Significatività ed analisi degli errori

- Quanto il **valore sperimentale** è vicino al **valore vero**?
  - Campioni replicati: si esclude a priori che venga eseguita una sola misurazione
- Quanto il valore sperimentale è riproducibile da operatori diversi in tempi diversi?
  - **Validazione**: accertare che una metodologia per un determinato campione dia risultati comparabili per la maggior parte degli analisti

- Fonti di errore
  - Errori nei prelievi di volume o nelle misure di massa
  - Contaminazione dei reagenti
  - Contaminazione incrociata

Valore medio e deviazione standard

- **Media aritmetica** $\bar{x} = \frac{\sum_i x_i}{N}$

  x_i : risultato della i-esima misurazione
  N : numero totale delle misurazioni

- Deviazione dalla media $d_i = x_i - \bar{x}$
- Varianza $\sigma = \frac{\sum_i (x_i - \bar{x})^2}{N-1}$
- **Deviazione standard** $s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N-1}}$

  è la grandezza statisticamente più significativa per valutare la riproducibilità di una serie di misurazioni

- Deviazione standard relativa $RSD = \frac{s}{\bar{x}}$

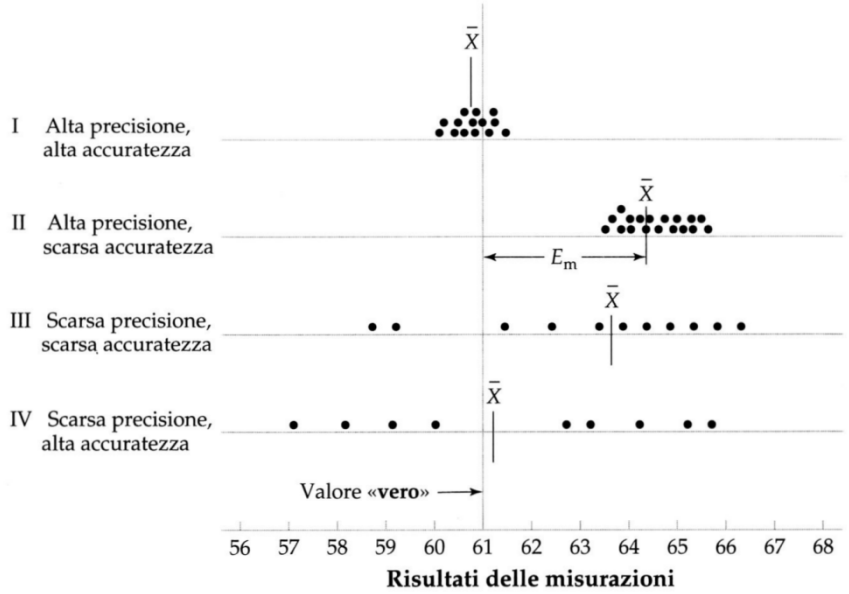
  numero puro adimensionale

- Deviazione standard relativa percentuale $RSD\% = \frac{s}{\bar{x}} \times 100$

  viene anche detta coefficiente di variazione

Precisione ed accuratezza

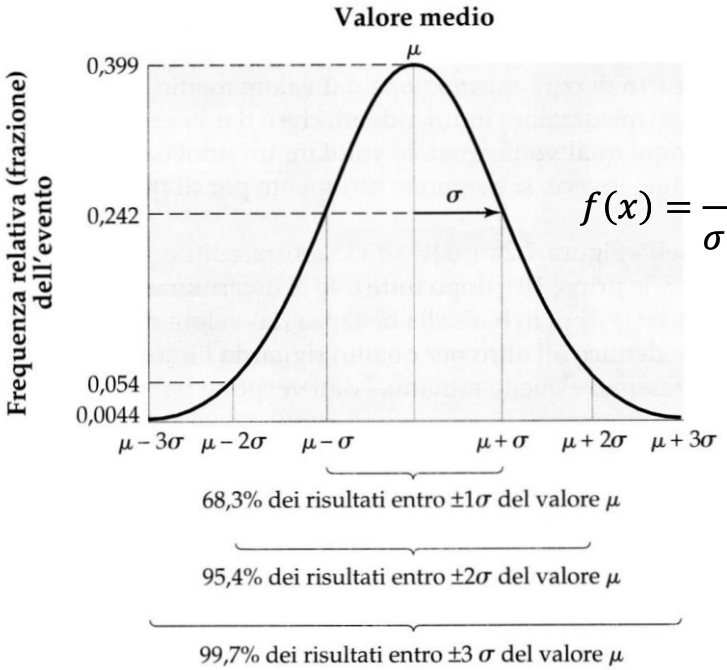
- **Precisione:** si riferisce alla ripetibilità di una misura; quando i dati sperimentali sono vicini fra loro si dice che le misure sono caratterizzate da una elevata precisione
- **Accuratezza:** si riferisce alla vicinanza della media delle misure rispetto al valore vero



