UNIVERSITÀ DEGLI STUDI DI TERAMO

CL in BIOTECNOLOGIE

Anno Accademico 2022/2023

CHIMICA ANALITICA

Statistica applicata ai metodi analitici

- E' impossibile effettuare una analisi chimica con risultati privi di incertezza, è quindi necessario determinare il grado di incertezza associato alla misura per ogni campione analizzato.

- Passaggio assolutamente necessario è stabilire quale è il massimo errore tollerabile nella misura!

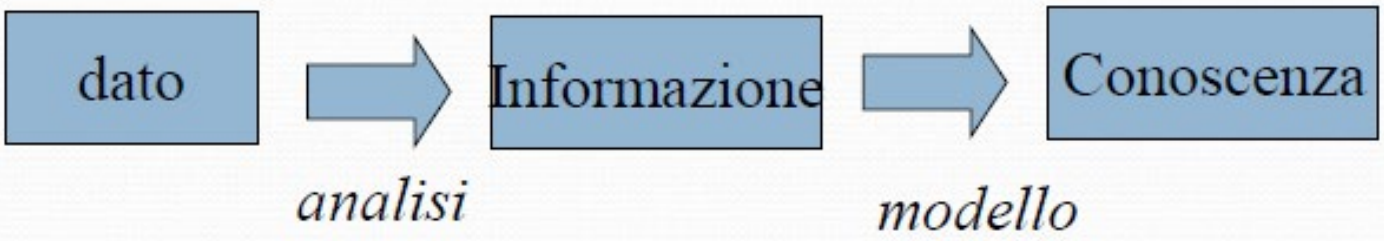
- Come si esprime una misura?

- I dati sono delle informazioni elementari che descrivono aspetti particolari di un fenomeno

dati di un individuo: Altezza, peso, colore pelle, concentrazione composti chimici nel sangue, composizione DNA, taglia abiti e calzature,…

- I dati possono essere qualitativi o quantitativi
- Di per se un dato non ha significato. E' necessaria una forma di analisi che correli il dato con qualche aspetto "significativo" del campione stesso in modo da aumentare la "conoscenza"

p.es. per dare senso alla composizione chimica del sangue è necessario un modello del corpo umano e delle azioni delle patologie.

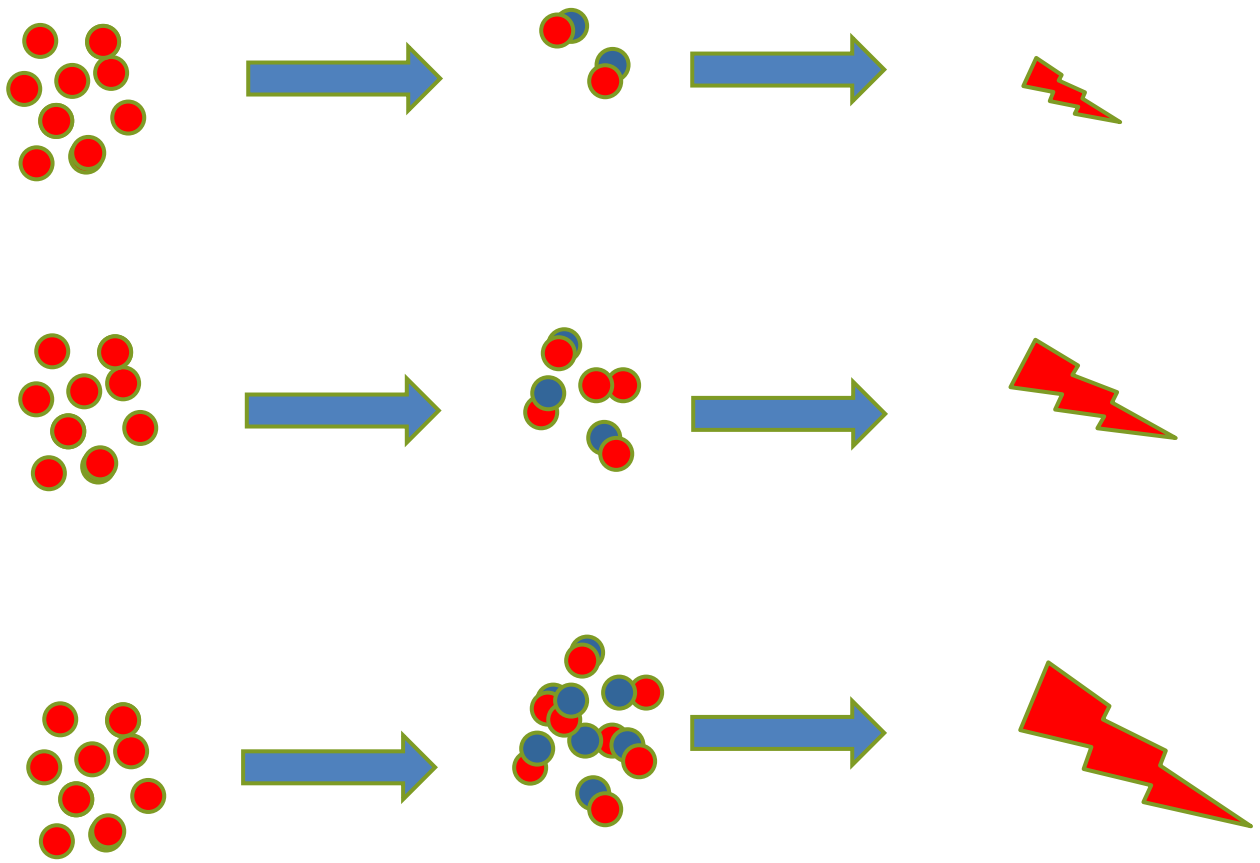


Tipologie di dati

- Quantitativi
  - Valore numerico ed unità di misura
    La temperatura dell'acqua è 400.0 K
  - I dati quantitativi sono la base della scienza galileiana e delle cosiddette "*hard sciences*": le discipline basate su dati rigorosi connessi tra loro da modelli matematici.

