldots < 1.42 < 1.5 < 2.$$

L'unico elemento comune a tutti questi intervalli è $\sqrt{2}$ e può esserlo *solo* perché irrazionale. Comunque si prenda un numero razionale come valore approssimato per $\sqrt{2}$, questo viene prima o poi a trovarsi all'esterno di un intervallo che contiene $\sqrt{2}$.

• L'ampiezza dell'intervallo in cui cade $\sqrt{2}$ ad ogni passo è un decimo di quella del passo precedente. Poiché in linea di principio si può applicare la procedura un numero di volte a piacere, allora è possibile approssimare $\sqrt{2}$ con un numero qualsivoglia di cifre significative esatte. Per avere 10 cifre è sufficiente ripetere il procedimento 10 volte.

2. Un secondo procedimento si fonda sempre sulla densità dei razionali sui reali ed è intuitivamente ancora più semplice: dato che $1 < \sqrt{2} < 2$, si può cercare di capire in quale dei due sottointervalli $(1, 3/2)$, $(3/2, 2)$ si trova $\sqrt{2}$. Si calcola il punto di mezzo dell'intervallo $(1, 2)$, si eleva al quadrato e si confronta con 2. È $(\frac{3}{2})^2 = \frac{9}{4} > 2$, quindi certamente $1 < \sqrt{2} < \frac{3}{2}$. Ora si ripete il procedimento: il punto di mezzo di $(1, \frac{3}{2})$ è $\frac{5}{4}$, per cui vale

$(\frac{5}{4})^2 = \frac{25}{16} < 2$, allora segue $\frac{5}{4} < \sqrt{2} < \frac{3}{2}$. Si può continuare ripetendo questi semplici calcoli indefinitamente ...

• La situazione è analoga all'esempio precedente: gli estremi dell'intervallo sono numeri razionali, ad ogni passo l'ampiezza dell'intervallo si riduce (qui si *dimezza*), gli intervalli sono incapsulati e il processo di successiva approssimazione non ha fine proprio perché $\sqrt{2}$ è un numero irrazionale.

3. L'algoritmo di Erone (Alessandria $\approx$ 130 dC, ma era già noto agli Assiro-Babilonesi) consente di calcolare molto rapidamente un gran numero di cifre della radice quadrata di un numero a.

Il procedimento calcola una nuova migliore approssimazione a partire dalla precedente:

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right)$$

Un algoritmo che 'richiama' se stesso si dice *ricorsivo*. Qui x_{n+1} è la *media aritmetica* delle quantità $x_n,\ a/x_n$; d'altra parte è:

$$\sqrt{a} = \sqrt{x_n \cdot \frac{a}{x_n}}$$

e cioé la radice cercata è la *media geometrica* delle stesse quantità. Si dimostra facilmente che $\forall a, b > 0$ vale:

$$\frac{a+b}{2} \geq \sqrt{ab}$$

valendo il segno di uguaglianza solo se $a = b$. Allora la media aritmetica di due numeri supera sempre la loro media geometrica. Quindi, almeno da x_1 in poi, sarà $x_n > \sqrt{a}$.

Per vedere all'opera l'algoritmo, si applica al calcola di $\sqrt{2}$.

È $1 < \sqrt{2} < 2$ e poniamo $x_0 = 1$ per la prima approssimazione. Otteniamo:

$$x_1 = \frac{1}{2}\left(1 + \frac{2}{1}\right) = \frac{3}{2} = 1.5;$$
$$x_2 = \frac{1}{2}\left(\frac{3}{2} + 2 \cdot \frac{2}{3}\right) = \frac{17}{12} \approx \underline{1.41}66\ldots$$
$$x_3 = \frac{1}{2}\left(\frac{17}{12} + 2 \cdot \frac{12}{17}\right) = \frac{577}{408} \approx \underline{1.414216}\ldots$$

dove sono state sottolineate le cifre corrette ($\sqrt{2} \approx 1.414213562$ con 10 cifre esatte).

• Vale $\frac{3}{2} > \frac{17}{12} > \ldots > a_i > \ldots > \sqrt{2}$, cioé almeno da x_1 in poi la *successione* dei numeri x_i è decrescente, pur restando ciascun termine sempre maggiore di $\sqrt{2}$. Questo perché nel punto (2) si

calcola la media aritmetica dei due numeri x_i e $2/x_i$, la cui media geometrica vale proprio $\sqrt{2}$.

Poiché la media aritmetica è sempre maggiore o uguale a quella geometrica, $(a+b)/2 \geq \sqrt{ab}$ (il segno di uguaglianza si ha solo quando $a = b$) e poiché gli x_i sono razionali e $\sqrt{2}$ è irrazionale, il caso di $x_i = \frac{2}{x_i}$ non si verifica mai. D'altra parte la media aritmetica dimezza l'intervallo $(\frac{2}{x_i}, x_i)$, e quindi gli x_i sono successive migliori approssimazioni razionali di $\sqrt{2}$.

In tutti questi esempi $\sqrt{2}$ è il valore-limite a cui tendono le successioni dei razionali, nei fatti è definito da queste successioni, ma *non* è un numero razionale, non sta in **Q**. La contraddizione sta nel fatto che i razionali, pur essendo densi, hanno delle lacune, non sono 'completi', non potendo rappresentare numeri irrazionali come $\sqrt{2}$. La completezza dei reali consiste appunto nel comprendere anche i nuovi numeri ottenuti attraverso procedimenti di successiva approssimazione. Nei reali quindi non vi sono lacune. Riassumendo

- i numeri reali sono un ampliamento dei razionali, ampliamento imposto dall'esistenza di nuovi numeri;
- gli irrazionali si ottengono e sono definiti come valori-limite di successioni di razionali;
- gli irrazionali si possono approssimare a volontà con i razionali; nei calcoli si usano queste approssimazioni, come si fa, ad esempio, quando si pone $\pi = 3.14$.
- I numeri reali sono il corpo numerico con cui si opera in trigonometria, con i logaritmi, con gli esponenziali, con le funzioni, la geometria analitica e in generale con l'analisi matematica.

Negli argomenti che seguono si intende senza altro avviso che si opera con i numeri reali.

Potenze con esponente razionale

Sia $a > 0$, ad esempio $a = 2$. Da:

$$2^1 = 2^{\frac{2}{2}} = 2^{\frac{3}{3}} = \cdots = 2^{\frac{10}{10}} = \cdots 2^{\frac{n}{n}}$$

si ottiene:

$$2 = (2^{\frac{1}{2}})^2 = (2^{\frac{1}{3}})^3 = \cdots = (2^{\frac{1}{10}})^{10} = \cdots = (2^{\frac{1}{n}})^n$$

Per la definizione di radice quadrata, cubica,..., decima,..., ennesima di un numero è:

$$2^{\frac{1}{2}} = \sqrt{2} \qquad 2^{\frac{1}{3}} = \sqrt[3]{2} \quad \cdots \quad 2^{\frac{1}{10}} = \sqrt[10]{2} \quad \cdots \quad 2^{\frac{1}{n}} = \sqrt[n]{2}$$

Da $a^1 = a^{\frac{n}{n}}$ discende quindi $a^{\frac{1}{n}} = \sqrt[n]{a}$, per $a > 0$.

Da $(a^m)^n = a^{mn}$ viene $(a^{\frac{1}{n}})^m = a^{\frac{m}{n}} = \sqrt[n]{a^m} = (\sqrt[n]{a})^m$, sempre per $a > 0$.