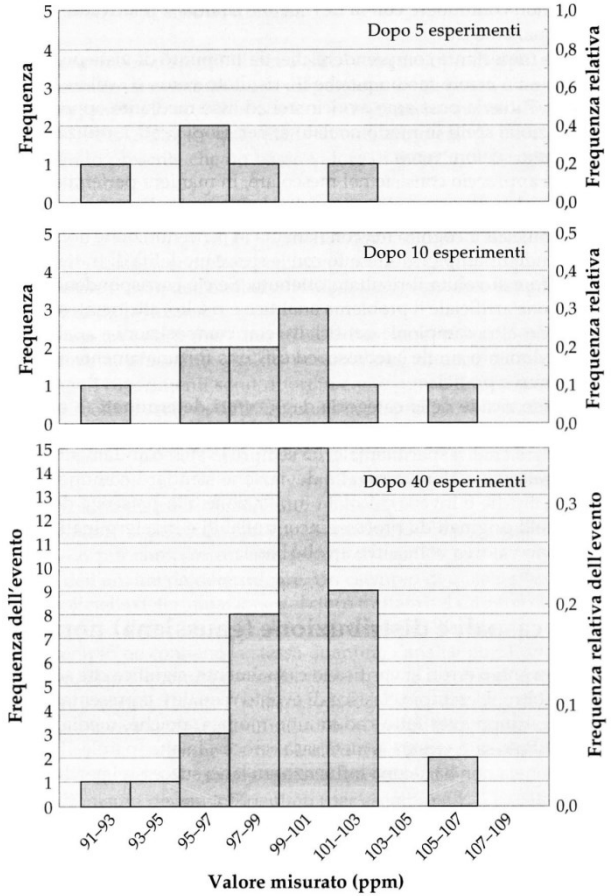
- **Errori sistematici:** sono la causa della differenza fra valore medio misurato e valore vero (accuratezza); sono errori determinati e talvolta controllabili
- **Errori casuali:** determinano la deviazione standard in quanto sono originati da processi incontrollabili ed indeterminati

Errori casuali e distribuzione gaussiana

- Quando gli errori sono casuali tendono a distribuirsi in modo caratteristico su entrambi i lati del valore medio
- Supponendo di avere a che fare con un numero infinito di misurazioni o replicati la distribuzione degli errori casuali intorno al valore medio può essere descritta dalla **curva gaussiana dell'errore** (curva di distribuzione normale)



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

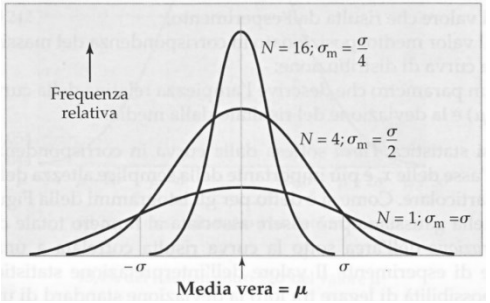


- Per un numero elevato di prove è possibile correlare la deviazione standard s di numero finito di prove, con σ riferito ad numero infinito di prove

Intervallo di fiducia

- Aumentando il numero di replicati si ha sempre maggiore certezza che i **risultati successivi** siano compresi nell'intervallo dei risultati fino a quel momento ottenuti

σ = deviazione standard del singolo risultato
σ_m = deviazione standard del valore medio