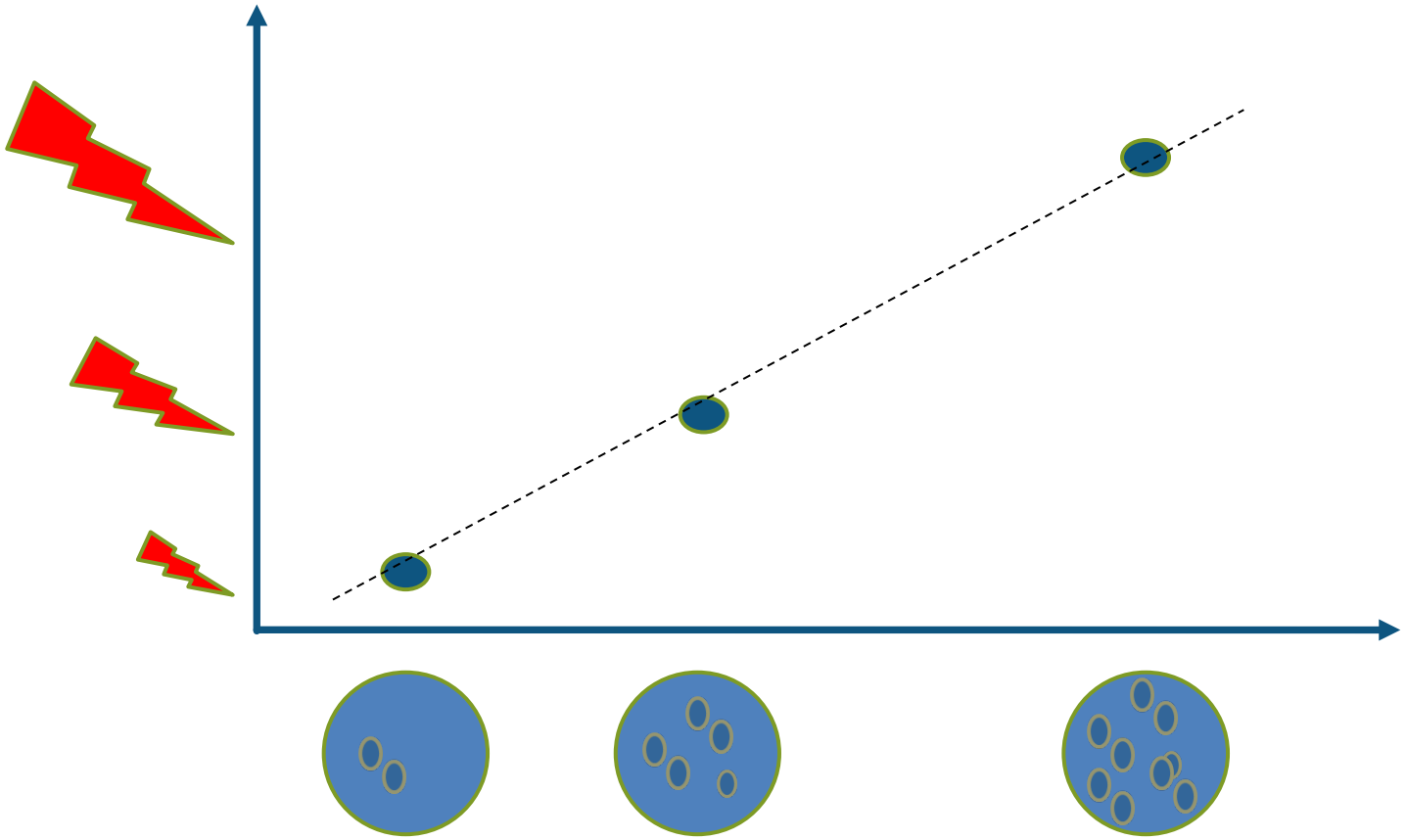
- Qualitativi
  - Etichette, descrittori, categorie
  - Generalmente sono espressi verbalmente
    "l' acqua è *calda*"
  - Dati standardizzabili e riproducibili con difficoltà (es. analisi sensoriale)

Criterio fondamentale dell'analisi di dati



Condizione necessaria è l'esistenza di una relazione (matematica) stechiometrica tra analita e «sorgente»

Dal segnale analitico alla curva di calibrazione



Le misure implicano SEMPRE errori e incertezze

l termine errore ha due significati leggermente differenti. Nel primo, «errore» rappresenta la differenza fra il valore misurato e il valore “vero” o valore “noto”. Nel secondo «errore» indica l'incertezza stimata in una misura o in un esperimento.

Errore rappresenta la differenza tra il valore «vero» e valore trovato

2. Errore indica l’incertezza stimata in una misura

E’ impossibile effettuare un’analisi priva di errori o incertezze.

Ciò che si può fare è: MINIMIZZARE GLI ERRORI E STIMARNE L’ENTITA’ CON UNA ACCURATEZZA ACCETTABILE

Allo scopo di migliorare l'affidabilità del metodo analitico e di ottenere informazioni circa la variabilità dei risultati, la procedura analitica viene solitamente ripetuta su diverse porzioni di un dato campione (replicati).

I risultati individuali di un insieme di misure sono raramente gli stessi, di solito il «valore centrale» viene usato come la “migliore” stima dei dati.
Perché effettuare misure replicate?
Il valore centrale di un insieme dovrebbe essere più affidabile di ciascun risultato individuale (media e mediana vengono usate come valore centrale di una serie di misure replicate)
La variazione dei dati dovrebbe fornire una misura dell'incertezza associata con il valore centrale.

Gli errori nelle analisi chimiche: MEDIA E MEDIANA

In chimica analitica si ricorre a misure replicate dello stesso campione per ottenere una serie di dati replicati attraverso i quali è possibile valutare l'incertezza legata alla misura e calcolare un valore medio.