Allora hanno senso espressioni come:

$$\frac{1}{\sqrt{2}} = 2^{-\frac{1}{2}} \quad \sqrt[3]{4} = 2^{\frac{2}{3}} \quad \dots$$

Esempio

$-3 = (-27)^{1/3} = (-27)^{2/6} = ((-27)^2)^{1/6} = 729^{1/6} = +3$, allora $-3 = +3$!!

Questa contraddizione obbliga a definire le potenze con esponente razionale *solo* per una base a positiva ($a > 0$).

Attenzione: $\sqrt[3]{-8} = -2$, ma $(-8)^{\frac{1}{3}}$ *non* è definita. Questo perché le due operazioni sono diverse: con la radice cubica di un numero si intende quel numero che elevato al cubo, ecc., mentre con $-8^{\frac{1}{3}}$ si estende la potenza ad esponenti razionali, con tutte, come si è appena visto, le proprietà dei numeri razionali. L'uguaglianza $x^{\frac{1}{3}} = \sqrt[3]{x}$ va letta nel senso che, se x è positivo, allora $x^{\frac{1}{3}}$ è uguale al valore della radice cubica del numero.

Teoria degli errori

Quando si misura una qualsiasi quantità inevitabilmente si associa alla misura un *errore*. Qui errore ha il significato di *incertezza*, *indeterminazione* e non quello di qualcosa di rimediabile o correggibile.

Scopo della teoria degli errori è fornire le basi matematiche per il trattamento degli errori di misura in diversi casi:

– quando si tratta di una sola misura, com'è nella maggioranza dei casi (*errore assoluto e relativo*);

– quando si tratta di indicare un valore per una gran massa di dati ricavati dalla misurazione dello stesso fenomeno (*distribuzione normale degli errori*); in questo caso risulta utile visualizzare con un grafico la distribuzione dei valori (*istogramma*);

– quando il risultato cercato dipende da una formula in cui vanno inseriti i valori misurati delle diverse grandezze che compaiono nella formula (*propagazione degli errori*);

– quando tra due grandezze x, y misurate contemporaneamente vi è una relazione lineare, per intendere, del tipo $y = mx+q$ (*best-fit*).

Errori nelle misure

La più semplice misura diretta, misurare la lunghezza di un segmento, consiste nel confrontare il segmento con le lunghezze-campione riportate sotto la forma di segni equidistanti marcati sulla stecca.

Nel compiere questa operazione è inevitabile la *stima* della posizione dell'estremo del segmento rispetto ad una coppia di segni contigui sulla riga. Allora altrettanto inevitabile è ottenere un risultato certo a meno della metà del minimo intervallo misurabile.

Possiamo ridurre l'ampiezza dell'intervallo di errore usando uno strumento di misura con una *risoluzione* maggiore (in questo caso con più divisioni per mm), ma il problema rimane, viene soltanto spostato ad un ordine di grandezza minore.

Esempio

Restando nel campo della misura di lunghezze, il nastro usato per misurare gli edifici (da 20 a 50 m) ha una risoluzione di 1/2 cm, la stecca da disegno risolve 1 mm, una buona stecca da officina 1/2 mm, il calibro risolve 1/20 mm, il comparatore 1/100 mm e i metri ottici delle macchine a controllo numerico 1 μm o meno.

Evidentemente l'intervallo di indeterminazione si riduce, ma rimane comunque presente e non ignorabile.

Legato alla presenza inevitabile degli errori è il concetto di tolleranza di lavorazione o del valore di componenti, ecc.

Questo tipo di errori, cioè le indeterminazioni di misura che si verificano casualmente, nel senso che è ugualmente probabile

sovra- o sotto-stimare la quantità misurata, si chiamano errori *accidentali*.

Errori sistematici

Se invece, per cattiva taratura dello strumento di misura (un metro 'corto'), o per l'errato procedimento di misura, ecc. oltre agli errori accidentali sempre presenti, accade che il valore misurato è in ciascuna misura sistematicamente maggiore (o minore) del valore della quantità da misurare si parla di errori *sistematici*.

Esempi

1. Uno degli errori sistematici più comuni misurando lunghezze è quello di allineamento e/o perpendicolarità. La misura della lunghezza di un rettangolo lungo e stretto (un nastro) dipende dall'allineamento della stecca rispetto al lato lungo, quindi ogni imperfetto allineamento comporta una lunghezza maggiore del vero; la misura della larghezza di un nastro dipende da come è inclinato il regolo: qualunque posizione diversa dalla perpendicolare introduce un errore sistematico e il nastro misura una larghezza maggiore di quella misurabile correttamente tenendo perpendicolare il regolo.

2. Un altro errore sistematico è quello di *parallasse*, che riguarda l'allineamento di un indice con la scala sottostante: ogni posizione diversa dalla perpendicolare introduce un errore, e per questo gli strumenti a indice di qualità hanno uno specchio per controllare la perpendicolarità. Questo problema di 'lettura' della scala viene in parte risolto con gli strumenti di misura digitali, in cui il valore viene presentato direttamente in forma numerica.

3. L'elevata precisione del calibro viene vanificata se la temperatura varia: il metallo si dilata e si introduce un errore sistematico. Infatti sul calibro è riportata la temperatura di taratura 20°C.

4. Molte (sempre di più) misure vengono condotte con strumenti tarati, cioé strumenti la cui accuratezza dipende da campioni interni o dal confronto con campioni esterni. Qui il fabbricante assegna la precisione in forma di percentuale di errore, soltanto però per misure condotte correttamente, ad una certa temperatura, ecc. Allora quando alcune di queste condizioni non sono soddisfatte, (anche per il puro e semplice invecchiamento dei componenti) si introduce nella misura anche un errore sistematico.

- Riassumendo:

– Appare chiaro che gli errori sistematici vanno accuratamente evitati (ri) facendo la misura dopo aver preso le opportune precauzioni (ad esempio, dopo aver verificato la taratura di uno strumento) che garantiscano che gli errori sistematici sianoeffettivamente trascurabili.

– È evidente la necessità di disporre di un metodo per scrivere il risultato tenendo conto degli errori accidentali.

Errore assoluto e percentuale

Se gli errori sistematici vanno eliminati, per gli errori accidentali si procede diversamente: poiché questi sono inevitabili e insiti nel fatto stesso di eseguire una misura, lo scopo della *teoria degli errori* è quello di stimarne la grandezza e dalla stima dell'errore ottenere una valutazione della attendibilità della misura.

In tal caso si vuole poter scrivere per la grandezza a:

$$a = a_m \pm \Delta a \quad \text{(unità di misura per } a\text{)}$$

dove a_m è il valore misurato di a e Δa è la stima dell'errore associato alla misura. Per semplicità di scrittura in seguito si indica con *a sia* la grandezza *che* il suo valore. Chiamiamo Δa errore *assoluto* su a.

La stima di Δa nel caso più semplice di una sola misura coincide con la semiampiezza del minimo intervallo misurabile.

Esempi

1. Misurando la temperatura ambiente con un termometro digitale (risoluzione 0.1 °C) si conviene che la temperatura misurata sia $\mathrm{T} = 23.2 \pm 0.05\,^{\circ}\mathrm{C}$, ovvero che la temperatura reale, peraltro sconosciuta, cada tra 23.15 e 23.25 °C.

