Errori di Misura e Trattamento Statistico dei dati (Richiami e Approfondimenti)

Fernando Palombo

http://www.mi.infn.it/~palombo/didattica/Lab-OFM/Lezione1.pdf

Introduzione

- Le nozioni di base di teoria degli errori e di trattamento statistico dei dati sperimentali sono state date nel corso del laboratorio del primo anno.

- Nelle due lezioni che io vi farò rivedremo alcuni argomenti della teoria degli errori e ne approfondiremo altri di statistica inferenziale.

- Queste due lezioni le potete trovare al link :

  http://idefix.mi.infn.it/~palombo/didattica/Lab-OFM

Misure Sperimentali

- Esempio di scrittura di una misura: L = (32.5 ± 0.3) cm
  Valore stimato (32.5) e incertezza (0.3) devono essere compatibili (stesso numero di cifre decimali !)

- Prima di scrivere la misura io analizzo il grado di sensibilità del mio strumento di misura (apparato sperimentale) e della analisi alla quantità che sto misurando e scrivo il valore stimato e l'incertezza in accordo con questa sensibilità

- Sbaglierei se scrivessi per esempio questa misura in uno dei seguenti modi:
  L = (32.53 ± 0.3) cm
  L = (32.5 ± 0.03) cm
  L = (32.534 ± 0.03) cm
  L = (32.53467245 ± 0.34297592) cm

Cifre Significative

- Il numero di cifre significative è il numero di tutte le cifre (zero incluso) a partire da destra sino all'ultima cifra diversa da zero a sinistra. Esempi:

Numero	Cifre significative
51.20	4
51.2	3
0.05	1
0.056	2
0.0563	3
0.2067	4

- Il numero di cifre significative di una misura deve essere determinato a partire dall'incertezza su questa misura; Il valore numerico della misura quindi va arrotondato opportunamente.

Notazione Scientifica e Ordini di Grandezza

- Supponiamo di avere la seguente misura sperimentale:

  $L = (0.00000235 \pm 0.00000013)$ m

- In notazione scientifica scrivo:

  $L = (2.35 \pm 0.13)\ 10^{-6}$ m $= (2.35 \pm 0.13)$ µm

- Questa notazione permette di evidenziare le cifre significative e soprattutto rende facilmente leggibili numeri molto grandi o molto piccoli:

  $4370000000000000 \rightarrow 4.37\ 10^{16}$

  $6397 = 6.397\ 10^{3}$

  $0.00000000000000000032 \rightarrow 3.2\ 10^{-20}$

- Il primo numero ha ordine di grandezza 10^{16}, il secondo 10^{4} (la mantissa è > 5) e il terzo 10^{-20}

Incertezza in Misura Singola

- In una misura singola l'incertezza di misura essenzialmente è legata alla sensibilità dello strumento, che è il più piccolo valore della grandezza che lo strumento riesce a distinguere.

- Se il righello ha sensibilità di 1 mm e la parte finale del foglio, di cui misuro la lunghezza, cade nell'intervallo [22.3 , 22.4] cm, noi diamo come misura il valore al centro dell'intervallo, 22.35 cm, ed incertezza pari a metà dell'incertezza di sensibilità, 0,05 cm:

  $L = (22.35 \pm 0.05)$ cm

- Talvolta (per cautelarsi) questa misura è data con errore pari alla sensibilità: $L = (22.3 \pm 0.1)$

- Se la misura fosse una misura indiretta, dovrei propagare l'incertezza.

Incertezza in Misure Ripetute

- È chiaro che se il mio strumento ha scarsa sensibilità le mie misure cadrebbero tutte entro la sensibilità dello strumento e quindi sarebbero tutte coincidenti. In questo caso come incertezza uso l'incertezza di sensibilità.
- In presenza di un strumento (rivelatore) di "adeguata sensibilità", tutte le misure in generale avranno valori diversi.
- Ciò succede a causa di incertezze nella misura di tipo casuale, sistematico o di altro tipo (ad esempio di tipo teorico).
- Incertezze casuali: sono ineliminabili e dovute alle mutevoli condizioni sperimentali in cui è fatta la misura. I valori misurati fluttuano casualmente in eccesso o in difetto.
- L'entità di questo tipo di incertezza è stimata con metodi statistici : → incertezza statistica

Incertezze Sistematiche

- Finita l'analisi statistica dei vostri dati, voi avete una misura insieme alla sua incertezza statistica $x_0 \pm \sigma_0$. Ora dovete stimare l'incertezza sistematica su questa misura.