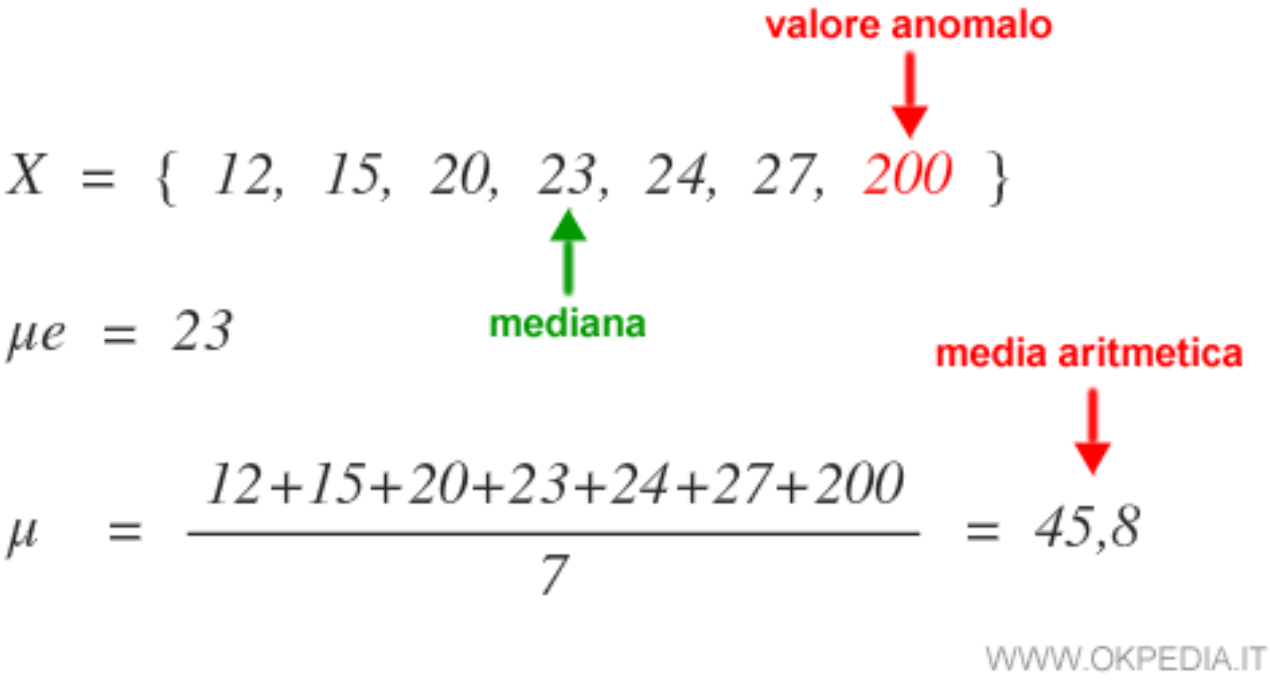
In genere si usa la media

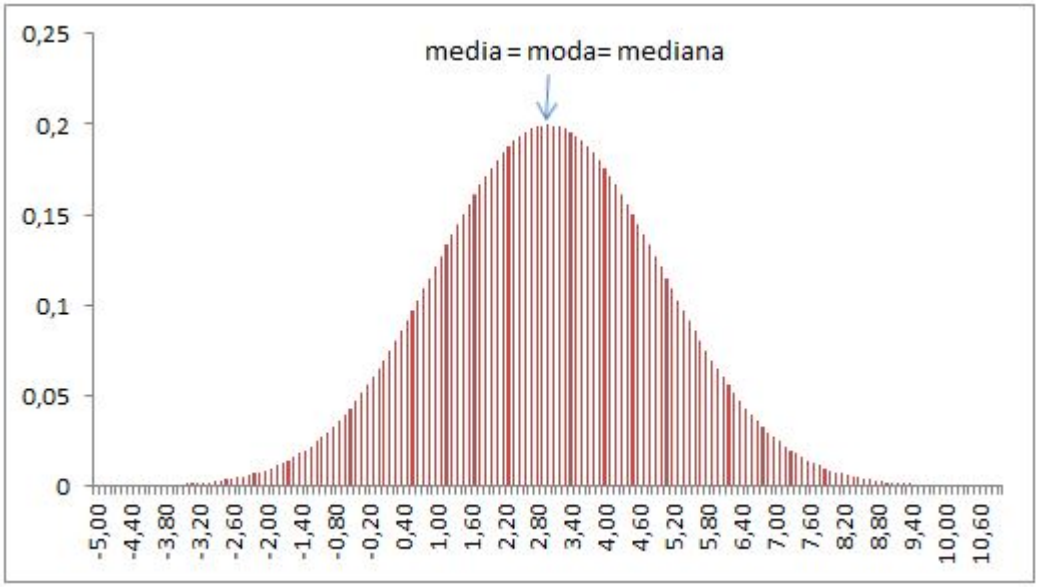
La media (media aritmetica) si ottiene dividendo la somma delle misure replicate per il numero delle stesse nell'insieme:

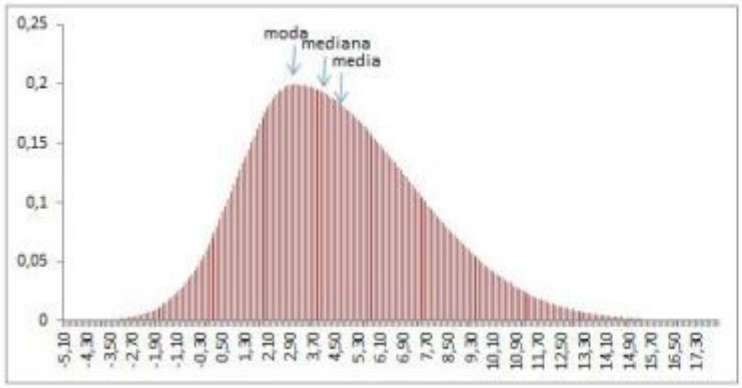
$$m = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

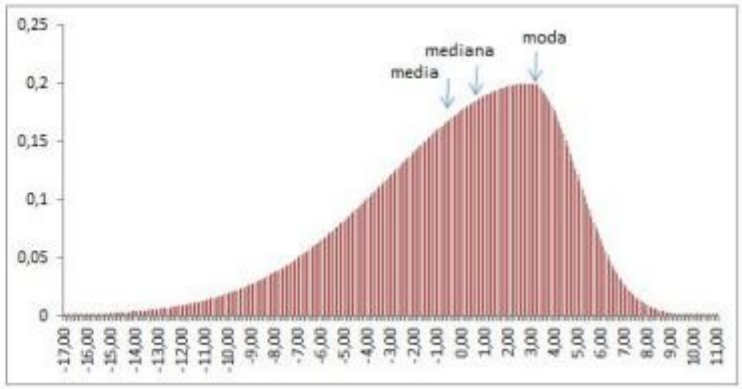
Oppure può essere utile calcolare la mediana:
La mediana è il risultato centrale quando i dati replicati sono ordinati in modo crescente o decrescente.

Per un numero dispari di dati, la mediana può essere ricavata disponendo in ordine i dati ed individuando il valore centrale, mentre nel caso di un numero pari viene usata la media della coppia centrale




media = moda= mediana
0,25
0,2
0,15
0,1
0,05
0
-5,00
-4,40
-3,80
-3,20
-2,60
-2,00
-1,40
-0,80
-0,20
0,40
1,00
1,60
2,20
2,80
3,40
4,00
4,60
5,20
5,80
6,40
7,00
7,60
8,20
8,80
9,40
10,00
10,60