Oltre all'errore assoluto si incontra più spesso l'errore *relativo* o *percentuale*, definito come:

$$\text{errore relativo} = \frac{\Delta a}{a}$$

ovvero l'errore assoluto posto in rapporto con a, misurato usando come unità di misura il valore di a.

2. L'incertezza di una misura di lunghezza è $\Delta a = 0.5$ mm. Se $a = 162$ mm allora $\frac{\Delta a}{a} = 0.5/162 \approx 0.3\%$, mentre se $a = 16$ mm è $\frac{\Delta a}{a} = 0.5/16 \approx 3\%$, 10 volte maggiore del precedente.

3. L'errore percentuale di una certa misura è dell'1%. Se si parla di una lunghezza di 10 cm ciò equivale all'errore assoluto di $0.01 \times 100 = 1$ mm , mentre se si parla di una pista di atletica lunga 100 m l'errore assoluto è di ben 1 m, del tutto inaccettabile.

L'errore relativo dà una indicazione quantitativa della accuratezza della misura indipendentemente dal valore misurato e il confronto degli errori relativi di due misure consente di stabilire quale è piú accurata. Inoltre, come si vedrà piú avanti, un elevato errore relativo in una delle misure che concorrono, mediante una formula, a determira un certo risultato, vanifica la minor incertezza delle altre misure.

Statistica

I metodi della statistica permettono di ricavare informazioni *verosimili* da qualsivoglia *esperimento*: un sondaggio di opinione, una misura di laboratorio, un controllo di qualità sui campioni di un prodotto, ecc. purché si disponga di un numero sufficiente di dati, che chiamiamo tutti indifferentemente *misure*.

Si ricorre al trattamento statistico dei dati nel caso di misure ripetute di una stessa quantità, fatto inusuale nelle comuni misurazioni, ma importante per ricavare dati che riassumono, ad esempio, l'incidenza dei comportamenti di una gran parte della popolazione.

Esempi

1. Nelle inchieste di mercato, nei sondaggi di opinione, ecc. un dato sempre richiesto è l'età dell'intervistato. Lo scopo più evidente è quello di porre in relazione l'età e le risposte, ma è anche importante che il campione così estratto dalla popolazione sia verosimile e quindi riproduca la piramide delle età della popolazione, un dato questo ben noto a chi si occupa di demografia.

L'età è un numero che varia, diciamo, tra 11 e 90, cioè può prendere 80 valori diversi, troppi! Si dice che i dati sono *dispersi*. In effetti vi è un eccesso di dettaglio. Allora, poiché molto di rado è significativo distinguere tra le opinioni, per dire, di un ventiduenne e di un ventitreenne, conviene *raggruppare* le misure in intervalli più ampi, in questo caso intervalli di 10 anni, le fasce di età $11-20$, $21-30$, ecc. In tal modo i dati diventano subito più comprensibili.

• Si dice *frequenza* n_i il numero di valori che cadono nell'i-esimo intervallo. La somma delle frequenze dà il numero n delle misure: $n = \sum n_i$. Con i dati raggruppati si forma una tabella, cosa fattibile a mano fino a un centinaio di misure, con adatti programmi di calcolo (foglio elettronico) per masse di dati più consistenti.

2. La tabella riporta i punteggi (il numero di risposte esatte) di un test con 10 domande.

punti	0	1	2	3	4	5	6	7	8	9	10	p_i
freq.	0	0	1	2	6	11	15	10	5	2	2	n_i

Qui gli intervalli sono tutti ampi 1 punto. In totale i test valutati sono

$$n = \sum n_i = 1 + 2 + 6 + 11 + 15 + 10 + 5 + 2 + 2 = 54.$$

È utile visualizzare i dati con un *istogramma*. Si assegna, per esempio, un'unità (un quadretto) sia al punteggio (asse X) che alla frequenza (asse Y). Per ogni punteggio si anneriscono verticalmente un numero di quadretti pari alla frequenza.

Per come sono costruiti, gli istogrammi rendono comprensibile a colpo d'occhio la distibuzione dei dati. Infatti

– l'area sotto l'istogramma è il numero di dati, in questo caso l'area sotto l'istogramma conta 54 quadretti.

– il valore centrale della distribuzione e quello con maggior frequenza sono facilmente identificabili.

– di solito, la forma dell'istogramma è all'incirca quella di una campana. Deviaziooni macroscopiche da questa forma, per esempio la presenza di due picchi devono far sospettare che si stiamo misurando due fenomeni e non uno solo.

Infine, per ottenere un ottimo istogramma non occorre essere dei disegnatori esperti, ma basta introdurre i dati in una tabella del foglio elettronico, di cui l'istogramma è una, tra le molte possibili, presentazione grafica dei dati.

Le frequenze *relative* (in percentuale rispetto a n) si calcolano come $nr_i = n_i/n$

p_i	0	1	2	3	4	5	6	7	8	9	10
n_i	0	0	1	2	6	11	15	10	5	2	2
$\frac{n_i}{n}(\%)$	0	0	1.8	3.7	11.1	20.3	27.8	18.5	9.3	3.7	3.7

Ora $\sum nr_i = 1.00$ (100% per le percentuali); il corrispondente istogramma è *normalizzato*, cioè l'area sottostante vale 1. Ciò torna utile per confrontare distribuzioni ottenute da campioni di diversa dimensione.

La frequenza relativa rappresenta la frequenza per un campione normalizzato (ad esempio, riferito a 100 parti percentuali) e perciò reso *indipendente* dalle sue effettive dimensioni.

• Adatto ad una rappresentazione in termini di frequenze relative è il grafico a torta, dove all'angolo-giro 2π corrisponde il 100%. Anche questa presentazione grafica è standard nei fogli elettronici.

Spesso nelle presentazioni le misure vengono raggruppate in intervalli di ampiezza diversa. Nell'esempio 2 possiamo assegnare una valutazione complessiva in questo modo

punteggio	uscite	frequenza	valutazione
0 .. 3	3	5.6%	D
4 .. 5	17	31.5%	C
6 .. 8	30	55.6%	B
9 .. 10	4	7.4%	A

In questo caso, nel disegnare l'istogramma occorre tener conto che l'area sotto l'istogramma deve rappresentare il 100% delle misure e, rispettivamente, l'area di ciascuna colonna la percentuale corrispondente a quell'intervallo.

Tenendo l'intervallo di un quadretto per il punteggio, 100 quadretti per l'area sotto l'istogramma, l'altezza h_i della colonna si calcola (in quadretti!) come

$$h_i = \frac{\text{freq. rel} \times 100}{\text{larghezza della colonna}}.$$

valutazione	intervallo	ampiezza	altezza	quadretti
D	0 .. 3	2	$100 \times 0.056/4$	≈ 1.4
C	4 .. 5	2	$100 \times 0.315/2$	≈ 15.7
B	6 .. 8	3	$100 \times 0.556/3$	≈ 18.5
A	9 .. 10	2	$100 \times 0.074/2$	≈ 3.7

e l'area totale sotto il grafico è appunto

$4 \times 1.4 + 2 \times 15.7 + 3 \times 18.5 + 2 \times 3.7 \approx 100$ quadretti.

