# ELEMENTI DI STATISTICA

## 1. Misure ed errori

In un'analisi chimica si misurano dei valori chimico-fisici di svariate grandezze, tuttavia **ogni misura comporta sempre una incertezza**, dovuta alla presenza non eliminabile di errori nella misura; ne consegue che il valore che si ottiene non è il valore vero ma è una sua stima. Scopo dell'analista è far si che il valore misurato sia ragionevolmente vicino al valore vero, riducendo entro valori tollerabili il margine di incertezza, cioè l'errore.
Ovviamente tale incertezza dipende dal numero di misure effettuate: infatti, generalmente, si effettuano più misure della stessa grandezza, dette **repliche**; effettuando infinite misure si potrebbe pensare di compensare i molteplici errori ma in pratica il numero di misure è limitato.
La **statistica** offre gli strumenti matematici per valutare, partendo da un numero limitato di risultati:
- quanto il risultato finale di un'analisi sia rappresentativo, cioè esprima effettivamente la grandezza misurata
- quanto sia affidabile, cioè sia ragionevolmente vicino al valore vero

Qualsiasi misura di grandezze chimico-fisiche comporta **2 tipi di errori**:
- errori sistematici (o determinati), legati alle capacità dell'operatore, all'organizzazione del laboratorio, alla taratura degli strumenti, al metodo analitico adottato. Questi errori possono essere facilmente eliminati eliminando le sostanze interferenti, migliorando la pulizia dei recipienti, tarando periodicamente gli apparecchi di misura. Devono comunque essere assenti da qualsiasi misura e quindi anche da un'analisi chimica
- errori casuali (o indeterminati o random): sono casuali, non prevedibili e non eliminabili, dovuti per esempio a piccole fluttuazioni dell'ambiente del laboratorio (piccole variazioni di temperatura, umidità, ecc.). E' proprio l'esistenza di questi errori che fa si che il valore misurato non sia quello vero ma una sua stima. La loro influenza sui risultati può essere analizzata mediante metodi statistici applicati ai dati raccolti mediante una serie ripetuta di misure.

Nella statistica si usano alcuni **termini specifici**:
- **valore vero (μ)**: si intende il valore ottenuto da un analista esperto in perfette condizioni operative (assenza di errori sistematici) che effettui un numero infinito di prove analitiche usando metodi appropriati e strumenti efficienti. Naturalmente si tratta di una situazione ipotetica, non realizzabile in pratica: in linea teorica, la **media aritmetica** $\bar{x}$ di questa serie infinita di dati viene fatta coincidere col valore vero, in quanto vi è la massima probabilità che gli errori casuali si compensino
- **popolazione**: è l'insieme degli infiniti dati analitici ottenuti nell'ipotetica situazione prima descritta; in altre parole la media aritmetica di una popolazione di dati ha la massima probabilità di coincidere col valore vero e perciò viene fatta convenzionalmente coincidere con esso
- **campione**: è un insieme limitato di dati estratto dall'intera popolazione (infinita) di dati possibili; lo scopo dell'analista è quello di estrarre dalla popolazione un campione significativo (che rappresenti cioè la popolazione). Poiché il numero di analisi effettuate su un campione sarà necessariamente limitato, la media aritmetica di un campione non coincide col valore vero ma la statistica permette di valutare quanto si avvicini e quindi di valutare l'efficacia del metodo analitico seguito.

## 2. Esattezza e accuratezza

Entrambe indicano quanto un singolo dato (accuratezza) o la media aritmetica di una serie di dati (esattezza), si avvicinano al valore vero μ; sono entrambe espresse in termini di errore (o scarto).
Avendo la seguente serie di dati: $x_1$, $x_2$,...$x_n$, indicando con $x_i$ un generico valore della serie, con $\bar{x}$ il valore medio della serie, con μ il valore vero, si possono definire i diversi tipi di errore:

errore assoluto: è definito mediante l'accuratezza o l'esattezza

$$E_{ass} = x_i - \mu \quad \text{(accuratezza)} \qquad \text{ovvero} \quad E_{ass} = \bar{x} - \mu \quad \text{(esattezza)}$$

$$\bar{x} = \frac{\sum_1^n x_i}{n}$$

La sommatoria che compare nella definizione del valore medio ha il significato di sommare tutti i termini della serie da 1 a n

errore relativo: anch'esso è definito in termini di accuratezza ed esattezza

$$E_{rel} = \frac{x_i - \mu}{\mu} \quad \text{(accuratezza)} \quad \text{ovvero} \quad E_{rel} = \frac{\bar{x} - \mu}{\mu} \quad \text{(esattezza)}$$

errore relativo percentuale: è definito sempre come accuratezza ed esattezza

$$E_{rel\%} = \frac{x_i - \mu}{\mu} \cdot 100 \quad \text{(accuratezza)} \qquad \text{ovvero} \quad E_{rel\%} = \frac{\bar{x} - \mu}{\mu} \cdot 100 \quad \text{(esattezza)}$$

