Facoltà: : BioScienze e Tecnologie Agro-Alimentari e Ambientali

Denominazione Corso di Laurea: Biotecnologie Avanzate (Laurea Magistrale)

**Corso: Statistica e bioinformatica per le biotecnologie**

MODULO:

Chemometria applicata (5 CFU, 40 ore)

Docente: Marcello Mascini

(mmascini@unite.it)

Il Docente e' disponibile per chiarimenti al termine della lezione o su richiesta via mail

**1UD Richiami di Statistica univariata (1 CFU = 8 ore).**

Dati, informazioni, modelli; Tipi di dati; Rappresentazione analitica dei dati; Calibrazione e regressione. Probabilità e densità di probabilità; Media e varianza; La distribuzione normale; Metodo dei minimi quadrati; Regressione polinomiale; Regressione non-lineare; Metodo del chi^2; La validazione del modello. Analisi della varianza (ANOVA).

# Introduzione: dati e strumenti

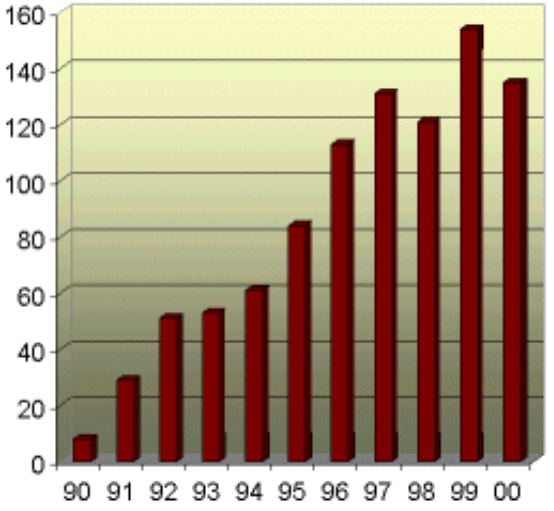
Dati univariati e multivariati

Gli errori di misura e l’analisi dei dati

Parametri caratteristici degli strumenti di misura

# Chemometria

- L'applicazione dei metodi matematici e statistici alla analisi dei dati di chimica analitica è detta chemometria
- Dalla fine degli anni '70 è iniziato la sperimentazione sui metodi multivariati
  - Pionieri:
    - B. Kowalski, S. Wold, P. Massart, P. Lindgren, Geladi
- L'obiettivo principale è il trattamento dei dati di strumenti multidimensionali come spettrometri e gas-cromatografi
- Il trattamento dei dati multivariati è stato sviluppato all'inizio per studiare fenomeni economici e psicometrici
  - Predizione degli andamenti macroeconomici da vari "indicatori"
  - La "misura" dell'intelligenza

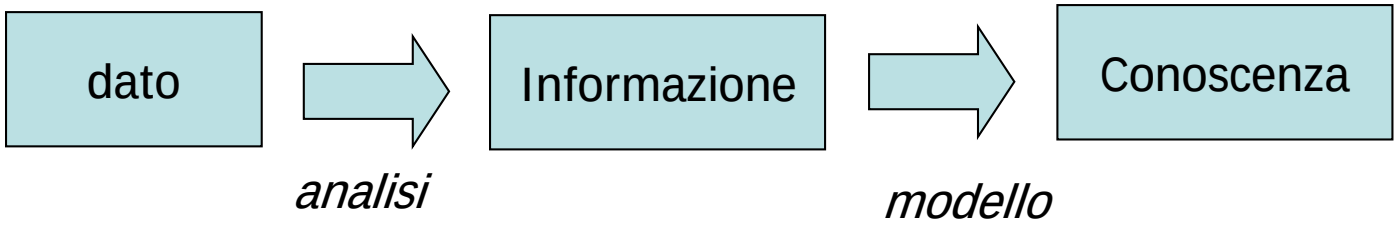



Riviste specializzate
- Chemometrics and Intelligent Laboratory systems
- Journal of Chemometrics

Articoli appaiono anche in:
- Analytical Chemistry
- Analytica Chimica Acta
- Trends in Analytical Chemistry
- J. computer aided molecular design
- ……

# I dati

- I dati sono delle informazioni elementari che descrivono aspetti particolari di un fenomeno.
  - Esempio:
    - 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"
  - Esempio: per dare senso alla composizione chimica del sangue è necessario un modello del corpo umano e delle azioni delle patologie.



# Tipologie di dati

- Quantitativi (hard)
  - 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 (soft)
  - Etichette, descrittori, categorie
  - Generalmente sono espressi verbalmente
    - "l' acqua è *calda*"
  - Dati difficilmente standardizzabili e riproducibili (es. analisi sensoriale)
    - *Fuzzy logics: tentativo di rendere quantitativi dei dati espressi verbalmente*
- Dati discreti:
  - Range limitato e valori pre-definiti
- Dati continui
  - Range limitato ma valori continui
    - I limiti strumentali possono dar luogo a discretizzazioni
      - Esempio conversione Analogico-Digitale

# Dati Univariati

- Molti procedimenti analitici producono dati univariati in cui cioè il dato sperimentale dipende da una sola variabile
  - Misura di una singola variabile incognita
  - Controllare le interferenze
  - Tenere costanti tutte le condizioni sperimentali tranne la variable target
  - Richiede una preparazione elaborata del campione per isolare solo la variabile da misurare

- Un dato univariato è espresso con uno scalare e una unità di misura.
  - Esempio:
    - La misura di una resistenza elettrica è 100KΩ
    - Il peso di una mela è 80g
    - La concentrazione di $K^+$ in un acqua è 1.02 mg/l

# Dati multivariati

- Alcuni strumenti producono una grande quantità di dati per campione singolo
- La maggior parte dei fenomeni sono intrinsecamente complessi e modellabili solo considerando una molteplicità di indicatori.
- *L'analisi multivariata è quindi necessaria per lo studio dei fenomeni reali e per l'interpretazione completa dei dati sperimentali*

# Dati multivariati

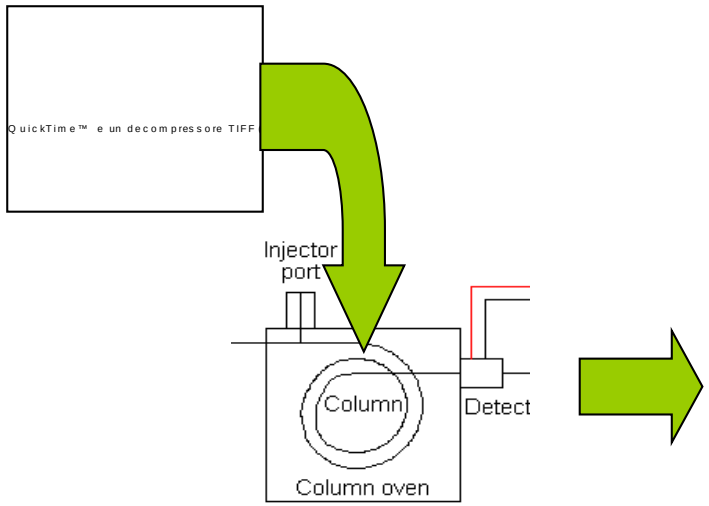
- Si ha un dato multivariato quando l'applicazione di una misura ad un campione produce una sequenza ordinata di grandezze scalari
  - L'ordine è relativo al significato fisico della misura stessa
- Oppure quando un fenomeno è descritto da un insieme di descrittori o attributi

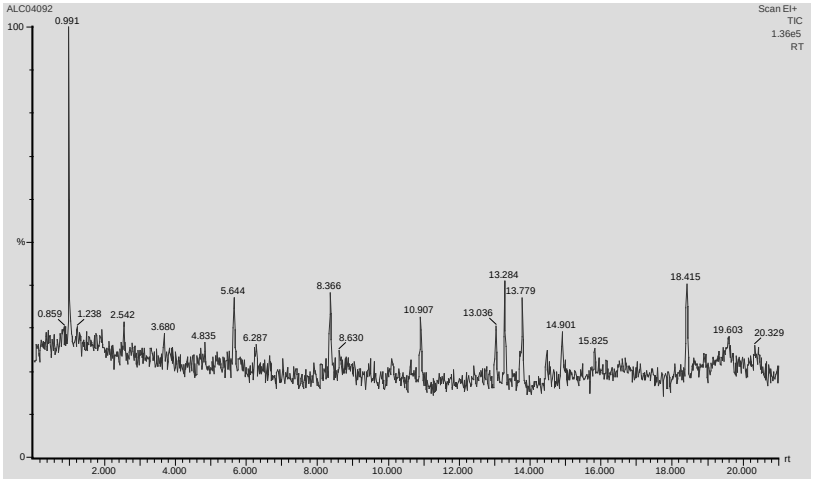
- Sorgenti di dati multivariati:
  - Strumenti o tecniche di misura che intrinsecamente forniscono dati multivariati
  - Lo studio di "fenomeni" o "campioni" complessi" richiede la collezione di più misure in un dato multivariato

- Matematicamente, una sequenza ordinata di numeri è un vettore. Ad ogni misura multivariata corrisponde quindi un vettore in un opportuno spazio vettoriale.

# Strumenti Multidimensionali 1

- **Gas cromatografia**

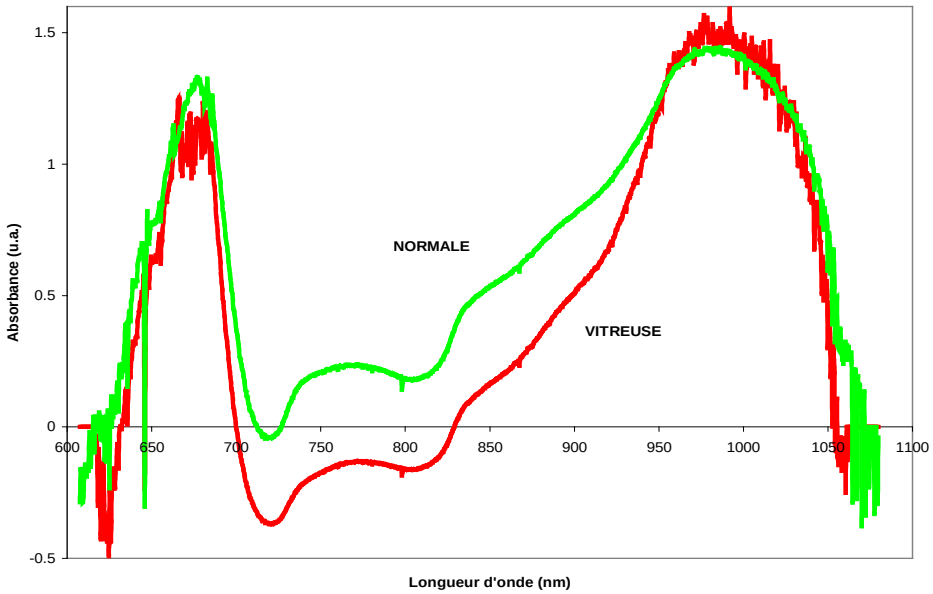


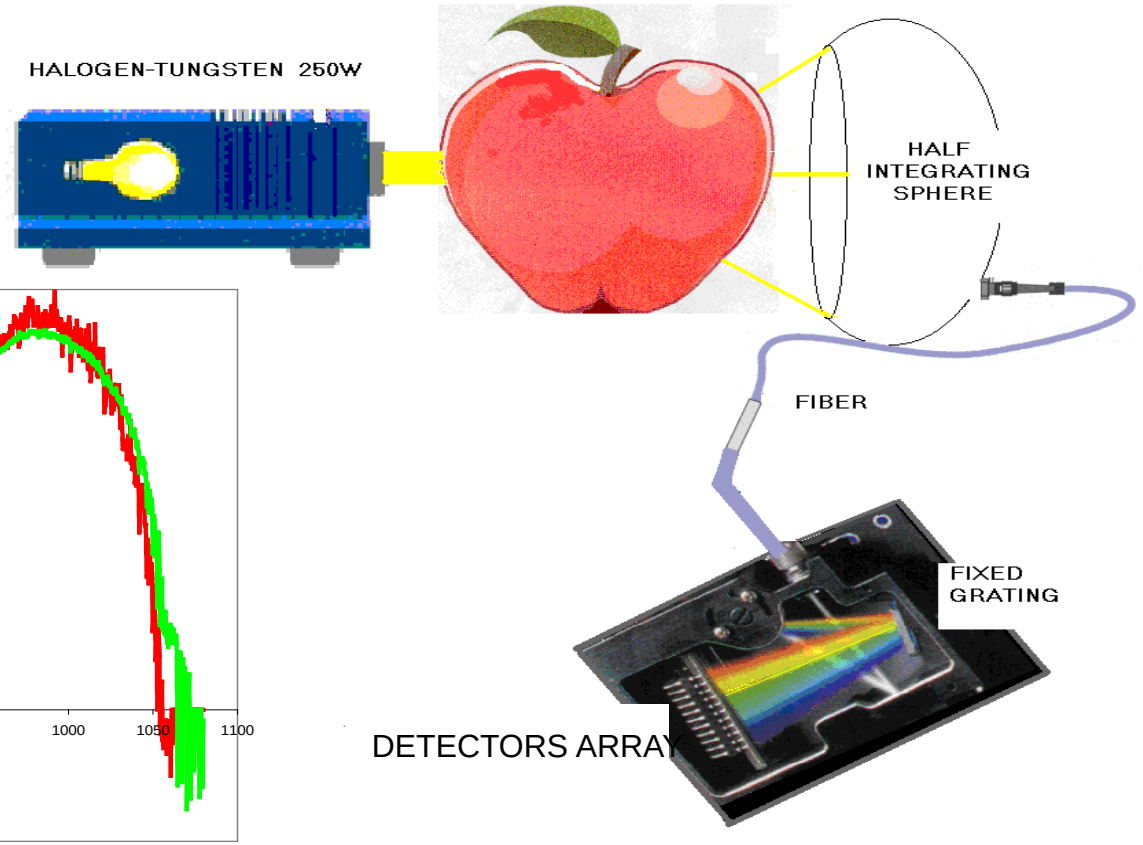




- Per ogni campione si ottiene uno spettro (intensità del segnale vs. tempo di eluizione)

# Strumenti Multidimensionali 2

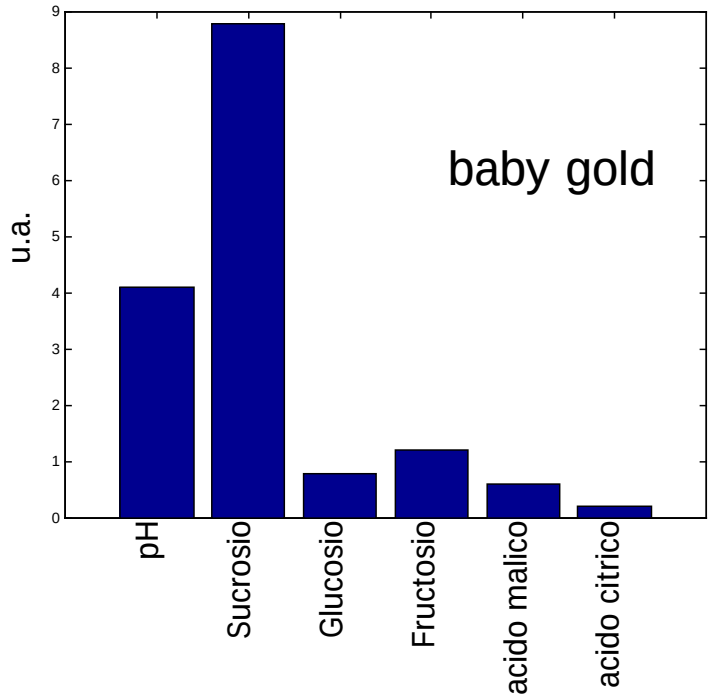
- **Spettroscopia**
  - Esempio: spettroscopia Vis/NIR di un frutto.

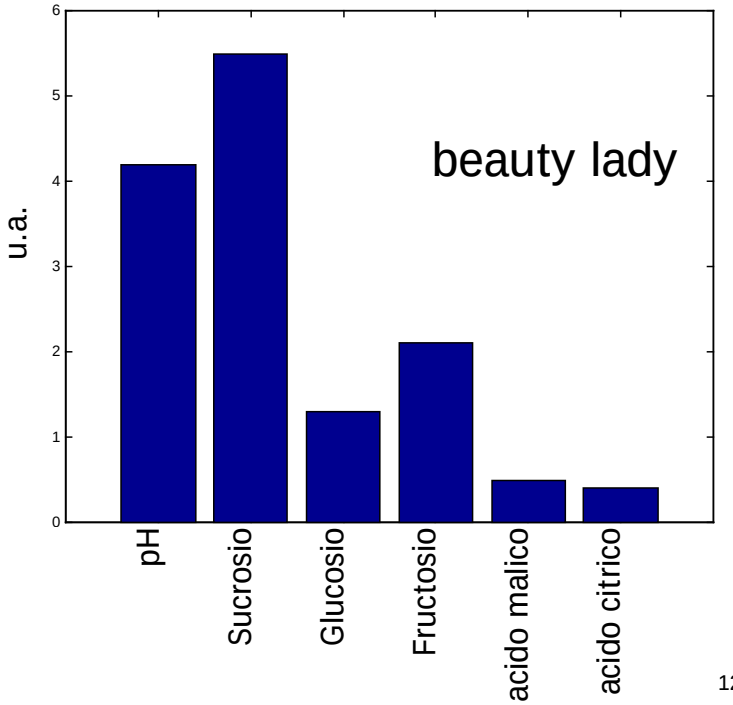


# Strumenti Multidimensionali 3

- **Array di Sensori**
  - Esempio: set di biosensori per la misura di pH, sucrosio, glucosio, fructosio, acido malico e acido citrico nelle pesche



# Sistemi Multidimensionali

- **Descrizione di fenomeni complessi**
  - Qualità alimenti (es. frutta)
    - Zuccheri, acidi, pH, etilene libero,...
  - Acque Minerali
    - pH, $CO_2$, Cl, Na, K, Mg,...
  - Coondizioni meteorologiche
    - T, RH, velocità del vento, pressione atmosferica,...
  - Condizione di un veicolo
    - velocità, accelerazione, livello carburante, olio freni, olio motore,...
- Ciascun indicatore proviene in genere da una misura indipendente
- L'insieme degli indicatori consente di attribuire il campione a classi generali (insiemi) di cui ogni campione è elemento
  - Il profilo degli indicatori si chiama **Pattern**
  - L'operazione che assegna il pattern all'insieme si chiama **Pattern Recognition**

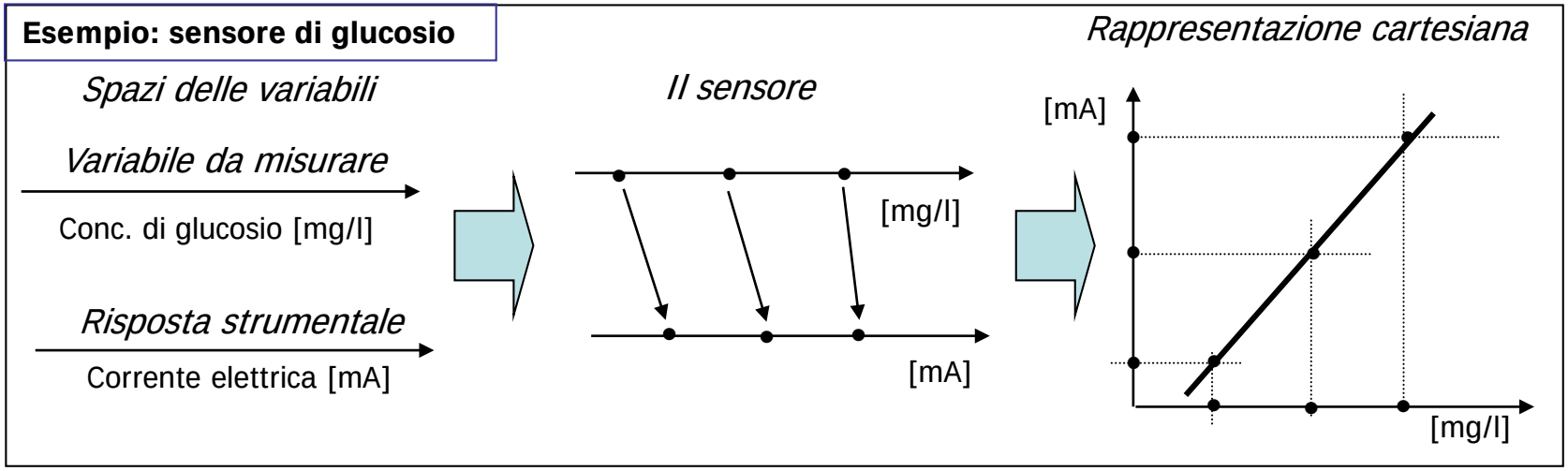
# Esempio:
# Acque Minerali

Pattern

Insieme

# Criterio fondamentale della analisi dati

- I dati sono rappresentati in spazi vettoriali euclidei
- Ad ogni osservabile viene fatta corrispondere una dimensione dello spazio ed è associato un vettore di base.
- Il sistema di riferimento dello spazio vettoriale è costituito da una base di vettori ortonormali pari al numero degli osservabili descritti.
- Questa assunzione è ovvia per i dati univariati:



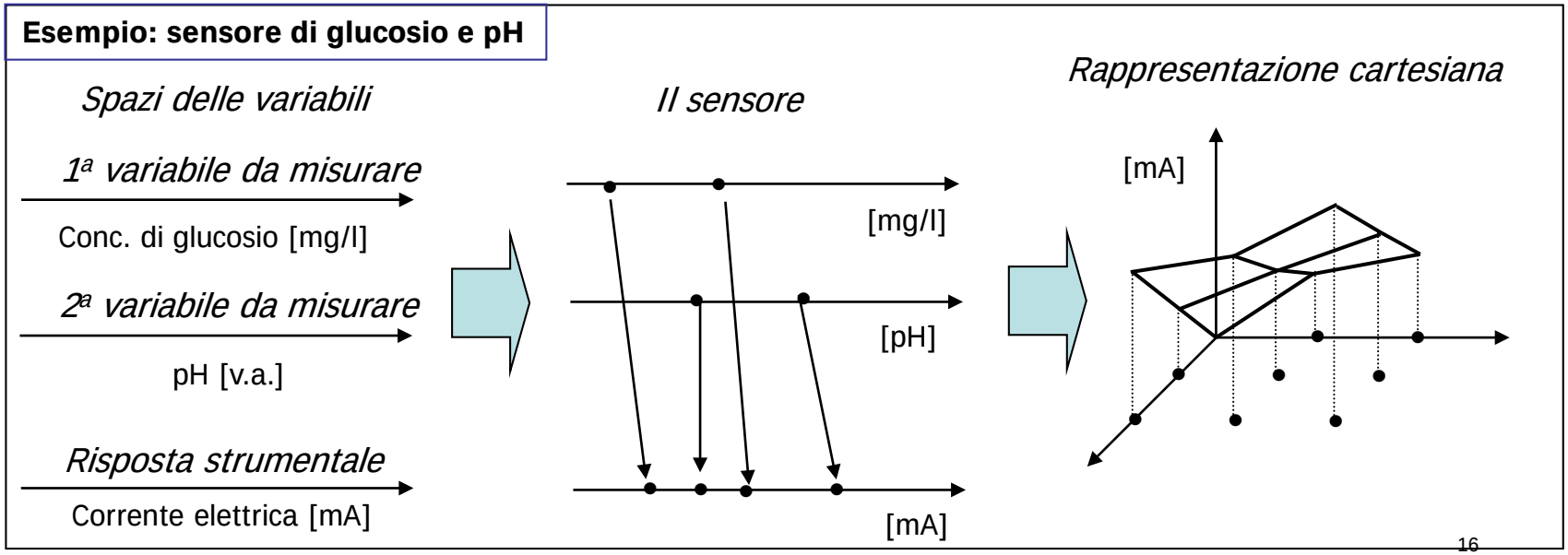

- Per dati multivariati vedremo che il nucleo della analisi dei dati consiste nella definizione di una nuova base ottenuta come combinazione lineare degli osservabili nella quale viene massimizzata “l’informazione” dei dati stessi.