Media e scarti

Aggiungendo alla tabella una riga con i prodotti $n_i \times p_i$:

punti p_i	0	1	2	3	4	5	6	7	8	9	10
freq. n_i	0	0	1	2	6	11	15	10	5	2	2
freq. $\times$ punti $n_i\, p_i$	0	0	2	6	24	55	90	70	40	18	20

si calcola il punteggio medio $\overline{p}$ (il numero di risposte esatte in *media*) come:

$$\begin{aligned}\overline{p} &= \frac{\sum n_i\, p_i}{n} \\ &= \frac{2+6+24+55+90+70+40+18+20}{54} = \frac{325}{54} \approx 6.0.\end{aligned}$$

Si aggiunge ora alla tabella una riga per gli gli scarti $\Delta p_i = p_i - \overline{p}$ calcolati rispetto al valor medio, e per il prodotto del loro quadrato per la corrispondente frequenza.

p_i	0	1	2	3	4	5	6	7	8	9	10
n_i	0	0	1	2	6	11	15	10	5	2	2
$n_i\, p_i$	0	0	2	6	24	55	90	70	40	18	20
Δp_i	-6	-5	-4	-3	-2	-1	0	1	2	3	4
$n_i(\Delta p_i)^2$	0	0	16	18	24	11	0	10	20	18	32

Infine si calcola lo *scarto quadratico medio* o *varianza*

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum n_i\,(\Delta p_i)^2}{n-1}} \\ &= \sqrt{\frac{16+18+24+11+10+20+18+32}{53}} \\ &= \sqrt{\frac{139}{53}} \approx 1.6.\end{aligned}$$

Se si assume che il valor medio sia il valore piú probabile, quello che riassume fedelmente in un solo dato il complesso delle 54 misure, la varianza è la stima dell'incertezza sul valor medio e, come si vedrà pi'u avanti, permette di stimare la verosimiglianza del risultato.

Media, moda, mediana, dispersione

Dalla tabella, oltre alla *media* o valor medio, si possono ottenere altre quantità caratteristiche di una distribuzione di misure, quantità che si possono riportare graficamente nell'istogramma.

• Si dice *moda* la misura che si presenta più volte delle altre, per cui la frequenza relativa è massima. Qui la frequenza massima 15 (28%) si ha per $p = 6$ e quindi la moda è 6.

• Si chiama *mediana* il valore centrale della distribuzione. Se vi è un numero pari di misure per la mediana si considera la media dei due valori centrali. In questo esempio le misure vanno da 2 a 10; il valore centrale, la mediana, è $(10 + 2)/2 = 6$.

• La *dispersione* è una misura di quanto i valori sono o meno addensati intorno al valor medio. Per essa si può considerare l'intervallo tra la misura minima e quella massima. Qui $10-2 = 8$. Più spesso si calcola come l'intervallo tra un quarto e tre quarti dei valori. In questo caso $7 - 2 = 5$.

Teoria degli errori

La giustificazione del metodo usato per ricavare il valore più probabile di una grandezza di cui si conoscano n valori misurati è una conseguenza dell'ipotesi che ciascuno dei valori misurati x_i si possa pensare scritto come:

$$x_i = x_v + \delta_i, \quad i = 1 \dots n$$

dove x_v è il valore 'vero' e δ_i è l'i-esima fluttuazione. Questa espressione è priva di utilità pratica, visto che le uniche quantità *note* sono i valori misurati x_i: valore 'vero' e fluttuazioni fanno parte del modello della misura, ma restano purtroppo ignoti. Si possono però fare delle ipotesi su come sono distribuite le fluttuazioni δ_i. Possiamo ragionevolmente aspettarci che:

a) siano più probabili fluttuazioni piccole in valore assoluto, ovvero che la probabilità $P(\delta)$ di ottenere valori misurati con una deviazione δ tenda a 0 al crescere della grandezza della deviazione, $P(\delta) \to 0$ quando $\delta \to \infty$;

b) la distribuzione delle fluttuazioni non dipenda dal loro segno, ovvero che la distribuzione sia simmetrica rispetto al valor 'vero'; $P(-\delta) = P(\delta)$;

c) sia massima la probabilità di incontrare fluttuazioni *nulle*, $P(\delta) = \max \iff \delta = 0$.

Da queste ipotesi viene che per un numero arbitrariamente grande di valori misurati la somma delle fluttuazioni δ_i tende a zero; allora per la media aritmetica dei valori misurati vale

$$\overline{x} = \frac{\sum x_i}{n} \to x_v \quad \text{quando } n \to \infty.$$

In realtà è disponibile un insieme *finito* di valori misurati; si può però considerare questo insieme come un *campione scelto a caso* dall'universo delle infinite misure possibili.

• Il fallimento di un sondaggio o di un'inchiesta di mercato dipende in massima parte dalla scelta di un campione non rappresentativo. Alcuni casi di sondaggi sul voto espresso (exit-poll) mostrano l'importanza di una accurata scelta del campione.

• Nell'esempio del test di matematica vengono considerati *tutti* i dati disponibili. Ovviamente non ha senso considerarne solo una parte. È altrettanto vero però che spesso è impossibile o estremamente costoso testare *tutta* la produzione. Nell'esempio 1 i 100 bulloni sono un *campione* di oggetti, così come le persone intervistate sul voto già espresso (si parla di un *sondaggio*) sono un campione estratto dalla popolazione dei votanti. Necessariamente i campioni sono finiti e, se non altro per ragioni economiche, limitati a $10^2 \div 10^4$ oggetti o individui. Fanno eccezione alcuni fenomeni (vita media, malattie, consumi, ecc.) per cui esistono dati e statistiche su periodi di 1-200 anni e su milioni di individui.

• Perché i risultati siano degni di fiducia occorre che i campioni riproducano la distribuzione di misure o di voti nell'insieme dei bulloni o della popolazione, ovvero che siano estratti *a caso* da questi insiemi. Esistono precisi protocolli per l'ottenimento di campioni rappresentativi. Essi fanno parte di quelle norme che consentono, attraverso il *controllo statistico di qualità*, di certificare il prodotto (oltre a migliorare la produzione).

• In questo caso per il campione valgono le proprietà a, b, c. Accettare questo punto di vista significa dire che se il campione è rappresentativo allora $\overline{x}$ è il valore che più si avvicina al valor vero, anzi si può assegnare come risultato della misura proprio $\overline{x}$. Chiamiamo questa quantità *valor medio*. A partire da questa calcoliamo le deviazioni rispetto al valor medio del campione, ovvero gli *scarti*:

$$\Delta x_i = x_i - \overline{x}.$$

Per come questi sono ottenuti vale

$$\sum \Delta x_i = 0\,.$$

• Dimostrazione: $\sum \Delta x_i = \sum (x_i - \overline{x}) = \sum x_i - n\overline{x} = 0$.

• Assegnare un valore (misura) ad una quantità implica la necessità di una stima sulla sua verosimiglianza, ovvero sulla incertezza della misura, l'*errore*. Un primo modo di assegnare l'errore consiste nel valutare la deviazione media intorno al valor medio, lo *scarto medio* Δx. Si considerano gli scarti in valore assoluto e se ne calcola la media aritmetica:

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

Più usata e significativa è però la quantità[1]

$$\sigma = \sqrt{\frac{\sum \Delta x_i^2}{n}}$$

detta *deviazione standard* oppure *scarto quadratico medio*. Si può dimostrare rigorosamente che questa quantità è in assoluto minima quando gli scarti sono calcolati come $\Delta x_i = x_i - \overline{x}$, ovvero si definisce il valor medio come media aritmetica dei valori misurati. Si può allora scrivere il risultato come

$$x = \overline{x} \pm \sigma \qquad \text{oppure} \quad x = \overline{x} \pm \Delta x.$$

La prima scrittura è quella usata universalmente; è possibile incontrare la seconda e anche altre forme di scrivere il risultato.