- Si assume che: $\sigma_m = \frac{\sigma}{\sqrt{N}}$

- L'**intervallo di fiducia** esprime quantitativamente quanto sopra esposto qualitativamente

$$\mu = \bar{x} \pm \frac{z}{\sqrt{N}}$$

$z = 1.96 \cdot \sigma_m$

quando σ è conosciuto, per un intervallo di fiducia del 95%; si osservi che $\pm 2 \cdot \sigma_m$ esprime una probabilità del 95.4% mentre $\pm 1.96 \cdot \sigma_m$ una del 95.0%

$z = t \cdot s$

quando σ non è conosciuto, si utilizza il fattore t di Student e si sostituisce s a σ; si osservi che all'aumentare del numero delle misurazioni non importa conoscere σ perché gli errori casuali tendono a compensarsi giungendo allo stesso limite

t di Student

N	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
1	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
100	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
∞	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

- In teoria delle probabilità la distribuzione di Student, o t di Student, è una distribuzione di probabilità continua che governa il rapporto tra due variabili aleatorie, la prima con distribuzione normale e la seconda il cui quadrato ha distribuzione chi quadrato.
- Questa distribuzione interviene nella stima della media di una popolazione che segue la distribuzione normale, e viene utilizzata negli omonimi test t di Student per la significatività e per gli intervalli di confidenza della differenza tra due medie.
- Il t di Student consente di valutare la significatività statistica

$$\pm t = \frac{\mu - \bar{x}}{s} \times \sqrt{N}$$

Differenza fra gruppi di dati

- Per ogni gruppo devono essere noti:

  $\bar{x}$: il valore medio

  s : la deviazione standard

  N : il numero di misurazioni effettuate

- Le deviazioni standard devono essere paragonabili; a questo provvede il test di Fisher volto a verificare l'ipotesi che due popolazioni normali abbiano la stessa varianza contro l'ipotesi alternativa che le varianze siano diverse; deve verificarsi che : $F_{tabulato} < F_{calcolato} = \frac{s_{maggiore}}{s_{minore}}$

$$\bar{x}_{combinata} = \frac{N_A \cdot \bar{x}_A + N_B \cdot \bar{x}_B}{N_A + N_B}$$

$$s_{combinata} = \sqrt{\frac{(N_A - 1) \cdot {s_A}^2 + (N_B - 1) \cdot {s_B}^2}{N_A + N_B - 2}}$$

	Gruppo A	Gruppo B
$\bar{x}$	4.35	4.47
s	0.07	0.05
N	6	8

$$\bar{x}_{combinata} = \frac{6 \cdot 4.35 + 8 \cdot 4.47}{6 + 8} = 4.42$$

$$s_{combinata} = \sqrt{\frac{(6-1) \cdot 0.07^2 + (8-1) \cdot 0.05^2}{6 + 8 - 2}} = 0.059$$

Differenza fra due medie

- Si può credere alle differenze apparenti tra i due gruppi di dati oppure sono da attribuire ad errori casuali?
- Si calcola il **t di Student** che permette di stabilire se le due medie sono realmente diverse

$$t_{calcolato} = \frac{|\bar{x}_A - \bar{x}_B|}{s_{combinata}} \times \sqrt{\frac{N_A \times N_B}{N_A + N_B}}$$

- Se il valore calcolato $t_{calcolato}$ è maggiore del $t_{tabulato}$ per N-2 gradi di libertà e per il livello di probabilità desiderato, si può concludere che **a quel livello di fiducia le medie sono diverse**

- E.g. determinazione del Cr in un corso d'acqua, in seguito a campionamenti effettuati prima e dopo una fuoriuscita di materiale inquinante

	Prima	Dopo
$\bar{x}$	0,95 ppb	1,10 ppb
s	0,05	0,08
N	5	6

$$s_{combinata} = \sqrt{\frac{(5-1)\cdot 0.05^2 + (6-1)\cdot 0.08^2}{5+6-2}} = 0.068$$

$$t_{calcolato} = \frac{|0{,}95 - 1{,}10|}{0{,}068} \times \sqrt{\frac{5 \times 6}{5+6}} = 3{,}64$$

$$t_{tabulato}\{N = 9; 95\%\} = 1{,}833$$

- quindi al 95% di probabilità i valori medi determinati sono diversi, ovvero il Cr è aumentato

Confronto fra metodi diversi (test t a coppie)