# Analisi Multivariata

- L'analisi multivariata consente di mettere in relazione tra loro vettori

$$y = f(x) \rightarrow \vec{y} = f(\vec{x})$$

- In particolare, poiché trattiamo di dati sperimentali, abbiamo a che fare con grandezze "stocastiche", quindi per trattare dati multivariati c'è bisogno di estendere la statistica univariata al caso multivariato.



# Il Problema Generale dell'Analisi Dati

- Come estrarre da una misura strumentale informazioni sul campione misurato.

- Nell'analisi univariata lo strumento fornisce un "output" per un unico "input"

$$y = k \cdot x$$

*y: risposta dello strumento*
*x: sollecitazione del campione*
*k: caratteristica dello strumento*

- Lo scopo dello strumento è dato y come posso ricavare x (valore incognito)? Attraverso la conoscenza di k; Come conosco k? Attraverso la calibrazione.
- Calibrare lo strumento vuol dire esporlo a sollecitazioni (x) note, per cui misurando l'output y posso ricavare il valore di k e quindi rendere lo strumento utilizzabile.

# Calibrazione

- Ogni sensore è descritto da una funzione caratteristica che mette in relazione la grandezza d'uscita (segnale V) con la grandezza alla quale il sensore è sensibile (misurando x)

$$V = f(x)$$

- Nei casi più semplici, *f* è lineare
  - Es. strain gauge:

$$V = k \cdot \varepsilon$$

    - V: segnale; $\varepsilon$: sollecitazione (strain); k: parametro funzionale del sensore

- Il sensore è utilizzabile, cioè dal segnale si può stimare il misurando, solo quando sono noti sia la funzione caratteristica che I parametri funzionali.
- La stima dei parametri funzionali può essere ottenuta **solo** calibrando il sensore, cioè attraverso una serie di misure sperimentali ed applicando una **regressione statistica**.

# Concetti fondamentali della teoria della misura

- **Errori di misura: ripetendo più volte la “stessa misura” si ottengono risultati diversi.**
- Per “stessa misura” si intende l’esposizione del sensore allo stesso misurando, le condizioni ambientali non sotto controllo possono variare, e per effetto della cross-selettività influenzare la risposta del sensore (o della catena di trasduzione).
- Fluttuazioni della risposta: la **media aritmetica** è la quantità che meglio rappresenta la misura. Più grande è il numero di misure ripetute più affidabile e significativa è la rappresentazione del valor vero della media aritmetica.

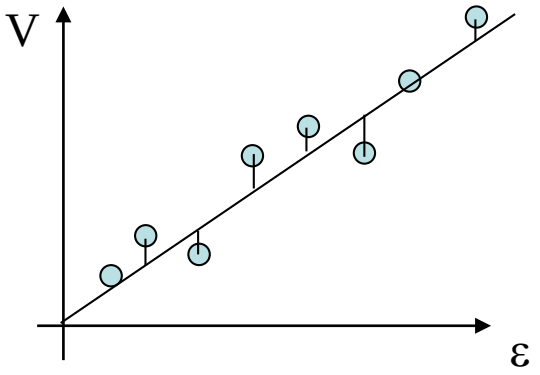
# Stima dei parametri funzionali

- La forma funzionale della caratteristica deve essere imposta a-priori
- **Le deviazioni tra forma funzionale e dati sperimentali vengono interpretate come errori di misura.**
- Per calibrare il sensore lo si deve sottoporre a sollecitazioni note. Quindi si deve essere in grado di generare valori noti del misurando con grande precisione (standards)
- Da questa precisione dipende la bontà della calibrazione e quindi la bontà delle misure che potrò eseguire con il sensore.

# Stima dei parametri funzionali esempio: strain gauge

Misurando: strain ($\Delta L/L$)
Output: tensione elettrica
Forma funzionale: lineare
Parametro funzionale: k (fattore di gauge, sensibilità)

$$V = k \cdot \varepsilon$$




calibrazione

V: noto;
ε: nota
k: ignoto

$$k = \frac{V}{\varepsilon}$$

misura

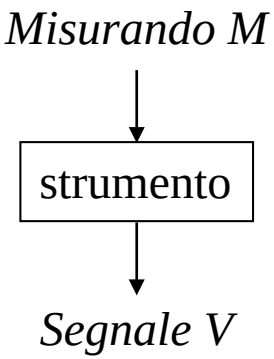
V: noto;
ε: ignota
k: nota

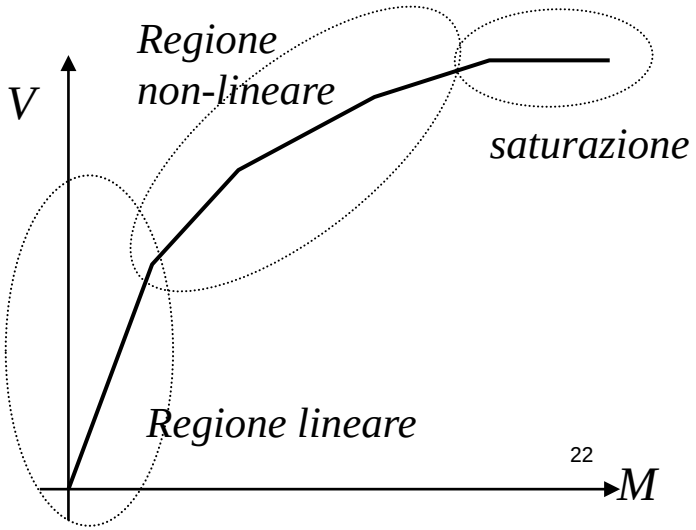
$$\varepsilon = \frac{V}{k}$$

Esistono gli errori di misura: Quindi non è possibile applicare le formule a lato ma serve un a teoria statistica (regressione) che minimizzi l'errore nella stima di k.

# I parametri caratteristici degli strumenti: 1. La curva di risposta

- Formalmente uno strumento descrive un mapping dallo spazio del misurando allo spazio del segnale d'uscita.
- Se questi spazi hanno dimensione 1, il sensore è rappresentabile attraverso una funzione *V=f(M).*
- Questa funzione è detta risposta I/O o caratteristica del sensore e rappresenta il parametro fondamentale per caratterizzare un sensore.
- *La conoscenza della curva di risposta permette di usare il sensore come strumento di misura: dalla misura di V si evince una stima del misurando M*
- La curva di risposta si ottiene attraverso un processo di calibrazione.

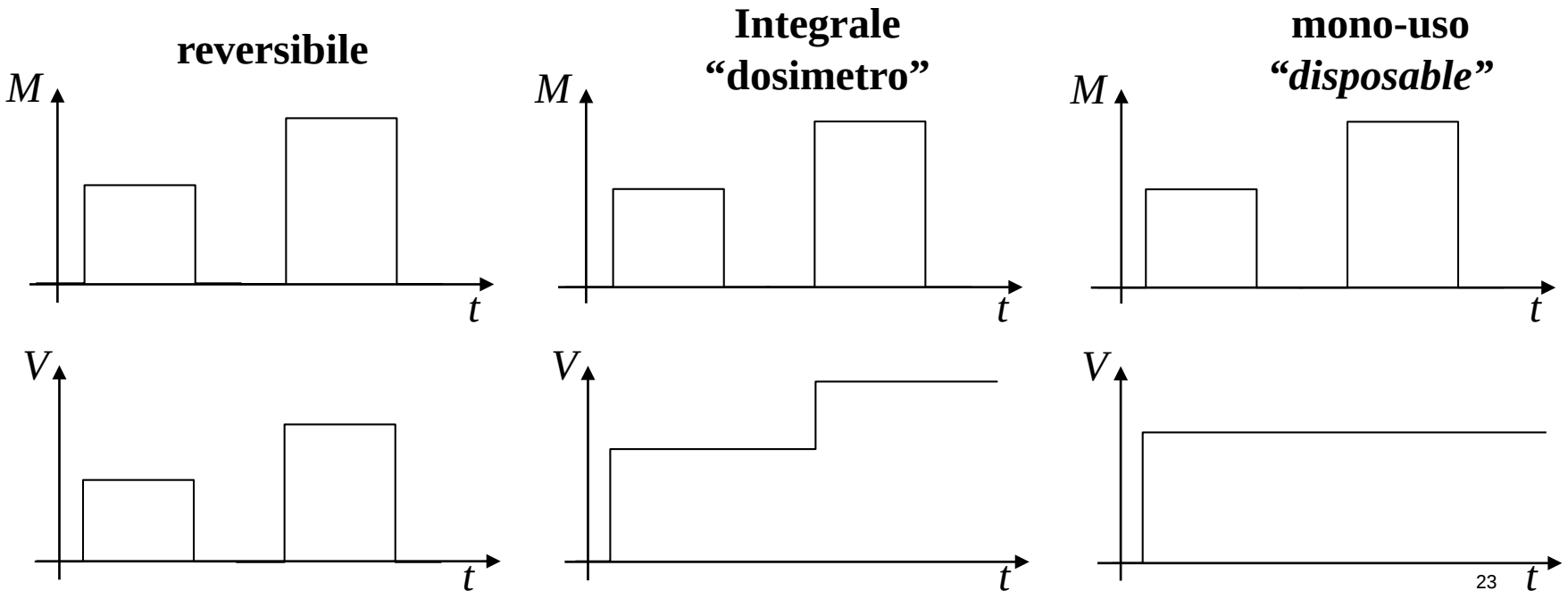


# I parametri caratteristici degli strumenti: 2. Reversibilità

- La reversibilità esprime la capacità dello strumento di misura di seguire, con una dinamica tipica dello strumento stesso, le variazioni del misurando.
- In particolare, uno strumento è reversibile se al cessare della sollecitazione del misurando la risposta si annulla.

---

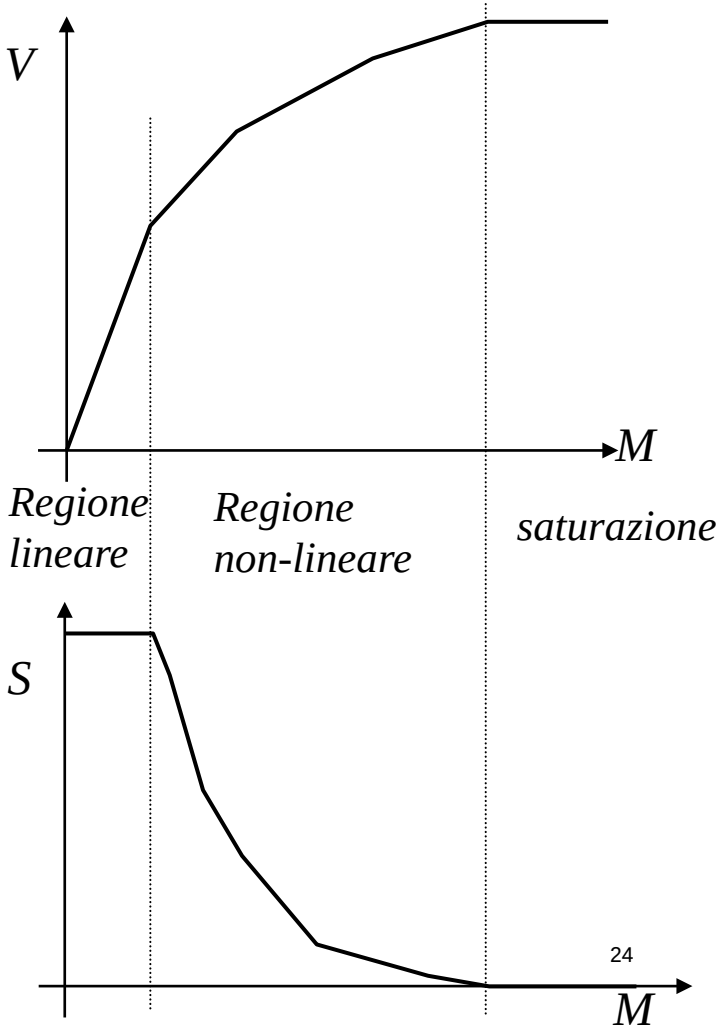


# I parametri caratteristici degli strumenti: 3. Sensibilità

- La sensibilità è definita come il rapporto tra la variazione del segnale e la variazione del misurando.
- Definisce la capacità dello strumento di misura di seguire le variazioni del misurando
- Matematicamente, si esprime come la derivata della curva di risposta dello strumento

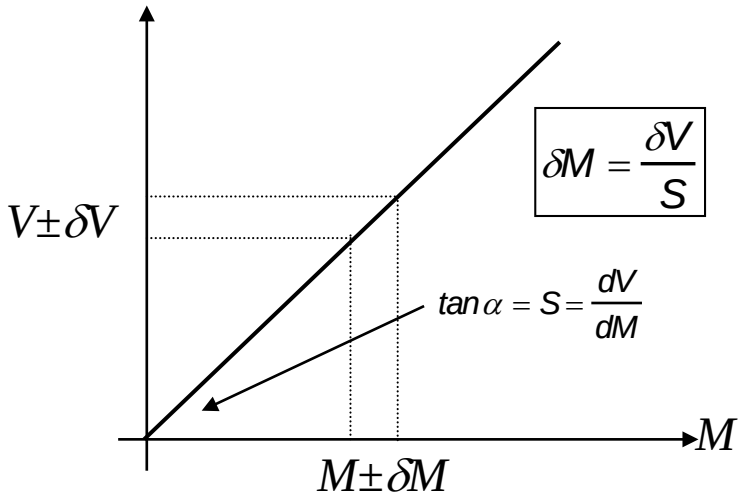
$$S = \frac{dV}{dM}$$

- Nella regione di non linearità, S è funzione del misurando.
- Nella regione di linearità S è massima, perciò sono massime le prestazioni dello strumento



# I parametri caratteristici degli strumenti: 5. Risoluzione

- La risoluzione è legata all'esistenza degli errori di misura e del rumore.
- Per questo motivo, il segnale del sensore non è una grandezza deterministica ma ha una componente aleatoria: V±δV. Dove δV esprime tutti gli errori di misura
- δV è limitato inferiormente dal rumore elettronico del segnale V.
- La risoluzione esprime come l'incertezza δV si traduce in una incertezza δM sulla misura del misurando.
- Nella regione lineare:

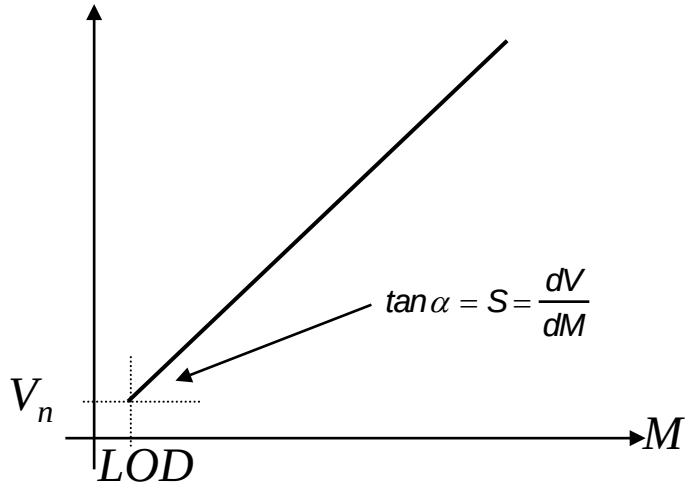



$$resolution = \lim_{V_{out} \to V_{noise}} \frac{V_{out}}{S} = \frac{V_{noise}}{S}$$

- La risoluzione dipende dalla sensibilità.
- In strumenti con sensibilità più alta gli errori di misura influiscono di meno sulla stima del misurando.
- La definizione vale anche per strumenti non lineari, se nell'intervallo δV la curva è assimilabile ad una retta.

# I parametri caratteristici degli strumenti: 6. Limite di rivelazione

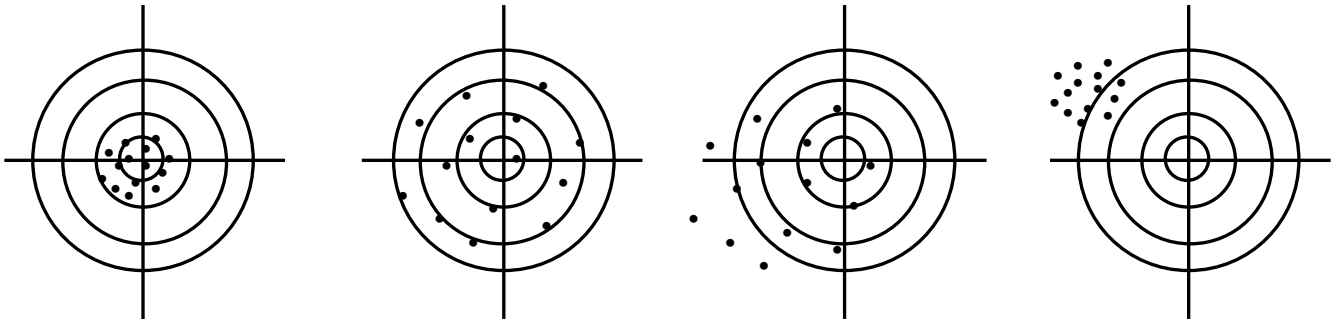
- La risoluzione calcolata per un segnale uguale a 0, definisce il *limit of detection* (LOD) dello strumento.
- La definizione traduce il fatto che non può esservi misura inferiore al suo errore. Quando l'errore di misura raggiunge il suo limite inferiore, il rumore elettronico $V_n$, si ha il limite di rivelazione teorico.
- Si definisce un $LOD_{convenzionale}$= (3 o 9)*LOD.




$$LOD = \frac{V_n}{S}$$

# I parametri caratteristici degli strumenti: 7. Accuratezza e Riproducibilità

- Accuratezza: capacità di un sistema di misura di fornire un valore del misurando uguale al valor vero (ignoto)
- Riproducibilità: capacità di uno strumento di fornire lo stesso segnale a parità di condizioni ambientali.
- Sono grandezze statistiche: date N misure, il valor medio è relativo alla accuratezza e la varianza alla riproducibilità.



# Analisi Statistica dei Dati

Frequenza e Probabilità

Distribuzione di probabilità

Media e varianza

Regressione statistica

Test del $\chi^2$

Correlazione lineare

Analisi della varianza

Media e varianza multivariata

Gaussiana Multivariata

# Analisi dati: deterministica o statistica

- Il dato in quanto risultato di una misura è un'entità stocastica
  - Questo è un concetto fondamentale delle scienze sperimentali
  - Ripetendo N volte "la stessa misura" si ottengono N valori differenti
  - Se non si ottengono valori differenti lo strumento di misura ha una risoluzione non adeguata.
- Esempio:
  - Se si misura 5 volte la lunghezza di un asse con diversi strumenti si ottiene:

| *Strumento* | *Risoluzione* | *Misure [cm]* |
|---|---|---|
| Metro da sarta | 1 cm | 120, 120, 120, 120, 120 |
| Metro da falegname | 1 mm | 119.8, 119.9, 120.1, 120.0, 120.2 |
| Calibro | 0.1 mm | 119.84, 120.31, 119.83, 120.10, 120. 34 |
| Micrometro | 0.01 mm | 119.712, 120.032, 119.879, 120.320, 119.982 |
| Interferometro laser | 0.5 µm | 119.9389, 120.0032, 119.8643, 119.3492, 120.2010 |

All'aumentare della risoluzione strumentale il dato diventa da deterministico stocastico

# Analisi Statistica

- La statistica è quella parte della matematica che mette in relazione grandezze rappresentate da distribuzioni di probabilità.
- Esempio:
  - La misura dell'asse precedente non ha un valore vero assoluto, ma appartiene ad una distribuzione che fissa la probabilità che eseguita una misura il valore sia compreso in un certo intervallo.
- Il concetto di probabilità è il concetto fondamentale della statistica. Essa è definita come:

$$Probabilità = \frac{casi\ \ favorevoli}{casi\ \ possibili}$$

  - Esempio: la probabilità tirando un dado "non truccato" di avere un valore è 1 (caso favorevole)/6 (casi possibili).