• Il metodo delineato consente anche di verificare l'effettiva verosimiglianza del campione dei valori misurati e cioé valutare quanto la distribuzione dei valori misurati è in accordo con la distribuzione teorica ricavata con i metodi del calcolo delle probabilità .

• Nella distribuzione teorica lo scarto è una variabile continua; si ottiene una funzione di distribuzione[2] del tipo:

$$\mathrm{erf}(z) = \frac{h}{\sqrt{\pi}}\, e^{-h^2 z^2}$$

chiamata *gaussiana* o *funzione-errore*; z è lo scarto e h è il parametro di precisione della distribuzione. La funzione $\mathrm{erf}(z)$ gode delle proprietà:

a') $\lim_{z\to\infty} \mathrm{erf}(z) = 0$: sono più probabili fluttuazioni piccole in valore assoluto;

b') $\mathrm{erf}(-z) = \mathrm{erf}(+z)$: la distribuzione è simmetrica rispetto a $z = 0$ ($\mathrm{erf}(x)$ è una funzione pari).

c') $\mathrm{erf}(0) = \frac{h}{\sqrt{\pi}}$ è il massimo assoluto per la distribuzione.

Qui $\mathrm{erf}(0) \propto h$: tanto più grande è h, tanto più alto risulta il picco intorno allo zero. Ora h dipende dalla deviazione standard σ attraverso

$$h = \frac{\sqrt{2}}{\sigma}, \quad \text{ovvero} \quad h\sigma = \sqrt{2},$$

[1] Al denominatore si usa anche scrivere $n-1$. La giustificazione sta nel fatto che, poiché $\sum \Delta x_i = 0$ vi sono $n-1$ e non n valori indipendenti. La differenza acquista significato solo per n piccolo.

[2] Ottenuta da De Moivre nel 1733 passando al limite per una variazione continua degli scarti nella distribuzione binomiale, cosí chiamata perchè ottenuta dalla formula del binomio di Newton

quindi tanto minore è l'incertezza sul valor medio, tanto maggiore è la precisione della misura. Si nota come le proprietà a', b', c' coincidano con quelle poste come ipotesi per il modello delle misure. Inoltre $\mathrm{erf}(z)$ è tale che

d) $\int_{-\infty}^{+\infty} \mathrm{erf}(z)\, dz = 1$, la funzione è normalizzata.

• La questione importante è verificare il buon accordo tra la distribuzione teorica e quella ottenuta dalla misura: è infatti questa somiglianza che autorizza a ritenere il valor medio il più probabile per il risultato della misura. Un controllo molto semplice riguarda il rapporto $\frac{\sigma}{\Delta}$: teoricamente deve valere 1.25 per qualsiasi valore di h, cioè comunque sia (o meno) precisa la misura. Un'altro verifica consiste nel contare gli scarti che cadono in intervalli simmetrici ampi 2σ, 4σ, 6σ intorno all'origine.

z/σ	$\int_{-z}^{+z}$	scarti in %
1	0.68	68%
2	0.95	95%
3	0.995	99.5%

3. Gli intervalli dei valori 'normali' per vari esami clinici sono ottenuti da un campione di popolazione in buono stato di salute generale, senza patologie. Essi tuttavia si riferiscono al 95% dei casi possibili, ovvero si è preso l'intervallo ampio 2σ intorno al valor medio. Ciò significa che esiste un 5% della popolazione che, pur non presentando patologie (individui sani), presenta 'valori clinici', per esempio il numero di globuli rossi, fuori dall'intervallo accettato come normale.

• Si può verificare anche visivamente l'accordo tra le due distribuzioni. sovrapponendo all'istogramma dei valori il grafico della distribuzione teorica. Per ottenere quest'ultimo innanzitutto occorre cambiare la forma della curva ponendo $h = \frac{\sqrt{2}}{\sigma}$, dove σ è la deviazione standard del caso. Cambiato il profilo, occorre mettere materiale: l'area sotto la curva deve divenire n volte più grande e deve esser suddivisa in tratti ampi Δz.

Qui Δz è l'ampiezza degli intervalli in cui è suddivisa l'ascissa dell'istogramma. Allora l'ordinata corrispondente sulla curva vale

$$n \times \mathrm{erf}(z) \times \Delta z = \frac{n\sqrt{2}}{\sqrt{\pi}\,\sigma} e^{-\frac{2}{\sigma^2} z^2}, \quad \Delta z = A e^{-B z^2},$$

$$\text{dove} \quad A = \frac{n\sqrt{2}}{\sqrt{\pi}\sigma} \Delta z, \quad B = \frac{2}{\sigma^2}$$

Qui le quantità $n,\ \Delta z,\ \sigma$ son tutte note. Si traccia ora il grafico per punti sovrapposto all'istogramma.

4. Un campione di 100 chiodi di lunghezza nominale $l = 35.0$ mm viene misurato con il calibro e risulta così composto:

l_i	34.6	34.7	34.8	34.9	35.0	35.1	35.2	35.3	mm
n_i	5	12	8	25	26	14	6	4	

NB: i chiodi sono stati misurati con un calibro. La loro lunghezza perciò dovrebbe essere nota con una incertezza di 0.05 mm. In questo modo però la dispersione per sole 100 misure risulta eccessiva. Allora si raggruppano le misure che cadono in una intervallo ampio 0.1 mm. In tal modo per l vi sono 8 valori diversi e l'istogramma risulta significativo (figura).

Prepariamo una tabella con colonne per i,n_i, l_i, Δl_i, $(\Delta l_i)^2$, $n_i|\Delta l_i|$, $n_i(\Delta l_i)^2$ e calcoliamo il valor medio. Spesso i valori misurati sono stati raggruppati e ordinati come nella tabella dove $\sum n_i = n$. In questo caso il valor medio si calcola come:

$$\overline{x} = \frac{\sum x_i n_i}{\sum n_i}$$

È $\bar{l} = 34.9\,4$. Riempiamo la tabella.

i	n_i	l_i	Δl_i	$(\Delta l_i)^2$	$n_i\|\Delta l_i\|$	$n_i(\Delta l_i)^2$
1	5	34.6	-0.3	0.09	1.5	0.45
2	12	34.7	-0.2	0.04	2.4	0.48
3	8	34.8	-0.1	0.01	0.8	0.08
4	25	34.9	0.0	0.0	0.0	0.0
5	26	35.0	$+0.1$	0.01	2.6	0.26
6	14	35.1	$+0.2$	0.04	2.8	0.56
7	6	35.2	$+0.3$	0.09	1.8	0.54
8	4	35.3	$+0.4$	0.16	1.6	0.64

È $\Delta l \approx 0.14$ e $\sigma_l \approx 0.17$, il rapporto $\frac{\Delta}{\sigma}$ vale ≈ 1.21, abbastanza vicino al teorico 1.25. Inoltre nell'intervallo $[-\sigma,\ \sigma]$ cadono 25+25+3+3 = 65 misure mentre dovrebbero essere 68 (68%), in $[-2\sigma,\ 2\sigma]$ cadono 65+14+42+5 = 96 (95%), in $[-3\sigma,\ 3\sigma]$ cadono 100 misure (99.5%).

