Appunti di statistica ed analisi dei dati

Indice generale

Analisi dei dati

Una parte essenziale del pensiero scientifico è la prova di una ipotesi (vedi il metodo scientifico).
Provare un'ipotesi vuol dire predisporre un esperimento nel quale occorre fare delle misure per arrivare a confermare o confutare l'ipotesi.
In sostanza si cerca di misurare qualche "costante di proporzionalità" e confrontare il valore misurato con quello atteso.

Dal momento che non è possibile dare una valore esatto di una grandezza, bisogna tener conto di due fatti:

1. devo fare le misure con la più **alta precisione** possibile (scegliere gli strumenti di misura adeguati, ripetere più volte la misura, ecc..) per eliminare e ridurre errori sistematici, grossolani e casuali.

2. Devo stabilire un **criterio** che mi permetta di **confrontare** i valori misurati con quelli attesi.

Esempio:
Voglio misurare l'accelerazione di gravità sulla superficie della Terra g:

A) cerco un esperimento nel quale si possa misurare g, ad esempio il periodo di oscillazione di un pendolo

B) faccio le misure più volte e con i mezzi messi a disposizione della statistica trovo un valore di $g_{misurato}$= 9,7±0,2 m/s^2 : l'errore della misura è del 0,2/9,7 = 0,02 = 2% → abbiamo eseguito le misure nel modo corretto (una percentuale superiore al 10%-20% ci porterebbe a dire che le misure non sono buone e occorrerebbe ripetere l'esperimento con strumenti più precisi o fare più misure)

C) lo confronto con il valore atteso di g_{atteso}= 9,81 m/s^2 :

(criterio semplificato)

- trovo la "distanza" della mia misura da quella attesa: |9,7-9,81| = 0,11
- se la "distanza" trovata è minore dell'errore della misura allora posso affermare di aver misurato un valore concorde con il valore di g atteso (0,11< 0,2)

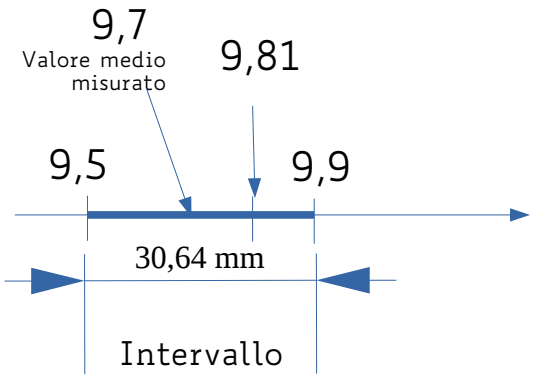
(criterio più sofisticato)

- trovo sempre la "distanza"
- divido la "distanza" per l'errore della misura 0,11/0,2 = 0,55
- valuto la probabilità, con il numero di misure effettuate, di trovare il valore misurato alla "distanza" calcolata nell'ipotesi che i due valori non siano compatibili; se questa probabilità è inferiore al 5% allora dico che è bassa la probabilità che siano incompatibili, quindi sono compatibili. Questo calcolo avviene utilizzando diverse "distribuzioni di probabilità" (argomento non trattato in questi appunti), ma a parte il calcolo, il ragionamento è lo stesso. Il 5% è la soglia di compatibilità stabilita a priori.

Un altro modo di interpretare il **criterio semplificato** del punto C è il seguente:

$g_{misurato}$= 9,7±0,2 m/s^2 identifica un intervallo da 9,7-0,2 = **9,5** a 9,7 + 0,2= **9,9**

Se il valore atteso cade in questo intervallo allora viene confermata l'ipotesi, altrimenti i valori misurati non sono compatibili con i valori attesi. Più stretto è l'intervallo, e maggiore sarà la precisione della nostra ipotesi[1].