L'errore assoluto di una singola misura (accuratezza), indicato con $E_i$ ed espresso da $E_i = x_i - \mu$ può essere scritto e calcolato in un modo particolare; dalla formula precedente sommiamo e sottraiamo il valore $\bar{x}$:

$$E_i = x_i - \mu + \bar{x} - \bar{x} = (x_i - \bar{x}) + (\bar{x} - \mu)$$

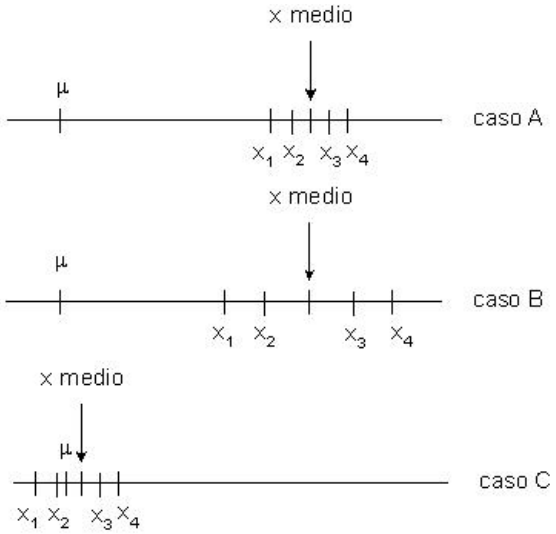
Il primo termine $(x_i - \bar{x})$ rappresenta l'errore assoluto casuale essendo riferito al singolo dato, mentre il secondo termine $(\bar{x} - \mu)$ rappresenta l'errore assoluto sistematico, essendo riferito all'intera media della serie di dati. Quindi l'errore $E_i$ commesso in una singola misura è la somma dell'errore casuale e dell'eventuale errore sistematico. La serie di misure è corretta quando è presente solo il primo termine mentre il secondo deve essere nullo.

## 3. Precisione, accuratezza, deviazione

Indica l'accordo di una serie di dati tra loro ed è sempre riferita alla serie e non alla singola misura; viene espressa come "deviazione" dalla media aritmetica ed è una misura della dispersione dei dati cioè alla distribuzione dei dati ottenuti all'interno della serie. Viene valutata mediante:

$precisione = x_i - \bar{x}$ e quindi coincide con l'errore assoluto casuale $E_i$

Poiché vi è una certa tendenza a confondere accuratezza e precisione, si deve notare che si può essere precisi ma non accurati e viceversa: per esempio la presenza di un errore sistematico non eliminato comporta una bassa accuratezza (i dati ottenuti sono lontani dal valore vero) ma possono essere invece poco dispersi, cioè vicini al valore medio, se la precisione è stata elevata. Gli schemi seguenti chiariscono quanto detto in precedenza:




Il caso A rappresenta una serie di misure precise (piccola deviazione o scarto dal valore medio) ma poco accurate (grande errore in quanto il valore medio è lontano dal valore vero); probabilmente un errore sistematico ha alterato tutte le misure effettuate, che sono poco disperse tra loro e quindi danno l'illusione di aver ottenuto un risultato valido.
Il caso B è ancora peggiore: le misure non sono né precise né accurate.
Il caso C è quello che dovrebbe verificarsi per ogni analisi chimica: si tratta di misure precise ed accurate, visto che il valore medio quasi coincide con il valore vero. In questo caso i risultati ottenuti sono senz'altro significativi.

Per esprimere la precisione e quindi la dispersione dei dati di una serie si possono utilizzare vari parametri statistici:

deviazione: è la differenza tra ogni singolo valore della serie ed il valore medio:

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

nuovamente quindi compare la definizione di precisione, ovvero di errore assoluto casuale $E_i$. Da notare come vi siano varie definizioni che vengono espresse dalla stessa relazione, che possono senz'altro indurre confusione nella terminologia: l'errore assoluto, lo scarto, la precisione e la deviazione coincidono!

deviazione media: è la media aritmetica dei valori assoluti (senza il segno) delle deviazioni della serie di n dati

$\bar{d} = \frac{\sum_1^n |d_i|}{n}$ il valore assoluto di $d_i$ è necessario per eliminare i segni negativi di alcune deviazioni

La deviazione media fornisce indicazioni sulla precisione ottenuta nella serie di misure, tanto è vero che spesso viene utilizzata nella presentazione dei risultati, utilizzando la notazione:

$x = \bar{x} \pm \bar{d}$ dove x è il risultato finale della serie di misure

deviazione relativa percentuale: viene espressa mediante

$$d_{\%} = \frac{\bar{d} \cdot 100}{\bar{x}}$$

definita come la deviazione media della singola misura quando il valore medio è 100. Viene utilizzata per confrontare la precisione di due diverse serie di misure; per esempio se in una prima serie si ha $d_{\%}$ = 1,5% mentre in una seconda serie si ha $d_{\%}$ = 0,2% allora la seconda serie ha un grado di precisione superiore alla prima.