# Probabilità e Frequenza

- Ogni grandezza misurabile è caratterizzata da una distribuzione di probabilità i cui parametri sono stimabili attraverso misure sperimentali
- All'aumentare delle misure sperimentali i valori stimati approssimano i valori reali.
- Lo stimatore, calcolato cioè a partire dai dati sperimentali, della probabilità è la frequenza.

$$Frequenza = \frac{N_{favorevoli}}{N_{totali}}$$

$$Probabilità = \lim_{N_{totali} \to \infty} Frequenza$$

# Distribuzioni di probabilità (PDF)

- Data una grandezza la distribuzione di probabilità, PDF(x), è una funzione del valore della grandezza estessa a tutti i valori possibili (da -∞ a +∞).
- Per ogni valore essa definisce la probabilità di osservare il valore stesso.
- Poiché ogni strumento di misura ha una incertezza intrinseca derivante dal metodo di misura il risultato di una misura è espresso come intervallo x±Δx. Per cui la probabilità della misura è la somma delle probabilità di tutti i valori compresi tra x-Δx e x+Δx:

$$p = \int_{x-\Delta x}^{x+\Delta x} PDF\left(x\right) \cdot dx$$

- Ovviamente, poiché la misura esiste si ha le seguente regola di normalizzazione della PDF

$$\int_{-\infty}^{+\infty} PDF\left(x\right) \cdot dx = 1$$

# Media e Varianza

- La media e la varianza sono due grandezze fondamentali per descrivere una variabile statistica indipendentemente dalla forma funzionale della PDF.
- La media m è quel valore particolare di x per il quale si ottiene:

$$\int_{-\infty}^{m} PDF(x) \cdot dx = \int_{m}^{+\infty} PDF(x) \cdot dx$$

- La varianza ($\sigma^2$) è una grandezza che definisce “l’ampiezza” della PDF, definisce cioè la dispersione possibile dei valori della misura di x.
  - In pratica, la probabilità di x è concentrata in un range di valori la cui ampiezza è definita dalla varianza.
  - La varianza è definita attraverso la PDF come:

$$\sigma^2 = \int_{-\infty}^{+\infty} (x - m)^2 \, f(x) dx$$

# Media Aritmetica come miglior stima del valore medio

- Data una sequenza di misure affette da errore si dimostra che la media aritmetica è una buona stima del valore aspettato della misura stessa, cioè al valore medio della PDF della variabile misurata.
- Al crescere del numero delle misure la media aritmetica converge verso il valore medio:

$$m = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} x_i$$

- La prova di questo teorema costituisce la cosiddetta "*diseguaglianza di Tchebychev*" la quale stabilisce che data una variabilie casuale x, avente una distribuzione di probabilità generica con valore aspettato m e varianza $\sigma^2$ finiti si ha:

$$P\left(\left|x - m\right| \geq K\sigma\right) \leq \frac{1}{K^2}$$

La ***disuguaglianza di Chebyshev*** (che prende nome da Pafnuty Chebyshev) individua un limite superiore per la probabilità che una variabile casuale sia più distante di un certo valore dalla sua media.

# Dispersione quadratica media come miglior stima della varianza

- Data una sequenza di misure di x, la miglior stima della varianza della PDF che genera le misure è data dalla dispersione quadratica media, ovvero dalla medie del quadrato della differenza tra le singole misure ed il valore medio.

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$

- Si noti che la media è effettuata dividendo per N-1. Questo perche il valor medio di x è stato calcolato dai dati stessi per cui il numero di dati indipendenti è diminuito di 1
  - N-1: gradi di libertà
- La radice della varianza è detta *deviazione standard* ed è spesso usata per esprimere l'incertezza di una misura sperimentale.

# Esempio:
# gradi Brix in Pesche

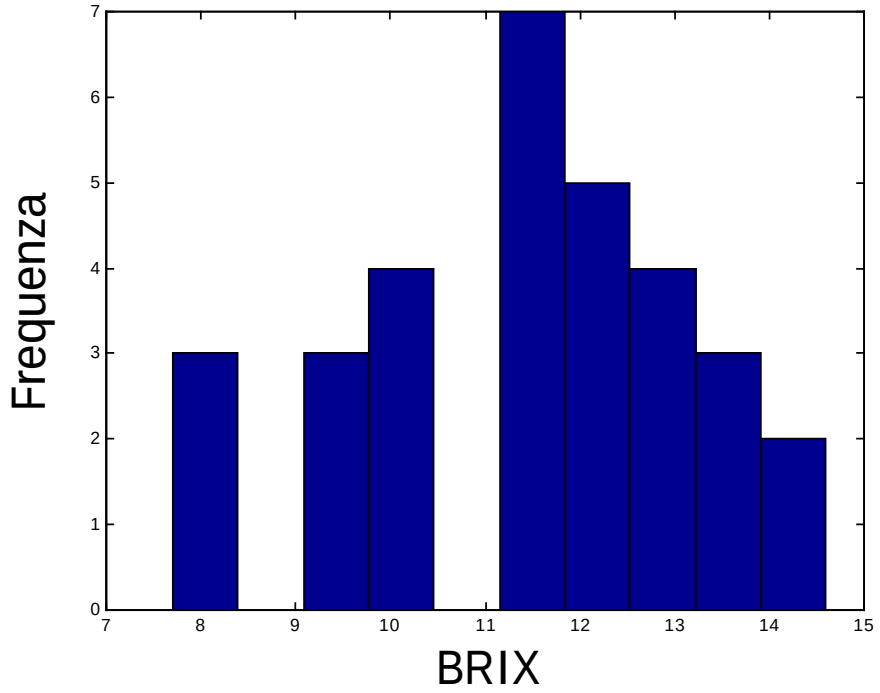
- Supponiamo di avere misurato il contenuto in gradi brix di una popolazione di 31 pesche.
- 9.80 8.00 10.20 9.10 11.30 12.80 11.30 11.20
  9.90 7.70 11.50 11.90 11.60 9.20 8.30
  11.90 10.40 9.50 11.20 13.20 14.60 13.70
  13.20 13.70 13.20 13.70 11.50 12.50 14.30
  12.50 12.10

$$media = \frac{1}{31}\sum_{i=1}^{31} x_i = 11.45$$

$$\sigma^2 = \frac{1}{30}\sum_{i=1}^{N}\left(x_i - 11.45\right)^2 = 3.51$$

# istogramma

- L’istogramma è una rappresentazione della frequenza con la quale in una esperimento si è verificato un valore compreso in un intervallo $x \pm \Delta x$
- Poichè la frequenza approssima la probabilità, l’istogramma è una approssimazione della PDF

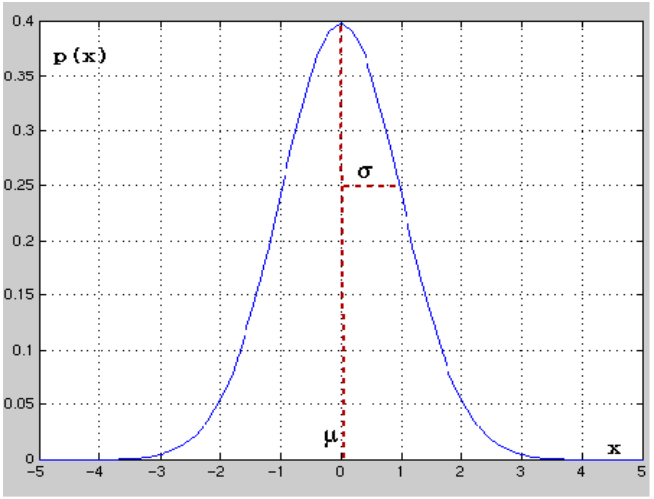


# Distribuzione Normale

a.k.a. *gaussiana*

- La distribuzione di Gauss è la PDF più importante soprattutto per quanto riguarda le misure sperimentali.
- La distribuzione ha la seguente forma funzionale

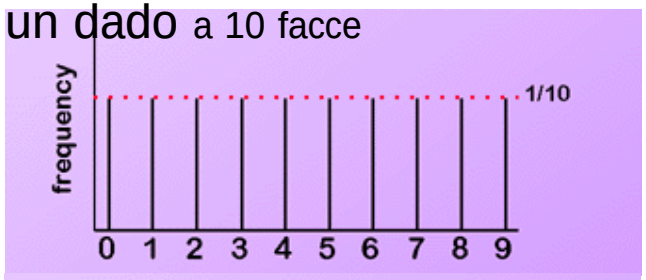
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[ -\frac{(x-\mu)^2}{2\sigma^2} \right]$$




- Dove μ è il valor medio e $\sigma^2$ la varianza
- La funzione è simmetrica attorno al valor medio
- Il valore di $\sigma$ determina la larghezza della curva.

# PDF normale come somma di processi casuali elementari

Frequenza uscite
da un dado a 10 facce



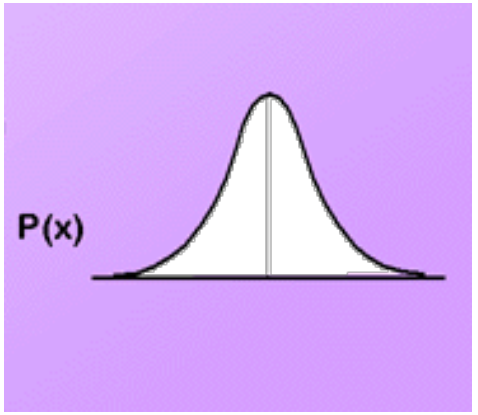

Due dadi


Tre dadi


Quattro dadi

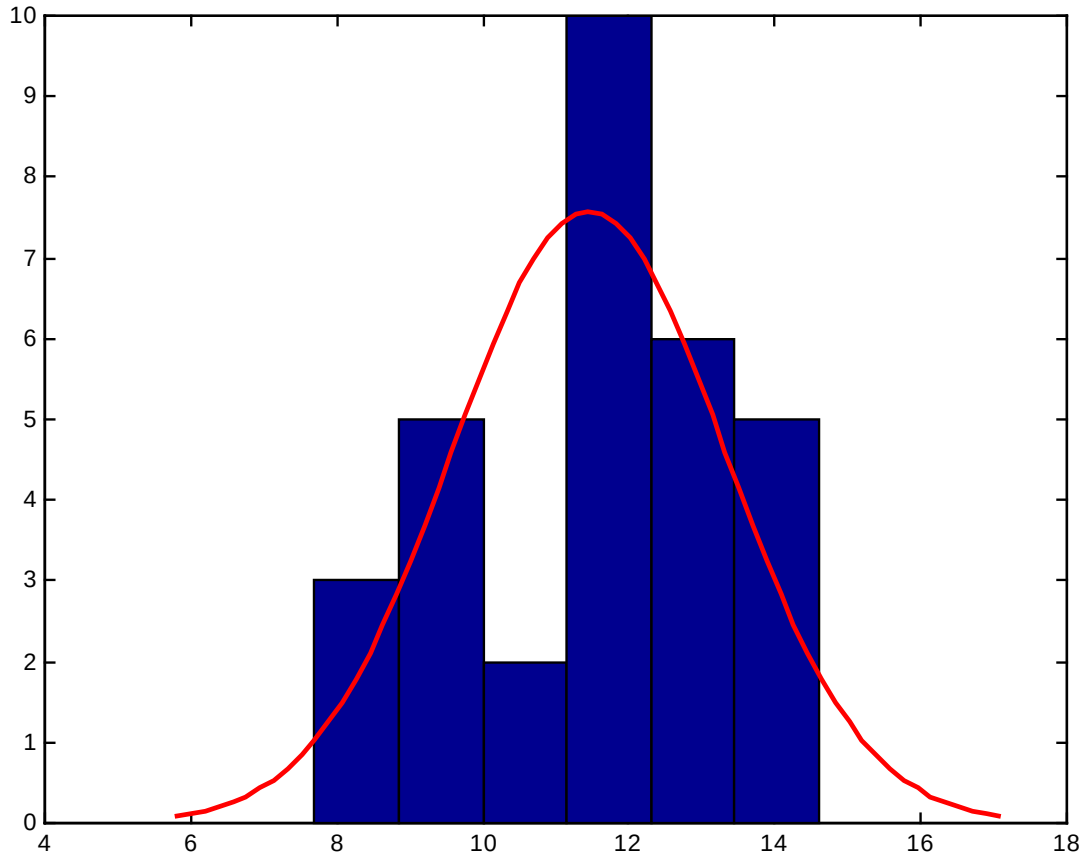

Un fenomeno è distribuito normalmente quando è costituito da un numero molto elevato (tendente a infinito) di processi casuali.

$N_{dadi} \Rightarrow \infty$



# Gaussiana ed istogramma

Dati Brix nelle pesche

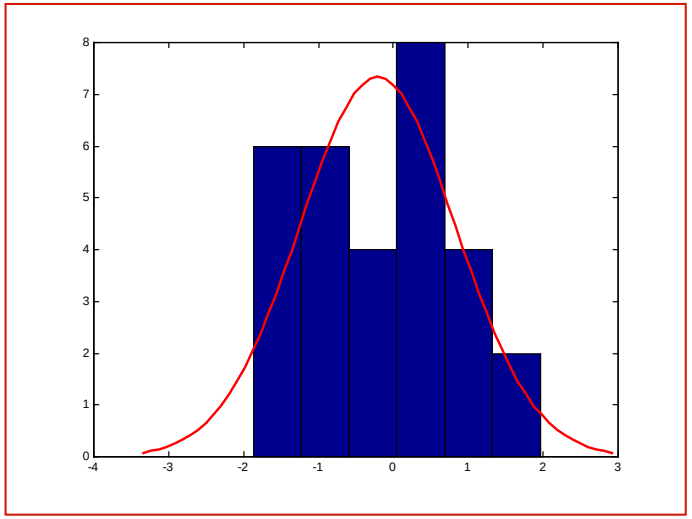


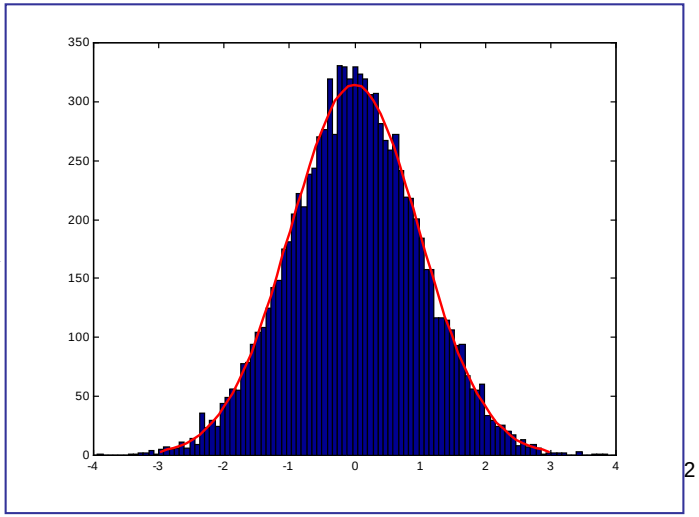
# Gaussiana e dati sperimentali

- Data una grandezza misurabile con distribuzione gaussiana quante repliche sono necessarie per ottenere una stima ottima della guassiana?
- Consideriamo una distribuzione gaussiana con media nulla e varianza unitaria ed “estraiamo” dati da questa distribuzione per ogni set di dati stimiamo la media e la varianza.
- Estrazione numeri casuali
  - MATLAB funzione **randn**
    - RANDN  Normally distributed random numbers.
    - RANDN(N) is an N-by-N matrix with random entries, chosen from
    - a normal distribution with mean zero and variance one.

# Numero di dati e valori stimati

| $N_{campioni}$ | Media | varianza |
|---|---|---|
| 3 | 0.54 | 0.22 |
| 6 | 0.25 | 1.58 |
| 10 | -0.34 | 0.95 |
| 20 | -0.39 | 0.81 |
| 30 | -0.21 | 1.08 |
| 100 | 0.00 | 0.90 |
| 1000 | 0.03 | 1.04 |
| 10000 | 0.00 | 0.98 |



# La variabile ridotta

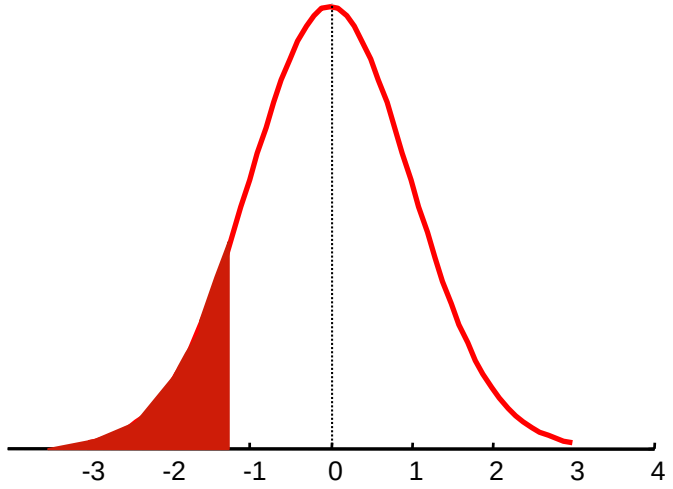
- Per facilitare l'uso della distribuzione normale conviene ridurre la variabile eventuale ad una variabile ridotta con media nulla e varianza unitaria
- In questo modo il calcolo della probabilità diventa indipendente dalla particolare natura della variabile e possono essere usate delle tabelle generali
- Per ridurre la variabile è necessario conoscere il valore medio (μ) e la varianza ($\sigma$) della variabile stessa.

$$u = \frac{x - \mu}{\sigma}$$

# Uso della variabile ridotta

- Una volta ridotta la variabile la probabilità è indipendente dal significato della variabile stessa.
- La tabella da il valore della probabilità (area) da ±u a ±∞
  - La curva è simmetrica

| \|u\| | area | \|u\| | area |
|---|---|---|---|
| 0.0 | 0.5000 | 2.0 | 0.0227 |
| 0.2 | 0.4207 | 2.2 | 0.0139 |
| 0.4 | 0.3446 | 2.4 | 0.0082 |
| 0.6 | 0.2743 | 2.6 | 0.0047 |
| 0.8 | 0.2119 | 2.8 | 0.0026 |
| 1.0 | 0.1587 | 3.0 | $1.3 \times 10^{-3}$ |
| 1.2 | 0.1151 | 4.0 | $3.2 \times 10^{-5}$ |
| 1.4 | 0.0808 | 6.0 | $9.9 \times 10^{-10}$ |
| 1.6 | 0.0548 | 8.0 | $6.2 \times 10^{-16}$ |
| 1.8 | 0.0359 | 10.0 | $7.6 \times 10^{-24}$ |



# Esempio dell'uso della distribuzione normale

- La statistica relativa al chilometraggio di una produzione di pneumatici è la seguente
  - $\mu$=58000 Km; $\sigma$=10000 Km
- Quale kilometraggio deve essere garantito per avere meno del 5% di pneumatici da rimpiazzare

- Nella tabella precedente un area di circa 0.05 la si ottiene per u=-1.6
  - Ovviamente il valore deve essere minore del valore medio
- Utilizzando la equazione della variabile ridotta si ottiene:

$$u = \frac{x - \mu}{\sigma} \Rightarrow x = \sigma \cdot u + \mu = 10000 \cdot \left(-1.6\right) + 58000 \Rightarrow x = 42000\,Km$$

# Altro esempio

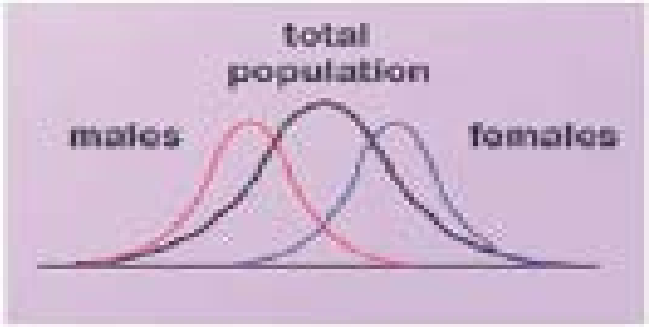
- Un elettrodo per il monitoraggio del pH viene fornito dal produttore con la seguente statistica relativa al tempo di vita
  - $\mu$=8000 ore; $\sigma$=200 ore
- Se l'elettrodo deve essere sostituito dopo 7200 ore si può affermare che l'elettrodo era difettoso?

- Si calcola la variabile ridotta $u = \frac{x - \mu}{\sigma} = \frac{7200 - 8000}{200} = -4.0$

- Nella tabella per tale valore di u si ottiene una probabilita P=3.2*$10^{-5}$
- Quindi solo lo 0.0032% degli elettrodi si guasta dopo 7200 ore quindi l'elettrodo era sicuramente difettoso.

# Cautele nell'uso di piccoli insiemi di dati

- Quando si usano insieme ridotti di dati bisogna considerare se il campione sia rappresentativo della popolazione
  - I valori devono essere veramente casuali
- Ad esempio si deve evitare di usare "sottopopolazioni"




- Una set di dati "mal-campionati" si dice *biased*
- Bisogna distinguere se il bias è dovuto ad una cattivo campionamento o se invece c'è una differenza sostanziale nei dati.
- Servono quindi degli strumenti per valutare la esistenza del bias.