La funzione distribuzione calcolata per questo esempio è

$$nf(z)\Delta z = \frac{100 \cdot 4.18}{\sqrt{\pi}} e^{-(4.18z)^2} \cdot 0.1 \approx 23.6 e^{-17.5z^2}\,.$$

dove $n = 100$, $\Delta z = 0.1$, $h = \frac{\sqrt{2}}{2\sigma} \approx \frac{0.707}{0.169} = 4.18$. Si calcolano i valori corrispondenti al centro degli intervalli in cui è suddiviso l'istogramma e si sovrappone il grafico così ottenuto all'istogramma.

NB: I dati ottenuti portano ad assegnare una cifra significativa in più al valore finale. Per 100 misure la precisione aumenta di un fattore $\sqrt{100} = 10$ e allora si scriverà

$$\bar{l} = 34.94 \pm 0.17 \quad (\sigma) \quad \text{oppure} \quad \bar{l} = 34.94 \pm 0.14 \quad (\Delta)\,.$$

Questo risultato può sorprendere, ma è dovuto a come è definito σ: la varianza decresce, a parità di somme degli scarti, come $1/\sqrt{n}$. Intuitivamente, più misure vengono eseguite, minore è l'incertezza sul valore medio.

Esercizi

1. Si lanciano 400 volte due dadi ottenendo i punteggi:

punti	12	11	10	9	8	7	6	5	4	3	2
uscite	13	21	39	40	61	67	55	41	27	23	13

Trovare punteggio medio, scarto medio e deviazione standard.

2. Un campione di 200 lampadine viene testato (lasciandole accese finché non 'bruciano') allo scopo di stabilire la loro vita media. I risultati dell'esperimento sono dati in tabella. Trovare le varie quantità significative per questo esperimento.

durata		lampadine
da	a	n°
600	800	2
800	1000	7
1000	1200	29
1200	1400	54
1400	1600	63
1600	1800	34
1800	2000	8
2000	2200	3

Si tiene il valore centrale degli intervalli di 200 ore. Riportare il lavoro su foglio elettronico.

3. Raggruppare i dati dell'esempio 2 in due soli intervalli, presentando i risultati come 'sufficienti e meglio', voto 6 o più, 'insufficienti o peggio', voto 5 o meno. Tracciare l'istogramma.

Propagazione degli errori

Si parla di *propagazione* degli errori quando sia necessario valutare l'incertezza del risultato di una misura *indiretta*.

Esempio

La procedura per misurare l'area di un rettangolo consiste nel misurare con la stecca i lati e calcolare l'area applicando Area = base×altezza. La misura dell'area non è diretta, ma indiretta: si ottiene applicando una operazione arimetica a due misure dirette di lunghezza. Si considerano come dirette anche le misure effettuate con strumenti tarati, ma una misura ottenuta, per esempio, come rapporto tra due misure con strumenti tarati è ovviamente ancora indiretta.

Il problema dell'esempio è valutare quantitativamente quanto gli errori nelle misure dei lati influiscano sul valore dell'incertezza sull'area.

Una essenziale ipotesi semplificativa è considerare gli errori nelle misure come *indipendenti*, il che è verosimile quando si misurano i lati di un rettangolo, ma non in tutti i casi. Generalizzando si danno formule variamente complicate in cui una grandezza dipende da altre tramite operazioni.

Per valutare l'errore risultante dalla propagazione degli errori da ciascuna misura al risultato di una 'formula' anche complicata, occorre ricordare che questa non fa che descrivere una certa successione delle quattro operazioni dell'aritmetica. Quindi innanzitutto si vogliano esaminare da questo punto di vista le quattro solite operazioni dell'aritmetica.

Risolti questi casi semplici e ottenute le valutazioni dell'errore per essi, la trattazione di casi più complicati si riduce ad una applicazione ripetuta delle regole trovate per le quattro operazioni.

La stima di una quantità si esprime scrivendo $x = x_0 \pm \Delta x$ intendendo che x appartiene all'intervallo $[x_0 - \Delta x, x_0 + \Delta x]$, ovvero che $x_0 - \Delta x \le x \le x_0 + \Delta x$. Nel caso di una quantità misurata, convenzionalmente l'ultima cifra non è significativa e Δx misura l'incertezza. Si scrive $x = 12.3 \pm 0.05$ intendendo che x è uno dei valori nell'intervallo $[12.25, 12.35]$: appunto l'ultima cifra (3) non è significativa e le si assegna una incertezza di mezza divisione.

Si vogliono ricavare le regole per ottenere il risultato di $[a, b] \diamond [c, d]$, dove $\diamond$ è una delle quattro operazioni $+, -, \times, \div$ e $[a, b]$, $[c, d]$ sono due qualsiasi intervalli. Allora, se $x \in [a, b]$ e $y \in [c, d]$, si definisce

$$[a, b] \diamond [c, d] := \{x \diamond y \mid x \in [a, b], y \in [c, d]\}.$$

Qui $\{\ldots\}$ sta per l'insieme di tutti i valori $x \diamond y$ ottenuti con l'operazione $\diamond$. Per addizione e sottrazione valgono

	Esempio
$[a, b] + [c, d] = [a + c, b + d]$	$[2, 3] + [5, 6] = [7, 9]$
$-[a, b] = [-b, -a]$	$-[2, 3] = [-3, -2]$
$[a, b] - [c, d] = [a - d, b - c]$	$[5, 6] - [2, 3] = [2, 4]$

Per ottenere gli estremi dell'intervallo nel caso di moltiplicazione e divisione si devono ordinare i risultati delle operazioni sui numeri a, b, c, d.

$$\begin{aligned}[a, b] \times [c, d] &= [\min\{ac, ad, bc, bd\}, \max\{ac, ad, bc, bd\}] \\ [a, b]/[c, d] &= [\min\{a/c, a/d, b/c, b/d\}, \max\{a/c, a/d, b/c, b/d\}]\end{aligned}$$

Allora, ad esempio,

$[2, 3] \times [5, 6] = [10, 18]$, ma $[-3, -2] \times [5, 6] = [-18, -10]$;
$[5, 6] \div [2, 3] = [5/3, 6/2] = [5/3, 3]$, mentre
$[5, 6] \div [-3, -2] = [-6/2, -5/3] = [-3, -5/3]$.

- La divisione richiede attenzione: il divieto di dividere per zero continua a valere e quindi, perché l'operazione sia definita, l'intervallo $[c, d]$ *non* deve contenere lo zero, ovvero $c \cdot d > 0$

L'aritmetica degli intervalli comprende i numeri ordinari dato che si può scrivere $n = [n, n]$ e riottenere le operazioni per i numeri reali. In particolare se $[a, b] = [1, 1] = 1$, si ha per il reciproco di $[c, d]$

$$\frac{1}{[c,d]} = \left[\frac{1}{d}, \frac{1}{c}\right], \quad c \cdot d > 0.$$

Ad esempio $1/[2, 3] = [1/3, 1/2]$ e $1/[-3, -2] = [-1/2, -1/3]$.

Il calcolo con gli intervalli si può estendere alle funzioni, scrivendo

$$f([a, b]) := \{f(x)|x \in [a, b]\},$$

ovvero

$$f : [a, b] \mapsto [\inf f(x), \sup f(x)], \quad x \in [a, b]$$

Ad esempio, $\sqrt{[1.2, 1.7]} = [1.1, 1.3]$, $\sin([1, 5, 1.6]) = [\sin(1.5), 1]$.