## 4. Cifre significative

Il dato analitico è un numero che deriva da una misura sperimentale mentre il risultato di una prova è un numero che deriva dai dati analitici dopo elaborazioni numeriche e grafiche.
Il dato analitico deve quindi essere registrato in modo da contenere esclusivamente cifre significative, coerenti con la precisione degli strumenti utilizzati e col metodo seguito. In particolare si devono riportare tutte le cifre note con certezza più la prima incerta, indicando eventualmente di fianco l'intervallo di certezza; in mancanza di altre indicazioni l'ultima cifra deve essere considerata incerta a meno di (+) o (-) una unità. Alcuni esempi:

| | |
|---|---|
| bilancia digitale con precisione di + 0,1 mg | 4,0057 + 0,0001 g |
| potenziometro digitale con precisione di + 1 mV | 434 + 1 mV |
| buretta con divisioni da 0,05 ml e tolleranza di + 0,03 ml | 5,25 + 0,03 ml |

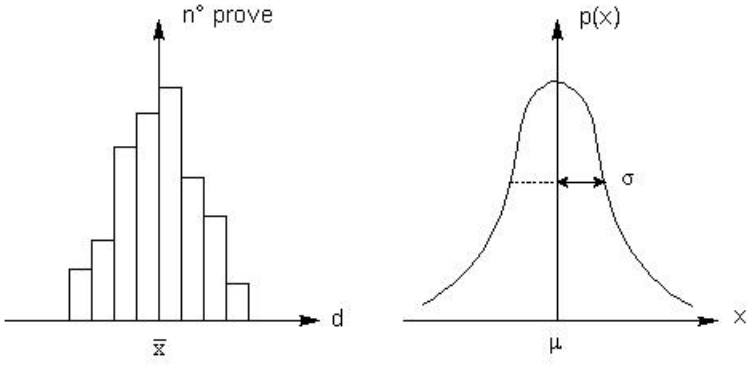
Per quanto riguarda il trattamento degli zeri:
- gli zeri che compaiono in mezzo a cifre significative sono significativi; per esempio 3,0046 ha 5 cifre significative;
- gli zeri iniziali non sono significativi a meno che si trovino a destra della virgola nei logaritmi; per esempio 0,0102 ha 3 cifre significative e può essere meglio scritto mediante la notazione scientifica $1{,}02 \cdot 10^{-2}$; al contrario il log(1,02) è 0,049 che ha 3 cifre significative;
- gli zeri al fondo di un numero sono di solito significativi: per es. 5,00 ha 3 cifre significative; se in un calcolo aritmetico si ottiene il valore 5000 ed è necessario che abbia solo 2 cifre significative allora va espresso come $5{,}0 \cdot 10^3$ in modo da evidenziare la precisione.

Durante i calcoli si combinano numeri con un diverso numero di cifre significative: il risultato finale, generalmente, deve avere lo stesso numero di cifre significative del numero che ne aveva di meno. Le cifre non significative devono essere eliminate dai calcoli mediante arrotondamento secondo le seguenti regole:
- l'ultima cifra significativa viene aumentata di 1 se quella successiva è maggiore di 5
- viene invece lasciata inalterata se quella successiva è minore di 5
- se è uguale a 5 si approssima al numero pari più vicino.

Alcuni esempi: 2,38 → 2,4 2,31 → 2,3 2,25 → 2,2 2,35 → 2,4

## 5. Distribuzione statistica dei dati

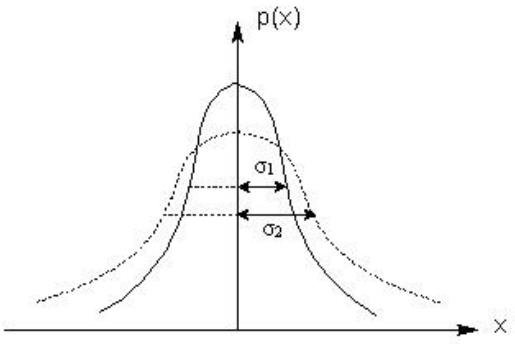
Le analisi di laboratorio spesso producono serie di dati dalle quali, con metodi statistici, deve essere dedotto un unico significativo risultato analitico. La statistica ha sviluppato numerosi strumenti per il trattamento delle serie di dati, che si basano sulla distribuzione statistica (cioè casuale) della serie di dati ottenuti: la presenza di errori statistici non eliminabili produce infatti risultati diversi per misure ripetute della stessa grandezza.
Supponendo di disporre di una serie di $x_1, x_2 \ldots x_n$ dati, si calcola il valor medio $\bar{x}$, si calcolano le n deviazioni dal valor medio e quindi si produce l'istogramma delle deviazioni d, suddividendo le deviazioni positive e negative in uguali intervalli e quindi riportando il numero di valori della serie (n° prove) compresi in ogni intervallo:



Si ottiene un istogramma: gli intervalli più popolati sono quelli attorno al valore nullo di deviazione, cioè la maggior parte dei dati si dispone attorno al valor medio, che si troverà all'origine degli assi: ciò perché, in presenza di soli errori casuali, è più probabile compiere un piccolo errore mentre la probabilità di compiere errori molto grandi diminuisce drasticamente e quindi diminuisce anche il numero di prove con grandi deviazioni.
Si supponga ora di aumentare il numero di dati fino ad avere una serie infinita e di restringere gli intervalli di deviazione fino a renderli infinitesimi: l'istogramma delle deviazioni si trasformerà in una curva continua detta **curva di Gauss**, con la tipica forma a campana (detta appunto distribuzione normale o gaussiana) ed un andamento simmetrico con le code che si estendono all'infinito: in questo caso sull'asse orizzontale si avrebbero infiniti intervalli di deviazione, cioè l'infinita serie di dati x (popolazione) mentre sull'asse verticale è presente una funzione detta densità di probabilità p(x), avente la seguente complessa espressione matematica:

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

Nella curva di Gauss all'origine degli assi è presente il valore vero μ che corrisponde al massimo della curva di probabilità; il termine σ che compare nella funzione di Gauss è detta **deviazione standard** e rappresenta la semi-larghezza della distribuzione riferita ai punti di flesso (dove la curva di Gauss cambia concavità); si ha infatti che i due punti di flesso sono individuati da $x = \mu \mp \sigma$
In pratica la curva di Gauss rappresenta la probabilità di distribuzione in una serie infinita di misure di una grandezza fisica: ovviamente il più probabile sarà il valore vero, mentre sempre meno probabili saranno i valori che mano a mano si allontanano da quest'ultimo. Rappresenta quindi la **distribuzione dell'errore casuale** nella suddetta serie infinita. Anche nelle serie finite di misure che si possono effettuare in laboratorio, l'errore commesso deve seguire la curva di Gauss per essere effettivamente casuale e di altro genere.
L'area totale sottesa dalla curva ha un valore unitario (cioè vale 1) in quanto rappresenta una probabilità unitaria (100%) e quindi la certezza che una qualsiasi misura ricada in questo intervallo.




Si possono avere più distribuzioni aventi larghezza ed altezza diversa: ciò è correlabile alla precisione della serie di misure. Una serie di misure precise fornirà una curva alta e stretta, una serie di misure poco precise produrrà una curva larga e bassa. Tuttavia l'area sottesa da tutte le infinite curve sarà sempre la stessa e pari ad 1.
Vi è quindi la necessità di esprimere questa situazione per evidenziare la precisione della serie di misure: si ricorre alla deviazione standard σ che graficamente rappresenta appunto la semi-larghezza della distribuzione nei punti di flesso.
Nel disegna a fianco la distribuzione con $\sigma_1$ rappresenta una serie di misure più precisa rispetto a quella più larga con $\sigma_2$

E' chiara quindi l'importanza della deviazione standard σ nella statistica: è un buon indicatore della precisione. La deviazione standard (o deviazione quadratica media) si può calcolare mediante le seguenti equazioni:

$$\sigma = \sqrt{\frac{\sum_1^n d_i^2}{n}}$$

dove $d_i$ è la deviazione della misura i-esima ed n è il numero di prove. Questa è detta deviazione standard effettiva ma non è calcolabile in pratica perché è riferita ad una serie infinita di misure (popolazione) che, ovviamente, non si possono realizzare poiché sarebbe necessario un tempo infinito. Vista l'equazione a fianco, σ può essere definita come la radice quadrata della media del quadrato degli scarti

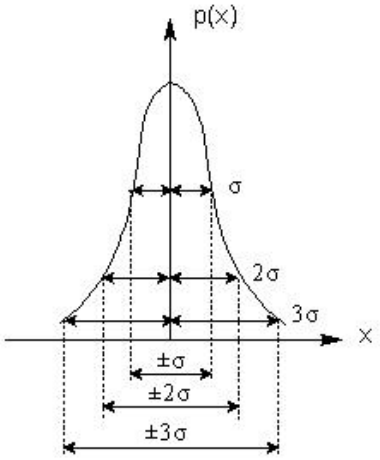
$$s = \sqrt{\frac{\sum_1^n d_i^2}{n-1}}$$

s è detta deviazione standard stimata e viene usata nel trattamento dei dati analitici in quanto il numero di prove è limitato (campione)

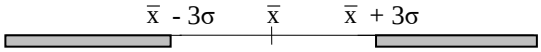
Si utilizza s al posto di σ per indicare che, avendo a disposizione una serie limitata di dati (cioè un campione e non una popolazione), non si può calcolare la vera deviazione standard ma solo fare una sua stima. Al crescere del numero di dati della serie (tendente ad infinito), in assenza di errori sistematici il valor medio $\bar{x}$ tende al valore vero μ, la deviazione standard stimata s tende alla deviazione standard effettiva σ.

## 6. Limiti di attendibilità

Una caratteristica matematica della curva di Gauss, che rappresenta la distribuzione degli errori casuali, è quella che l'area sottesa all'intera curva rappresenta la percentuale del 100% di trovare il valore di una misura, cioè la certezza. Infatti il valore di una qualsiasi grandezza è certamente compreso tra $+\infty$ e $-\infty$, limiti tra i quali si estende la curva di Gauss, che si avvicina all'asse orizzontale ma non lo tocca mai.
Un problema che si può manifestare quando si effettua una serie di misure è la presenza di uno o più valori anomali (aberranti) molto diversi dagli altri; si deve prendere la decisione se accettarlo o scartarlo quando si calcola il valore medio. Il criterio di accettabilità o meno deriva dalla stessa distribuzione di Gauss.