- Questa incertezza sistematica dipende dall'apparato sperimentale, dall'interazione dello sperimentatore (studente) con l'apparato, dalla strategia di analisi (selezione dei dati, tipo di fit, ecc) e talvolta anche dalla numerosità dei dati.

- Bisogna indagare ogni operazione di misura per stabilire se essa può contribuire ad un errore sistematico e valutare le sorgenti di incertezze sistematiche.

- Alcune di queste sorgenti sono aspettate, altre sono del tutto inaspettate e forse altre non sono trovate (ma potrebbero essere trovate in future analisi degli stessi dati).

Incertezze Sistematiche

- Le incertezze sistematiche generalmente si cercano con tecniche di simulazione Monte Carlo, con campioni di controllo, variando i valori assunti per certe quantità sperimentali entro l'incertezza del rivelatore, variando di ±σ le quantità che intervengono nella misura ma che sono note con incertezza, ecc.

- Stimiamo la differenza Δx sistematica tra la misura che ottengo x_0 e il valore che dovrei ottenere x_{vero} (per esempio perché già noto o perché così preso nella simulazione). La misura ha un bias $\Delta x = x_0 - x_{vero}$ che va corretto.

- Se il bias Δx è valutato con alta precisione allora semplicemente correggo la misura x_0 per il bias e non considero più questa sorgente di incertezza sistematica.

- Se Δx è incerto correggo la misura per questo bias e associo una incertezza sistematica (per esempio uguale a $|\Delta x|/2$ oppure a $|\Delta x|$) che tenga conto dell'**Incertezza sulla correzione** del bias).

- Se non riesco a correggere il bias, questo apparirà tutto come incertezza sistematica. Le incertezze scovate e valutate si sommano in quadratura per dare l'incertezza sistematica totale sulla misura fatta.

Roger Barlow

Systematic Errors

Systematic Error:
reproducible inaccuracy introduced by faulty equipment, calibration, or technique

Bevington

Systematic effects is a general category which includes effects such as background, scanning efficiency, energy resolution, angle resolution, variation of couner efficiency with beam position and energy, dead time, etc. The uncertainty in the estimation of such as systematic effect is called a *systematic error*

Orear

Error=mistake?

Error=uncertainty?

Incertezze Sistematiche

- Stimata l'incertezza sistematica la misura è data così: $\mathbf{x_{mis} \pm \sigma_{stat} \pm \sigma_{sist}}$
  dove x_{mis} è la misura x_0 dopo la correzione per le incertezza sistematiche.
  σ_{stat} è l'incertezza statistica e σ_{sist} rappresenta l'incertezza sulle correzioni degli effetti sistematici.

- Talvolta la misura è data con un'unica incertezza somma in quadratura dell'incertezza statistica e di quella sistematica:

$$x_{mis} \pm \sigma \qquad \text{con} \quad \sigma^2 = \sigma_{stat}^2 + \sigma_{sist}^2$$

- Il primo tipo di scrittura è preferito quando si vuole avere mostrare il peso relativo dei due tipi di incertezza. Quindi:
  - è inutile continuare a prendere dati se l'incertezza è dominata dall'incertezza sistematica;
  - chi vorrà migliorare la nostra misura potrà avere una idea della statistica necessaria e del livello di controllo dei sistematici richiesto.

Incertezze, Errori e Sbagli

- Nelle slide precedenti ho parlato di incertezze statistiche e sistematiche. Molto spesso si usa parlare di errori statistici e errori sistematici.

- I termini incertezza ed errore nelle analisi statistiche (in particolare dei fisici) sono da considerarsi sinonimi. Di fatto si usano indifferentemente (anche nella stessa pubblicazione).