- Permette di confrontare due diversi metodi analitici applicati allo stesso insieme di campioni
- Si calcola il **t di Student** dalla differenza media tra i risultati ottenuti con i due metodi

$$t_{calcolato} = \frac{\overline{D}}{s_D} \times \sqrt{N}$$

$$s_D = \sqrt{\frac{\sum_i (d_i - \overline{D})^2}{N-1}}$$

- Se il valore calcolato $t_{calcolato}$ è minore del $t_{tabulato}$ per il livello di probabilità desiderato, si può concludere che **a quel livello di fiducia i metodi danno risultati statisticamente coincidenti**

Campione	Metodo 1	Metodo 2	d_i
A	3,34	3,36	-0,02
B	5,19	5,13	0,06
C	3,06	3,05	0,01
D	9,33	9,43	-0,10
E	3,80	3,83	-0,03
F	7,47	7,55	-0,08
		Valore medio $\overline{D}$	-0,027

- E.g. valutazione di un metodo alternativo di preparazione del campione

$$s_D = \sqrt{\frac{(d_i + 0{,}027)}{5}} = 0{,}058$$

$$t_{calcolato} = \frac{|-0{,}027|}{0{,}058} \times \sqrt{6} = 1{,}14$$

$$t_{tabulato}\{N = 6; 95\%\} = 1{,}943$$

- quindi al 95% di probabilità i due metodi danno risultati coincidenti, ovvero non ci sono differenze significative

Confronto di una media sperimentale con un valore vero

- Permette di confrontare il valore medio ottenuto sperimentalmente con il valore vero ed è utile per validare una tecnica analitica: si utilizza l'errore medio $E_m = \left|\bar{x}_{sperimentale} - x_{vero}\right|$

$$t_{calcolato} = \frac{E_m}{s_{combinata}} \times \sqrt{N_{totali}}$$

$$s_{combinata} = \sqrt{\frac{(N_A - 1) \cdot {s_A}^2 + (N_B - 1) \cdot {s_B}^2}{N_A + N_B - 2}}$$

- E.g. determinazione del contenuto di Ni in una lega metallica dove il valore vero è pari a 4.44; valutare se i valori medi ottenuti da due laboratori sono statisticamente (al 95%) diversi dal vero

	Laboratorio A	Laboratorio B
$\bar{x}$	4.35	4.47
s	0.07	0.05
N	6	8

$$\bar{x}_{combinata} = \frac{6 \cdot 4.35 + 8 \cdot 4.47}{6 + 8} = 4.42$$

$$s_{combinata} = \sqrt{\frac{(6-1) \cdot 0.07^2 + (8-1) \cdot 0.05^2}{6 + 8 - 2}} = 0.059$$

$$t_{calcolato} = \frac{|4.42 - 4.44|}{0{,}059} \times \sqrt{14} = 1{,}27$$

$$t_{tabulato}\{N = 12; 95\%\} = 1{,}782$$

- essendo $t_{calcolato}$ più basso si può concludere che i due valori sono statisticamente identici.

Propagazione dell'incertezza

- Maggiore è il numero delle operazioni previste dal metodo analitico maggiore è l'incertezza sul risultato perché ogni passaggio contribuisce a rendere impreciso e/o inaccurato il risultato finale
- La propagazione dell'errore è calcolata attraverso il differenziale totale della funzione algebrica impiegata per calcolare il contenuto di analita in un campione

$$\Delta F = \frac{\partial F}{\partial a_1}\Delta a_1 + \frac{\partial F}{\partial a_2}\Delta a_2 + \cdots + \frac{\partial F}{\partial aN}\Delta aN$$

- E.g. in una determinazione il risultato è dato da:

$$concentrazione = \frac{X^2 \cdot Y}{Z} = F$$

l'errore assoluto è dato da:

$$\Delta\left[\frac{X^2 \cdot Y}{Z}\right] = 2\frac{X \cdot Y}{Z}\Delta X + \frac{X^2}{Z}\Delta Y - \frac{X^2 \cdot Y}{Z^2}\Delta Z = 2F\frac{\Delta X}{X} + F\frac{\Delta Y}{Y} - F\frac{\Delta Z}{Z}$$

mentre l'errore relativo è dato da:

$$\frac{\Delta F}{F} = 2\frac{\Delta X}{X} + \frac{\Delta Y}{Y} - \frac{\Delta Z}{Z}$$

