Cenni di Statistica

- Introduzione
- **Descrivere i dati** (statistica descrittiva)
  - Media aritmetica
  - Varianza
  - Deviazione standard
  - Correlazione
- **Distribuzioni teoriche**
  - La legge dei grandi numeri
  - Il valore di aspettazione
  - La distribuzione di Gauss
- **Errori**
  - *Teorema del limite centrale*
  - *Media pesata*
  - *Propagazione degli errori*
  - *Errori di misura casuali e sistematici*

Bibliografia

- Queste trasparenze sono state in larga parte scritte dal Prof. Riccardo Brugnera

- J.R. Taylor, *Introduzione all'analisi degli errori*, Ed. Zanichelli

- Giampaolo Mistura "Guida all'uso dei Metodi Statistici nelle Scienze Fisiche", Corso di Laurea in Scienza dei Materiali

Introduzione

- La statistica è uno strumento di analisi dati.
- In una scienza sperimentale (fisica, biologia, ...) si progettano e si eseguono degli esperimenti, alla fine si analizzano e si interpretano i risultati. Per fare questo si usano argomenti statistici e calcoli matematici.
- Le leggi fondamentali della scienza non trattano di statistica e di errori. La legge di Newton sulla gravitazione è:

$$F = \frac{GMm}{r^2}$$

... è scritto esattamente 2 e non, per esempio: 2.000± 0.012

Introduzione

Ma da dove vengono le leggi? Per esempio nel caso della legge di Newton dalle osservazioni astronomiche (dettagliate e accurate) fatte da Tycho Brahe e da altri.

Pertanto:

Quando si studia una scienza non serve la statistica ; nel momento in cui si inizia a fare scienza e si desidera sapere che cosa vogliono dire realmente le misure allora la statistica diventa una faccenda di importanza vitale.

Descrivere i dati

- Tutto inizia con i dati: insieme di misure di base dalle quali si vuole estrarre qualche informazione utile.

- Tipi di dati: dati quantitativi o numerici (sono scritti come numeri), dati qualitativi o non numerici. Qui ci interesseremo solo di dati numerici.

- I dati numerici possono poi essere discreti (rappresentati da numeri interi) oppure continui (da numeri reali).

Descrivere i dati

Esperimento:

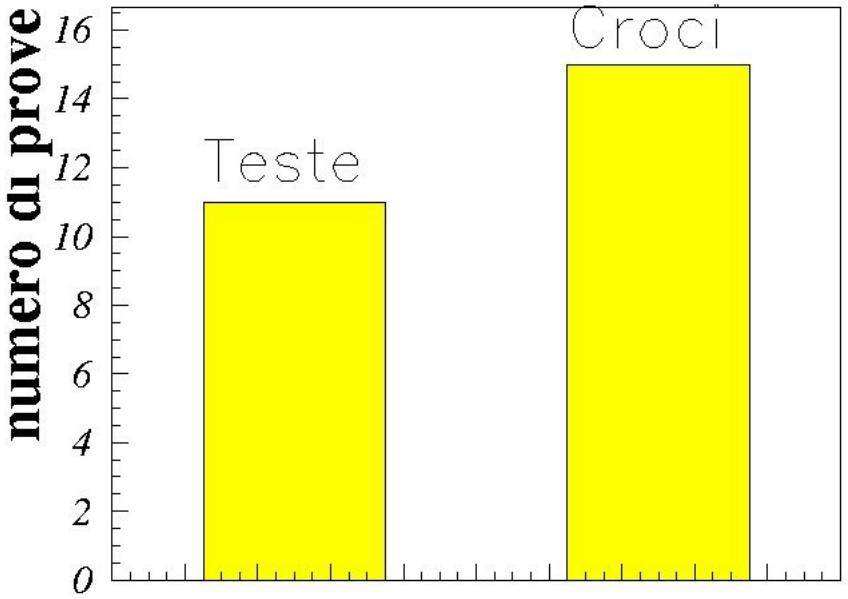
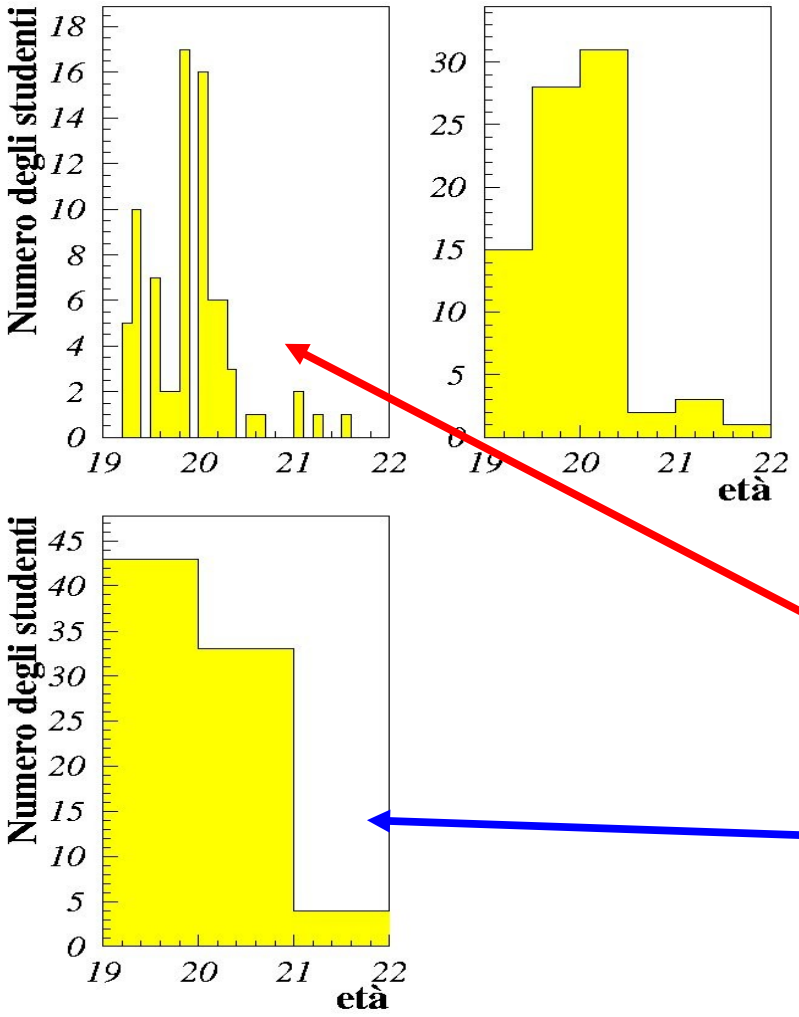