- Naturalmente nella presa dati o nell'analisi statistica dei dati ci possono essere **sbagli** (e talvolta ci sono). Ma questi non sono incertezza sistematiche ma sbagli che vanno scovati ed eliminati.

- Uno sbaglio (non riconosciuto) può dare origine ad una distorsione della misura sperimentale. Questo distorsione viene coperto con un errore sistematico.

Trattamento Statistico dei Dati

- La statistica è un ramo della matematica applicata.

- Tecniche statistiche per estrarre informazioni dai dati sperimentali sono oggi di base in ogni settore dell'attività umana.

- Le tecniche statistiche sono numerose e il loro utilizzo dipende dal settore di applicazione.

- Statistica Descrittiva e Statistica Inferenziale

Statistica Descrittiva

- Si occupa della classificazione e sintesi delle informazioni relative ad un determinato campione di dati. In modo conciso si sintetizzano i dati con pochi numeri o grafici.

- La sintesi porta alla perdita di una parte dell'informazione. Bisogna scegliere di volta in volta la parte di informazione che interessa, eliminando quella ritenuta non necessaria.

- Gli strumenti utilizzati sono essenzialmente di tre tipi:
  - Tabelle
  - Grafici (come istogrammi, diagrammi a barre, a torta, ecc)
  - Indici sintetici: come quelli di posizione (come media,mediana, moda), varianza, deviazione standard, ecc.

Statistica Inferenziale

- È detta popolazione la totalità degli elementi oggetto dell'indagine. Campione è un numero finito di elementi presi da una popolazione.

- Spesso l'analisi estesa all'intera popolazione è impossibile o poco pratica. Si pensi al controllo di qualità che spesso è distruttivo, o alla analisi su un campione di qualcosa che si vuole applicare a tutta la popolazione.

- La **statistica inferenziale** utilizza il campione di dati per fare previsioni di tipo probabilistico sulla popolazione da cui il campione è tratto.

- È senza dubbio la parte di statistica di maggiore interesse.

- Le aree principali dell'inferenza statistica sono la stima dei parametri e la verifica delle ipotesi.

Inferenza Induttiva

- L'inferenza statistica (induttiva) permette di attribuire alla popolazione il risultato ottenuto sul campione.

- L'inferenza induttiva è quindi il passaggio dal particolare (misura sul campione) al generale (proprietà della popolazione). Questa generalizzazione non è mai assolutamente certa!

- L'analisi statistica permette di associare un grado di incertezza ad ogni inferenza induttiva.

- Più il campione (casuale) è numeroso, minore è l'incertezza statistica dell'inferenza fatta.

- L'inferenza statistica che consideriamo qui è quella **frequentista** (detta talvolta anche classica). Questa si basa sul presupposto che una misura sia ripetibile e quindi sul concetti di probabilità come frequenza relativa.

Variabili Casuali

- Una variabile è detta casuale (o aleatoria) se assume un valore reale distinto per ogni elemento dello spazio campione.

- Noi associamo alla variabile casuale la distribuzione di probabilità secondo la quale la variabile casuale assume i valori possibili. Noi caratterizziamo i dati tramite queste distribuzioni di probabilità.

- Alcune di queste distribuzioni (Poissoniana, Gaussiana, del χ^2 ..) sono comunissime nei fenomeni fisici.

- Una variabile casuale può essere a valori discreti, a valori continui o a valori sia discreti che continui

- Vedremo ora brevemente come possiamo descrivere i nostri dati sperimentali

- Per semplicità assumiamo che la variabile assuma valori continui. Il passaggio a variabili discrete è abbastanza semplice (integrali → sommatorie)

Funzione Densità di Probabilità

Funzione densità di probabilità con una sola variabile casuale

- Faccio una misura x di una variabile casuale continua X i cui valori possono andare da -∞ a +∞ . Sia F(x_0) la probabilità che la mia misura x (variabile casuale) sia minore di x_0. Quindi F(-∞)= 0, F(+∞) = 1 e la probabilità che x sia compreso tra x e x + dx è:

$$P(x \in [x, x+dx]) = F(x+dx) - F(x)$$

- Si definisce funzione densità di probabilità (p.d.f.) la funzione

  f(x) = dF(x)/dx

- Quindi $P(x \in [x, x+dx]) = F(x+dx) - F(x) = f(x)dx$

  con $\int_{-\infty}^{+\infty} f(x)dx = 1$ (integrale di normalizzazione)