Come scrivere i risultati delle operazioni su intervalli nel consueto modo? Si passa da una notazione all'altra con

$$x = \frac{a+b}{2}, \quad \Delta x = \frac{b-a}{2}, \quad \text{dove } a = x - \Delta x, \quad b = x + \Delta x.$$

Esempio

Si calcola il volume $V = abc$ di un parallelepipedo di lati $a = 12.3 \pm 0.1\,\text{cm}$, $b = 19.7 \pm 0.1\,\text{cm}$, $c = 38.1 \pm 0.1\,\text{cm}$. Allora $a \in [12.2, 12.4]$, $b \in [19.6, 19.8]$, $c \in [38.0, 38.2]$. Risulta $abc \in [12.2 \times 19.6 \times 38.0, 12.4 \times 19.8 \times 38.2]$, ovvero $abc \in [9\,080, 9\,380]$, e finalmente $abc = 9\,230 \pm 150\,\text{cm}^3$.

Esercizio. Scrivere una formula per la divisione se a, b, c, d sono tutte quantità positive.

Per operare con gli errori relativi nel caso di somma o sottrazione si applica

$$\left|\frac{\Delta(x \pm y)}{x \pm y}\right| \leq \frac{|\Delta x| + |\Delta y|}{|x \pm y|}$$

• La sottrazione risulta fonte di errori relativi particolarmente grandi quando si opera tra due numeri a, b che siano confrontabili. In questo caso il denominatore può essere dello stesso ordine di grandezza del numeratore e l'errore relativo essere anche del 50 – 100%. Queste eventualità vanno evitate accuratamente, ad esempio avendo l'accortezza di condurre le operazioni in una successione opportuna. Ad esempio, si sommano *separatamente* le varie quantità positive e negative evitando di sottrarre direttamente quantità tra loro confrontabili.

Nel caso di moltiplicazione/divisione si appplica

$$\left|\frac{\Delta(x\cdot y)}{x\cdot y}\right| \leq \left|\frac{\Delta x}{x}\right| + \left|\frac{\Delta y}{y}\right| + \text{termini di ordine sup.}$$
$$\left|\frac{\Delta(x/y)}{x/y}\right| \leq \left|\frac{\Delta x}{x}\right| + \left|\frac{\Delta y}{y}\right| + \text{termini di ordine sup.}$$

Si dà la dimostrazione del risultato per la moltiplicazione, mentre non si riporta quella per la divisione, decisamente più complicata.

$$\begin{aligned}\Delta(xy)| =&|(x\pm\Delta x)(y\pm\Delta y)-xy|\\ =&|\pm x\Delta y\pm y\Delta x\pm\Delta x\Delta y|\\ \leq&|x\Delta y|+|y\Delta x|+|\Delta x\Delta y|.\end{aligned}$$

Si divide per $|xy|$

$$\begin{aligned}\left|\frac{\Delta(xy)}{xy}\right| \leq& \left|\frac{x\Delta y}{xy}\right| + \left|\frac{y\Delta x}{xy}\right| + \left|\frac{\Delta x\Delta y}{xy}\right|\\ \leq& \left|\frac{\Delta y}{y}\right| + \left|\frac{\Delta x}{x}\right| + \left|\frac{\Delta x\Delta y}{xy}\right|\end{aligned}.$$

L'ultimo termine è di ordine superiore ($\Delta x\Delta y$) rispetto agli altri termini e quindi può essere trascurato.

In effetti se, ad esempio, $\Delta x/x \approx \Delta y/y = 0.01$ (1%), è $\Delta x\Delta y/xy \approx 0.0001 = 1\times 10^{-4}$, trascurabile rispetto agli altri termini che sono dell'ordine di 1×10^{-2}.

Esempio

Trovare l'errore relativo sul volume della sfera quando il raggio r è noto con un errore relativo pari all'1%, ovvero $\Delta r/r = 0.01$.

Poiché $V = \frac{4}{3}\pi r^3$ si ha

$$\frac{\Delta V}{V} = \frac{\Delta r}{r} + \frac{\Delta r}{r} + \frac{\Delta r}{r} = 3\frac{\Delta r}{r} = 0.003 = 3\%.$$

- Allora, se una quantità z dipende da altre quantità secondo una formula del tipo $z = t^l\cdot x^m\cdot y^n\cdots$, si ricava l'errore relativo su z come

$$\frac{\Delta z}{z} = |l|\frac{\Delta t}{t} + |m|\frac{\Delta x}{x} + |n|\frac{\Delta y}{y} + \ldots$$

Poichè le formule sommano gli errori relativi con il loro esponente, il termine che influenza maggiormente l'errore relativo sul risultato è quello con l'errore relativo piú grande, cioè quello di minor precisione. Quindi è necessario condurre tutte le misure necessarie con un grado di precisione confrontabile, ovvero con errori relativi simili in grandezza.

Regressione lineare

Le scoperte della fisica derivano dall'osservazione diretta dei fenomeni (naturali o piú spesso riprodotti in laboratorio, *esperimenti*), dalla misura delle grandezze in gioco e dalla successiva interpretazione e formalizzazione delle regolarità che si presentano nei fenomeni osservati.

Cosí tutti hanno certamente visto una tavola flettersi sotto il peso applicato nel centro, ma pochi hanno dato di questo fenomeno una interpretazione razionale, in grado non solo di interpretarlo, ma di renderlo prevedibile e, ad esempio, usarlo per misurare una forza.

Ora, nell'interpretare i fenomeni della fisica e nel rappresentarli con formule, spesso la relazione tra le due grandezze misurate è *lineare*, e, dette x e y le due grandezze in gioco, la relazione si scrive nella nota forma $y = mx + q$. La relazione lineare esprime il fatto che la variazione Δy di una certa grandezza fisica è *direttamente proporzionale* alla variazione Δx di un'altra grandezza fisica (*linearità*). Ad esempio, la variazione Δh della freccia di una sbarra caricata ad un estremo con un peso p (*flessione*) è direttamente proporzionale alle variazione Δp della forza peso applicata (*elasticità*, legge di Hook); oppure la variazione della lunghezza di un regolo metallico è direttamente proporzionale alla variazione di temperatura (*dilatazione lineare*).

La relazione $y = mx + q$ è caratterizzata dalle costanti m e q che, in generale, varieranno in funzione della sostanza in esame. Quindi lo scopo dell'esperimento è anche quello di quantificare m e q e, ad esempio, stabilire così che l'acciaio e il rame sono entrambi elastici, ma che l'acciaio è piú adatto per fare molle.

I dati

Dall'esperimento si ottiene un certo insieme di $2n$ dati sperimentali (*misure*), che in generalità si indicano come $P(x_i, y_i)\, i = 1, \ldots, n$, nella forma di n coppie di valori misurati delle grandezze sotto osservazione. Ad esempio, per la dilatazione dei solidi x_i potrà corrispondere alla i-esima temperatura misurata e y_i alla determinazione della lunghezza del regolo metallico.

È assai naturale disporre i punti su un piano XOY, osservando poi come i punti siano *solo approssimativamente* allineati, ovviamente a causa delle inevitabili fluttuazioni (errori) nella determinazione del valore delle grandezze. Il problema che si pone è come ricavare i parametri quantitativi dell'esperimento, cioè le costanti m e q che caratterizzano la retta, ovvero la relazione tra le grandezze fisiche.