Grafico a barre (bar-chart)

Usato sia per dati qualitativi che quantitativi.
I dati vengono raccolti in intervalli (bins).
Il numero degli eventi è proporzionale all'altezza della barra.

Descrivere i dati



Istogramma

Usato per dati quantitativi

I dati vengono raccolti in bins.

Il numero di dati è proporzionale all'area sottesa.

La scelta dei bins deve essere fatta con oculatezza:

bins troppo stretti --> pochi eventi per bins; idealmente ci dovrebbero essere 10 eventi per bins, se ve ne sono di più tanto meglio;

però bins troppo larghi fanno perdere i dettagli.

Descrivere i dati

La media aritmetica

Se si vuole descrivere i dati con un solo numero, quello migliore è certamente la media aritmetica.

Se ci sono N elementi nell'insieme di dati:

$$\{x_1, x_2, x_3, \ldots, x_N\}$$

allora il valore medio di x è:

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

Nel medesimo modo si può calcolare il valore medio di una qualsiasi funzione f(x):

$$\bar{f} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

Descrivere i dati

La media aritmetica

Se i dati sono stati suddivisi in bin e il bin j corrisponde ad un valore x_j e contiene n_j dati, allora i valori medi si scrivono:

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{N} n_j x_j$$

$$\bar{f} = \frac{1}{N}\sum_{j=1}^{N} n_j f(x_j)$$

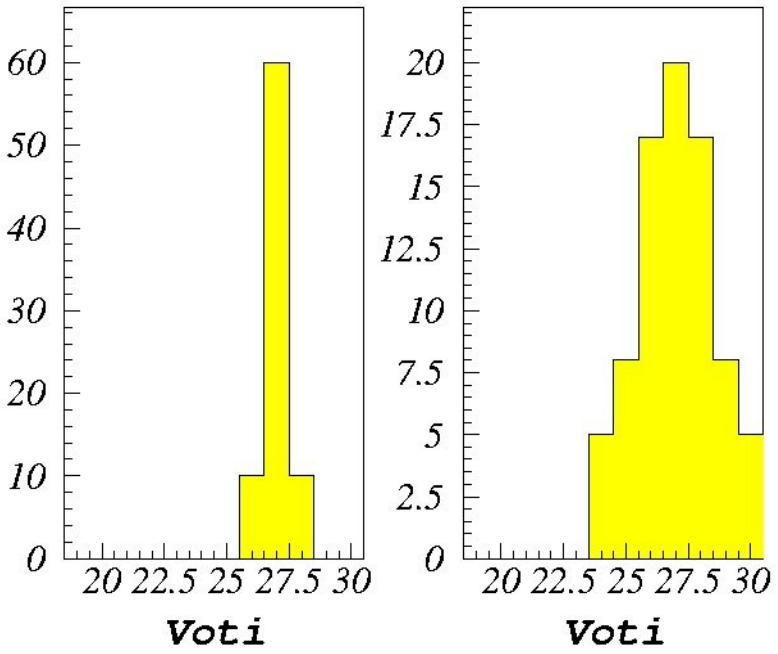
Attenzione: se qualche operazione di arrotondamento è stata fatta, allora la media dai bin è meno accurata rispetto ad una media su elementi non arrotondati,