# Controllo del bias per una variabile normale

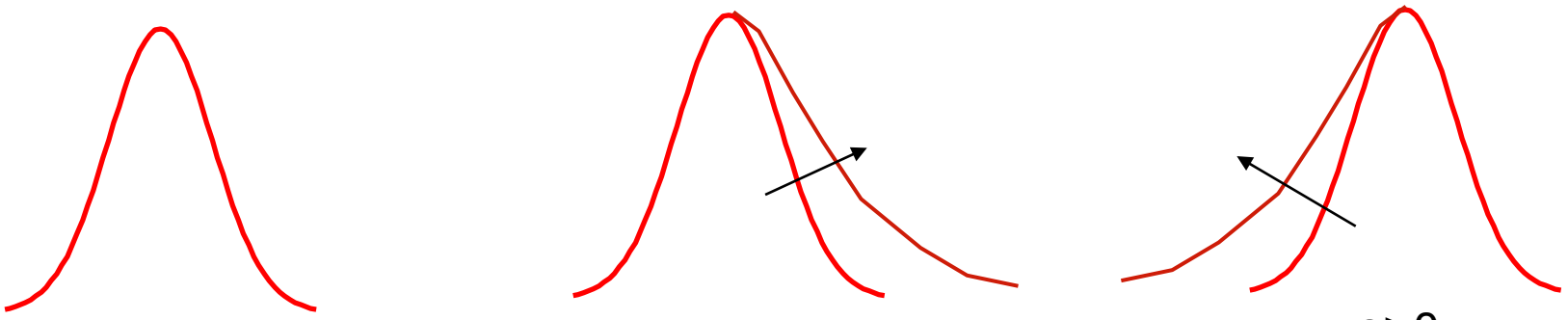
- I seguenti tre valori devono coincidere:
  - Nei limiti sperimentali
  - Media
    - Media numerica dei valori
  - Mediana
    - Valore centrale dei dati
      - È una valore vero se N è dispari e la media dei due valori centrali se N è pari
  - Moda
    - Il valore più frequente

- Esempio: brix in 31 pesche
  - 9.80; 8.00; 10.20; 9.10; 11.30 12.80 11.30 11.20 9.90
    7.70 11.50 11.90 11.60 9.20 8.30
    11.90 10.40 9.50 11.20 13.20 14.60
    13.70 13.20 13.80 13.20 13.60 11.50
    12.50 14.30 12.50 12.10
  - Media=11.45; Mediana=11.50; Moda=11.50
    - Il campionamento non è biased

# Lo skewness

- La distribuzione normale è simmetrica attorno al valore medio
- Lo skewness è una quantità che descrive la asimmetria di una distribuzione di dati
  - S=0 è la condizione di gaussianità

$$s = \frac{1}{N \cdot \sigma^3} \sum_{i=1}^{N} (x - \mu)^3$$




Per I valori di gradi brix nelle pesche s=-0.30
Quindi c'è un sospetto bias oppure la distribuzione è non gaussiana

# Distribuzione normale e analisi dati sperimentali

- Esistono due teoremi che rendono onnipresente la distribuzione di Gauss nell'analisi dei dati sperimentali
- Teorema del Limite Centrale:
  - Date n variabili casuali, ciascuna delle quali distribuita con legge arbitraria, ma aventi tutte valori aspettati dello stesso ordine di grandezza e varianze finite e dello stesso ordine di grandezza, si ha che la variabile casuale costruita con qualunque combinazione lineare di dette variabili ha una distribuzione che tende a quella normale al crescere di n.
- Teorema di Gauss:
  - un errore di misura è il risultato di molte cause indipendenti, ciascuna delle quali produce un errore piccolo, con segno a caso e dello stesso ordine di grandezza di tutti gli altri.
  - Se cioè l'errore di misura può essere considerato come dato da una combinazione lineare, in particolare della somma, di un numero molto grande di errori piccoli, si può applicare il teorema del limite centrale ed affermare che gli errori di misura seguono essi stessi una distribuzione di probabilità normale

# Conseguenza del Teorema di Gauss

- se l'errore e, definito come differenza tra valore misurato e valore aspettato (si ricordi ignoto) $\varepsilon$=x-m, è dato dalla somma degli $\varepsilon_i$ prodotti dalle n cause indipendenti:

$$\varepsilon = \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_n$$

con

$$E(\varepsilon_i) = 0 \quad ; \quad \sigma^2(\varepsilon_i) = E(\varepsilon_i)^2 \quad ; \quad E(\varepsilon) = 0$$

da cui

$$\sigma^2(\varepsilon) = \sum \sigma^2(\varepsilon_i)$$

- Per il teorema del limite centrale la distribuzione di $\varepsilon$ è: $$f(\varepsilon) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$$
- Che può essere scritta come:

$$f(\varepsilon) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$

- **In conclusione, le misure seguono una distribuzione normale centrata intorno al valore aspettato m e con varianza uguale a quella dell'errore.**

# Regressione statistica

- Chiamiamo regressione la stima parametri da dati sperimentali
- La regressione è utilizzata o per calcolare parametri fisico/chimici da misure sperimentali o per calibrare uno strumento di misura
  - Di modo che dato un segnale dello strumento siamo in grado di determinare il valore della grandezza misurata
- La regressione è una operazione statistica a causa della presenza degli errori di misura
- Gli errori di misura rendono le variabili misurate delle grandezze stocastiche dotate di una distribuzione di probabilità per cui il valore realmente osservato è solo una “occorrenza” del fenomeno.
- Il metodo generale per la regressione è il metodo dei minimi quadrati
  - La soluzione ottimale è quella che rende minimo l’errore quadratico.

# Metodo dei Minimi Quadrati I
## Least Squares

- Siano *y* ed *x* due grandezze misurabili in relazione funzionale tra loro

$$y = g(x; k_1, \ldots, k_m)$$

  - Esempio: polinomio $y = k_1 + k_1 x + k_3 x^2 + \ldots + k_n x^{n-1}$

- Supponiamo di aver eseguito n misure $x_i, y_i$
- **Se non ci sono errori di misura, la relazione funzionale *y=g(x)* è esatta, quindi ci sono n equazioni per m incognite**

$$y_i = g(x_i; k_1, \ldots, k_m)$$

- N=m il sistema e risolvibile
- N>m le equazioni eccedenti n=m sono delle identità

# Metodo dei Minimi Quadrati II
## Least Squares

- Errori di misura $\Rightarrow$ la relazione funzionale non è esatta

$$y_i = g\left(\left(x + e_{x_i}\right); k_1, \ldots, k_m\right) + e_{y_i}$$

- e: variabile aleatoria a media nulla
- Ipotesi 1: $e_{y_i} >> e_{x_i}$
  - L'errore di misura alla risposta del sensore è molto più grande dell'errore con cui è nota la sollecitazione (Non è sempre vero!)

- Criterio dei Minimi Quadrati:
- Tra I valori possibili dei parametri k si scelgono quelli che rendono massima la probabilità degli eventi osservati.

# Metodo dei Minimi Quadrati III

Least Squares

- Ipotesi 2: la variabile la variabile *y* ha una distribuzione gaussiana con valor medio *g(x)* e varianza $\sigma$
  - La probabilità di ogni misura è quindi:

$$DP(y_i) = \frac{1}{\sqrt{2\pi}\sigma_i} exp\left[ -\frac{(y_i - g(x_i))^2}{2\sigma_i^2} \right]$$

- Ipotesi 3: Gli n eventi osservati sono indipendenti tra loro (non sempre vero!).

# Metodo dei Minimi Quadrati IV
## Least Squares

- La probabilità totale dell'insieme dei dati è quindi il prodotto della probabilità dei singoli eventi:
  - La probabilità totale si chiama funzione di verosimiglianza (L)

$$DP\left(y_1 \ldots y_n\right) = \prod_{i=1}^{n} DP\left(y_i, \vec{k}\right) = L\left(y_i, \vec{k}\right)$$

$$L\left(y_1 \ldots y_n, \vec{k}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} exp\left[-\frac{\left(y_i - g\left(x_i, \vec{k}\right)\right)^2}{2\sigma_i^2}\right]$$

- tesi fondamentale dei minimi quadrati: il set di dati ha la probabilità massima
  - La probabilità massima si ottiene minimizzando la somma degli scarti tra valore ipotetico e misure.

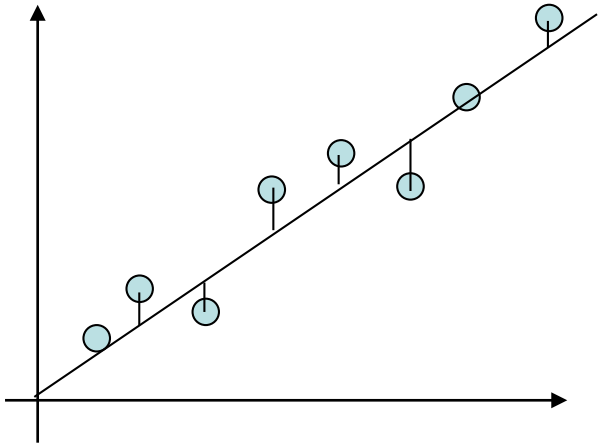
$$max\, DP \Rightarrow max\, L \Rightarrow min \sum_{i=1}^{n} \frac{\left(y_i - g\left(x_i, \vec{k}\right)\right)^2}{2\sigma_i^2}$$

# Metodo dei Minimi Quadrati V

Least Squares

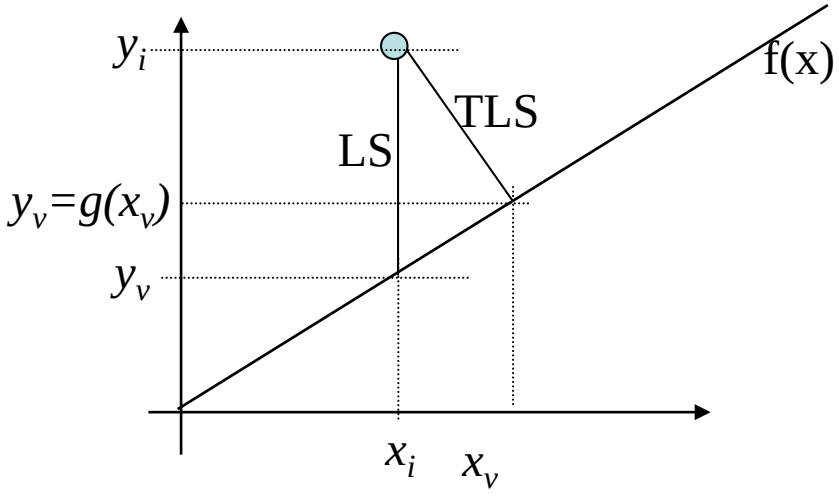
- Ipotesi 4: tutte le misure hanno uguale varianza $\sigma_i = \sigma$
- La condizione di minimo si ottiene annullando la derivata rispetto ai singoli parametri funzionali $\Rightarrow$ sistema di m equazione in m incognite.

$$\begin{cases} \dfrac{\partial}{\partial k_1} \displaystyle\sum_{i=1}^{n} \left( y_i - g(x_i, k_1 \ldots k_m) \right)^2 \\ \ldots \\ \dfrac{\partial}{\partial k_m} \displaystyle\sum_{i=1}^{n} \left( y_i - g(x_i, k_1 \ldots k_m) \right)^2 \end{cases}$$



# Total Least Squares

- Se l'ipotesi 1 non è verificata ma $e_{y_i} \approx e_{x_i}$
- Bisogna minimizzare l'errore complessivo tra la misura e la funzione



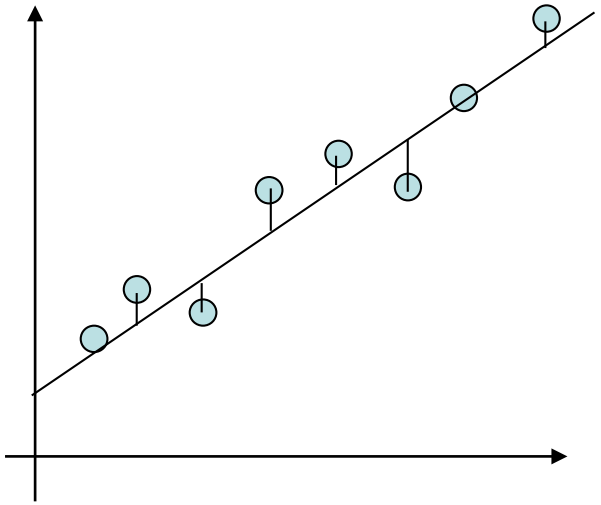

$$LS \Rightarrow \left(y_i - g\left(x_i\right)\right)$$

$$TLS \Rightarrow \sqrt{\left(y_i - g\left(x_v\right)\right)^2 + \left(x_i - x_v\right)^2}$$

# Minimi quadrati
# soluzione del caso lineare

$$y_i = a \cdot x_i + b$$

$$\begin{cases} \dfrac{\partial}{\partial a}\displaystyle\sum_{i=1}^{n}\left(y_i - a\cdot x_i + b\right)^2 = 0 \\ \dfrac{\partial}{\partial b}\displaystyle\sum_{i=1}^{n}\left(y_i - a\cdot x_i + b\right)^2 = 0 \end{cases} \Rightarrow \begin{cases} a = \dfrac{\displaystyle\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\displaystyle\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \\ b = \dfrac{\displaystyle\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i y_i}{\displaystyle\sum_{i=1}^{n}\left(x_i - y_i\right)^2} \end{cases}$$



dove $\bar{x} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{n} x_i$; $\bar{y} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{n} y_i$;

# La legge di propagazione degli errori

- La varianza della somma di variabili indipendenti è uguale alla somma delle varianze. Non esiste una legge simile per il prodotto o il rapporto tra variabili indipendenti
- In realtà capita di dover calcolare una grandezza utilizzando grandezze misurate:
- Qual'è l'errore di queste grandezze misurate?
- Si può dimostrare che data una funzione y=h(x) dove x è una grandezza misurata con media m e deviazione standard $\sigma$, la devizione standard di y è data da:

$$\sigma[y] = \left| \left( \frac{dy}{dx} \right)_{x=m} \right| \cdot \sigma(x)$$

- Per funzioni di n variabili $x_i$ ognuna con deviazione standard $\sigma_i$ si ha:

$$\sigma[y] = \sqrt{\sum_{i=1}^{n} \left( \frac{\partial y}{\partial x} \right)_{x=m}^{2} \cdot \sigma^2(x_i)}$$

# esempio

- Calcolo del volume di un parallelepipedo

$$x = 10 \ \ cm \quad \sigma_x = 0.1 \ \ cm$$

$$y = 15 \ \ cm \quad \sigma_y = 0.1 \ \ cm$$

$$z = 20 \ \ cm \quad \sigma_z = 0.1 \ \ cm$$

$$V = x \cdot y \cdot z = 10 \cdot 15 \cdot 20 = 3000 \ \ cm^2$$

$$\sigma_V = \left|\frac{\partial V}{\partial x}\right| \cdot \sigma_x + \left|\frac{\partial V}{\partial y}\right| \cdot \sigma_y + \left|\frac{\partial V}{\partial z}\right| \cdot \sigma_z = y \cdot z \cdot \sigma_x + x \cdot z \cdot \sigma_y + x \cdot y \cdot \sigma_z =$$

$$= 15 \cdot 20 \cdot 0.1 + 10 \cdot 20 \cdot 0.1 + 10 \cdot 15 \cdot 0.1 = 30 + 20 + 15 = 65 \ \ cm^2$$

# Errore della stima con i minimi quadrati

- Per stimare la deviazione standard di a e b possiamo usare la regola della propagazione degli errori ricordando che l'unica variabile affetta da errori è y e che le $\sigma_i$ sono tutte uguali.

$$\sigma_a = \sqrt{\sum_{i=1}^{n} \left( \frac{\partial a}{\partial y_i} \right)^2 \sigma_i^2} = \sigma \cdot \sqrt{\sum_{i=1}^{n} \left( \frac{\partial a}{\partial y_i} \right)^2}$$

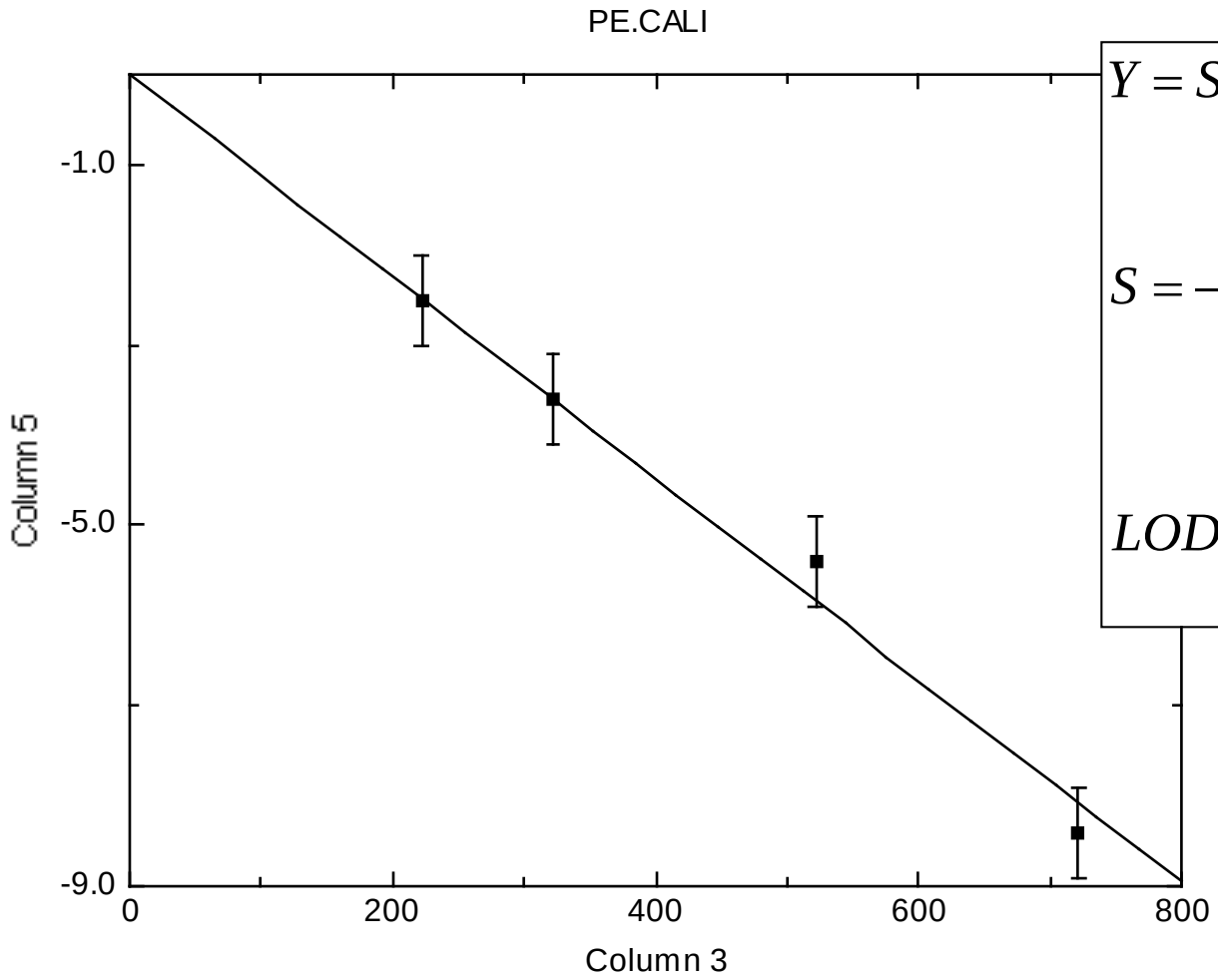
$$\sigma_b = \sqrt{\sum_{i=1}^{n} \left( \frac{\partial b}{\partial y_i} \right)^2 \sigma_i^2} = \sigma \cdot \sqrt{\sum_{i=1}^{n} \left( \frac{\partial b}{\partial y_i} \right)^2}$$

# Esempio: calibrazione di un sensore di metanolo

- Sensore di gas sensibile a vapori di metanolo
- La concentrazione di metanolo viene stimata calcolando la parte volatile di una quantità di metanolo tenuta a temperatura nota e costante. La concentrazione satura è poi miscelata con azoto da un sistema di flussimetri.
- La legge di regressione lineare non contiene l'intercetta
  - Non posso avere un segnale del sensore con concentrazione nulla di vapore.
- Ipotesi dei minimi quadrati

| | | | |
|---|---|---|---|
| 1 | L'errore su $y$ è molto maggiore dell'errore su $x$ | SI | La stabilità della temperatura e le prestazioni dei flussimetri sono sufficienti |
| 2 | $Y$ è distribuita normalmente | ≈ SI | Essendo misure sperimentali obbediscono probabilmente al teorema di Gauss |
| 3 | Gli eventi osservati sono indipendenti | SI | Controllare l'assenza di effetto memoria |
| 4 | Le misure hanno varianza uguale | SI | Le misure sono svolte in condizioni simili e sempre con gli stessi sensori |

# Esito della calibrazione



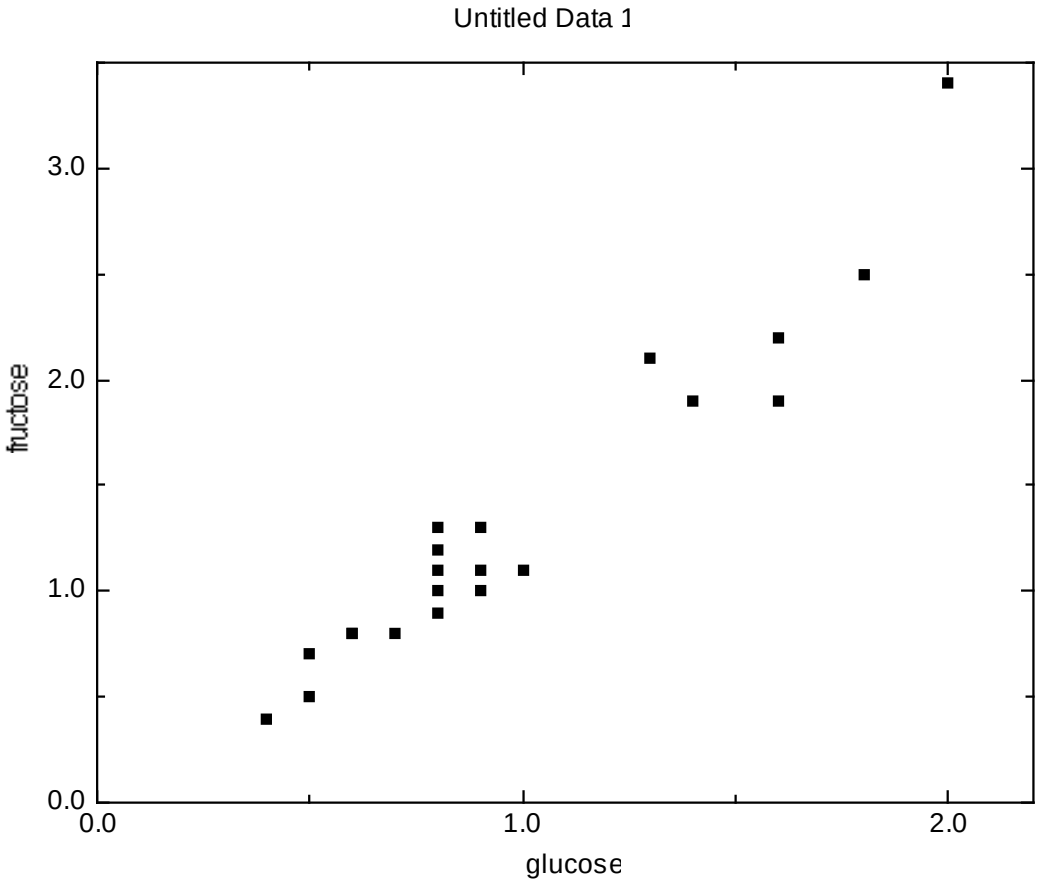

$$Y = S \cdot C \qquad \sigma_y = 0.5 \ \ Hz$$

$$S = -\left(11.1 \pm 0.8\right)\frac{mHz}{ppm}$$

$$LOD = \frac{\sigma_y}{|S|} = \frac{0.5}{11.1 \cdot 10^{-3}} = 45 \ \ ppm$$

# Relazione Glucosio-Fruttosio in una popolazione di pesche

- Natura dei dati:
  - 21 pesche e nettarine
  - Concentrazione di glucosio e fruttosio misurate con opportuni biosensori
- Le concentrazioni dei due zuccheri sono chiaramente “correlate”
  - Nel senso che le due quantità “co-variano”: se l’una cresce cresce anche l’altra e viceversa.
- Esiste una legge lineare che rappresenta questa relazione?
- A cosa attribuiamo le deviazioni dalla retta?