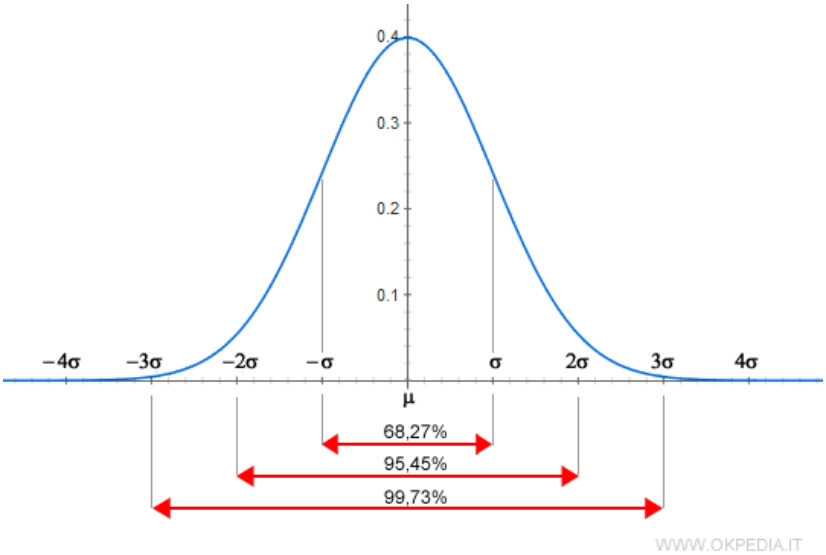
1 Quello che si continua a fare oggi, oltre a confermare una misura, è ridurre sempre più l'intervallo aumentando le misure, riducendo gli errori, ecc…

Calcolo della miglior stima di una serie di misure

Durante un esperimento vengono fatte più misure. Perché?
La risposta è semplice: per avere un errore più piccolo occorre ripetere più volte la misura. In realtà questo dipende da come viene calcolato l'errore.

Ogni volta che si compiono misure, si suppone che il valore più probabile sia il **valore medio** e che la maggior parte delle misure siano vicino a tale valore (*distribuzione normale* o *di Gauss*: è una rappresentazione della frequenza delle misure rispetto al valore medio, una campana, in figura)

μ è il valore medio e σ è l'errore calcolato dalle misure.
L'intervallo tra μ-σ e μ+σ rappresenta l'intervallo nel quale, sono



concentrate il 68,27% delle misure, tra μ-2σ e μ+2σ ci sono il 95,45% delle misure effettuate e così via.

Ma come si calcolano μ e σ?

Come si calcola μ

Valore medio μ di una serie di misure si calcola con la **classica media**:

$\mu = \bar{X} = X_{medio} = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$, con Xi le misure, N il numero di misure totali.

A volte viene usata la formula della **media pesata**:

$\mu = \bar{X} = X_{medio} = \frac{X_1 P_1 + X_2 P_2 + ... + X_N P_N}{P_1 + P_2 + ... + P_N} = \frac{\sum_{i=1}^{N} X_i P_i}{\sum_{i=1}^{N} P_i}$, dove Pi sono i pesi.

Esempio: voglio calcolare la media tra i voti dello scritto/orale che hanno un peso del 60% e i voti della relazione di laboratorio che hanno un peso pari al 40%:

Voto	6	5	6	6
Peso	60%	60%	40%	40%

$\mu = \bar{X} = X_{medio} = \frac{6*60 + 5*60 + 6*40 + 6*40}{60 + 60 + 40 + 40} = \frac{1140}{200} = 5{,}7$

I pesi sono arbitrari, potevo ad esempio scegliere 2 e 1 (il primo voto ha un peso doppio rispetto al secondo).