Descrivere i dati

Misurare la dispersione: la varianza (variance)

La media descrive tutti i dati con un solo numero. Può essere utile, ma a volte potrebbe indurre in errore.

Esempio



Distribuzione dei voti di due gruppi di 80 studenti, entrambi i campioni hanno come media 27. Però i dati sono distribuiti in modo differente.

E' necessario avere un numero che misuri la dispersione dei dati attorno alla media.

Descrivere i dati

Misurare la dispersione: la varianza (variance)

Si potrebbe usare la deviazione media dalla media, ma:

$$\frac{1}{N}\sum_i \left(x_i - \bar{x}\right) = \frac{1}{N}\sum_i x_i - \frac{1}{N}\sum_i \bar{x} = \bar{x} - \bar{x} = 0$$

perchè le deviazioni positive cancellano quelle negative.
Per fermare queste cancellazioni basta elevare al quadrato.
Ecco la variabile che misura la dispersione, essa è chiamata **varianza di x:**

$$V(x) = \frac{1}{N}\sum_i \left(x_i - \bar{x}\right)^2$$

per una funzione f qualsiasi:

$$V(f) = \frac{1}{N}\sum_i \left(f(x_i) - \bar{f}\right)^2$$

Descrivere i dati

Misurare la dispersione: la varianza (variance)

V(x) può essere manipolata in questa maniera per arrivare ad una formula più semplice:

$$V(x) = \frac{1}{N}\sum_{i=1,N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1,N}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \frac{1}{N}\sum_{i=1,N}x_i^2 - \frac{1}{N}\sum_{i=1,N}2x_i\bar{x} + \frac{1}{N}\sum_{i=1,N}\bar{x}^2$$

$$= \frac{1}{N}\sum_{i=1,N}x_i^2 - \frac{2\bar{x}}{N}\sum_{i=1,N}x_i + \frac{\bar{x}^2}{N}\sum_{i=1,N}1$$

$$= \overline{x^2} - 2\bar{x}^2 + \bar{x}^2 = \overline{x^2} - \bar{x}^2$$

Descrivere i dati

Misurare la dispersione: la varianza (variance)

Così si ottiene la formula fondamentale:

$$V(x) = \overline{x^2} - \overline{x}^2$$

Descrivere i dati

La deviazione standard (standard deviation)

Si definisce deviazione standard la radice quadrata della varianza e si indica con il simbolo σ. Può essere espressa in varie maniere equivalenti:

$$\sigma = \sqrt{V(x)}$$

$$\sigma = \sqrt{\overline{x^2} - \overline{x}^2}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_i x_i^2 - \left(\frac{1}{N}\sum_i x_i\right)^2}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_i \left(x_i - \overline{x}\right)^2}$$

Descrivere i dati

La deviazione standard (standard deviation)

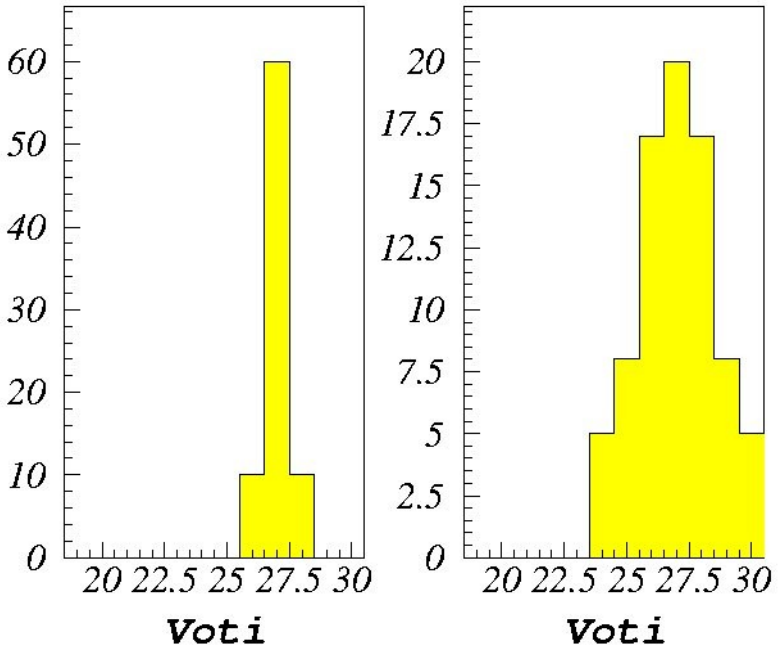
σ rappresenta una ragionevole quantità per un particolare dato di differire dalla media. Di solito non ci si sorprende se un dato differisce dalla media di 1 o 2 deviazione standard, c'e' da sospettare se un dato invece differisce di 3 o più σ .