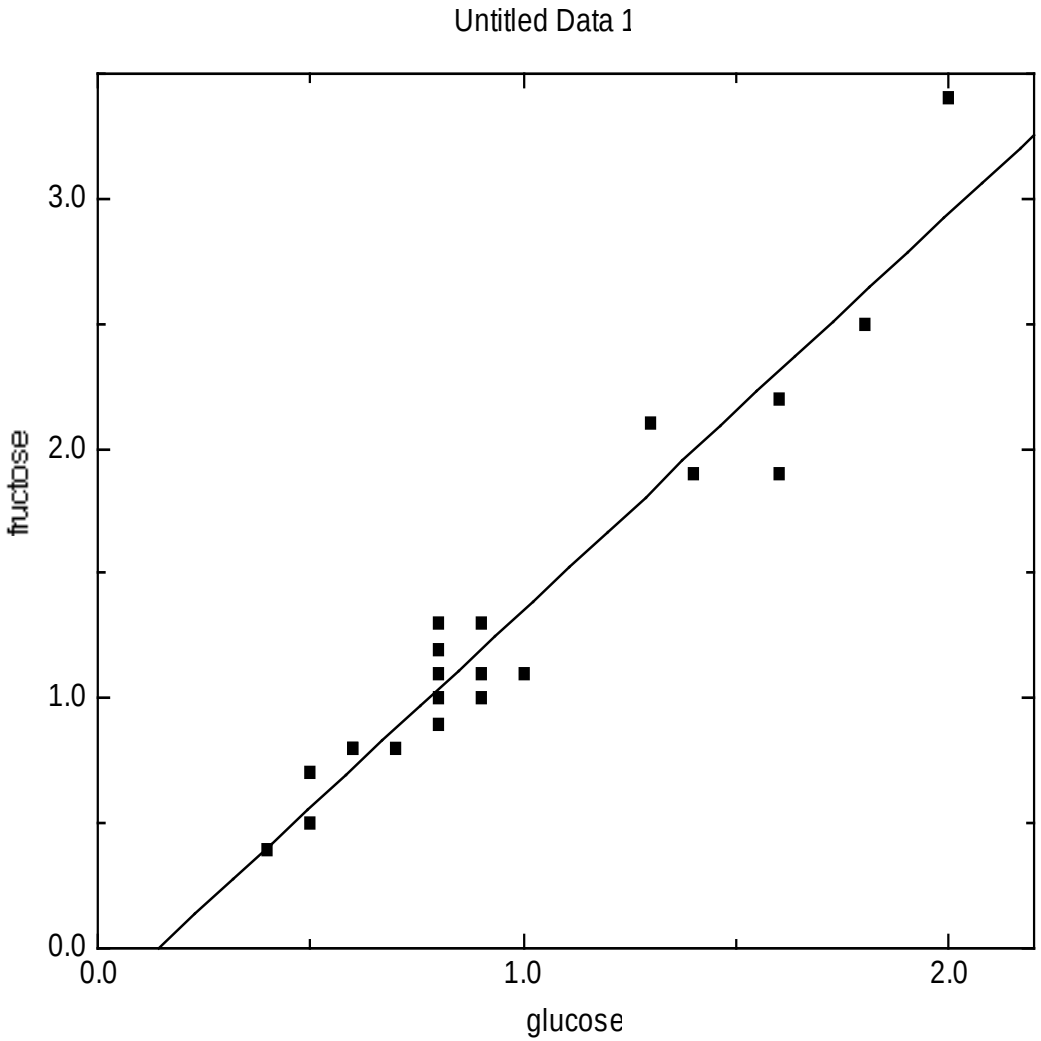


# Le ipotesi dei minimi quadrati sono verificate?

| | | | |
|---|---|---|---|
| 1 | L'errore su $y$ è molto maggiore dell'errore su $x$ | NO | Entrambe le quantità discendono da misure con sensori simili |
| 2 | $Y$ è distribuita normalmente | ≈ SI | Essendo misure sperimentali obbediscono probabilmente al teorema di Gauss |
| 3 | Gli eventi osservati sono indipendenti | SI | Le misure sono indipendenti (attenzione all'effetto memoria dei sensori) |
| 4 | Le misure hanno varianza uguale | SI | Le misure sono svolte in condizioni simili e sempre con gli stessi sensori |

Nonostante la prima ipotesi non sia verificata il metodo può essere applicato considerando però che la stima è "meno robusta"

# Retta di regressione
# legge completa




$$\left[fructose\right] = A \cdot \left[glucose\right] + B$$

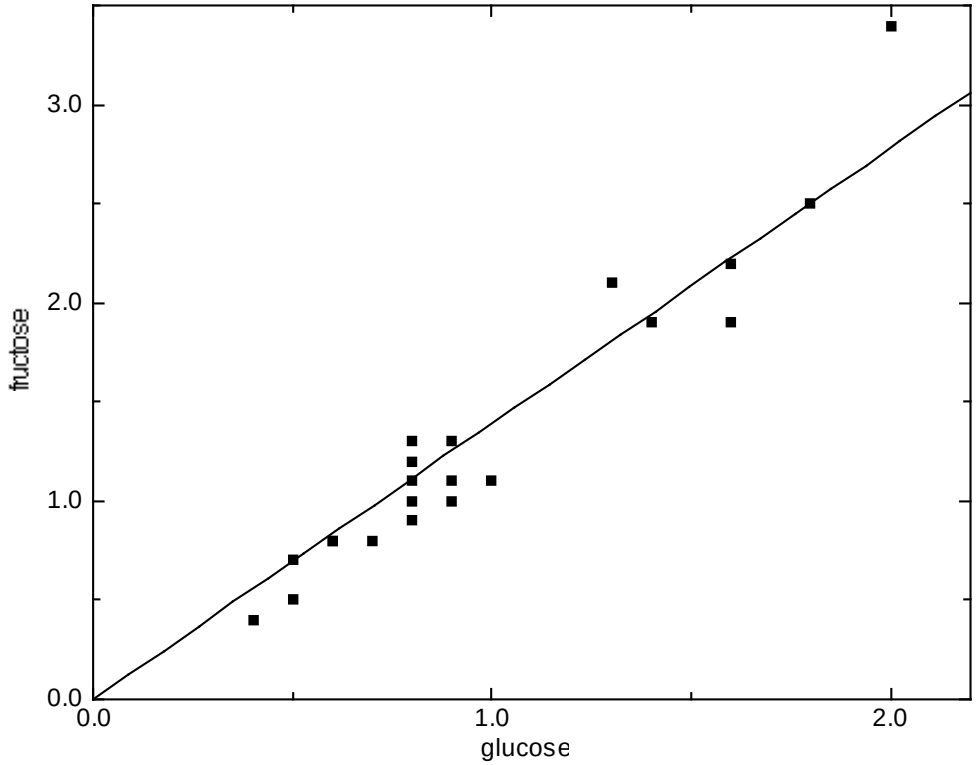
$$A = 1.58 \pm 0.16$$

$$B = -0.22 \pm 0.17$$

- La regressione è un'operazione statistica, i risultati sono quindi di natura statistica.
- Anche se in seguito non insisteremo su questo punto è importante sapere che la legge di regressione è sempre statistica

# Retta di regressione (intercetta?)

- L'intercetta ha senso o no?
- Nel caso in esempio ha senso fisico avere fruttosio in assenza di glucosio?
- Se no bisogna correggere la retta di regressione eliminando l'intercetta




$$[fructose] = A \cdot [glucose]$$

$$A = 1.39 \pm 0.07$$

# Errori di misura ed errori del modello

- La legge lineare è stata determinata ma le misure deviano dalla legge stessa. Perché?
  - Errori di misura: esistono sempre tanto più in questo caso in cui si sono usati sensori e non strumenti di alta precisione
- La motivazione più importante è che la legge lineare è di per se approssimata in quanto:
  - La relazione glucosio-fruttosio può non essere lineare
  - La quantità di fruttosio non dipende solo dal glucosio ma da tanti altri parametri chimici, fisici e biologici per cui la relazione esatta contiene molte più variabili
  - Tutte queste "inesattezze" del modello se sono distribuite casualmente possono essere trattate come errori di misura.
  - In questo contesto la prima ipotesi dei minimi quadrati può considerarsi soddisfatta perché:

$$[Fr] + \Delta[Fr] = K \cdot \left([Gl] + \Delta[Gl]\right) + \Delta Mod$$

Δ[Fr] e Δ[Gl] sono gli errori di misura dei rispettivi sensori
ΔMod è l'errore del modello lineare
Quindi l'errore globale di [Fr] è (Δ[Fr]+ΔMod) e poiché gli errori di misura sono verosimilmente simili la prima ipotesi dei minimi quadrati è soddisfatta

# Test del $\chi^2$

- Il test del $\chi^2$ è un procedimento statistico che consente di assegnare una certa probabilità alla ipotesi relativa alla forma funzionale che mette in relazione due variabili statistiche.
- Tale test si basa su una variabile detta $\chi^2$ e sulla sua distribuzione.
- Date $\nu$ variabili casuali indipendenti (x) con valori medi (m) e varianza ($\sigma^2$) si definisce la variabile $\chi^2$
  - Ha la stessa forma funzionale del termine che viene minimizzato nel metodo dei minimi quadrati

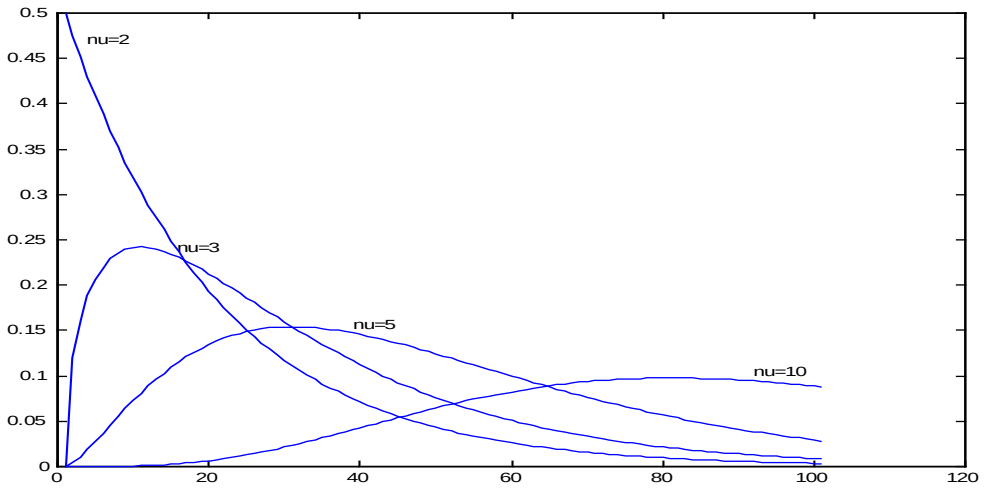
$$\chi^2 = \sum_{k=1}^{\nu} \frac{\left(x_k - m_k\right)^2}{\sigma_k}$$

# Distribuzione del $\chi^2$

- Si dimostra che la variabile $\chi^2$ segue la seguente PDF

$$f\left(\chi^2\right)=\frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)}e^{-\frac{\chi^2}{2}}\left(\chi^2\right)^{\frac{\nu}{2}-1} \quad con \quad \Gamma\left(\frac{\nu}{2}\right)=\left(\frac{\nu}{2}-1\right)!$$

- La distribuzione dipende solo dal parametro $\nu$ detto gradi di libertà
- Per $\nu$=1 la distribuzione diverge per $\nu \Rightarrow \infty$ converge alla distribuzione normale
- Il massimo della distribuzione si ottiene per $\chi^2 = \nu/2$



# Uso del test del $\chi^2$

- Supponiamo di avere eseguito la regressione di un insieme di dati con varie forme funzionali
- Per ogni forma funzionale si calcola il corrispondente valore del $\chi_o^2$ con la seguente relazione
  - In pratica il $\chi_o^2$ è il valore finale dello scarti quadratico medio tra dati ed ipotesi

$$\chi^2 = \sum_{k=1}^{N} \frac{\left(y_k - f\left(x_k; a_1, a_2, \ldots, a_n\right)\right)^2}{\sigma_k^2}$$

- I gradi di libertà si calcolano analogamente al caso della varianza

$$\nu = N - n_{param}$$

  - Dove N è il numero dei campioni e $n_{param}$ il numero dei parametri stimati
- Si calcola il valore della probabilità di avere un $\chi^2$ maggiore di $\chi_o^2$, tanto più bassa è tale probabilità tanto meno attendibile è l'ipotesi

$$\alpha = \int_{\chi_0^2}^{\infty} f\left(\chi^2\right) d\chi^2$$

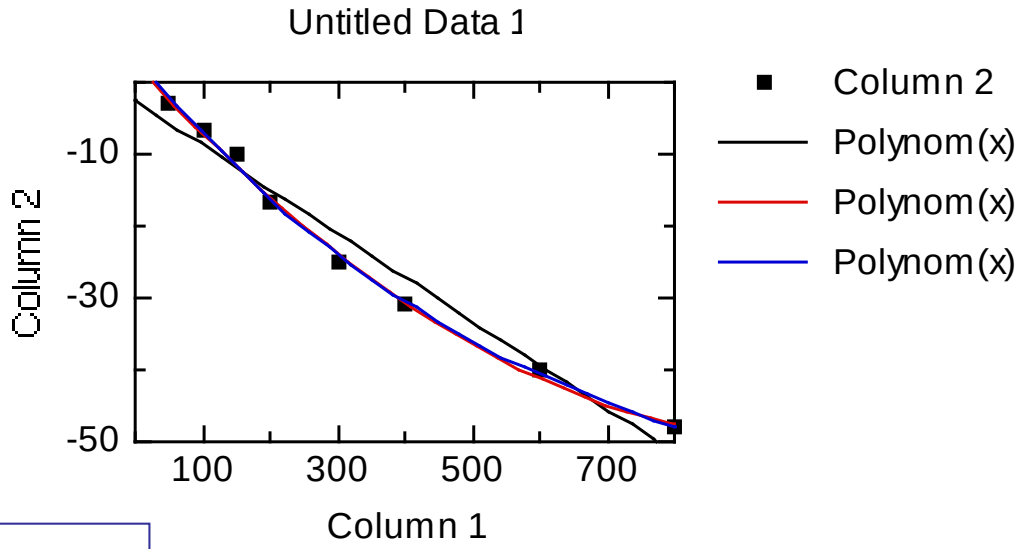
# Esempio 1

- Supponiamo di avere le seguenti 8 coppie di dati (x,y) e di seguire una regressione lineare, quadratica e cubica

$$y = s + k \cdot x$$
$$s = -2.60; \quad k = -6.12 \cdot 10^{-2}$$
$$\chi_0^2 = 60.42; \quad \nu = 8 - 2 = 6 \Rightarrow \alpha = 0.01$$

$$y = s + k_1 \cdot x + k_2 \cdot x^2$$
$$s = 2.74; \quad k_1 = -0.10; \quad k_2 = 5.06 \cdot 10^{-5};$$
$$\chi_0^2 = 6.97; \quad \nu = 8 - 3 = 5 \Rightarrow \alpha = 0.25$$

$$y = s + k_1 \cdot x + k_2 \cdot x^2 + k_3 \cdot x^3$$
$$s = 3.55; \quad k_1 = -0.11; \quad k_2 = 8.39 \cdot 10^{-5}; \quad k_3 = -2.63 \cdot 10^{-8}$$
$$\chi_0^2 = 6.42; \quad \nu = 8 - 4 = 4 \Rightarrow \alpha = 0.20$$



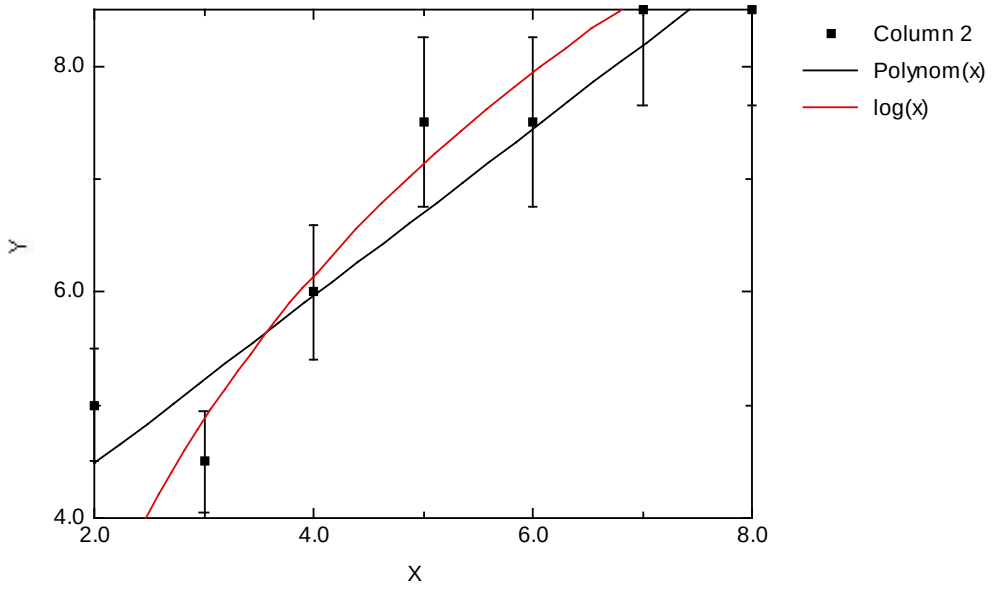

- La distribuzione quadratica è la più probabile

# Esempio 2

- 8 coppie di dati in cui i dati y sono affetti dal 10% di errore di misura
- Supponiamo di testare una relazione lineare e una logaritmica

$$y = s + k \cdot x$$
$$s = 2.99; \quad k = 0.74$$
$$\chi_0^2 = 5.17; \quad \nu = 8 - 2 = 6 \Rightarrow \alpha = 0.50$$

$$y = k \cdot \log x$$
$$k = 4.43$$
$$\chi_0^2 = 16.91; \quad \nu = 8 - 1 = 7 \Rightarrow \alpha = 0.025$$




La relazione lineare è molto più probabile della relazione logaritmica

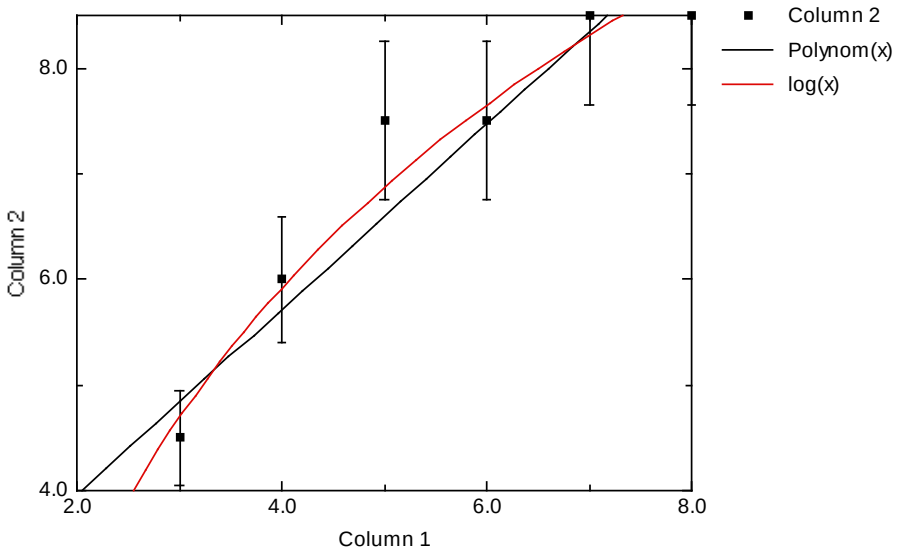
**MA...**

# Esempio 2: segue

- Supponiamo di dubitare della bontà della misura relativa a x=2
  - Oppure supponiamo di non avere eseguito tale misura!

$$y = s + k \cdot x$$
$$s = 2.20; \quad k = 0.87$$
$$\chi_0^2 = 3.01; \quad \nu = 8 - 2 = 6 \Rightarrow \alpha = 0.80$$

$$y = k \cdot \log x$$
$$k = 4.27$$
$$\chi_0^2 = 1.18; \quad \nu = 8 - 1 = 7 \Rightarrow \alpha = 0.99$$




I valori numerici dei parametri non cambiano apprezzabilmente
Ma le probabilità aumentano molto e addirittura l'andamento logaritmico è certo al 99%.

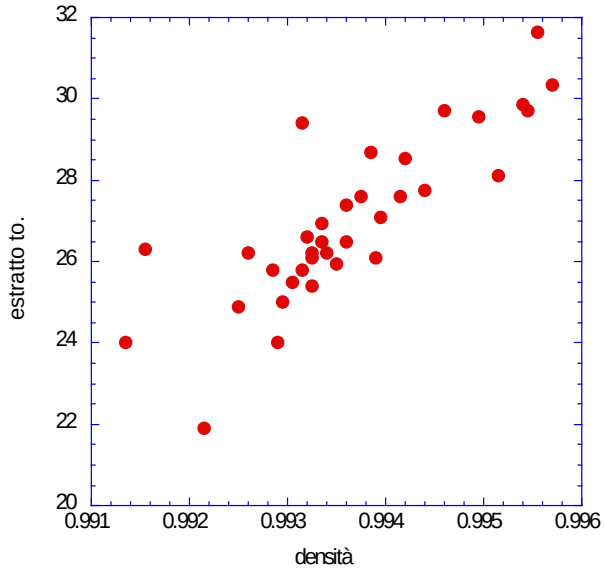
**La certezza statistica non sostituisce mai la fondatezza di un modello fisico.**

# Correlazione Lineare

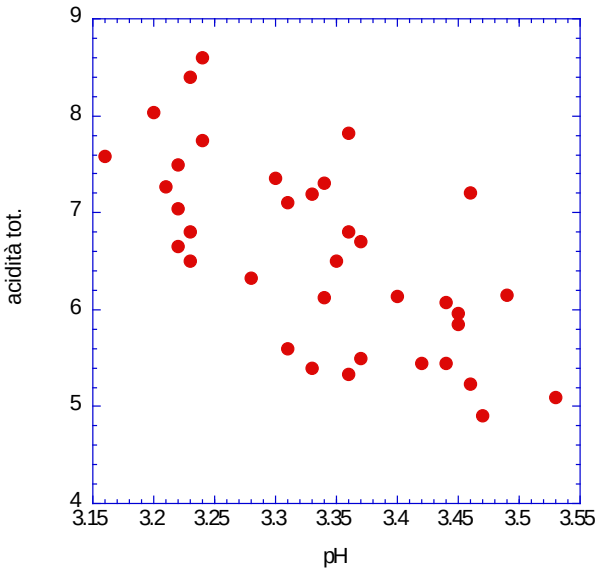
- Due grandezze Y e X sono dette correlate se dall'andamento di una è possibile dedurre l'andamento dell'altra.
  - In pratica, se una grandezza aumenta l'altra aumenta di conseguenza e viceversa. Si dice anche che le due grandezze "co-variano".
  - La correlazione può anche essere negativa, nel senso che all'aumentare di una grandezza l'altra "tende" a diminuire.
  - La correlazione implica l'andamento generale di una grandezza rispetto all'altra.
- Se due grandezze sono correlate tra loro è possibile determinare un modello di regressione tanto migliore tanto più elevata è la correlazione.
- La correlazione parziale indica che oltre a dipendere dalla variabile X, la variabile Y dipende da altre grandezze non determinate.