Il metodo dei minimi quadrati (*best fit*)

Ovviamente le rette che passano per l'insieme di punti sono in numero infinito, e, in assenza di un criterio di scelta, stabilire quale

retta si *adatta* meglio all'insieme dei dati sperimentali, risulta del tutto aleatorio.

Il criterio comunemente accettato cerca la retta che rende *minima* la somma dei quadrati delle distanze dei punti P_i dalla retta $y = mx + q$ (*minimi quadrati*).

Ricordando l'espressione per le distanza punto-retta si ha per il quadrato della i-esima distanza

$$d^2_{\perp i} = \frac{(y_i - mx_i - q)^2}{1 + m^2} \qquad (1.0$$

e per la somma dei quadrati[1]

$$f(m,q) = \sum_{i=1}^{n} d^2_{\perp i} =$$
$$= \frac{\sum y_i^2 - 2m \sum x_i y_i + m^2 \sum x_i^2 - 2q \sum y_i + 2mq \sum x_i + nq^2}{1 + m^2}$$

La somma dei quadrati delle distanze è una funzione delle due variabili m, q: si ricordi che i termini noti sono dati dalle coppie di valori (x_i, y_i) delle misure sperimentali.

Applicando la tecnica standard si cerca la coppia m, q tale che $\frac{\partial f}{\partial m} = 0$ e $\frac{\partial f}{\partial q} = 0$. Questo si traduce in un sistema di due equazioni in due incognite che sarebbe facilmente risolubile se non fosse che la presenza di $1+m^2$ al denominatore fa sí che la derivata parziale rispetto a m contenga termini in m^2 e m^3. Ciò rende assai pesante la ricerca della soluzione per m e ovviamente per q.

Una semplificazione necessaria

Si introduce allora una semplificazione, considerando l'errore concentrato soltanto nella misura y_i: la misura x_i diventa *esatta*.

Se la distanza punto-retta diventa la distanza lungo la direzione Y, si ha infatti per il suo quadrato

$$d^2_{yi} = (y_i - mx_i - q)^2 \qquad (2.0$$

e, confrontando con la formula esatta (1), al denominatore si è posto $m = 0$; conseguentemente per la somma[2] dei quadrati si ha

$$f_y(m,q) = \sum_{i=1}^{n} d^2_{y,i} = \sum y_i^2 - 2m \sum x_i y_i +$$
$$+ m^2 \sum x_i^2 - 2q \sum y_i + 2mq \sum x_i + nq^2$$

[1] È $\sum_1^n q^2 = nq^2$.

[2] Nel seguito tutte le sommatorie si intendono estese da $i = 1$ a $i = n$

Le derivate parziali rispetto a m, q sono

$$\frac{\partial f_y}{\partial m} = -2\sum x_i y_i + 2m\sum x_i^2 + 2q\sum x_i$$
$$\frac{\partial f_y}{\partial q} = -2\sum y_i + 2m\sum x_i - 2nq$$

Eguagliando a 0 si ha il sistema

$$m\sum x_i^2 + q\sum x_i = \sum x_i y_i$$
$$m\sum x_i + nq = \sum y_i$$

Dalla seconda equazione si ricava $q = (\sum y_i - m\sum x_i)/n$, che, sostituito nella prima equazione, da'

$$nm\sum x_i^2 + \sum x_i \sum y_i - m\sum x_i \sum x_i = n\sum x_i y_i$$

Ora si separa m

$$m\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right) = n\sum x_i y_i - \sum x_i \sum y_i$$

e si trova

$$m = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \qquad (3$$

Analogamente da $m = (\sum y_i - nq)/\sum x_i$ si ricava

$$\sum y_i \sum x_i^2 - nq\sum x_i^2 + q\sum x_i \sum x_i = \sum x_i \sum x_i y_i$$

e infine

$$q = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \qquad (4$$

Le espressioni per m, q appaiono a prima vista pesanti in termini di calcolo, ma si tratta infine di calcolare le quattro sommatorie $\sum x_i$, $\sum y_i$, $\sum x_i^2$ e $\sum x_i y_i$ e poi eseguire le operazioni indicate.

Estensione agli esponenziali

I dati sperimentali che pongono in relazione le quantità x_i, y_i attraverso funzioni del tipo $y = Ae^{bx}$ (esponenziale) possono essere trattati con il metodo appena delineato purché si considerino i logaritmi; prendendo il logaritmo naturale di entrambo i lati

$$y_i = Ae^{bx_i} \quad \text{diviene}$$
$$\ln y_i = bx_i + \ln A \quad \text{cioè}$$
$$y_i' = mx_i + q \quad \text{dove} \quad y_i' = \ln y_i,\ q = \ln A,\ m = b)$$

Ricavati i valori per m e q, si pone $A = e^q$, mentre $b = m$.

Calcoli

La regressione lineare è un trattamento standard implementato in qualsiasi foglio elettronico e richiede soltanto la introduzione di un certo numero di coppie di valori. Di solito il foglio elettronico può produrre anche un grafico che mostra le coppie di valori come punti sul piano XOY e la retta che meglio si adatta a quell'insieme dei punti. Questa presentazione è utile per aver farsi un'idea su come la retta passa tra i punti e della loro distribuzione nel piano, oltre a fare sempre effetto quando venga inserita in relazioni, articoli, tesi di laurea.

Discussione

I dati x_i, y_i provengono da misure e sono inevitabilente affetti da errore come qualsiasi altro dato sperimentale. Per ciascuno di essi si può stimare l'errore relativo in base a considerazioni sulla precisione degli strumenti impiegati nella misura. Il valore vero della coppia $P_i(x_i, y_i)$ cade quindi in un certo punto interno a a una regione rettangolare centrata su $P_i(x_i, y_i)$, ampia $2\Delta x_i$ e alta $2\Delta y_i$ dove $\Delta x_i = \frac{\Delta x}{x} x_i$, $\Delta y_i = \frac{\Delta y}{y} x_i$ e gli errori relativi $\frac{\Delta x}{x}$, $\frac{\Delta y}{y}$ si suppongono costanti su tutto l'intervallo di misura.

Con la semplificazione introdotta dalla (2) l'errore è concentrato solo sulla grandezza y_i e occorre aumentarlo per tener conto dell'errore su x_i. Sono possibili due atteggiamenti: si sostituisce all'errore relativo su y_i la somma degli errori relativi su y_i e su x_i, oppure, considerando che vi sia una sorta di compensazione statistica, stimare l'errore relativo su y_i come la radice della somma dei quadrati degli errori relativi su x_i e y_i.

La semplificazione introdotta con la (2) risulta tanto piú "buona" quanto è minore m, cioè tanto piú la direzione della retta $y = mx + q$ si avvicina alla direzione dell'asse X, cioè per $m < 1$. Infatti poiché $m^2 < m < 1$, la quantità al denominatore si approssima a 1, cioè nei fatti si pone $1 + m^2 \approx 1$.

Per le situazioni in cui $m > 1$ si può considerare la distanza X. In tal caso $x = m'y + q'$, e la distanza è $d_{x\,i} = (x_i - x)^2 = (x_i - m'y_i - q')^2$, una scrittura analoga a quella ottenuta per $y = mx + q$, per cui basterà scambiare il ruolo di x e y nelle formule per m e q per ottenere m' e q'. La retta nella consueta forma $y = mx + q$ non è altro che l'inversa della retta $x = m'y + q'$ e bastano pochi semplici passaggi per ricavarla.