NOTA:

Gli scienziati sperimentali amano lavorare con σ invece che con V(x), perché ha le stesse dimensioni di x.
Gli statistici tendono ad usare V(x), perché é più semplice da manipolare di σ.

Descrivere i dati

La deviazione standard (standard deviation)



Esempio

Ritornando ai due istogrammi precedenti riguardanti i risultati di due gruppi di studenti. Nel primo istogramma la media era 27 con una varianza di 0.25 e una deviazione standard di 0.5.

Nel secondo istogramma la media é sempre 27 ma la varianza é 2.35 e una deviazione standard di 1.53

Descrivere i dati

Più di una variabile: covarianza (covariance)

Ci sono casi in cui ogni dato é costituito da più numeri, per esempio di una particella in moto si puó registrare la posizione e il tempo, in tal maniera i dati sono costituiti dall'insieme delle coppie (x,t). Oppure data una classe di studenti si potrebbe registrare l'altezza, il peso, il QI, la forza fisica, ogni studente è adesso individuato da 4 valori.

Che relazioni ci sono tra queste quantità?

Supponiamo di avere un campione di dati costituito da coppie di numeri: $\{(x_1,y_1),(x_2,y_2), \ldots (x_n,y_n)\}$

Si puo' certamente calcolare $\bar{x}, \bar{y}$ e poi

$V(x), V(y), \sigma_x, \sigma_y$. Pero' i dati contengono

maggiori informazioni: le due variabili x e y sono indipendenti

o dipendono l'una dall'altra?

Descrivere i dati

Più di una variabile: covarianza (covariance)

Questa dipendenza è descritta dalla covarianza tra x e y così definita:

$$\mathrm{cov}(x,y) = \frac{1}{N}\sum_{i=1,N}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) =$$

$$= \overline{\left(x - \bar{x}\right)\left(y - \bar{y}\right)} = \overline{xy} - \bar{x}\bar{y}$$

Se cov(x,y)=0, allora x e y sono scorrelati.

Descrivere i dati

Più di una variabile: correlazione (correlation)

quantita' adimensionale

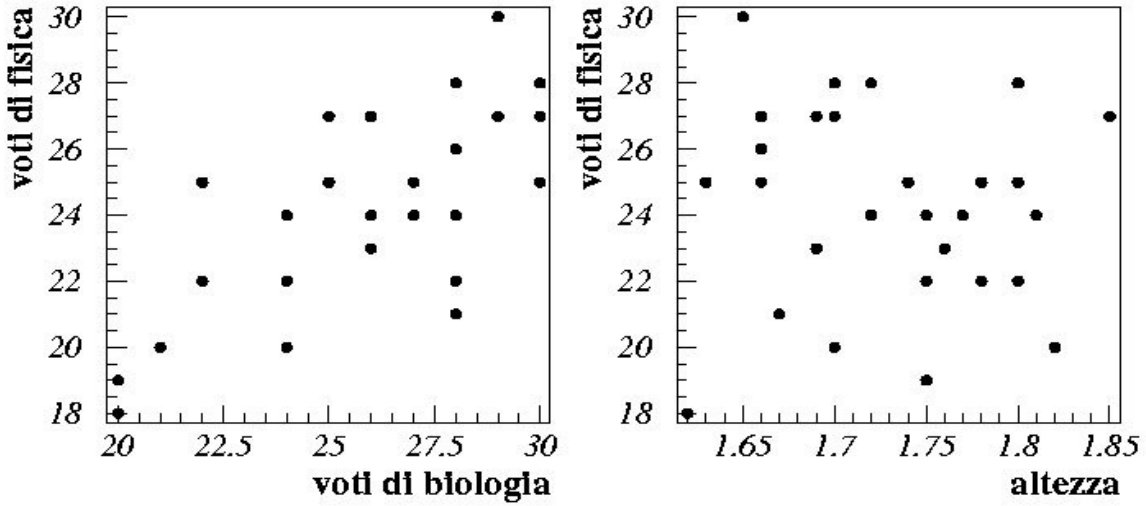
$$\rho = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y} \quad .$$

coefficiente di correlazione

Descrivere i dati

Più di una variabile: correlazione (correlation)

Per una correlazione negativa, un x più grande implica un y più piccolo.
Se la correlazione è 1 (o -1) allora x e y sono completamente correlate: se si conosce il valore di x, il valore di y è noto.