# Esempio di correlazione di quantità nel vino

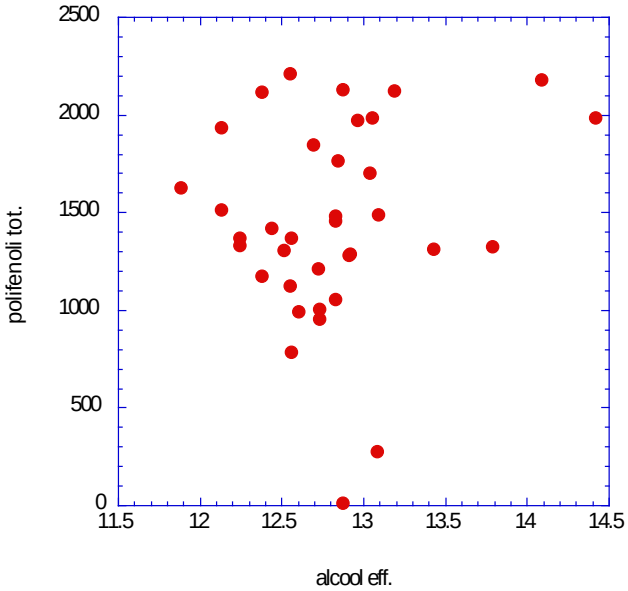
## Correlazione positiva




## Correlazione negativa



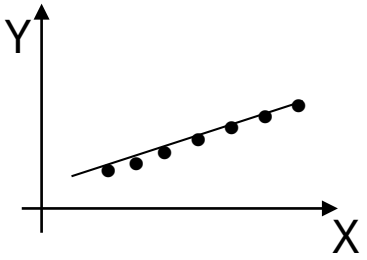

## Non correlazione



# Correlazione lineare

- La correlazione è legata alla regressione lineare.
- In particolare una serie di misure relative a due grandezze assolutamente correlate sono definite come:

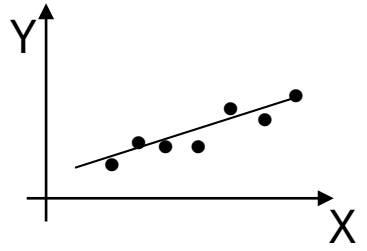
$$Y_i = k \cdot X_i$$




- Nel caso di correlazione parziale si può scrivere:

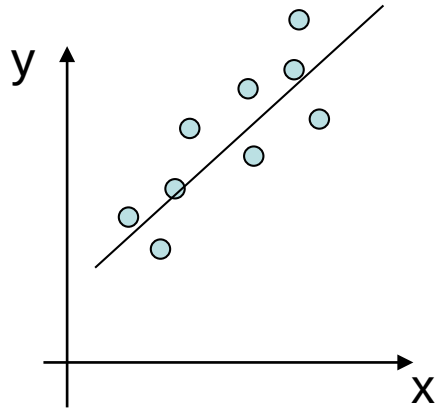
$$Y_i = k \cdot X_i + e_i$$

  - Esattamente come nel caso della regressione lineare




- La correlazione lineare esprime la deviazione dalla legge di linearità tra Y e X.

# Coefficiente di correlazione lineare




$$y = m \cdot x + b$$

$$x = m' \cdot y + b' \Rightarrow y = \frac{x}{m'} - \frac{b'}{m'}$$

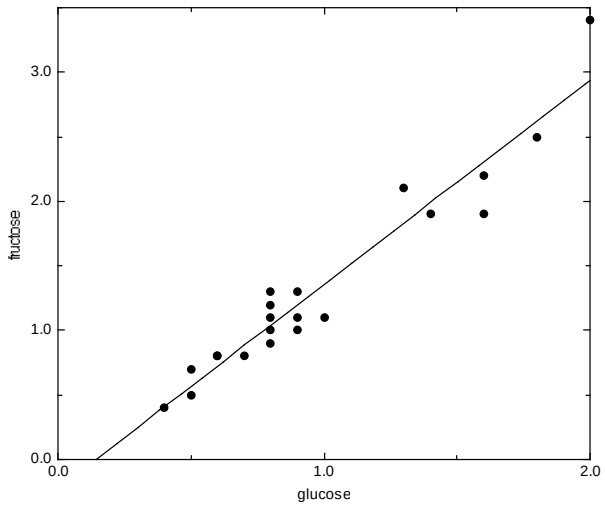
Correlazione perfetta 1/m'=m
Non correlazione mm'=0  correlazione mm'=1

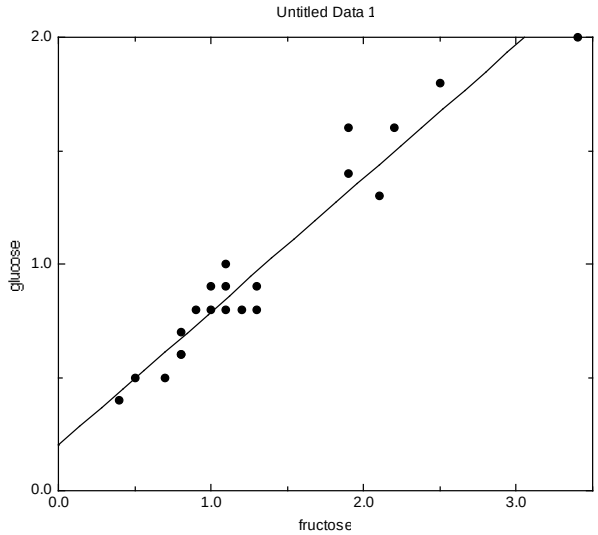
- Dati N coppie di valori x e y si definisce dai minimi quadrati lineari il seguente coefficiente

$$r = \sqrt{m \cdot m'} = \frac{N\sum_{i=1}^{N} x_i \cdot y_i - \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} y_i}{\left[N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\right]^{\frac{1}{2}} \cdot \left[N\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2\right]^{\frac{1}{2}}}$$

# esempio

## Glucosio-Fruttosio



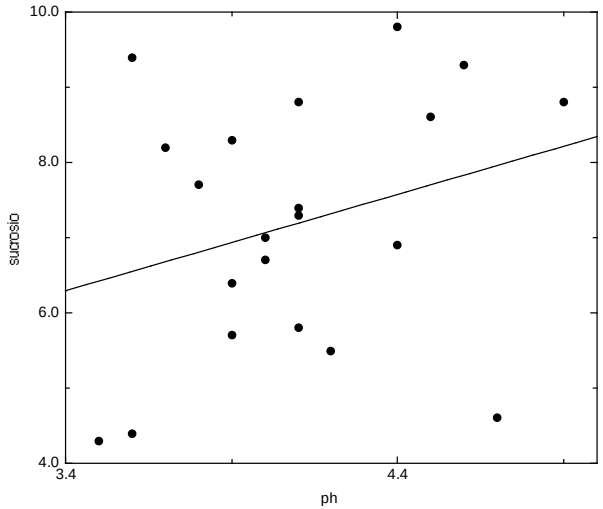



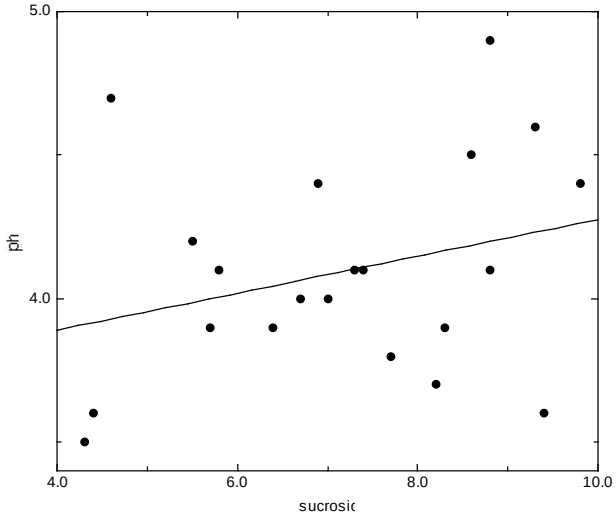

$$F = 1.58 \cdot G - 0.22$$

$$G = 0.58 \cdot F + 0.20$$

$$r = \sqrt{1.58 \cdot 0.58} = 0.95$$

## pH-sucrosio




$$S = 1.28 \cdot pH + 1.94$$

$$pH = 0.06 \cdot S + 0.63$$

$$r = \sqrt{1.28 \cdot 0.06} = 0.27$$

# Coefficienti di correlazione di un insieme di dati

- Definendo r a meno della radice quadrata si ottiene il coefficiente C.
- Cosideriamo una popolazione di pesche delle quali si siano misurate le concentrazioni delle seguente 5 quantità:
  - Acidità totale, antociani, gradi brix, carotene e clorofilla
- Supponiamo di voler studiare la correlazione reciproca tra queste quantità
- L'insieme delle correlazioni lo possiamo rappresentare in una tabella 5*5

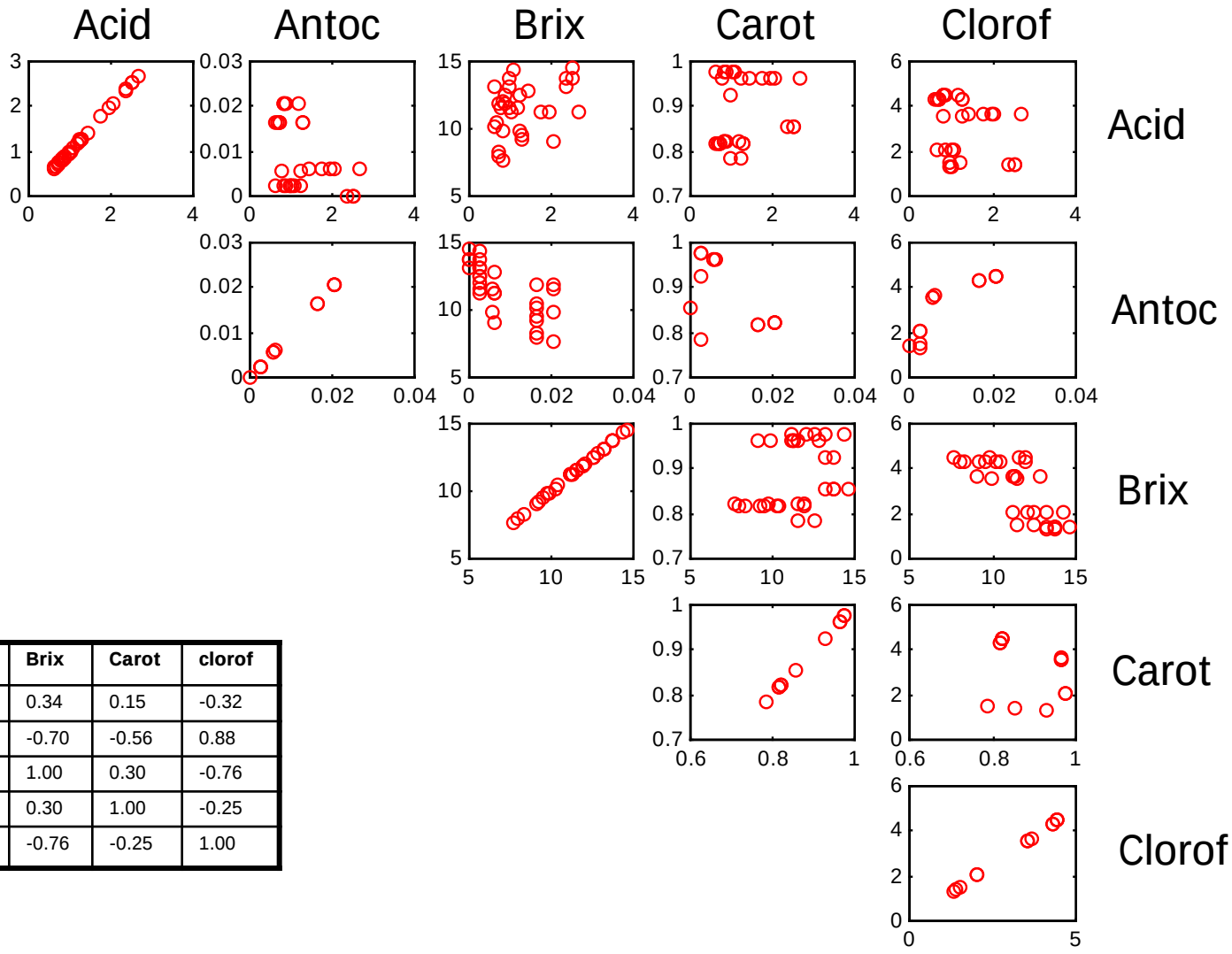
| | **Acid** | **Antoc** | **Brix** | **Carot** | **clorof** |
|---|---|---|---|---|---|
| **acid** | 1.00 | -0.48 | 0.34 | 0.15 | -0.32 |
| **antoc** | -0.48 | 1.00 | -0.70 | -0.56 | 0.88 |
| **brix** | 0.34 | -0.70 | 1.00 | 0.30 | -0.76 |
| **carot** | 0.15 | -0.56 | 0.30 | 1.00 | -0.25 |
| **clorof** | -0.32 | 0.88 | -0.76 | -0.25 | 1.00 |

# La matrice di correlazione

| | Acid | Antoc | Brix | Carot | clorof |
|---|---|---|---|---|---|
| **acid** | 1.00 | -0.48 | 0.34 | 0.15 | -0.32 |
| **antoc** | -0.48 | 1.00 | -0.70 | -0.56 | 0.88 |
| **brix** | 0.34 | -0.70 | 1.00 | 0.30 | -0.76 |
| **carot** | 0.15 | -0.56 | 0.30 | 1.00 | -0.25 |
| **clorof** | -0.32 | 0.88 | -0.76 | -0.25 | 1.00 |

- La matrice è simmetrica
  - Y rispetto a X ha la stessa correlazione di X rispetto a Y
- I valori sono contenuti tra -1 (anticorrelazione) e 1 (correlazione)
- Non esistono praticamente mai grandezze assolutamente correlate (c=1) ne scorrelate (c=0) ma la correlazione è quasi sempre parziale.
  - Quindi stabilire se due grandezze sono o no correlate dipende dal contesto e dall'applicazione.

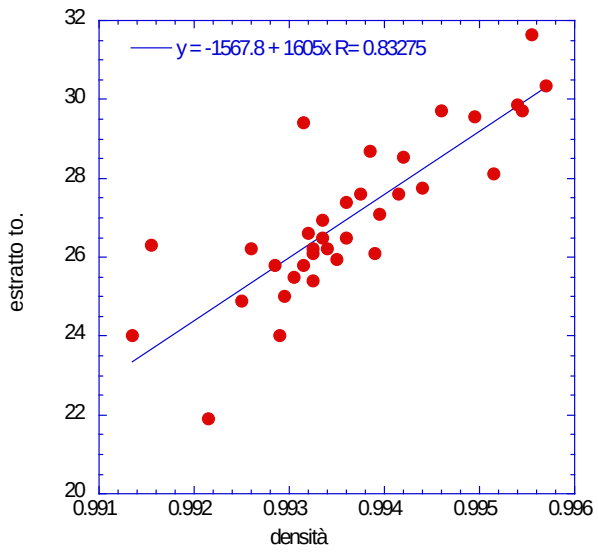
# Grafici di correlazione

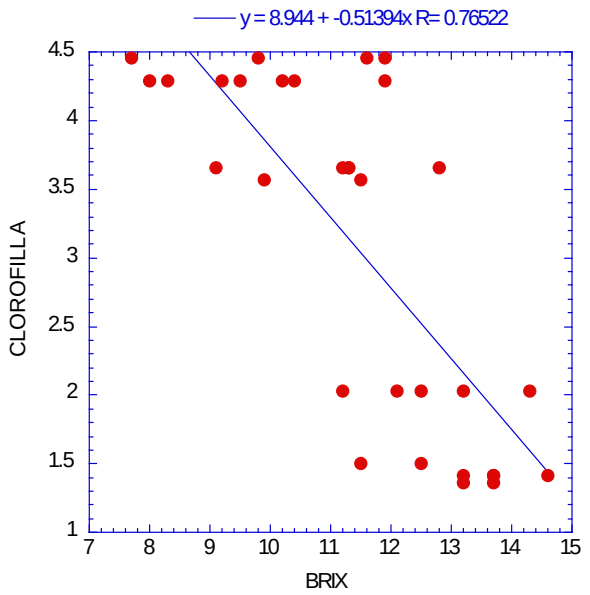



| | Acid | Antoc | Brix | Carot | clorof |
|---|---|---|---|---|---|
| **acid** | 1.00 | -0.48 | 0.34 | 0.15 | -0.32 |
| **antoc** | -0.48 | 1.00 | -0.70 | -0.56 | 0.88 |
| **brix** | 0.34 | -0.70 | 1.00 | 0.30 | -0.76 |
| **carot** | 0.15 | -0.56 | 0.30 | 1.00 | -0.25 |
| **clorof** | -0.32 | 0.88 | -0.76 | -0.25 | 1.00 |

# Il coefficiente di linearità: R

- Il coefficiente di correlazione esprime la linearità tra due grandezze.
- E' dotato di segno in modo da distinguere tra correlazione e anticorrelazione
- Due coefficienti di correlazione c=C e c=-C esprimono la stessa linearità a meno del segno del termine di proporzionalità
- Per questo motivo la radice quadrata del coefficiente di correlazione si usa per descrivere la bontà di un regressione lineare
- Tale termine è indicato come R



# Analisi della varianza

- Finora abbiamo assunto che la varianza degli osservabili proviene da un'unica sorgente casuale, cioè che i dati sperimentali osservati siano generati da una unica PDF generalmente gaussiana
- Questa assunzione non è verosimile
  - Ci possono essere molte sorgenti della varianza nel senso che data una popolazione di campioni ci sono molte distribuzioni possibili
    - Esempio: misura dei parametri della frutta: possibili sorgenti di varianza: cultivar, stato di maturazione, esposizione dei frutti, fluttuazione dello strumento di misura,...
- Quando le sorgenti della varianza sono indipendenti e non correlate le varianze sono additive

$$\sigma^2_{total} = \sigma^2_1 + \sigma^2_2 + \ldots + \sigma^2_k$$

# Determinazione delle sorgenti di varianza

- Consideriamo un esempio dove ci dovrebbero essere solo due sorgenti di varianza
  - Serie di 4 campioni ed ogni campione è misurato in tre repliche

| Campione | Repliche | Media |
|---|---|---|
| 1 | 15.9; 16.1; 16.3 | 16.1 |
| 2 | 14.9; 15.1; 15.3 | 15.1 |
| 3 | 15.8; 15.8; 15.8 | 15.8 |
| 4 | 16.2; 16.0; 15.9 | 16.0 |

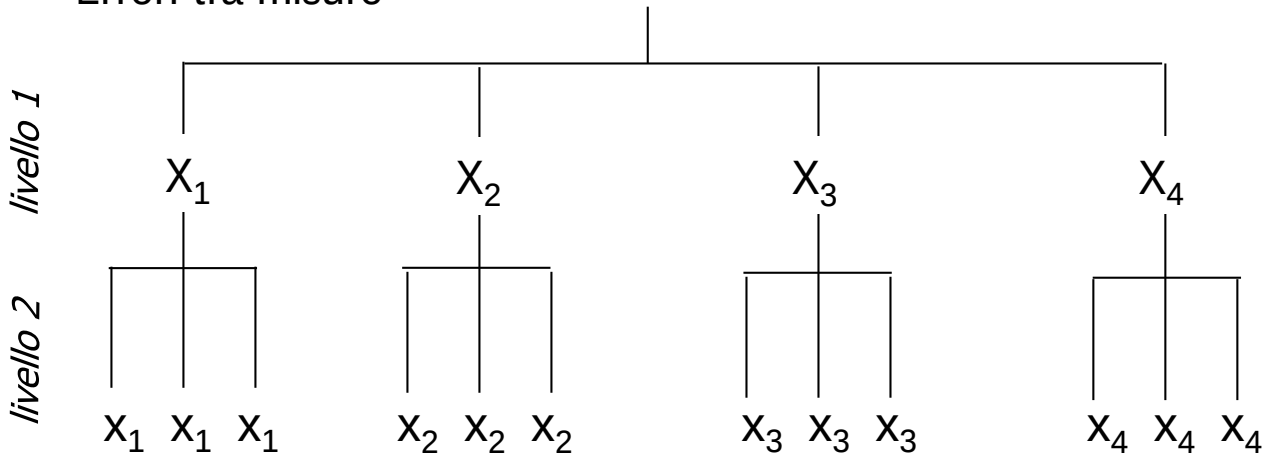
- Si vuole determinare se le medie ottenute sono dovute alla varianza del metodo di misura o a reali differenze tra i campioni

# Modello a due livelli

- $\sigma^2_{between}$: varianza tra campioni
- $\sigma^2_{within}$: varianza del metodo analitico

$$\sigma^2_{total} = \sigma^2_{between} + \sigma^2_{within}$$

- Ci sono due sorgenti potenziali di varianza
  - I campioni possono essere veramente differenti
  - Errori tra misure




Livello 1: da un idea della variabilità del campione
Livello 2: informazioni circa la variabilità del metodo

# Semplice analisi della varianza

- Varianza totale

$$\sigma_T^2 = \frac{\sum_i \left(x_i - \bar{x}_T\right)}{df_T}$$

$x_T$: media totale
$df_T$: totale misure-1

- Varianza tra le repliche

$$\sigma_{within}^2 = \frac{s_1^2 + s_2^2 + \ldots}{df_{within}}$$

- $df_{within}$: numero di dati-numero campioni
- S1 varianza dello scarto per campione

•Varianza tra i campioni

$$\sigma_{between}^2 = \frac{\sum_i n_r \left(\bar{x}_S - \bar{x}_T\right)^2}{df_S}$$

- $n_{r:}$ numero di repliche per campione
- $x_S$: media per ogni campione
- $df_S$: numero di campioni-1

# calcolo

| Campione | Repliche | Media | varianza |
|---|---|---|---|
| 1 | 15.9; 16.1; 16.3 | 16.1 | 0.040 |
| 2 | 14.9; 15.1; 15.3 | 15.1 | 0.040 |
| 3 | 15.8; 15.8; 15.8 | 15.8 | 0.000 |
| 4 | 16.2; 16.0; 15.9 | 16.0 | 0.023 |

$$\bar{x}_T = 15.75$$

$$\sigma_T^2 = \frac{\sum_{i=1}^{12}\left(x_i - \bar{x}_T\right)}{df_T} = \frac{2.04}{11} = 0.185$$

$$\sigma_{between}^2 = \frac{\sum_{i=1}^{4} n_r\left(\bar{x}_S - \bar{x}_T\right)^2}{df_S} = \frac{\sum_{i=1}^{4} 3\left(\bar{x}_S - 15.75\right)^2}{4-1} = \frac{1.83}{3} = 0.610$$

$$\sigma_{within}^2 = \frac{s_1^2 + s_2^2 + \ldots}{df_{within}} = \frac{\sum_{i=1}^{4}\sum_{n=1}^{3}\left(x_n - \bar{x}_i\right)^2}{12-4} = \frac{0.08 + 0.08 + 0 + 0.05}{8} = \frac{0.21}{8} = 0.026$$



$$2.04 = 1.83 + 0.21$$

# Valori calcolati

| Sorgente | df | $\sigma^2$ |
|---|---|---|
| Totale | 11 | 0.189 |
| Campione(between) | 3 | 0.610 |
| Repliche (within) | 8 | 0.026 |

Per determinare se ci sono differenze significative tra le due sorgenti di varianza si può usare l'**F test**

$$F = \frac{\sigma^2_{max}}{\sigma^2_{min}}$$

Se F è molto differente da 1 allora le popolazioni sono statisticamente differenti.
La distribuzione F quantifica la differenza in termini di probabilità
F viene confrontata con Fc relativo alla dimensione del set per verificare se la differenza è significativa

# F test

$$F = \frac{\sigma^2_{\max}}{\sigma^2_{\min}} = \frac{\sigma^2_{between}}{\sigma^2_{within}} = \frac{0.610}{0.026} = 23.46$$

Il valore di Fc per dfmax=3 e dfmin=8 è 4.07 quindi i campioni considerati sono statisticamente differenti.