moda
mediana
media
0,25
0,2
0,15
0,1
0,05
0

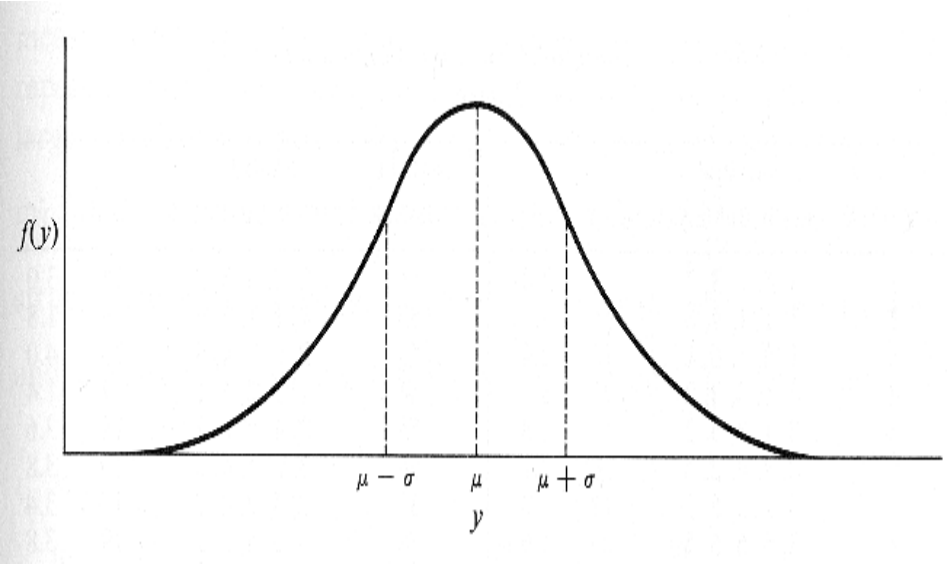

media
mediana
moda
0,25
0,2
0,15
0,1
0,05
0

La Funzione Gaussiana

- *Una **funzione gaussiana** è una funzione della seguente forma:*

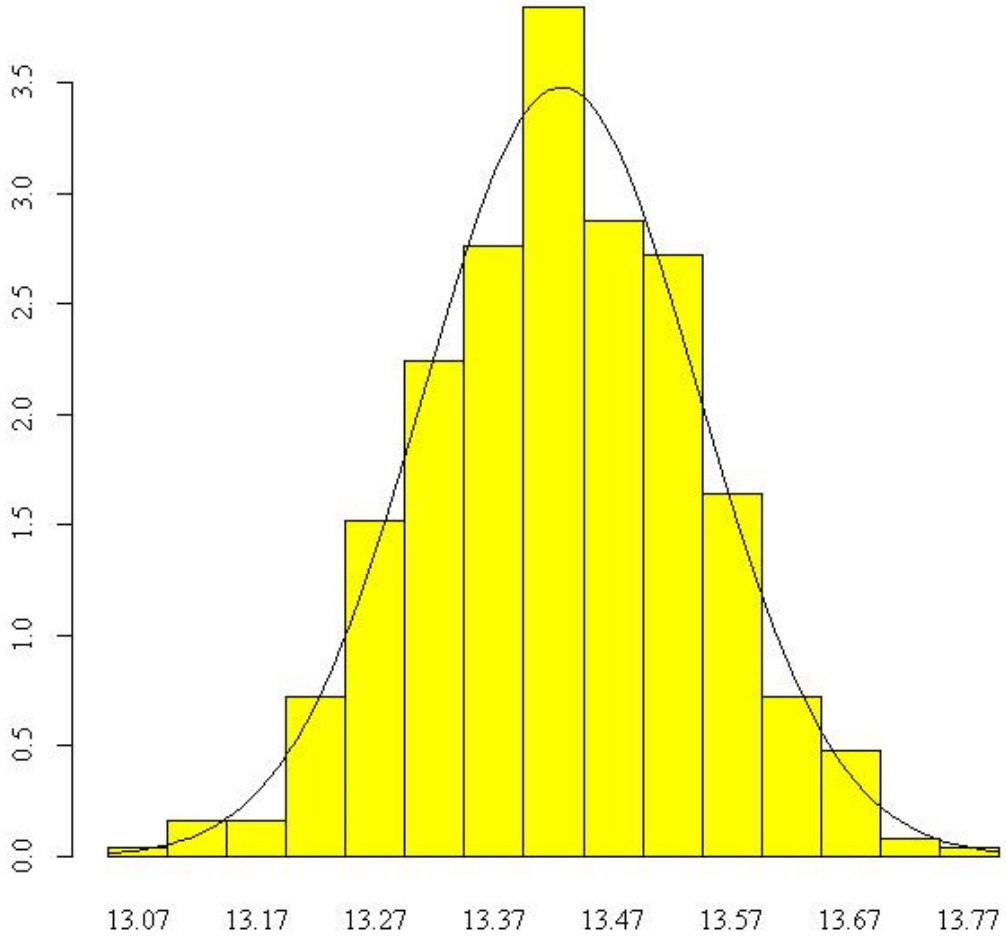
$$f(x) = ae^{-(x-b)^2/c^2}$$

$$f(y) = \frac{e^{\frac{-(y-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\,\sigma}, \qquad (-\infty < y < \infty),$$



dove μ e σ rappresentano la popolazione media e lo scarto quadratico medio (o deviazione standard). L'equazione della funzione di densità è costruita in modo tale che l'area sottesa alla curva rappresenti la probabilità. Perciò, l'area totale è uguale a 1.

La distribuzione normale



*La **deviazione standard***

*La **deviazione standard** o **scarto quadratico medio** è una misura della variabilità di una variabile casuale ed ha la stessa unità di misura dei valori osservati.*

In pratica misura la dispersione dei dati intorno al valore atteso.

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$$

dove $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *è la media aritmetica.*

*La **deviazione standard***

- Se si conosce solo un campione della popolazione, si sostituisce il fattore 1 / *n* con 1 / (*n* − 1), ottenendo come nuova definizione:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

A partire dalla deviazione standard si definisce anche la **deviazione standard relativa** come il rapporto tra σx e la media aritmetica dei valori:

$$\mathrm{RSD} = \sigma_r = \frac{\sigma_x}{\overline{x}}$$

Fonti di incertezza

- Incertezza di ripetibilità espressa come precisione del metodo (Contributo di Categoria A)
- Incertezza di taratura strumentale (Contr. Cat B)
- Incertezza di pesata (Contr. Cat B)
- Incertezza del volume finale (Contr. Cat B)
- Incertezza di preparazione standard interno (Contr. Cat B)
- Incertezza della soluzione di taratura (Contr. Cat B)

Regressione e correlazione

Esistono molti metodi di inferenza statistica che si riferiscono ad una sola variabile statistica.

Obiettivo della lezione: studio della relazione tra due variabili.

Tecniche oggetto di studio:

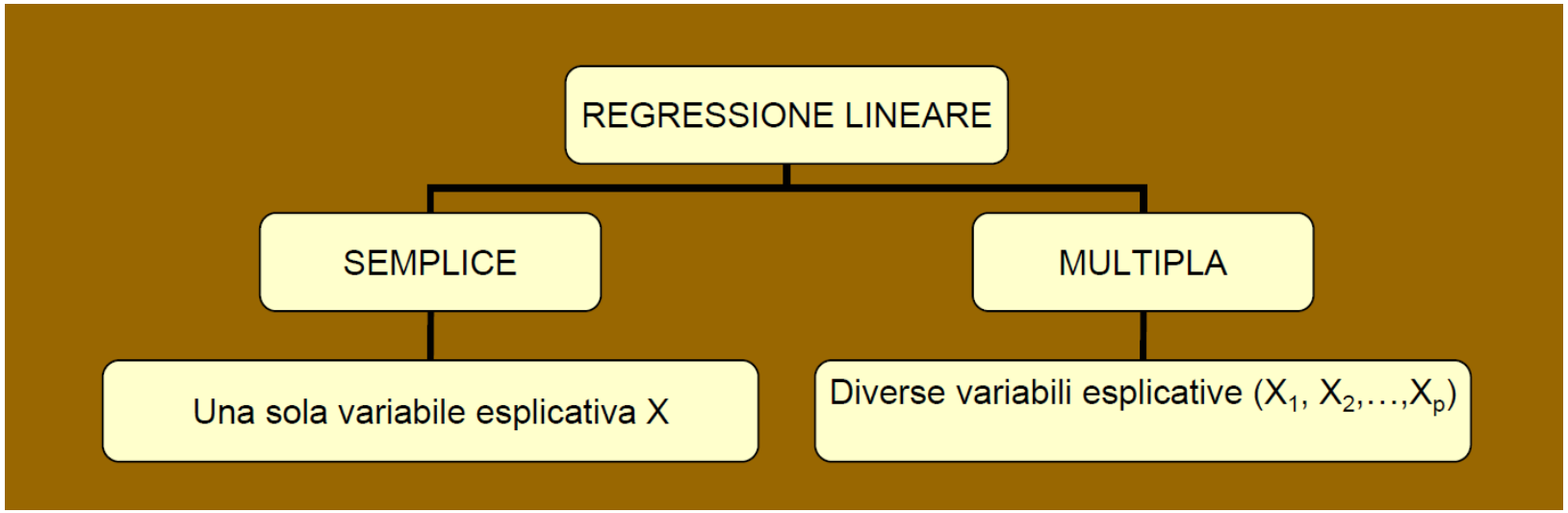
Regressione → Costruire un modello attraverso cui prevedere i valori di una variabile dipendente o risposta (quantitativa) a partire dai valori di una o più variabili indipendenti o esplicative

Correlazione → Studio della associazione tra variabili quantitative

Regressione lineare

Solitamente nel modello di regressione si indica con

- Y la variabile dipendente
- X la variabile esplicativa



Precisione e Accuratezza

I termini *precisione* e *accuratezza*sono messi in relazione con gli errori casuali e sistematici.