Funzione Distribuzione Cumulativa

- La distribuzione di probabilità di una variabile casuale X può essere caratterizzata dalla sua funzione di distribuzione cumulativa (c.d.f.) F(x) così definita:

$$F(x) = \int_{-\infty}^{x} f(u)du$$

dove f(x) è la p.d.f.. La c.d.f. F(x) è la probabilità che la variabile casuale X assuma un valore minore o uguale ad x

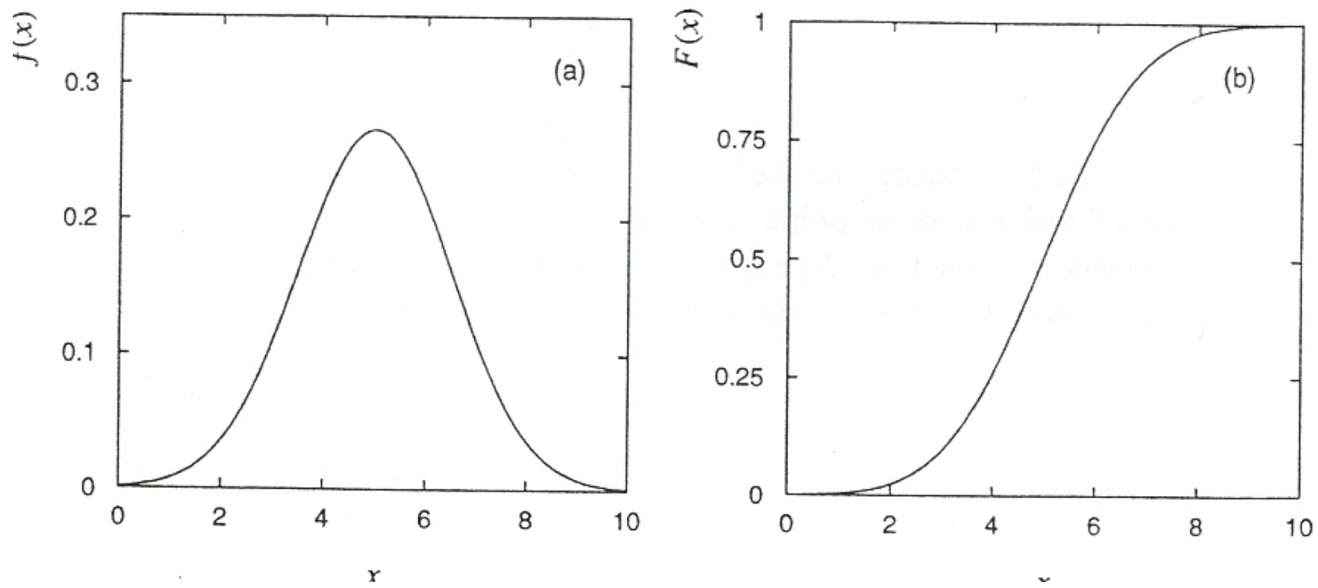
- Supponendo che F(x) sia una funzione strettamente crescente, allora c'è corrispondenza biunivoca tra il valore della variabile x e il valore assunto dalla c.d.f. F(x). → La funzione F(x) può essere invertita.

- Si definisce punto α (o anche quantile di ordine α) x_α il valore della variabile x per il quale si ha: $F(x_\alpha) = \alpha$ con $0 \leq \alpha \leq 1$.

- Il quantile è l'inverso della distribuzione cumulativa: **$x_\alpha = F^{-1}(\alpha)$**

- Il quantile più usato è quello di ordine 0.5, detto mediana. Questo quantile (la mediana) divide la distribuzione in due parti uguali.

p.d.f. e c.d.f.



p.d.f. e corrispondente c.d.f.