Si può infatti dimostrare che l'area sottesa da ±1σ rappresenta una probabilità del 63,8% di trovare un valore misurato, quella sottesa da ±2σ rappresenta una probabilità del 95% ed infine quella sottesa da ±3σ rappresenta una probabilità del 99,7% (in pratica la certezza): quindi è estremamente improbabile che, con una corretta distribuzione gaussiana dell'errore casuale, una misura presenti una deviazione standard superiore a 3σ.
Facendo quindi coincidere $\bar{x}$ con μ (approssimazione abbastanza ragionevole in assenza di errori sistematici per un determinato campione di dati), si può stabilire un criterio per determinare i limiti di attendibilità (o di affidabilità, o di confidenza) cioè un criterio per decidere se accettare o scartare un valore ottenuto da una misura nettamente diverso da tutti gli altri della serie: vengono accettati tutti quei valori sperimentali compresi nell'intervallo: ($\bar{x}$ - 3σ) limite inferiore di attendibilità e ($\bar{x}$ + 3σ) limite superiore di attendibilità; vengono invece scartati perché affetti probabilmente da qualche errore sistematico tutti quei valori derivanti da misure sperimentali che sono al di fuori di questo intervallo, con il 99,7% di probabilità di prendere una decisione corretta.




valori accettabili

Naturalmente σ non può essere determinato perché la serie di misure produce un campione e non una popolazione di valori. Pertanto sono stati elaborati vari test statistici, tutti basati sul concetto di limite di attendibilità, che permettono di prendere decisioni sulla serie di risultati analitici senza calcolare effettivamente σ ma utilizzando altri parametri statistici.

## 7. Test di Dixon (o Q-test)

Spesso in campo analitico si presente il seguente problema: un dato di una serie di misure risulta anomalo, cioè molto lontano dagli altri; deve essere accettato o scartato perché probabilmente affetto da errore sistematico? La risposta è che deve essere scartato se si trova al di fuori del limite di attendibilità (o di affidabilità o di confidenza). Infatti valori di una serie di misure al di fuori dell'intervallo di affidabilità devono essere scartati perché altererebbero in modo eccessivo il valore medio.
Sono stati elaborati vari test per decidere se accettare o scartare valori di una serie con deviazione elevata rispetto al valore medio; uno dei più usati è il **test di Dixon (o Q-test)**. Innanzitutto si ordinano i valori della serie in senso crescente: $x_1$, $x_2$ ... $x_n$ dal più piccolo al più grande
- se il dato dubbio è il primo della serie, cioè il più basso si calcola il rapporto r definito come:

$$r = \frac{x_2 - x_1}{x_n - x_1}$$

dove $x_1$ è il primo dato della serie ordinata, $x_2$ è il secondo e $x_n$ è l'ultimo. Si confronta il valore di r con i valori del quoziente critico Q (tabulato con una probabilità prefissata, ad esempio del 90%): se r < Q il dato viene accettato con una sicurezza del 90% altrimenti viene scartato
- se il dato dubbio è l'ultimo, cioè il più alto, si calcola r mediante:

$$r = \frac{x_n - x_{(n-1)}}{x_n - x_1}$$

dove $x_{(n-1)}$ è il penultimo dato della serie ordinata. Si confronta il valore di r con i valori di Q critico tabulato: anche in questo caso se r < Q il dato viene accettato con una sicurezza del 90%.

Si procede in questo modo sui dati della serie ordinata fino a trovare, alle due estremità, due valori che superano il test di Dixon e quindi sono da accettare; gli altri valori, interni alla serie, saranno anch'essi da accettare e con tali valori si può procedere alle elaborazioni statistiche (calcolo della media, ecc.).
I valori del quoziente Q critico di Dixon sono raccolti nella seguente tabella:

| numero di prove | Q critico (90% di sicurezza) |
|---|---|
| 3 | 0,94 |
| 4 | 0,76 |
| 5 | 0,64 |
| 6 | 0,56 |
| 7 | 0,51 |
| 8 | 0,47 |
| 9 | 0,44 |
| 10 | 0,41 |

## 8. Coefficiente t di Student

Avendo a disposizione una serie infinita di misure, prive di errori sistematici, il valore medio della serie $\bar{x}$ coincide con il valore vero μ; in realtà questa situazione non si verifica mai, perché quando si effettuano delle misure si dispone di un campione e non dell'intera popolazione; quindi $\bar{x}$ è solo una stima del valore vero, tanto più corretta quanto più è elevato il numero delle misure effettuate.
La statistica permette di valutare, con un grado di probabilità prefissata, l'intervallo di confidenza, cioè l'intervallo di valori attorno al valore medio in cui è situato il valore vero, mediante la relazione di Student:

$$\mu = \bar{x} \pm \frac{t \cdot s}{\sqrt{n}}$$

dove μ è il valore vero, $\bar{x}$ è il valore medio, t è il coefficiente di Student opportunamente tabellato, s è la deviazione standard stimata, n è il numero delle prove effettuate

Il coefficiente t di Student (un matematico che si chiamava in realtà W.S. Gosset) è tabellato in funzione del numero di prove e dell'intervallo di probabilità che viene scelto:

| | grado di probabilità | | |
|---|---|---|---|
| numero di prove | 90% | 95% | 99% |
| 2 | 6,31 | 12,7 | 63,66 |
| 3 | 2,92 | 4,3 | 9,92 |
| 4 | 2,35 | 3,18 | 5,84 |
| 5 | 2,13 | 2,78 | 4,60 |
| 6 | 2,02 | 2,57 | 4,03 |
| 7 | 1,94 | 2,45 | 3,71 |
| 8 | 1,90 | 2,36 | 3,50 |
| 9 | 1,86 | 2,31 | 3,36 |
| 10 | 1,83 | 2,26 | 3,25 |
| 11 | 1,81 | 2,23 | 3,17 |
| 12 | 1,80 | 2,20 | 3,11 |
| 13 | 1,78 | 2,18 | 3,06 |
| 14 | 1,77 | 2,16 | 3,01 |
| 15 | 1,76 | 2,14 | 2,98 |
| 16 | 1,75 | 2,13 | 2,95 |
| 20 | 1,72 | 2,09 | 2,84 |

## 9. Presentazione dei risultati

Nella pratica analitica di solito da una serie di misure si ottiene una serie di dati; se un dato risultasse anomalo, ovvero fosse troppo piccolo o troppo grande rispetto agli altri si deve utilizzare un test statistico (come il Q-test di Dixon) per valutare se accettare o scartare il dato. Quindi si calcola il valore medio $\bar{x}$, che in assenza di errori sistematici ha buone possibilità di essere vicino al valore vero μ e quindi può essere fornito come valore finale dell'analisi.