- Una misura è tanto più *precisa* quanto più i singoli valori misurati in condizioni di *ripetitibilità* si concentrano intorno alla media della serie di misure effettuate.
  - La variabilità dei risultati viene quantificata nella deviazione standard.
  - Si preferisce quantificare la precisione con il coefficiente di variazione, in genere espresso in percentuale.

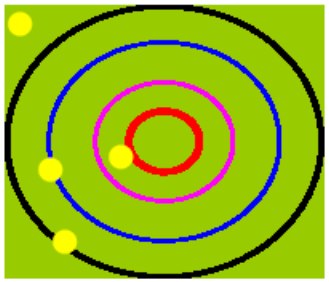
- L'*accuratezza* esprime invece l'assenza di errori sistematici nella misura:
  - una misura è tanto più accurata quanto più la media delle misure si approssima al valore vero della grandezza.
  - Anche l'accuratezza è spesso espressa come rapporto fra l'errore sistematico e il valore della grandezza.

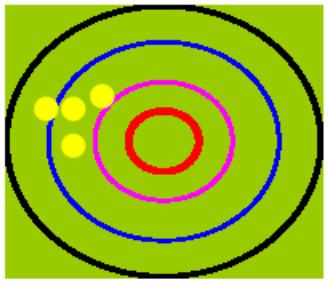
Precisione: bontà dell'accordo tra i risultati di misurazioni successive.

Esattezza*: bontà dell'accordo tra il risultato, x_i, o il valore medio dei risultati di un'analisi, ed il valore vero o supposto tale, x_t.

Gli errori possono essere **errori casuali** o **errori sistematici**. Quelli casuali influenzano la precisione, quelli sistematici l'esattezza. Gli errori casuali influenzano la precisione, quelli sistematici l'esattezza.

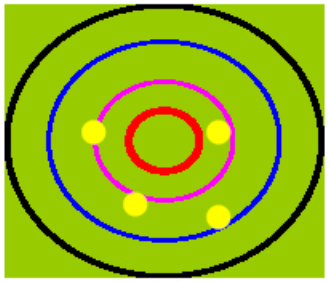
Né esatto né preciso

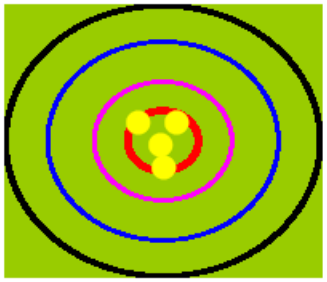




Non esatto ma preciso

Esatto ma non preciso





Esatto e preciso

influiscono sulla

Errori che portano ad una sovrastima (o sottostima) del valore vero:

errori sistematici

accuratezza della misura:
capacità di dare una risposta vicina al valore vero.

influiscono sulla

Errori che portano a stime in parte superiori in parte inferiori al valore vero:

errori casuali

precisione della misura:
ripetibilità
(nello stesso esperimento)
riproducibilità
(in esperimenti diversi)

Gli errori sistematici e casuali possono verificarsi indipendentemente, ed essere associati a diversi stadi dell'esperimento.

Gli errori sistematici dipendono da cause che agiscono secondo leggi definite.

- Errori metodologici
  (effetti di umidità e temperatura sulla pesata, effetti di svuotamento dello strumento di analisi volumetrica, errori di indicatore,…..)
- Errori legati all'accuratezza dello strumento (tolleranza ammessa nella calibrazione degli strumenti)
- Errori umani sistematici (astigmatismo, daltonismo,…..)

ERRORI SISTEMATICI

Gli errori *sistematici*, o *bias*, sono errori che possono essere individuati e quindi devono essere corretti.

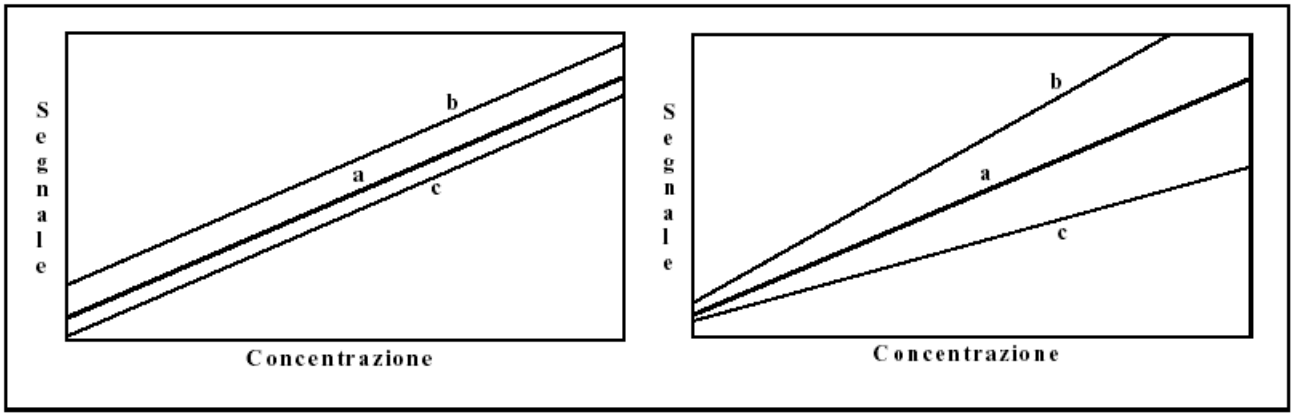
Gli errori *sistematici strumentali* sono dovuti a inesatta calibrazione o utilizzazione impropria della vetreria e degli strumenti di misura, all'uso di strumenti non idonei, ecc.

Gli errori *sistematici di metodo* sono dovuti a un comportamento non ideale di reattivi e reazioni, o all'uso di condizioni sperimentali non idonee (formazione di composti più o meno solubili del previsto, tempi di calcinazione inadeguati, ecc).

Gli errori *sistematici personali* sono dovuti a distrazione o ignoranza della corretta procedura (bolle d'aria nel beccuccio della buretta, errori di parallasse ecc.). Gli errori sistematici personali sono talvolta connessi a difetti fisici o a veri e propri pregiudizi inconsci (tendenza a terminare la titolazione dopo aver aggiunto un volume il più possibile confrontabile con quello ottenuto in titolazioni precedenti, oppure calcolato teoricamente, ecc.).

$$\text{bias} = \mu - x_t$$

Gli errori *sistematici* possono essere *costanti* o *proporzionali*.