Valori di Aspettazione

- Media (aritmetica) di una variabile casuale
  Sia f(x) la p.d.f. della variabile casuale X. Il valore di aspettazione E[x] di questa variabile è definito da

$$E[x] = \int_{-\infty}^{+\infty} x f(x) dx = \mu$$

  μ è detto anche media della popolazione (o semplicemente media). È la media aritmetica (quella più usata).

- Media (aritmetica) di una funzione di variabile casuale
  Sia a(x) una funzione della variabile casuale X distribuita secondo una p.d.f. f(x) → il valore di aspettazione della variabile a è:

$$E[a] = \int_{-\infty}^{+\infty} a(x) f(x) dx$$

Valori di Aspettazione

- Momento n-esimo (momento di ordine n)
  Il momento n-esimo di X è definito da:
  [$\mu_1' = \mu$ (media aritmetica)]

$$E[x^n] = \int_{-\infty}^{+\infty} x^n f(x)dx = \mu_n'$$

- Momento Centrale n-esimo (o di ordine n)
  è definito dalla quantità:

$$E[(x - E[x])^n] = \int_{-\infty}^{+\infty} (x - \mu)^n f(x)dx$$

- Il momento centrale secondo (o di ordine 2) è la **varianza** di x ed è indicato con σ^2 oppure con V[x] :

$$E[(x - E[x])^2] = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)dx = \sigma^2 = V[x]$$

- Si noti che $V[x] = E[x^2] - \mu^2$ (media dei quadrati meno quadrato della media)

- La deviazione standard σ è definita come la radice quadrata della varianza

Matrice di Covarianza

- Consideriamo eventi che coinvolgono più variabili casuali. Consideriamo il caso di due variabili casuali, X e Y.
- È necessario tenere conto dell'eventuale correlazione che lega le due variabili. Questa correlazione è detta covarianza V_{xy} [cov(x,y)], definita da:

$$V_{xy} = E[(x - \mu_x)(y - \mu_y)] = E[xy] - \mu_x \mu_y$$

$$V_{xy} = \int_{-\infty}^{+\infty} xy f(x,y) dx dy - \mu_x \mu_y$$

con $\mu_x = E[x]$, $\mu_y = E[y]$

- La matrice di covarianza , detta anche matrice degli errori, è una matrice simmetrica ($V_{xy} = V_{yx}$)

- Gli elementi diagonali della matrice sono le varianze delle variabili ($V_{xx} = \sigma^2_x$ $V_{yy} = \sigma^2_y$)

Coefficiente di Correlazione

- Il coefficiente di correlazione tra le due variabili X e Y è definito da :