Cifre significative

- La convenzione generale è che l'errore di una misura è rappresentato dall'errore esclusivamente sull'ultima cifra; se e.g. è noto che il risultato di una analisi abbia una precisione relativa dello 0.1% non avrebbe senso esprimerlo come 23.4522687 ppm ma andrebbe riportato considerando che 23.4522687 × 0.001 = 0.02 e quindi: 23.45 ± 0.02 ppm
- L'arrotondamento di un risultato deriva dall'analisi dell'errore e può essere così riassunto:
  - Non si devono mantenere cifre oltre la prima incerta
  - Se il numero situato oltre l'ultima cifra significativa risulta minore di 5 si mantiene il numero inalterato
  - Se la cifra su cui ricade l'incertezza è maggiore o uguale a 6 si arrotonda per eccesso
  - Se la prima cifra dopo l'ultima significativa è un 5 si arrotonda alla cifra pari più vicina
- Nelle operazioni matematiche si utilizzano tutte le cifre a disposizione e alla fine si mantiene un numero di decimali pari a quello del termine col minor numero di cifre dopo la virgola

$$\begin{array}{r} 21.2 \\ 3.035 \\ 0.12 \\ \hline 24.355 \end{array} \rightarrow 24.4$$

$$2.0 \times 43 = 86 \pm 4$$

$$2.0 \pm 0.1 \rightarrow 5\%$$

$$\begin{array}{ccccc} 31.1 & \times & 0.063 & \times & 98.9 = 193.77477 \\ \uparrow & & \uparrow & & \uparrow \\ 1/300 & & 1/60 & & 1/100 \end{array}$$

$$\rightarrow 1.9 \times 10^2$$

Dati discordi

- Un dato che differisca in maniera evidente da tutti gli altri (**outlier**) può essere ignorato oppure va considerato al pari degli altri?
- La risposta a questa domanda viene data sia su **base statistica** che da una **valutazione soggettiva**

7.06	7.04	7.19	7.10	7.02	7.09

- Nella serie di misurazioni riportata, il valore 7.19 sembrerebbe essere discordante; per decidere se scartare il dato si ricorre al **test di Dixon**

$$Q_n = \frac{|x_{outlier} - x_i|}{x_{max} - x_{min}}$$

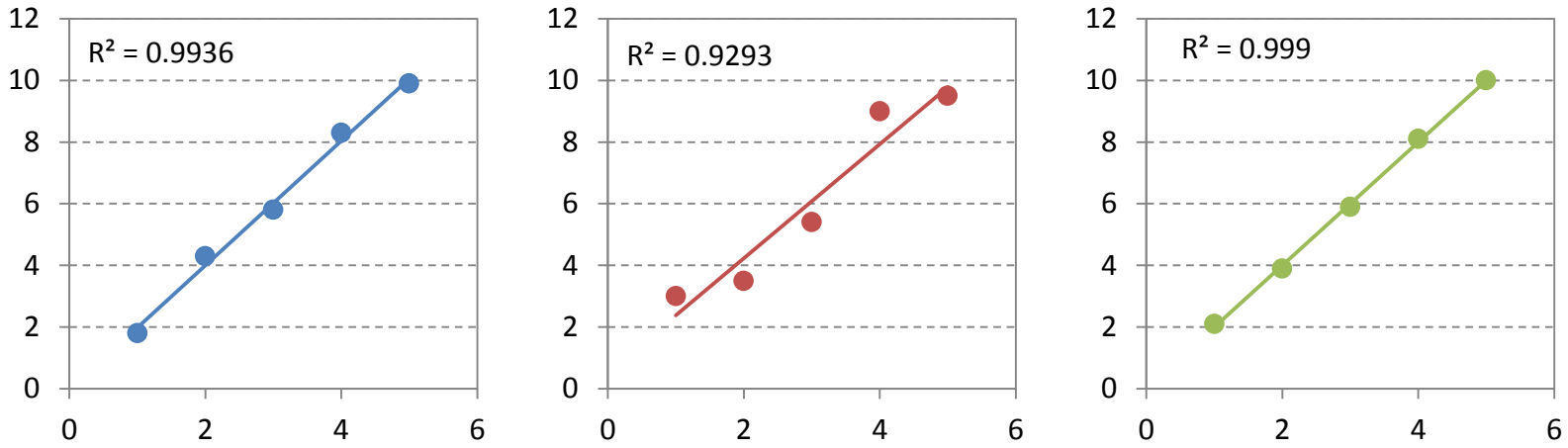
$$\frac{\text{differenza fra il valore sospetto e quello ad esso più vicino}}{\text{differenza tra il valore massimo ed il valore minimo}}$$