C'è correlazione positiva tra i risultati di esami di Fisica e quelli in Biologia.
Non c'e' correlazione tra i risultati in Fisica e l'altezza degli studenti.

Distribuzioni teoriche

Una semplice distribuzione:

- La probabilitá, per il lancio di una moneta, che esca testa (T) vale ½, lo stesso vale per croce (C).
- La probabilitá che lanciando quattro monete escano 4 T, vale:

$P(4) = (½)^4 = 1/16$ *(una sola possibilitá: TTTT)*

- La probabilitá che escano, in un lancio qualsiasi di 4 monete, 3T e 1C vale:

$P(3) = 4 \times 1/16 = ¼$ *(4 possibilitá: TTTC, TTCT, TCTT, CTTT)*

- La probabilitá che escano, in un lancio qualsiasi di 4 monete, 2T e 2C vale:

$P(2) = 6 \times 1/16 = 3/8$ (4 *possibilitá: TTCC, CCTT, CTCT, TCCT, CTTC, TCTC)*

- La probabilitá che escano, in un lancio qualsiasi di 4 monete, 1T e 3C vale:

$P(1)=P(3) = 4 \times 1/16 = ¼$ *(4 possibilitá: CCCT, CCTC, CTCC, TCCC)*

- Similmente la probabilitá che ci siano 4 croci vale: $P(0) = P(4) = 1/16$ *(una sola possibilitá: CCCC)*

Distribuzioni teoriche

Se si sommano tutte le probabilitá sopra riportate (il che equivale a chiedersi quale è la probabilitá che qualcosa accada) si ottiene:

$$\sum_r P(r) = P(0) + P(1) + P(2) + P(3) + P(4) = \frac{16}{16} = 1$$

Pertanto indicando con r il numero di T (r=0,1,2,3,4) noi abbiamo una collezione di probabilitá P(r) = (1/16,1/4,3/8,1/4,1/16) che rappresentano la probabilitá che in un lancio di 4 monete (non truccate) appaiano r teste.
Questo è un esempio (semplice) di **distribuzione di probabilitá**

Distribuzioni teoriche

La legge dei grandi numeri

I calcoli appena fatti sono di tipo teorico. Hanno qualche corrispondenza con la realtá?

L'unica maniera di verificarlo è lanciare quattro monete.Quello che si scopre è che se il numero di prove è piccolo (per esempio 16) l'accordo non è buono anche se si intravede una certa rassomiglianza. L'accordo diventa via via migliore all'aumentare delle prove (160, 1600, 16000).

Si impara che:

- *La teoria predice un insieme di probabilitá.*
- *Le frequenze (n_i/N) dei dati osservati non si accordano molto bene con esse per N piccolo.*
- *Quando il numero dei dati N aumenta le fluttuazioni diminuiscono e le frequenze tendono alle probabilitá per $N \to \infty$.*

Questa è la ***legge dei grandi numeri.***

Distribuzioni teoriche

Valore di aspettazione

Se si conosce la distribuzione di probabilitá P(r) di un qualche processo si può calcolare il numero medio dei successi che ci si puó attendere:

$$\langle r \rangle = \sum_r rP(r)$$

È chiamato valore di aspettazione di r (a volte è indicato con E(r))
Per una funzione qualsiasi di r, f(r) si ha:

$$\langle f(r) \rangle = \sum_r f(r)P(r)$$

Distribuzioni teoriche

Valore di aspettazione

Ovvio parallelismo tra valore di aspettazione e media di un campione di dati. La prima è una somma su una distribuzione di probabilitá teorica la seconda è una somma su un campione di dati reali.
La legge dei grandi numeri assicura che se un campione di dati è descritto da una distribuzione teorica, allora quando N, la dimensione del campione di dati, cresce all'infinito:

$$\bar{f} \rightarrow \langle f \rangle$$

Distribuzioni teoriche

Valore di aspettazione

Finora abbiamo trattato variabili discrete, ma ci sono anche variabili continue: per esempio si può pensare di misurare le lunghezze di un grande numero di pezzi di corda distribuite in maniera casuale tra 10 cm e 12 cm. Se qualcuno chiede quanti pezzi di corda sono lunghi 11 cm, la risposta è nessuno. Infatti molto presubimilmente ce ne saranno alcuni con lunghezze comprese tra 10.9 e 11.1 ma nessuno esattamente a 11 cm.