Tuttavia il risultato finale della serie di misure, per essere significativo, deve anche esprimere l'accuratezza, la precisione, la dispersione dei dati ecc. e quindi si pone il problema della presentazione del risultato stesso. Vi sono due modi per esprimere il risultato numerico di una serie di misure:

$$x = \bar{x} \pm \bar{d}$$

$$\mu = \bar{x} \pm \frac{t \cdot s}{\sqrt{n}}$$

Da notare che la seconda espressione, che attualmente viene preferita alla prima, non è altro che la miglior valutazione del valore vero μ espressa dalla relazione di Student e quindi riassume i parametri statistici richiesti per la presentazione e valutazione del risultato ottenuto.
Pertanto nel referto analitico dovrebbero comparire il risultato della serie di analisi effettuate sul campione espresso nel seguente modo:

- risultato numerico: il valore medio $\bar{x}$
- incertezza: il termine che contiene il t di Student $\pm \frac{t \cdot s}{\sqrt{n}}$

Il valore medio ottenuto permetterà eventualmente di effettuare successivi calcoli ed elaborazioni dei dati mentre l'incertezza, espressa attraverso la relazione di Student, fornisce indicazioni sulla precisione dell'analisi.

## 10. Relazioni lineari tra due variabili (statistica bivariata)

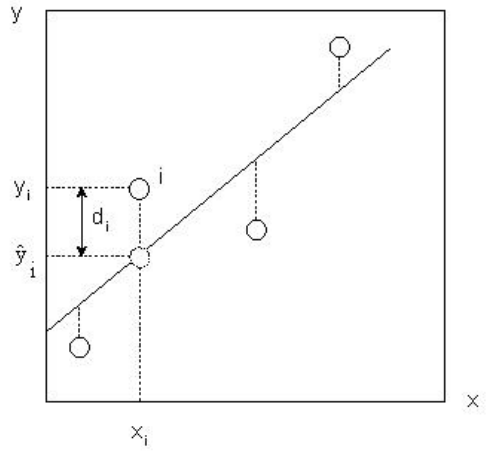
Finora è stata applicata l'analisi statistica monovariata, applicata cioè ad una singola serie di dati $x_1, ...x_n$; nella pratica analitica.
Al contrario, spesso si lavora con due serie di dati: ad esempio nella **costruzione di una retta di lavoro**, oltre a quella precedente ve ne sarà anche una del tipo $y_1, ...y_n$, dove la serie x rappresenta una variabile chimica (la concentrazione degli standard) in una serie di campioni e la serie y rappresenta una variabile strumentale (il segnale, cioè la risposta, prodotta da un apparecchio in corrispondenza di ogni valore di x). E' quindi necessario introdurre alcuni elementi di analisi statistica bivariata, che fornisce gli strumenti statistici utili al trattamento di questi casi.
In generale si può affermare che "tra due variabili vi è un legame" quando:

- le due variabili possono essere associate ad una legge del tipo y = f(x), con una corrispondenza biunivoca tra ogni valore di x ed un solo valore di y: è il caso della costruzione di una curva di lavoro. Se la relazione è di tipo lineare si otterrà una retta di lavoro, caso generalmente da preferire
- le due variabili possono essere correlate in modo generico, nel senso che non esiste una vera e propria legge conosciuta e rigorosa ma è possibile tracciare una curva di regressione che evidenzi un legame più o meno forte tra loro

### 10.1. Regressione

Lo studio della regressione riguarda l'analisi della **relazione di dipendenza tra due variabili**, delle quali una è indipendente mentre l'altra dipende dalla prima; tra le due esiste una relazione di dipendenza generale del tipo y = f(x) come ad es. nella costruzione di una curva di calibrazione, essendo x la variabile chimica (di solito la concentrazione) ed y quella strumentale (la risposta dello strumento ovvero il segnale).
Spesso la legge di dipendenza è quella di una retta del tipo: $y = b_1 \cdot x + b_0$ dove $b_1$ è il coefficiente angolare (o pendenza) e $b_0$ è il termine noto (o intercetta). Mediante standard a C nota si misura un qualche parametro analitico per mezzo di uno strumento, ottenendo due serie di valori: $x_1, ...x_n$ ed $y_1, ...y_n$ che, disposti in grafico, dovrebbero allinearsi per formare una retta. In realtà le misure comportano sempre errori casuali e quindi i punti non si dispongono perfettamente su di una retta: vi è quindi la necessità di adattare ai punti sperimentali la "retta migliore", ovvero quella più probabile, cioè di effettuare il "best fitting" (migliore adattamento). Può essere fatto "a occhio", cercando di passare più vicino possibile ai punti sperimentali o, meglio, utilizzando lo strumento matematico della regressione. La **regressione** quindi è un procedimento matematico in cui, a partire da dati sperimentali, si cerca l'equazione della funzione y = f(x) che meglio si adatta ai punti.
Nel caso in cui la legge di dipendenza sia una retta si parla di **regressione lineare** ed uno dei metodi più usati è quello dei **minimi quadrati**.
Si supponga di costruire una retta di lavoro avendo a disposizione le due serie di dati $x_1 \ldots x_n$ (contrazione degli standard) ed $y_1 \ldots y_n$ (segnale prodotto dall'apparecchio di misura per ogni standard); a causa degli errori casuali i punti, riportati in grafico, non si disporranno perfettamente lungo una retta ma fluttueranno casualmente attorno ad essa. Si suppone che la variabile indipendente x non sia affetta da errore (nessun errore relativo alla concentrazione degli standard) e che gli unici errori casuali siano relativi alle letture effettuate con lo strumento, cioè interessino la sola variabile y.