Effetto di errori sistematici costanti e proporzionali (positivi e negativi) sulla curva di calibrazione

Errori sistematici costanti: bias negativo dovuto a perdite per solubilità in gravimetria, bias positivo dovuto ad assorbimenti estranei in spettrofotometria, ecc.
Errori sistematici proporzionali: bias negativo dovuto a perdite di analita in seguito a estrazioni non efficienti, ecc.

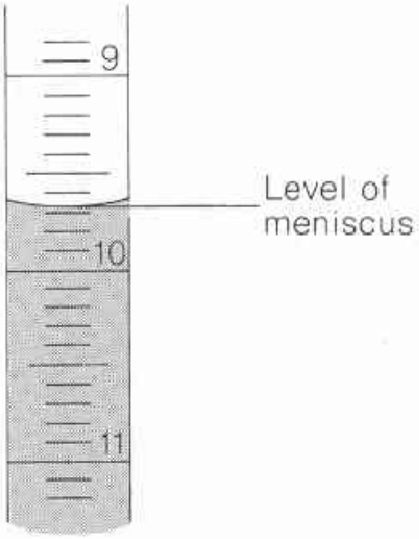
Gli errori sistematici possono essere identificati ed annullati mediante

- analisi di campioni standard, se disponibili;
- analisi del campione mediante un metodo indipendente, ovvero che prevede l'utilizzo di strumentazione di provata affidabilità o di riferimento;
- analisi del *bianco*, cioè di una soluzione contenente tutti i componenti presenti nel campione in esame eccetto l'analita di interesse; il bianco ideale è costituito dalla stessa matrice in cui è contenuto l'analita di interesse; l'analisi del bianco nelle titolazioni volumetriche consente, per esempio, di correggere l'errore connesso al volume di titolante necessario per far virare l'indicatore colorimetrico stesso;
- analisi di campioni contenenti un diverso ammontare della variabile misurata (per es. si pensi alla perdita connessa alla solubilità durante il lavaggio con volumi diversi di acque di lavaggio).

Gli errori casuali dipendono da numerose cause, legate al modo col quale viene effettuata la misura, che agiscono di volta in volta in modo diverso.

Es.: lettura di una misura da una scala graduata

buretta
con sensibilità = 0.1 ml



Lettura: 9.68 ml.

La stima a livello di 0.01 ml (1/10 della sensibilità dello strumento) e' soggetta ad errore.

Non possono venire eliminati. E' possibile minimizzarli, mediante un utilizzo accurato della tecnica di misura, misurarli e valutare il loro significato mediante analisi statistica su misure ripetute.

ERRORI CASUALI

Gli ***errori casuali*** (detti anche ***indeterminati*** o "random" in lingua inglese), causano una dispersione più o meno simmetrica dei dati intorno al valore medio.

Essi sono legati a fluttuazioni indefinite di una miriade di parametri sperimentali, quali temperatura, pH, pressione, umidità, punto d'arresto di una titolazione, forza ionica, ecc. oltre che alle tolleranze dei pesi delle bilance e della vetreria utilizzata per la misurazione di volumi e alle incertezze dei valori desunti dagli strumenti di misura.

Queste fluttuazioni avvengono anche cercando di lavorare con la massima cura.

Gli errori casuali non possono essere eliminati, anche se possono essere ridotti operando con cura.

Si definiscono:

campione = l'insieme delle misure in esame

popolazione = l'insieme di tutte le possibili misure

$\bar{x}$ = la media del campione

μ = la media della popolazione

s = la deviazione standard del campione

σ = la deviazione standard della popolazione

La distribuzione del campione è rappresentabile in modo discreto (istogramma).

La distribuzione della popolazione è rappresentabile con una curva continua (curva normale o Gaussiana).

Il modello di regressione

Per studiare la relazione tra due variabili è utile il diagramma di dispersione in cui si riportano i valori della variabile esplicativa X sull'asse delle ascisse e i valori della variabile dipendente Y sull'asse delle ordinate.

La relazione tra due variabili può essere espressa mediante funzioni matematiche più o meno complesse tramite un modello di regressione.

Il modello di regressione lineare semplice è adatto quando i valori delle variabili X e Y si distribuiscono lungo una retta nel diagramma di dispersione.

Il modello di regressione lineare semplice

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad (9.1)$$

dove

β_0 = l'intercetta per la popolazione

β_1 = l'inclinazione per la popolazione

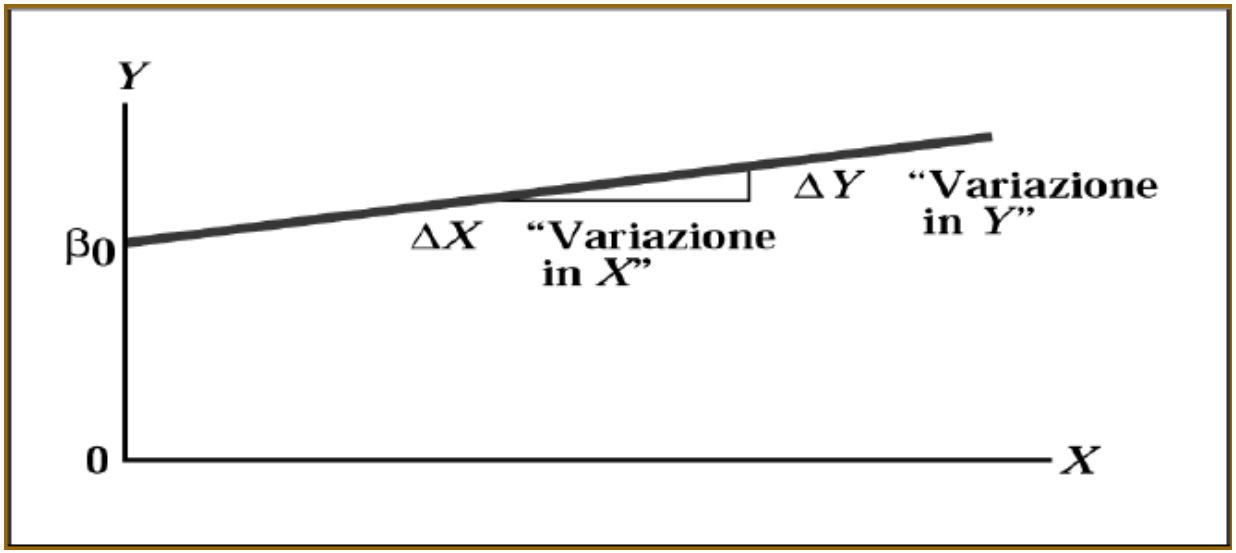
ϵ_i = l'errore casuale in Y corrispondente all'i–esima osservazione

Il modello di regressione

L'inclinazione β_1 indica come varia Y in corrispondenza di una variazione unitaria di X.