La probabilitá che x sia in un dato intervallo è invece una quantitá non nulla ed è sensato parlarne, ciò è descritto dalla **distribuzione di densitá di probabilitá, p(x):**

$$\texttt{Probabilita' che il risultato sia tra } x_1 \texttt{ e } x_2 = \int_{x_1}^{x_2} p(x)dx$$

Distribuzioni teoriche

Valore di aspettazione

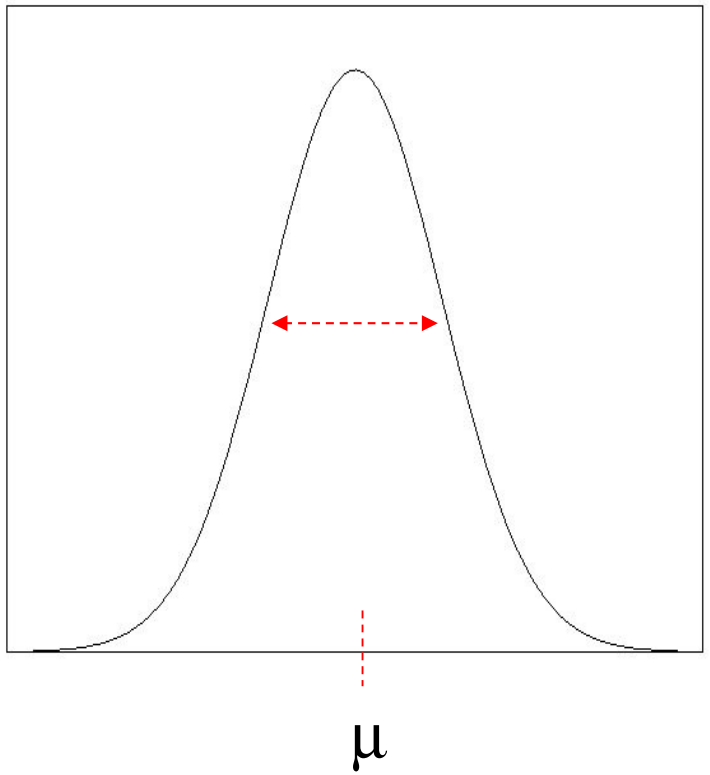
e quindi:

$$\langle x \rangle = \int_{-\infty}^{+\infty} x\, p(x) dx$$

$$\langle f(x) \rangle = \int_{-\infty}^{+\infty} f(x)\, p(x) dx$$

La distribuzione di Gauss

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



- La distribuzione di probabilità di Gauss (anche detta distribuzione gaussiana, o normale) è la più nota e utile delle distribuzioni.
- Curva a forma di campana.
- Centrata attorno al valore medio μ e simmetrica rispetto ad esso.
- La sua larghezza è regolata dal parametro σ, la deviazione standard della distribuzione. La distribuzione é larga se σ é grande, é stretta se σ é piccolo.

La distribuzione di probabilità di Gauss

Proprieta'

$$\int_{-\infty}^{+\infty} p(x,\mu,\sigma)dx = 1$$

$$\langle x \rangle \equiv \int_{-\infty}^{+\infty} x\, p(x,\mu,\sigma)dx = \mu$$

$$\langle (x-\mu)^2 \rangle \equiv \int_{-\infty}^{+\infty} (x-\mu)^2 p(x,\mu,\sigma)dx = \sigma^2$$

La distribuzione di Gauss

Sfortunatamente l'integrale indefinito della Gaussiana non può essere calcolato analiticamente, si devono usare tabelle o programmi per computer. Valori da ricordare nel caso:

$$p(y) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-(y-\mu)/\sigma}^{(y-\mu)/\sigma} e^{-(x-\mu)^2/2\sigma^2} dx$$

in cui si calcoli la Gaussiana tra quei due estremi simmetrici,
i.e. la probabilità che, se un evento è guidato dalla distribuzione di Gauss, esso giacerà entro un certo numero di deviazioni standard dalla media.