Per tracciare la retta più probabile si consideri il punto sperimentale generico i di coordinate ($x_i$, $y_i$): se tale punto si trovasse esattamente sulla retta avrebbe coordinate ($x_i$, $\hat{y}_i$) per cui il segmento $d_i$ costituisce lo scarto o deviazione, cioè l'errore casuale commesso per il punto i. E' intuitivo che la retta più probabile, cioè più rappresentativa delle due serie di dati tra loro correlate, sarà quella che nel complesso passerà più vicino a tutti i punti, in modo da rendere minime le deviazioni per tutti i punti: magari non toccherà nessun punto ma complessivamente renderà minima la somma delle deviazioni (scarti)




In senso matematico il principio è quello di rendere minima la funzione:

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

che rappresenta la sommatoria del quadrato degli scarti. Da notare che si rende minima la sommatoria del quadrato degli scarti e non la semplice sommatoria degli scarti, in modo da eliminare il segno negativo di alcune deviazioni. Per questo motivo questo tipo di regressione viene denominata "metodo dei minimi quadrati"

In pratica l'equazione della retta di regressione si ottiene mediante le relazioni seguenti:

$$y = b_1 \cdot x + b_0$$ equazione della retta di regressione

$$b_1 = \frac{\sum_1^n (x_i \cdot y_i) - \frac{(\sum_1^n x_i) \cdot (\sum_1^n y_i)}{n}}{\sum_1^n x_i^2 - \frac{(\sum_1^n x_i)^2}{n}} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$

Noti il coefficiente angolare $b_1$ e l'intercetta $b_0$ si può costruire la retta mediante due punti, che si possono determinare assegnando il valore delle loro x e calcolando le relative y. Ovviamente i vari calcoli possono essere semplificati ed automatizzati ricorrendo ad un PC e ad un software come un foglio di calcolo, per es. Excel, che permette con apposite formule di calcolare la pendenza ($b_1$) e l'intercetta ($b_0$) di una retta di regressione e di tracciarla direttamente sul grafico che rappresenta la retta.
Un altro importante parametro relativo alla regressione è il **coefficiente di correlazione r**, che rappresenta la "forza" della correlazione tra le due variabili, quella indipendente x e quella dipendente y. Viene utilizzato di solito per valutare se tra due variabili osservate esiste o meno una relazione di dipendenza:

$$r = \frac{\sum_1^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_1^n (x_i - \bar{x})^2 \cdot \sum_1^n (y_i - \bar{y})^2}}$$

I termini $\bar{x}$ e $\bar{y}$ sono i valori medi rispettivamente della serie x e della serie y di valori. Il coefficiente r:

- è una quantità adimensionale
- varia da -1 a +1 (+1 per una perfetta correlazione lineare positiva, -1 per una perfetta correlazione lineare negativa)

Quindi una valida retta di lavoro, dove è noto fin dall'inizio che le concentrazioni degli standard (x) ed i corrispondenti segnali prodotti dall'apparecchio di misura (y) sono correlati tra loro, deve avere un valore di r molto vicino ad 1.
Occorre infine fare alcune precisazioni sul metodo dei minimi quadrati, per evitare di arrivare talora a conclusioni errate:

- quando si usa il modello della regressione (in particolare quella lineare), si deve fare attenzione a non commettere errori di interpretazione: anche se due variabili x e y risultano talmente correlate da poter essere rappresentate da una retta di regressione ciò non significa che esista una relazione di causa-effetto, cioè che la variazione di x provochi la variazione di y e che per qualunque valore di x, usando la retta di regressione, si possa ricavare il corrispondente valore di y. La relazione causa-effetto può essere provata solo attraverso lo studio chimico-fisico del fenomeno e non solo con metodi statistici
- il modello di regressione lineare è valido solo nell'intervallo dei dati sperimentali, detto intervallo di linearità ma non si possono estrapolare i dati al di fuori di esso perché la relazione di linearità $y = f(x) = b_1 \cdot x + b_0$ potrebbe non essere più valida per quel particolare fenomeno allo studio

## 10.2. Esercizi

### Esercizio 1

La standardizzazione di una soluzione circa 0,1 N di HCl, condotta mediante 4 diverse titolazioni, ha fornito i seguenti risultati:
- 0,1024 N
- 0,1142 N
- 0,0925 N
- 0,0860 N