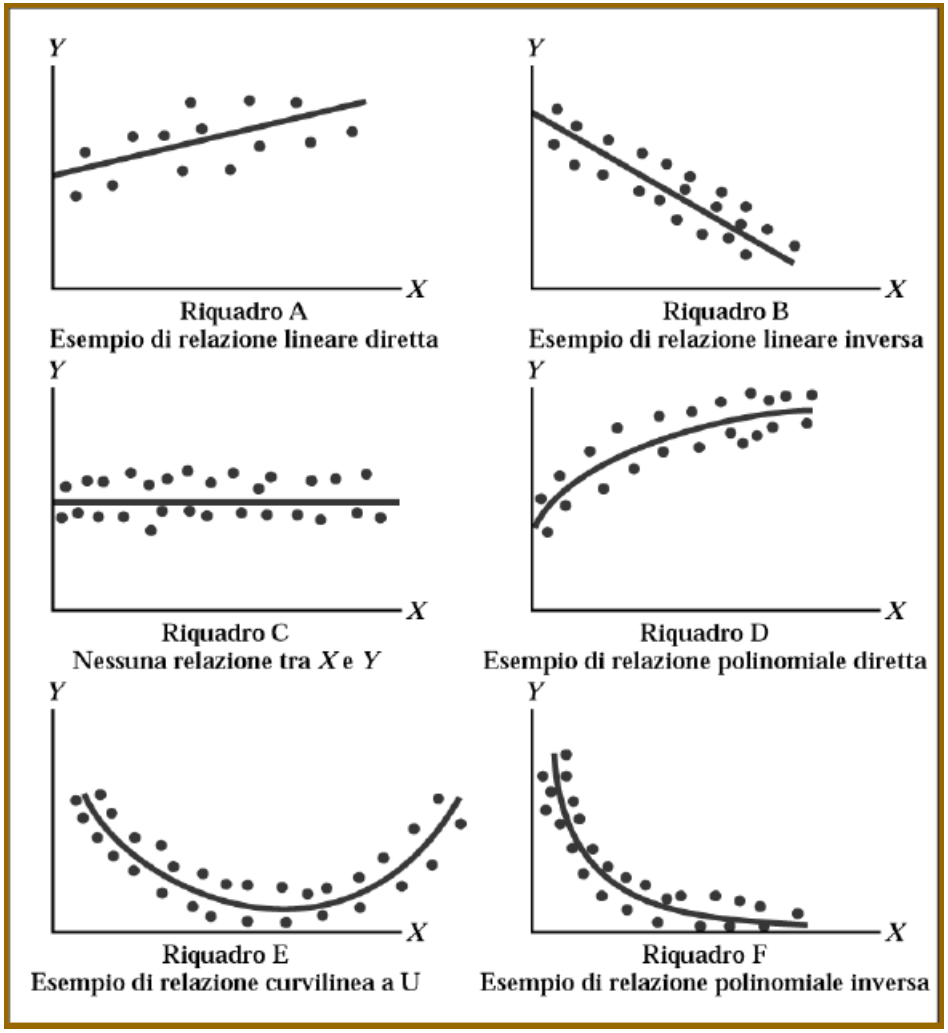
L'intercetta β_0 corrisponde al valore medio di Y quando X è uguale a 0.

Il segno di β_1 indica se la relazione lineare è positiva o negativa.



Il modello di regressione

La scelta del modello matematico appropriato è suggerita dal modo in cui si distribuiscono i valori delle due variabili nel diagramma di dispersione



Equazione della retta di regressione

Si dimostra che sotto certe ipotesi i parametri del modello β0 e β1 possono essere stimati ricorrendo ai dati del campione. Indichiamo con b0 e b1 le stime ottenute.

L'equazione campionaria del modello di regressione lineare

La previsione di Y in base al modello di regressione lineare è data dalla somma tra l'intercetta campionaria e il prodotto tra il valore di X e l'inclinazione campionaria

$$\hat{Y}_i = b_0 + b_1 X_i \qquad (9.2)$$

dove

$\hat{Y}_i$ = previsione di Y per l'osservazione i

X_i = valore di X per l'osservazione i

La regressione ha come obiettivo quello di individuare la retta che meglio si adatta ai dati.
Esistono vari modi per valutare la capacità di adattamento Il criterio più semplice è quello di valutare le differenze tra i valori osservati (Y_i) e i valori previsti ($\hat{Y}_i$)

Equazione della retta di regressione

La previsione di un valore di Y in corrispondenza di un certo valore di X può essere definita in due modi, in relazione all'intervallo di valori di X usati per stimare il modello:

- **interpolazione:** se la previsione di Y corrisponde ad un valore di X interno all'intervallo

- **estrapolazione:** se la previsione di Y corrisponde ad un valore di X che non cade nell'intervallo

Misure di variabilità

Le seguenti misure di variabilità consentono di valutare le capacità previsive del modello statistico proposto.

Variabilità totale
(somma totale dei quadrati) → variabilità di Y

Variabilità spiegata
(somma dei quadr. della regress.) → variabilità di $\hat{Y}$

Variabilità non spiegata
(somma dei quadr. degli errori) → variabilità dell'errore

Misure di variabilità

Le misure di variabilità nella regressione

Somma totale dei quadrati = somma dei quadrati della regressione + somma dei quadrati degli errori

$$SQT = SQR + SQE \tag{9.3}$$

La somma totale dei quadrati (*SQT*)

La somma totale dei quadrati (*SQT*) è data dalla somma dei quadrati delle differenze tra i valori osservati di Y e la loro media.

$$SQT = \text{variabilità totale} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{9.4}$$

La somma dei quadrati della regressione (*SQR*)

La somma dei quadrati della regressione (*SQR*) è data dalla somma dei quadrati delle differenze tra i valori previsti di Y e la media di Y.

$$SQR = \text{variabilità spiegata} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \tag{9.5}$$