Come si calcola σ

σ (sigma) rappresenta l'errore casuale delle misure.
Il modo più semplice per stimare tale errore è quello di calcolare la semidispersione: $\sigma = \sigma_{semidispersione} = \frac{X_{max} - X_{min}}{2}$, la differenza tra la misura massima e la misura minima trovata diviso 2.
In realtà, una stima migliore dell'errore è data dalla **deviazione standard** o **scarto quadratico medio (corretto)**:

$\sigma = \sigma_X = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N-1}}$ **,** vale a dire che si calcola la differenza tra valore medio e misura, si eleva al quadrato, si sommano tutte e si divide per N-1, poi si fa la radice.
In questo caso, σ_X rappresenta l'**errore medio di ogni singola misura**.
L'intervallo tra $\mu - \sigma_X$ e $\mu + \sigma_X$ rappresenta l'intervallo nel quale, sono concentrate il 68,27% delle misure o si può anche dire che ho il 68,27% di probabilità che rifacendo una nuova misura il valore sia compreso in tale intervallo e così via per gli altri intervalli.

Se σ_X è l'errore medio di ogni singola misura, qual'è allora l'errore associato al valore medio che tenga conto di tutte le misure effettuate?
E' la **deviazione standard della media**:

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{N}} = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{(N-1)N}}$$

Se aumento le misure N, l'errore diminuisce.

In definitiva, da una serie di misure, la miglior stima della grandezza misurata è data da :

$$X_{migliore} = \mu \pm \sigma_{\bar{X}}$$

Che in genere, per semplificare l'analisi, si riduce a quello noto:

$$X_{migliore} = \mu \pm \sigma_{semidispersione}$$

Esercizio:

in un esperimento si ottengono i seguenti risultati:

86, 85, 84, 89, 86

Completa la tabella:

N	
μ	
$\sigma_{\text{semidispersione}}$	
$X_{migliore} = \mu \pm \sigma_{semidispersione}$	
σ_X	
$\sigma_{\bar{X}}$	
$X_{migliore} = \mu \pm \sigma_{\bar{X}}$	

In quale caso, l'errore è minore?

Propagazione degli errori

Altra questione è quella della propagazione degli errori perché spesso ci troviamo a dover calcolare, con una formula, una costante partendo da misure affette da errori. Vediamo in pratica.

Esempio 1 (somma e sottrazione):

Dobbiamo calcolare la somma di due lunghezze A=100±3 cm e B=50±2 cm. I valori di A e B sono stati misurati più volte ed è stata calcolata la media e l'errore.

Qual è la miglior stima della misura C = A+B ?

Di sicuro, 100+50 = 150 è il miglior valore di C. Ma il suo errore?

Vale una **regola per le somme e sottrazioni**:

quando sommo/sottraggo due grandezze con errore, l'errore associato alla somma/sottrazione è dato dalla ***somma degli errori***

Quindi σ_C =3+2 = 5

Pertanto C = 150±5 cm

Esempio 2 (prodotto e divisione):

Dobbiamo calcolare l'area di un quadrato di base B=100±3 cm e altezza H=50±2 cm. I valori di H e B sono stati misurati più volte ed è stata calcolata la media e l'errore.

Quando calcoliamo l'area A=B*H, che errore dobbiamo associare ad A?

A= 100*50 = 5000 cm^2 è la miglior stima dell'area.

Per l'errore vale la **regola per le moltiplicazioni e divisioni**:

*quando moltiplico/divido due grandezze con errore, l'****errore relativo*** *associato al prodotto/divisione è dato dalla* ***somma degli errori relativi***

Nel nostro esempio:

errore percentuale = 3/100 + 2/50 = 0,07 = 7% ed il 7% di 5000 vale 5000*0,07 = 350

Quindi la miglior stima del prodotto A = 5000 ± 350 = 5000 ± 400 cm^2

L'ultimo passaggio è stato fatto approssimando 350.

L**e regole da ricordare:**

per la somma/sottrazione → somma degli errori

per la moltiplicazione/divisione → somma degli errori relativi

Esercizio:

A=200±2 B=150±10 C=300±5

Calcola

	Valore medio	Errore
A+C		
B-C		
A*C		
A/B		
A*B/C		