Determinare il titolo esatto della soluzione di HCl

Innanzitutto si verifica con il test di Dixon se tutti i valori ottenuti sono da accettare; si ordina la serie in senso crescente:
- 0,0860 N
- 0,0925 N
- 0,1024 N
- 0,1142 N

Si applica il test al valore più piccolo della serie, utilizzando la tabella del $Q_{critico}$ con il 90% di sicurezza:

$r = \frac{x_2 - x_1}{x_n - x_1} = \frac{0{,}0925 - 0{,}0860}{0{,}1142 - 0{,}0860} = 0{,}23$ $Q_{critico} = 0{,}76$ per n = 4 poiché r < Q il dato iniziale è da accettare

$r = \frac{x_n - x_{(n-1)}}{x_n - x_1} = \frac{0{,}1142 - 0{,}1024}{0{,}1142 - 0{,}0860} = 0{,}42$ $Q_{critico} = 0{,}76$ per n = 4 poiché r < Q il dato finale è da accettare

Pertanto tutti i dati della serie risultano corretti con il 90% di sicurezza.
Si può ora calcolare la media della serie:

$\bar{x} = \frac{\sum_1^n x_i}{n} = \frac{0{,}0860 + 0{,}0925 + 0{,}1024 + 0{,}1142}{4} = 0{,}0988\ N$

Per completare il risultato si calcola ora l'incertezza, espressa attraverso il coefficiente t di Student: per n = 4 si ha dalla tabella di Student, per una probabilità del 95%, t = 3,18

$s = \sqrt{\frac{\sum_1^n d_i^2}{n-1}} = \sqrt{\frac{(0{,}086 - 0{,}0988)^2 + (0{,}0925 - 0{,}0988)^2 + (0{,}1024 - 0{,}0988)^2 + (0{,}1142 - 0{,}0988)^2}{4-1}} = 0{,}0123$

$\frac{t \cdot s}{\sqrt{n}} = \frac{3{,}18 \cdot 0{,}0123}{\sqrt{4}} = 0{,}0196$

Il risultato finale della standardizzazione sarà pertanto: N = 0,0988 ± 0,02 e quindi vi sarà il 95% di probabilità che la N vera della soluzione sia compresa in questo intervallo.

### Esercizio 2

Si vuole costruire la retta di lavoro relativa all'analita X. Alcuni standard di analita X sono stati analizzati mediante un opportuno apparecchio di misura, che ha prodotto i seguenti valori:

| C (mg/l) | 0,10 | 0,20 | 0,30 | 0,40 | 0,50 | 0,60 |
|---|---|---|---|---|---|---|
| S | 0,124 | 0,236 | 0,330 | 0,480 | 0,610 | 0,715 |

Ricavare l'equazione della retta di regressione ed il valore del coefficiente di correlazione, tracciando infine il grafico di calibrazione.

L'equazione della retta di regressione sarà una generica retta del tipo: $S = b_1 \cdot C + b_0$ dove C è la concentrazione di ogni standard, S il corrispondente segnale prodotto dall'apparecchio durante l'analisi dello standard. Il coefficiente angolare (pendenza) $b_1$ e l'intercetta $b_0$ della retta si ricavano mediante il metodo dei minimi quadrati:

$$b_1 = \frac{\sum_1^n (x_i \cdot y_i) - \frac{(\sum_1^n x_i) \cdot (\sum_1^n y_i)}{n}}{\sum_1^n x_i^2 - \frac{(\sum_1^n x_i)^2}{n}} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$

Utilizzando un foglio di Excel basta inserire i dati in una tabella e quindi utilizzare le funzioni predefinite [pendenza] e [intercetta] per determinare l'equazione della retta.
Procedendo manualmente conviene calcolare prima tutte le sommatorie e solo alla fine utilizzare le equazioni. I risultati sono o seguenti:
- $b_1 = 1{,}208$ pendenza
- $b_0 = -0{,}007$ intercetta

Pertanto l'equazione della retta di regressione risulta: $S = 1{,}208 \cdot C - 0{,}007$ Per costruire la retta si individuano due punti calcolando il valore di S per ogni valore di C e quindi si traccia la retta congiungendo i due punti. Utilizzando Excel basta inserire il grafico relativo alla tabella dei dati iniziali.
Di seguito è riportata la retta di lavoro costruita con Excel:

| C | S |
|---|---|
| 0 | -0.007 |
| 0.10 | 0.124 |
| 0.20 | 0.236 |
| 0.30 | 0.330 |
| 0.40 | 0.480 |
| 0.50 | 0.610 |
| 0.60 | 0.715 |

| | |
|---|---|
| pendenza | 1.208 |
| intercetta | -0.007 |
| r | 0.998 |


Il calcolo del coefficiente di correlazione, effettuato sulla serie di valori con la funzione [correlazione] di Excel ha fornito il seguente risultato: r = 0,998 a conferma del fatto, anche visivo, che i dati sperimentali sono ottimamente correlati linearmente, cioè seguono quasi perfettamente la retta di regressione. Potrebbe voler dire (ma non vi è mai la sicurezza!) che i risultati ottenuti nell'analisi di campioni con l'analita X, a cui la retta è riferita, saranno accurati e precisi e quindi significativi.