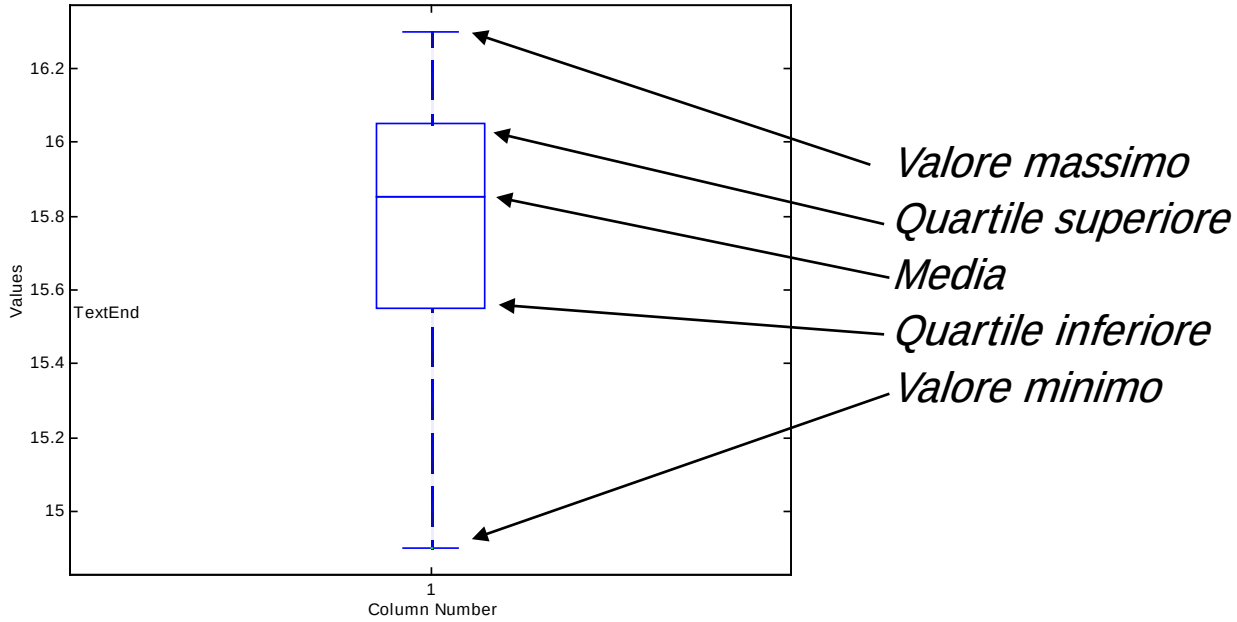
**Tabella dei valori critici di *F* test a una coda, 1-α= 0.95.**

*gradi di libertà*

| N➔ D⬇ | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 10 | 20 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 161.4 | 199.5 | 215.7 | 224.6 | 230.2 | 234.0 | 238.9 | 241.7 | 248.0 |
| 2 | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.37 | 19.40 | 16.45 |
| 3 | 10.13 | 9.552 | 9.277 | 9.117 | 9.013 | 8.941 | 8.845 | 8.786 | 8.660 |
| 4 | 7.709 | 6.944 | 6.951 | 6.388 | 6.256 | 6.163 | 6.041 | 5.964 | 5.803 |
| 5 | 6.608 | 5.786 | 5.409 | 5.192 | 5.050 | 4.950 | 4.818 | 4.735 | 4.558 |
| 6 | 5.987 | 5.143 | 4.757 | 4.534 | 4.387 | 4.284 | 4.147 | 4.060 | 3.874 |
| 8 | 5.318 | 4.459 | 4.066 | 3.838 | 3.687 | 3.581 | 3.438 | 3.347 | 3.150 |
| 10 | 4.965 | 4.103 | 3.708 | 3.478 | 3.326 | 3.217 | 3.072 | 2.978 | 2.774 |
| 15 | 4.543 | 3.682 | 3.287 | 3.056 | 2.901 | 2.790 | 2.641 | 2.544 | 2.328 |
| 20 | 4.351 | 3.493 | 3.098 | 2.866 | 2.711 | 2.599 | 2.447 | 2.348 | 2.124 |

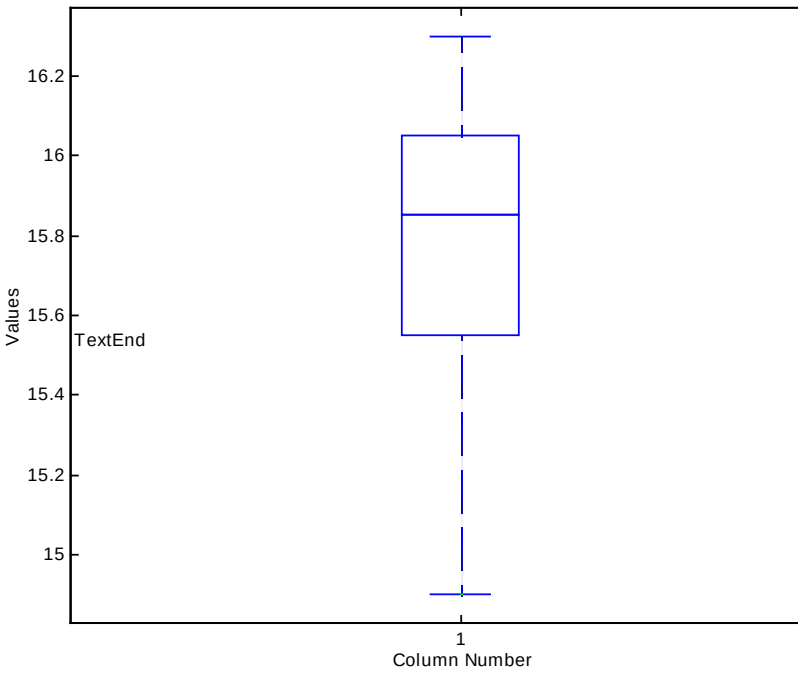
# Analisi grafica

- Box and whisker plot
  - Un tipo di grafico nel quale sono mostrati il quartile inferiore superiore e la media in un box e l'estensione dei dati
  - Il plot mostra anche gli outliers come dati che eccedono la distribuzione dei dati stessi

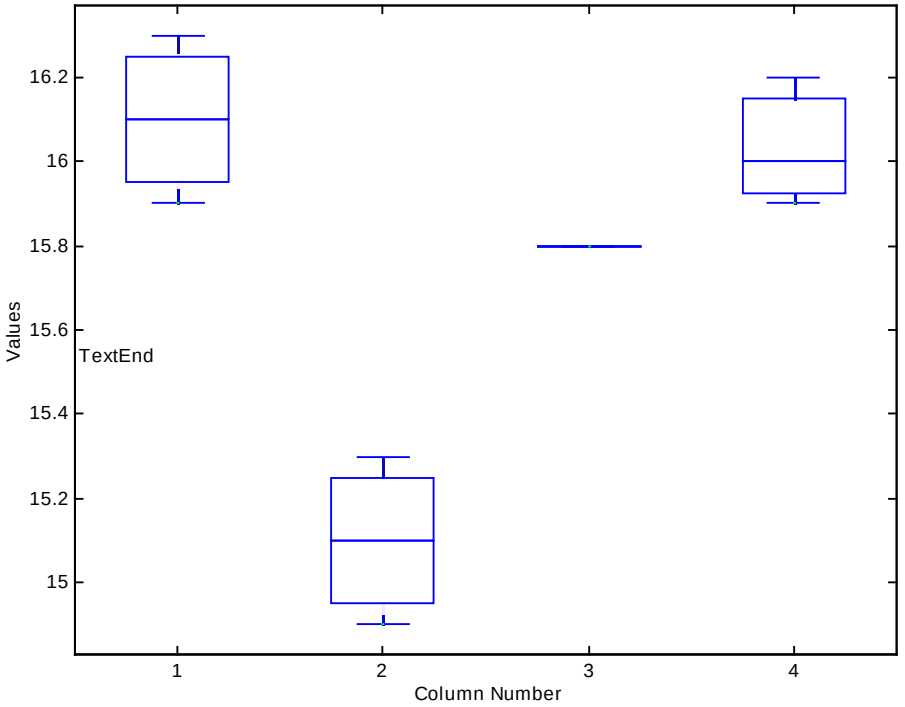


# Esempio analisi grafica

Tutti i dati




Per singolo campione




I campioni 2 e 3 appartegono a distribuzioni differenti rispetto a 1 e 4
I campioni 1 e 4 hanno distribuzioni sovrapposte.

# Distribuzioni di Probabilità multivariate

- Quale è la probabilità che in una pesca il contenuto di carotene sia 0.40±0.02 e di clorofiilla 4.31±0.23?
- Per rispondere a questa domanda non basta conoscere la probabilità della singola grandezza, infatti quella che si chiede è la probabilità congiunta.
  - Ci sono due possibilità:
    - Y e X sono indipendenti $\Rightarrow$ P(X,Y)=P(X)+P(Y)
    - Y e X non sono indipendenti $\Rightarrow$ P(X,Y)
  - Nel primo caso si usa il prodotto di due PDF monovariate
  - Nel secondo caso bisogna introdurre una PDF di due grandezze: si dice PDF bivariata.

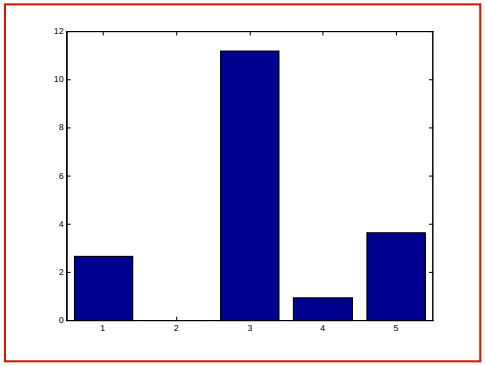
# I dati multivariati

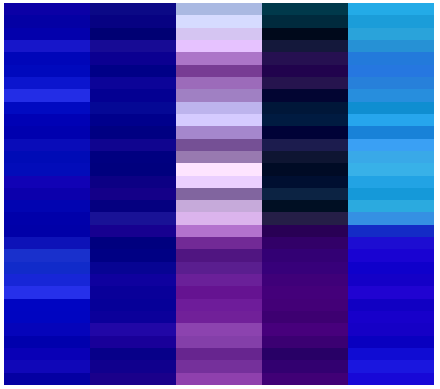
- La grandezza multivariata è una matematicamente un **vettore**
- L'insieme dei dati multivariati è una **matrice**
- Esempio:
  - Cosideriamo una popolazione di 20 pesche delle quali si siano misurate le concentrazioni delle seguente 5 quantità:
    - Acidità totale, antociani, gradi brix, carotene e clorofilla

**variabili** →

**campioni** ↓

| Acidità | Antociani | Brix | Carotene | Clorofilla |
|---|---|---|---|---|
| 0.8200 | 0.0206 | 9.8000 | 0.8231 | 4.4600 |
| 0.7300 | 0.0165 | 8.0000 | 0.8179 | 4.2900 |
| 0.6000 | 0.0165 | 10.2000 | 0.8179 | 4.2900 |
| 2.0400 | 0.0060 | 9.1000 | 0.9642 | 3.6600 |
| 1.7600 | 0.0060 | 11.3000 | 0.9642 | 3.6600 |
| 1.4200 | 0.0060 | 12.8000 | 0.9642 | 3.6600 |
| 1.9700 | 0.0060 | 11.3000 | 0.9642 | 3.6600 |
| 2.6800 | 0.0060 | 11.2000 | 0.9642 | 3.6600 |
| 1.2440 | 0.0057 | 9.9000 | 0.9634 | 3.5700 |
| 0.8300 | 0.0206 | 7.7000 | 0.8231 | 4.4600 |
| 0.7880 | 0.0057 | 11.5000 | 0.9634 | 3.5700 |
| 0.8600 | 0.0206 | 11.9000 | 0.8231 | 4.4600 |
| 1.1800 | 0.0206 | 11.6000 | 0.8231 | 4.4600 |
| 1.2700 | 0.0165 | 9.2000 | 0.8179 | 4.2900 |
| 0.7300 | 0.0165 | 8.3000 | 0.8179 | 4.2900 |
| 0.7200 | 0.0165 | 11.9000 | 0.8179 | 4.2900 |
| 0.6600 | 0.0165 | 10.4000 | 0.8179 | 4.2900 |
| 1.2600 | 0.0165 | 9.5000 | 0.8179 | 4.2900 |
| 1.0000 | 0.0025 | 11.2000 | 0.9756 | 2.0300 |
| 0.6400 | 0.0025 | 13.2000 | 0.9756 | 2.0300 |

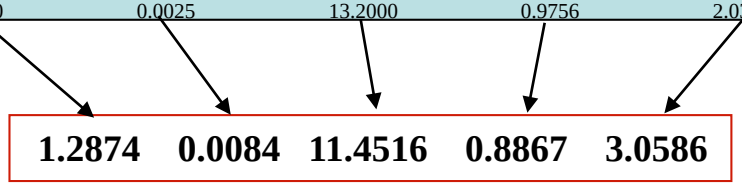


# Media multivariata

- Data un insieme di dati multivariati, si definisce media multivariata il vettore formato dalla media di ogni singola variabile.

| Acidità | Antociani | Brix | Carotene | Clorofilla |
|---|---|---|---|---|
| 0.8200 | 0.0206 | 9.8000 | 0.8231 | 4.4600 |
| 0.7300 | 0.0165 | 8.0000 | 0.8179 | 4.2900 |
| 0.6000 | 0.0165 | 10.2000 | 0.8179 | 4.2900 |
| 2.0400 | 0.0060 | 9.1000 | 0.9642 | 3.6600 |
| 1.7600 | 0.0060 | 11.3000 | 0.9642 | 3.6600 |
| 1.4200 | 0.0060 | 12.8000 | 0.9642 | 3.6600 |
| 1.9700 | 0.0060 | 11.3000 | 0.9642 | 3.6600 |
| 2.6800 | 0.0060 | 11.2000 | 0.9642 | 3.6600 |
| 1.2440 | 0.0057 | 9.9000 | 0.9634 | 3.5700 |
| 0.8300 | 0.0206 | 7.7000 | 0.8231 | 4.4600 |
| 0.7880 | 0.0057 | 11.5000 | 0.9634 | 3.5700 |
| 0.8600 | 0.0206 | 11.9000 | 0.8231 | 4.4600 |
| 1.1800 | 0.0206 | 11.6000 | 0.8231 | 4.4600 |
| 1.2700 | 0.0165 | 9.2000 | 0.8179 | 4.2900 |
| 0.7300 | 0.0165 | 8.3000 | 0.8179 | 4.2900 |
| 0.7200 | 0.0165 | 11.9000 | 0.8179 | 4.2900 |
| 0.6600 | 0.0165 | 10.4000 | 0.8179 | 4.2900 |
| 1.2600 | 0.0165 | 9.5000 | 0.8179 | 4.2900 |
| 1.0000 | 0.0025 | 11.2000 | 0.9756 | 2.0300 |
| 0.6400 | 0.0025 | 13.2000 | 0.9756 | 2.0300 |



# Varianza multivariata

- Per un insieme di dati multivariati, il concetto di varianza si trasforma in **matrice di covarianza**
- La matrice di covarianza è legata al concetto di correlazione e di matrice di correlazione.
- Dato un insieme di dati **x** la matrice di covarianza è definita come:

$$cov(X) = \Sigma = E\left[(x-m)^T \cdot (x-m)\right]$$

- La matrice di covarianza è quadrata ha dimensione pari al numero di variabili ed è simmetrica
- Gli elementi diagonali sono le varianze delle singole variabili
- Gli altri elementi sono proporzionali ai coefficienti di correlazione ($\rho$)

$$\Sigma_{ii} = \sigma_i^2 \quad ; \quad \Sigma_{ik} = \rho_{ik}\sigma_i\sigma_k$$

| | | | | |
|---|---|---|---|---|
| **0.4167** | -0.0023 | 0.4175 | 0.0072 | -0.2635 |
| -0.0023 | **0.0001** | -0.0099 | -0.0003 | 0.0084 |
| 0.4175 | -0.0099 | **3.5179** | 0.0409 | -1.8080 |
| 0.0072 | -0.0003 | 0.0409 | **0.0053** | -0.0236 |
| -0.2635 | 0.0084 | -1.8080 | -0.0236 | **1.5868** |

# Covarianza e Correlazione

- La matrice di covarianza può essere scritta come:

$$\Sigma = \Gamma \cdot R \cdot \Gamma = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & & \\ \dots & & \dots & \\ 0 & & & \sigma_n \end{bmatrix} \cdot \begin{bmatrix} 1 & \rho_{21} & \dots & \rho_{n1} \\ \rho_{12} & 1 & & \\ \dots & & \dots & \\ \rho_{1n} & & & 1 \end{bmatrix} \cdot \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & & \\ \dots & & \dots & \\ 0 & & & \sigma_n \end{bmatrix}$$

  - Dove R è la matrice di correlazione

- Dalla definizione di covarianza inoltre:

$$\Sigma = E\left[\left(X - m\right)^T \cdot \left(X - m\right)\right] = E\left[X^T X\right] - m^T m = S - m^T m$$

  - Dove S è la matrice di autocorrelazione

# Correlazione e Indipendenza

- Due variabili casuali X e Y sono incorrelate se:

$$E\left[X \cdot Y\right] = E\left[X\right] \cdot E\left[Y\right]$$

  - Questa condizione è detta anche indipendenza lineare

- Due variabili casuali X e Y sono indipendenti se:

$$P\left[X \cdot Y\right] = P\left[X\right] \cdot P\left[Y\right]$$

# PDF normale multivariata 1

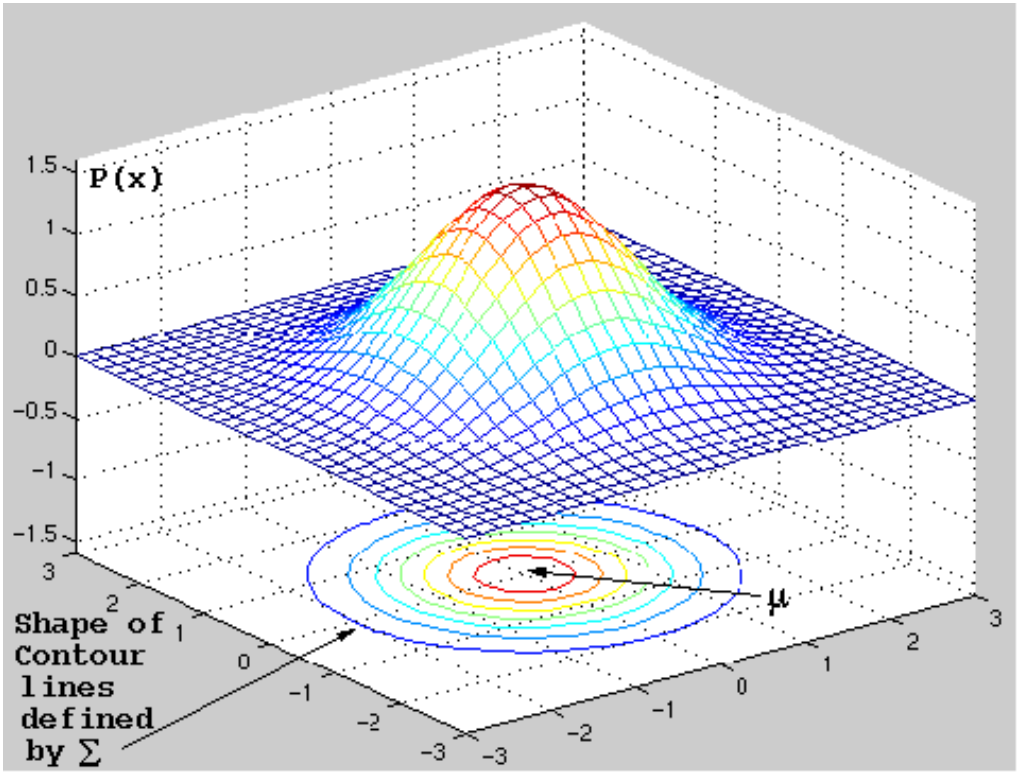
$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1d} \\ \vdots & & \vdots \\ \sigma_{d1} & \cdots & \sigma_{dd} \end{bmatrix}$$

$$p(x) = \frac{1}{\sqrt{2\Pi}\sqrt{\Sigma}} \exp\left[ \frac{-1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu) \right]$$

$$f_{\vec{X}}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n |C|}} \exp\left( -\frac{1}{2}(\vec{x}-\vec{\mu})^T C^{-1} (\vec{x}-\vec{\mu}) \right)$$

where n is the dimensionality of the space under consideration.