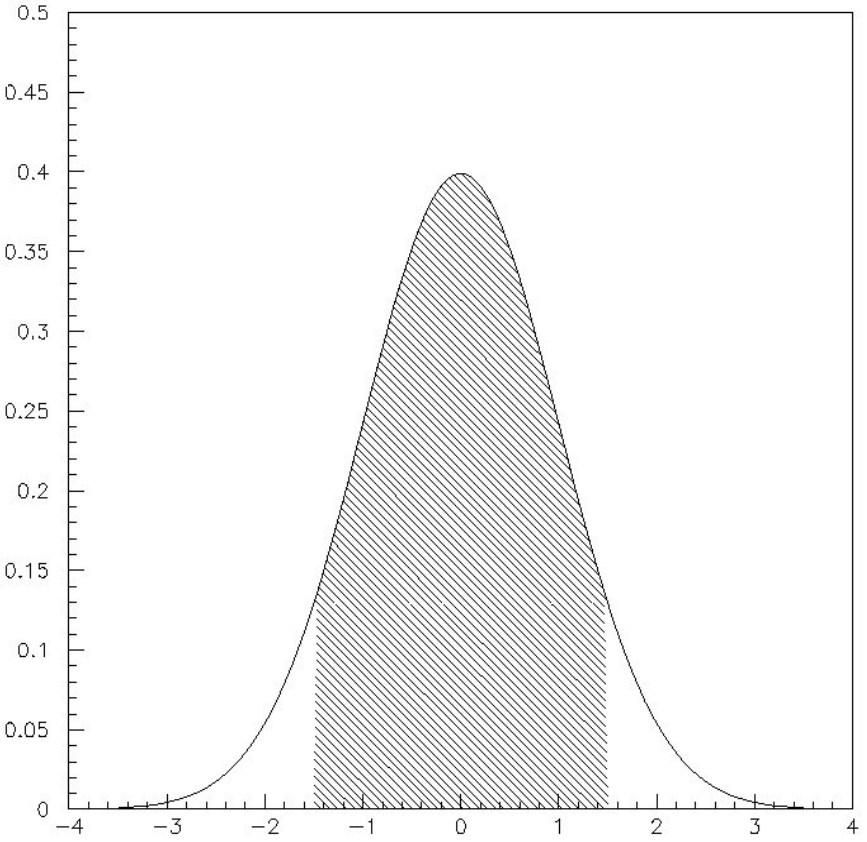
La distribuzione di Gauss

Si può calcolare che:

68.27% dell' area sta entro 1σ dalla media

95.45% sta entro 2σ

99.73% sta entro 3σ



Errori

Quando ad un risultato di una misura si associa un errore, questo va inteso come una deviazione standard gaussiana .

Esempio: se la lunghezza di una corda è: 12.3 ± 0.1 cm, questo vuol dire che io ho misurato la corda con uno strumento che dá delle risposte che differiscono dal valore vero per meno di 0.1 cm (1σ) il 68% delle volte, 0.2 cm ($2\ \sigma$) per il 95% delle volte e 0.3 cm ($3\ \sigma$) per il 99.7% delle volte.

Ma perchè risultati ed errori sono generalmente ben descritti dalla distribuzione gaussiana?

Errori

Non è un caso, esiste questo fantastico teorema detto

teorema del limite centrale:

N *variabili indipendenti,* x_i*, dove* $i=1,2,..,N$*, ciascuna variabile presa da una distribuzione di media* μ_i *e varianza* $\mathbf{V_i}$ *, la distribuzione per la funzione somma:*

$$S = \sum_{i=1}^{N} x_i$$

a) ha un valore di aspettazione: $\langle S \rangle = \sum_i \mu_i \equiv \mu_S$

b) ha varianza: $V(S) = \sum_i V_i = \sum_i \sigma_i^2 \equiv \sigma_S^2$

Teorema del limite centrale

$$S = \sum_{i=1}^{N} x_i$$

$$p(S) \underset{N \to \infty}{\longrightarrow} e^{-\frac{(x-\mu_S)^2}{2\sigma_S}}$$

Errori

Morale:

Una quantità prodotta dall'effetto cumulativo di molte variabili indipendenti sarà, almeno approssimativamente, gaussiana, indipendentemente dal tipo di distribuzioni seguite dalle variabili originarie.
Gli errori di misura si comportano proprio così.

Esempio:

Le altezze delle braccia, delle gambe degli uomini etc. sono ben descritte dalla distribuzione gaussiana perchè sono dovute agli effetti combinati di molti fattori genetici e ambientali.

Errori

Riassumendo:

Supponiamo di aver misurato una quantità un numero N di volte nelle stesse condizioni e con la stessa procedura, noi sappiamo:

a) la media aritmetica è la migliore stima del valore vero.

b) la larghezza della distribuzione è data dalla deviazione standard che rappresenta la misura di quanto una singola misura può discostarsi dalla media.

c) se la migliore informazione sul valore vero è dato dalla media aritmetica quale errore le si può associare? Si può dimostrare che:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$ DEVIAZIONE STANDARD DELLA MEDIA

cioè la deviazione standard della media varia come $1/\sqrt{N}$.