$$\rho_{xy} = \frac{V_{xy}}{\sigma_x \sigma_y}$$

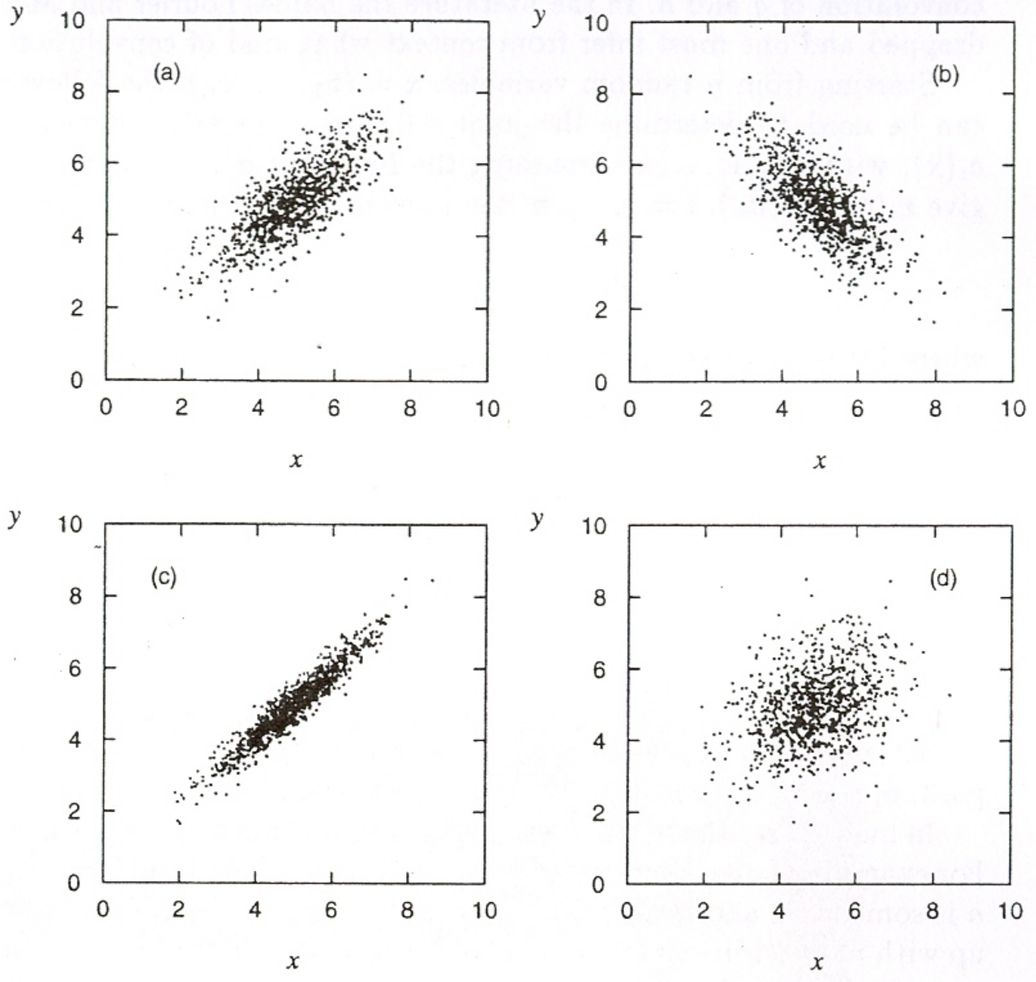
- È una quantità che assume valori tra -1 e +1.

- È una grandezza adimensionale (ma gli elementi della matrice di covarianza hanno dimensione!!). Misura il grado di correlazione **lineare** tra le due variabili.

- Due variabili indipendenti hanno grado di correlazione lineare uguale zero.

- Ma due variabili che hanno grado di correlazione lineare uguale a zero non necessariamente sono indipendenti. Possono infatti avere correlazioni di ordine superiore (quindi non lineari !) .

Coefficiente di Correlazione



Propagazione degli Errori

$$V[y] = \left(\frac{\partial y}{\partial x_1}\right)^2 V[x_1] + \left(\frac{\partial y}{\partial x_2}\right)^2 V[x_2] + 2\left(\frac{\partial y}{\partial x_1}\right)\left(\frac{\partial y}{\partial x_2}\right) V_{x_1 x_2}$$

$$\sigma_y^2 = \left(\frac{\partial y}{\partial x_1}\right)^2 \sigma_{x_1}^2 + \left(\frac{\partial y}{\partial x_2}\right)^2 \sigma_{x_2}^2 + 2\left(\frac{\partial y}{\partial x_1}\right)\left(\frac{\partial y}{\partial x_2}\right) \rho\sigma_{x_1}\sigma_{x_2}$$

dove i differenziali sono calcolati nei valori medi delle variabili.

- Se $Y = X_1 + X_2$ e le variabili X_1 e X_2 sono scorrelate, allora : $\sigma_y^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2$
  Gli errori si sommano in quadratura.

- Se $Y = X_1 \ X_2$ allora in generale si ha: $\frac{\sigma_y^2}{y^2} = \frac{\sigma_{x_1}^2}{x_1^2} + \frac{\sigma_{x_2}^2}{x_2^2} + 2\frac{V_{x_1 x_2}}{x_1 x_2}$

- Se le due variabili X_1 e X_2 sono scorrelate, allora si ha: $\frac{\sigma_y^2}{y^2} = \frac{\sigma_{x_1}^2}{x_1^2} + \frac{\sigma_{x_2}^2}{x_2^2}$
  In questo caso sono gli **errori relativi** che si sommano in quadratura.

- Questo risultato vale anche per il rapporto $Y = X_1/X_2$