I luoghi di punti isoprobabili sono delle forme quadratiche.
Ellissi per PDF bivariate.



# PDF normale multivariata 2

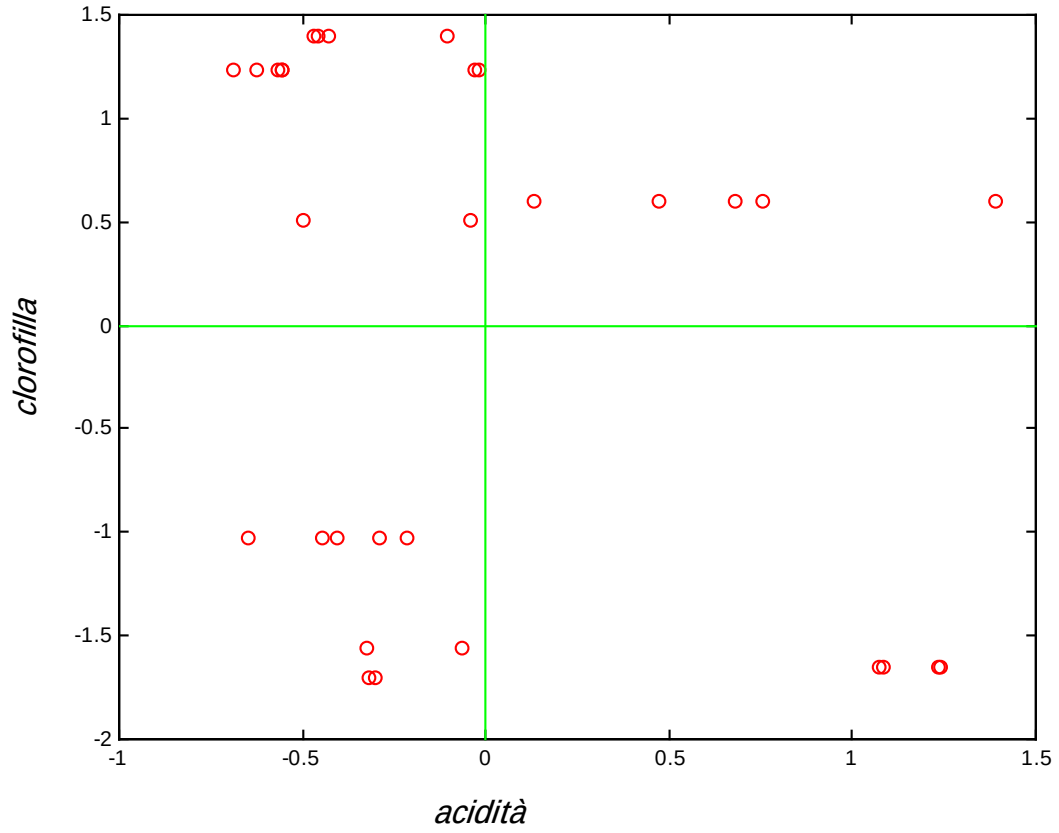
qualsiasi vettore aleatorio Y la cui densita' e' data da:

$$f_{\mathbf{Y}}(y) = \|\mathrm{C}|^{\frac{1}{2}}(2\pi)^{-p/2}\exp\left(-\frac{1}{2}\quad[y-\mu]\mathrm{C}\right)$$

(con C definita positiva di rango p ) e' distribuito secondo una normale multivariata di parametri μ e $\Sigma = \mathrm{C}^{-1}$.

# Esempio:
# distribuzione bivariata

Acidità e clorofilla in 31 pesche
Dati normalizzati a media nulla




$$\sum = \begin{bmatrix} 0.4167 & -0.2635 \\ -0.2635 & 1.5868 \end{bmatrix}$$

# Curve isoprobabilità

- Data una distribuzione bivariata normale i luoghi di punti (x,y) isoprobabili sono ellissi definite dalla seguente forma quadratica

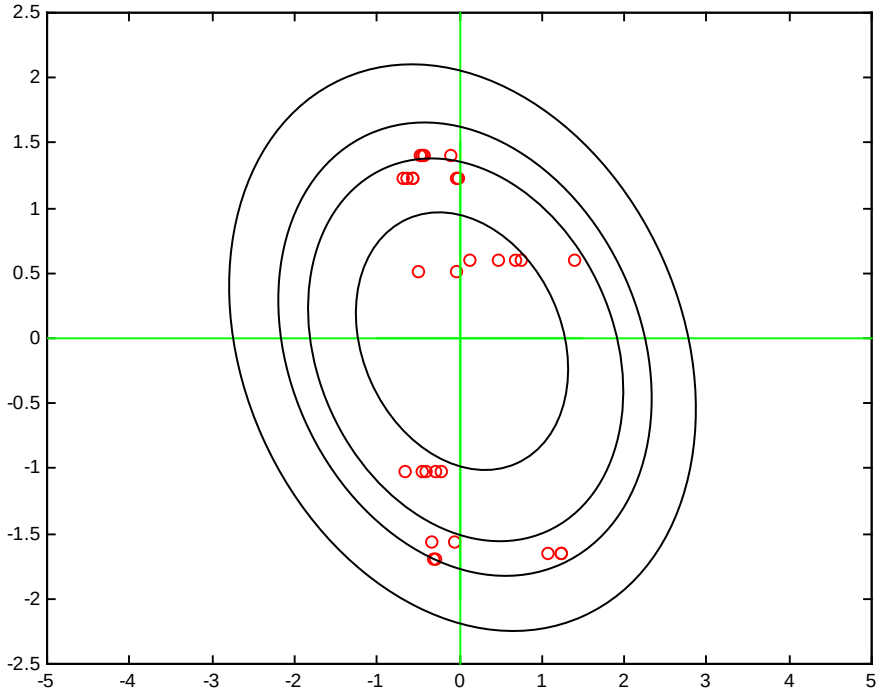
$$\begin{bmatrix} x-\bar{x} & y-\bar{y} \end{bmatrix} \cdot \begin{vmatrix} a & b \\ b & c \end{vmatrix} \cdot \begin{vmatrix} x-\bar{x} \\ y-\bar{y} \end{vmatrix} = k$$

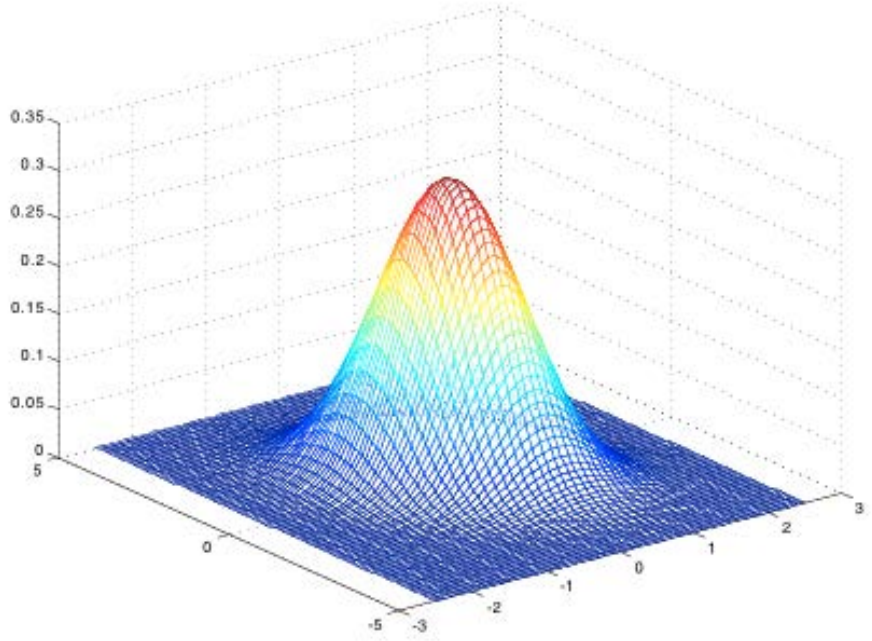
- Dove la matrice centrale è una matrice simmetrica come la matrice di covarianza
- Sviluppando la forma quadratica si ottiene l'equazione di una quadrica
  - Per semplicità consideriamo nulla la media

$$a \cdot x^2 + 2 \cdot b \cdot x \cdot y + c \cdot y^2 = k$$

- Per vari valori di k si ottengono le curve di livello

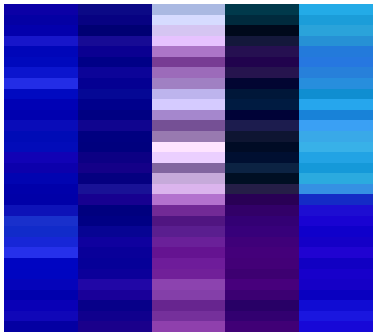
# Forma della distribuzione

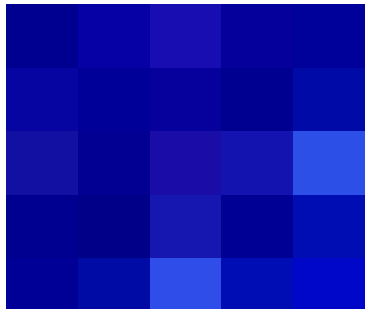


# Normalizzazioni e matrici di autocorrelazione

X

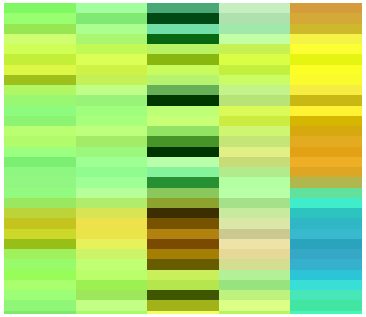


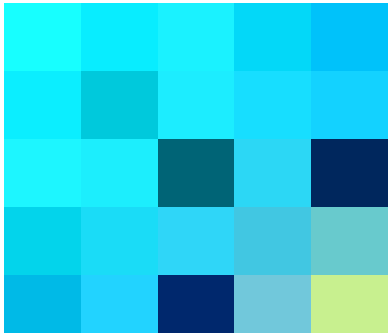


Dispersione

$X^T X$

A
Media nulla

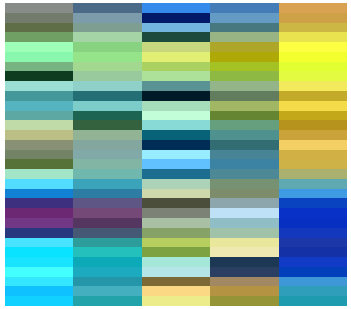




Covarianza

$A^T A$

Z
autoscaled



Correlazione

$Z^T Z$

# Il problema della Normalizzazione

- I dati possono essere “scalati” per:
  - Ridurre l’eccesso di informazione in una variabile
  - Eliminare le differenze numeriche tra variabili
  - Eliminare le differenze di unità di misura
- Due importanti normalizzazioni
  - Media nulla
    - Ogni variabile viene scalata attorno al valore di 0, in modo che la media di ogni colonna sia nulla
  - Media nulla e varianza unitaria
    - La varianza di ogni variabile viene resa pari a uno
    - Quest’operazione è anche detta (autoscaling)

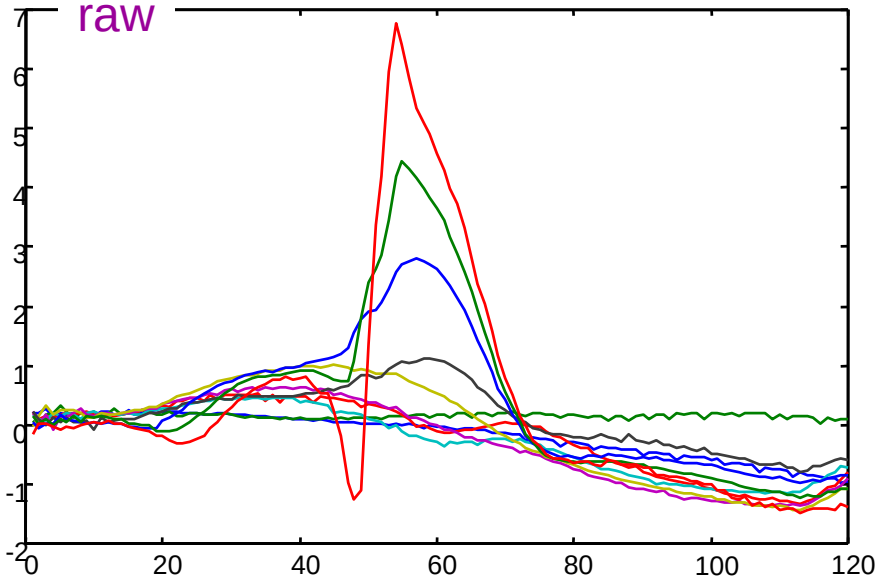
Matrice data $X$

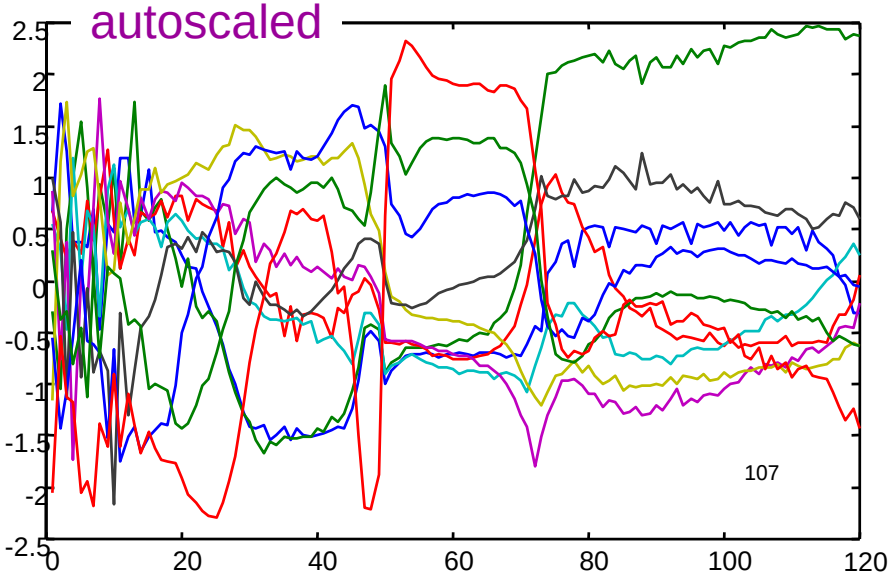
Media nulla $A_{ij} = X_{ij} - m_j$

Autoscaled $Z_{ij} = \frac{X_{ij} - m_j}{\sigma_j}$

# Il problema della normalizzazione

- Abbiamo già introdotto la normalizzazione come operazione che riduce le colonne (features) della matrice a media nulla o autoscalate (media nulla e varianza unitaria).
- L'uso dell'autoscaling da lo stesso peso ad ogni features, la cosa è buona se si è certi che ogni features indipendentemente dal suo valore numerico ha la stessa importanza nel problema.
- L'autoscaling diventa pericoloso quando una o più features sono rumorose oppure quando le relazioni numeriche tra features sono importanti
  - Caso tipico è quello degli spettri dove l'autoscaling distrugge completamente l'informazione



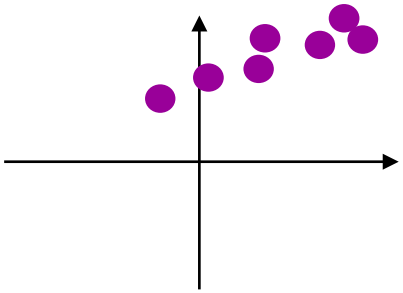



# Normalizzazione e Pattern Recognition

- L'autoscaling può modifica, anche profondamente, I risultati della pattern recognition

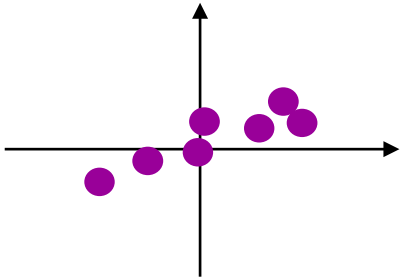
*raw*

$X$



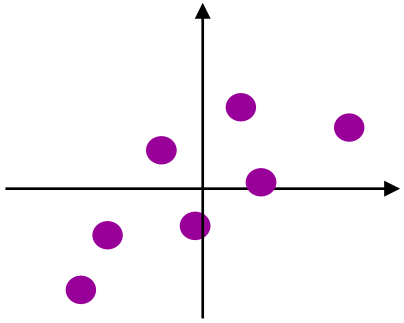
*centered*

$G = X - \mu$

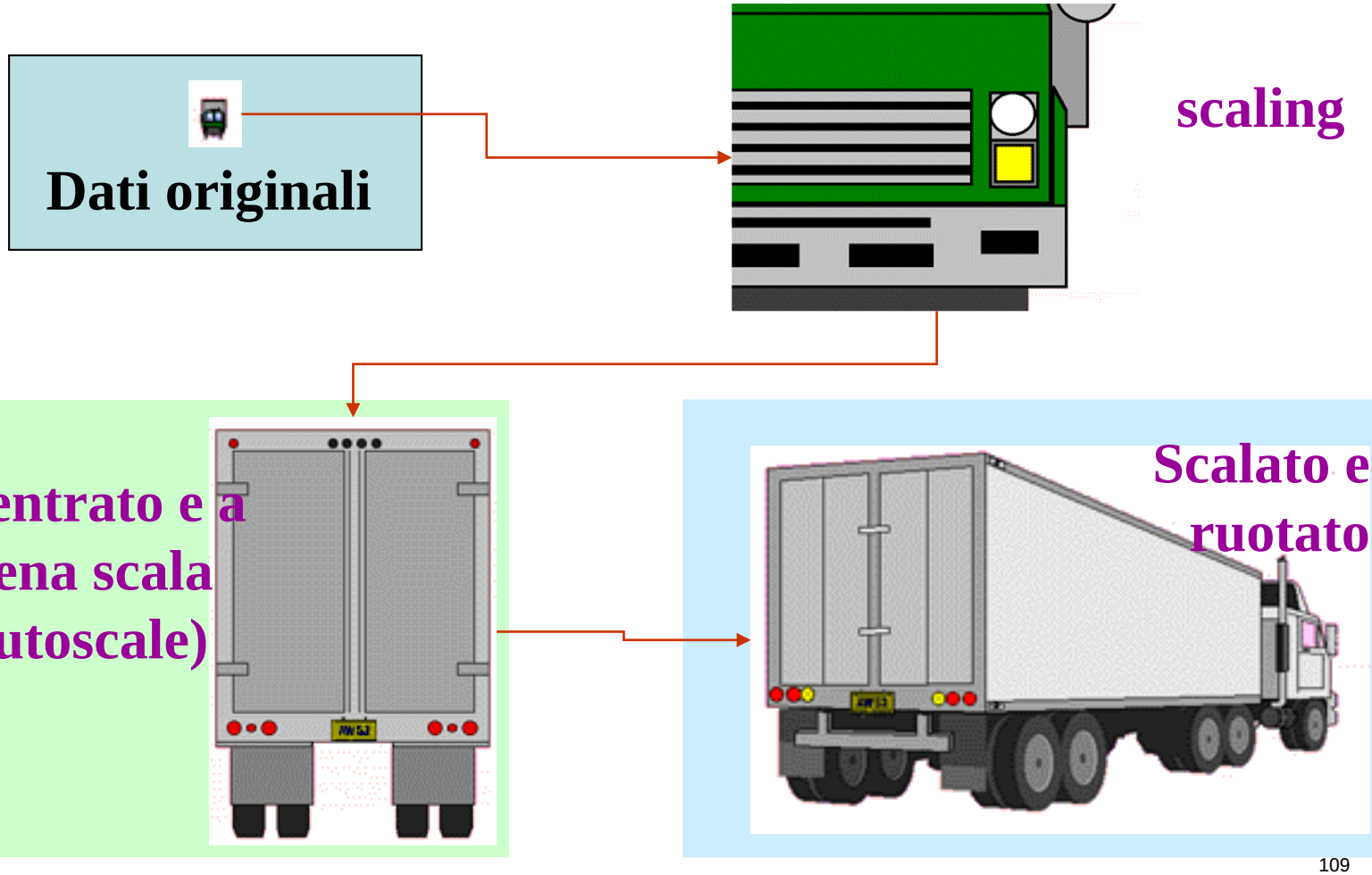


*autoscaled*

$$Z = \frac{X - \mu}{\sigma}$$



# Normalizzazione e rappresentazione: questione del punto di vista



# Spazi di rappresentazione

- Il problema della colinearità:
  - Le features utilizzate per descrivere un fenomeno non sono assolutamente scorrelate tra loro.
    - Esempio: peso ed altezza degli individui
  - A causa della variabilità del campione la correlazione naturale tra grandezze viene accentuata
    - La correlazione tra antociani e clorofilla è propria dei frutti ed è regolata dal processo di maturazione
- Come abbiamo già visto, la correlazione tra elementi del vettore delle features fa si che lo spazio delle features non sia “completamente” occupato ma che I pattern tendono a giacere in sottospazi
- E’ quindi opportuno studiare dei metodi di riduzione delle variabili che consentono:
  - Di individuare le variabili più significative
  - Di ridurre le dimensioni del problema
  - Di rappresentare pattern multidimensionali in spazi visualizzabili (2 o 3 D).