Se abbiamo 25 misure con deviazione standard σ, la deviazione standard della media sarà più piccola di un fattore $\sqrt{25} = 5$

Media ed errore pesati

Può accadere che una stessa quantitá venga misurata con metodi diversi aventi ognuno un differente errore σ_i . Come si possono combinare assieme tali risultati? Attraverso la **media pesata** in cui le misure con σ_i più piccoli pesano di più nella media:

$$\bar{x} = \frac{\sum x_i / \sigma_i^2}{\sum 1 / \sigma_i^2}$$

$$\sigma_{\bar{x}} = \sqrt{\frac{1}{\sum 1 / \sigma_i^2}}$$

Propagazione Errori

Come si combinano e si propagano gli errori?

Supponiamo che **f = ax+b**, misuriamo x con errore σ_x, a e b sono delle costanti. Che errore si fa su f?

$$\sigma_f = \left|\frac{df}{dx}\right|\sigma_x$$

Più in generale se **f= f(x,y)** e noi misuriamo x e y, l'errore su f vale:

$$\sigma_f = \sqrt{\left(\frac{df}{dx}\right)^2\sigma_x^2 + \left(\frac{df}{dy}\right)^2\sigma_y^2 + 2\left(\frac{df}{dx}\right)\left(\frac{df}{dy}\right)\rho_{xy}\sigma_x\sigma_x}$$

Se x ed y sono indipendenti $\rho=0$

$$\sigma_f = \sqrt{\left(\frac{df}{dx}\right)^2\sigma_x^2 + \left(\frac{df}{dy}\right)^2\sigma_y^2}$$

Errori percentuali

Dato il valor medio $\overline{x}$

E l'errore $\sigma_{\overline{x}}$

L'errore percentuale vale:

$$\sigma_{\overline{x}} / \overline{x}$$

Errori casuali

Gli errori di misura casuali risultano evidenti quando si ripete più volte la stessa misura in condizioni nominalmente eguali con un sensibilità sufficientemente spinta.
{Si definisce **sensibilità di uno strumento** (o del procedimento sperimentale in cui viene usato) la minima differenza apprezzabile tra il valore della grandezza da misurare e quella campione}.

Molte le cause:

1 Condizioni sperimentali fluttuanti in maniera non controllabile;
2 Disturbi estranei alla misura;
3 Grossolana interpolazione tra due divisioni successive nella scala dello strumento;
4 Definizione vaga della grandezza da misurare ....
Per questo tipo di errore ha senso ripetere le misure.

Errori sistematici

Gli errori sistematici tipicamente sono errori che falsano la misura sempre nell stesso modo: *per esempio se si fa una misura con un metro lungo 999 mm tutte le misure saranno errate per eccesso.*

Possibili cause:

1 Errori di calibrazione degli strumenti;

2 Errori personali: ... tipo errori personali di parallasse

3 Condizioni sperimentali: per esempio se uno strumento viene usato in condizioni sperimentali diverse da quelle per cui è stato calibrato, se nessuna correzione viene fatta ne risulta un errore sistematico

4 Tecnica imperfetta: La misura di viscosità mediante la legge di Poiseuille richiede la misura della quantità di liquido che emerge da un apparato in un dato intervallo di tempo.

Se una piccola quantità di liquido esce dal recipiente usato per raccoglierlo, ne risulta un errore sistematico.

Errori sistematici

È ovvio che ripetere le misure avendo degli errori sistematici di mezzo non porta ad alcun miglioramento. Bisogna scovarli, analizzando attentamente i dati raccolti, ripetendo le misure con strumenti oppure metodi diversi, se questo è possibile, oppure stimarne la loro entitá.

Una volta scoperti e valutati (se possibile) si devono correggere i dati per l'effetto degli errori sistematici e/o quotarli nell'analisi degli errori.

Interpolazione lineare

L'esperimento fornisce N dati e cioe' N punti di coordinate (x_i, y_i).

L'idea e' quella d'interpolare i dati dati con una funzione f=a+bx. Occore quindi stimare a e b .

Interpolazione lineare

Interpolare dati con una funzione f=a+bx

$$\hat{a} = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{\Delta}$$

$$\hat{b} = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{\Delta}$$

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

Interpolazione lineare

Errori su a e b

$$\sigma_b = \sigma_y \sqrt{N/\Delta}$$

$$\sigma_a = \sigma_y \sqrt{\sum x_i^2 / \Delta}$$