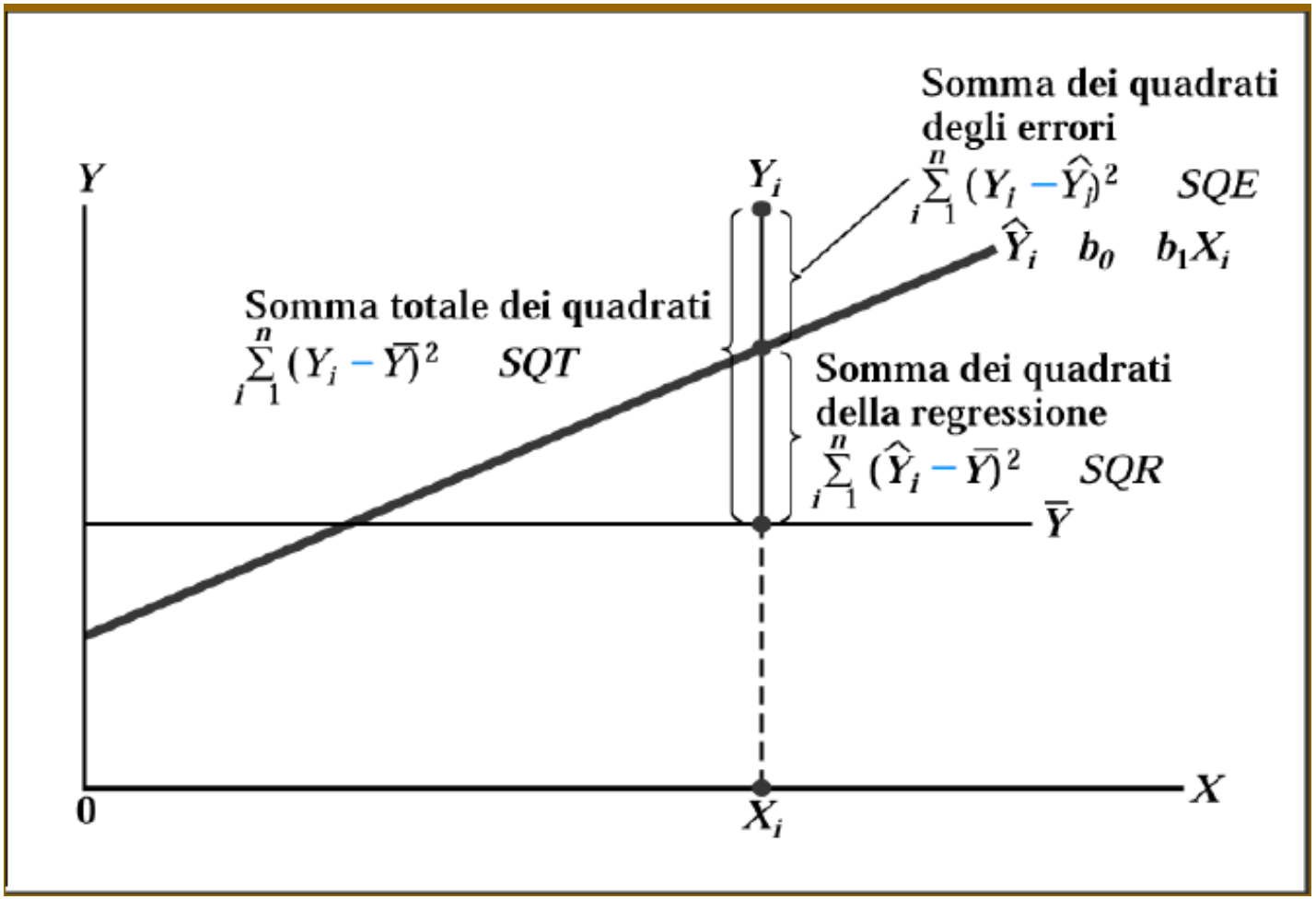
$$= SQT - SQE$$

La somma dei quadrati degli errori (*SQE*)

La somma dei quadrati degli errori (*SQE*) è data dalla somma dei quadrati delle differenze tra i valori osservati e i valori previsti di Y

$$SQE = \text{variabilità non spiegata} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{9.6}$$

Misure di variabilità



Misure di variabilità

Il coefficiente di determinazione è una misura utile per valutare il modello di regression. Esso misura la parte di variabilità di Y spiegata dalla variabile X nel modello di regressione.

Il coefficiente di determinazione

Il coefficiente di determinazione è dato dal rapporto tra la somma dei quadrati della regressione e la somma totale dei quadrati.

$$r^2 = \frac{SQR}{SQT} \qquad (9.7)$$

Analisi dei residui

Il residuo e_i è una stima dell'errore che commetto nel prevedere Yi tramite Ŷi.

Il residuo

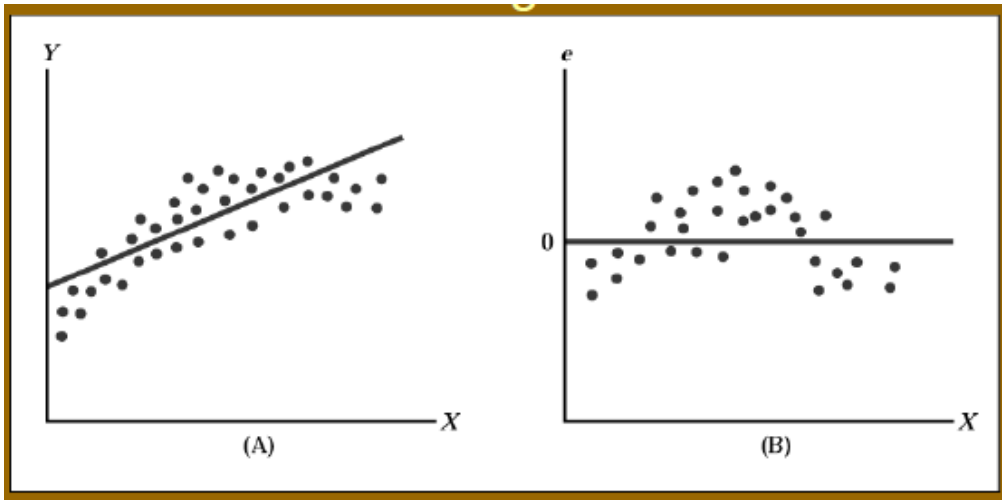
Il residuo è uguale alla differenza tra valore osservato e il valore previsto di Y:

$$e_i = Y_i - \hat{Y}_i \qquad (9.9)$$

Per stimare la capacità di adattamento ai dati della retta di regressione è opportuna una analisi grafica ⟹ grafico di dispersione dei residui (ordinate) e dei valori di X (ascisse).

Se si evidenzia una relazione particolare il modello non è adeguato.

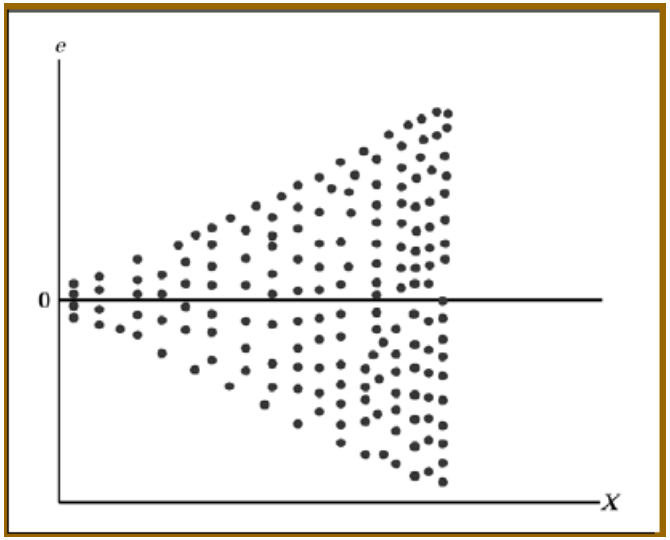
Nell'esempio a lato il modello di regressione lineare non sembra appropriato. Il grafico a destra evidenzia lo scarso adattamento ai dati del modello (lack of fit). Quindi il modello polinomiale è più appropriato.



Analisi dei residui

Valutazione delle ipotesi:

- Omoschedasticità: il grafico dei residui rispetto a X consente di stabilire anche se la variabilità degli errori varia a seconda dei valori di X



Il grafico in alto evidenzia ad esempio che la variabilità dei residui aumenta all'aumentare dei valori di X.