N	3	4	5	6	7	8	9	10
$Q_{90\%}$	0.941	0.765	0.642	0.560	0.507	0.468	0.437	0.412
$Q_{95\%}$	0.970	0.829	0.710	0.625	0.568	0.526	0.493	0.466
$Q_{99\%}$	0.994	0.926	0.821	0.740	0.680	0.634	0.598	0.568

- Nel caso in esame $Q_6 = \frac{7.19-7.10}{7.19-7.02} = 0.529$; quindi poiché Q_6=0.560 al 90% di fiducia il dato 7.19 non può essere scartato

Minimi quadrati

- Quando si ha a che fare con campioni reali non esistono casi in cui si abbia una relazione perfettamente lineare fra risposta strumentale e concentrazione, quindi si cerca di trovare quella che meglio interpola i punti sperimentali
- In pratica si opera ricercando la retta che rende minima la somma dei quadrati degli scarti; questo equivale a rendere minima la sommatoria: $\sum_i \left[y_{i\ (sperimentale)} - y_{i\ (retta)} \right]^2$



- La qualità della regressione viene valutata mediante il **coefficiente di correlazione R^2** che misura la frazione della variabilità delle osservazioni che siamo in grado di spiegare tramite il modello lineare; se la correlazione è perfetta $R^2=1$.
- L'R^2 non misura se effettivamente sussista una relazione (di qualsiasi tipo) tra le variabili, ma soltanto **fino a che punto un modello lineare consente di approssimare la realtà dei dati osservati**.

Analisi dei componenti principali

- L'analisi in componenti principali o PCA (principal component analysis) è una tecnica per la semplificazione dei dati utilizzata nell'ambito della **statistica multivariata**.
- Lo scopo primario di questa tecnica è la riduzione di un numero più o meno elevato di variabili (rappresentanti altrettante caratteristiche del fenomeno analizzato) in alcune variabili latenti. Ciò avviene tramite una **trasformazione lineare delle variabili che proietta quelle originarie in un nuovo sistema cartesiano** nel quale le variabili vengono ordinate in ordine decrescente di varianza: pertanto, la variabile con maggiore varianza viene proiettata sul primo asse, la seconda sul secondo asse e così via. La riduzione della complessità avviene limitandosi ad analizzare le principali (per varianza) tra le nuove variabili.

Campioni	Cu	Pb	Cr	As	Hg	PCB	BOD	TOC	…	p
1	1.34	4.50	0.02	9.81	5.12	1.11	5.91	4.44	…	0.01
2	2.18	3.12	0.15	2.87	1.09	5.10	4.73	9.12	…	0.02
3	0.16	8.19	0.83	3.14	0.08	2.84	7.29	3.68	…	0.03
…	…	…	…	…	…	…	…	…	…	…
n	1.52	0.22	0.04	7.34	1.56	0.77	8.27	9.97	…	0.00

- Il primo componente principale sarà quindi la combinazione lineare delle variabili:

$$Z_1 = a_{11} \cdot X_1 + a_{12} \cdot X_2 + a_{13} \cdot X_3 + \cdots + a_{1p} \cdot X_p$$

- E.g. : Z_1 = 0,45·[Cu] + 0,32·[Pb] - 0,18·[Cr] + 0,09·[As] - 0,87·[Hg] + … + 0,22·[p]

Analisi dei componenti principali

- Il secondo componente principale sarà anch'esso una combinazione lineare delle variabili, ovviamente con coefficienti diversi:

$$Z_2 = a_{21} \cdot X_1 + a_{22} \cdot X_2 + a_{23} \cdot X_3 + \cdots + a_{2p} \cdot X_p$$

- Ed analogamente il terzo, il quarto, fino a quello di indice n:

$$Z_n = a_{n1} \cdot X_1 + a_{n2} \cdot X_2 + a_{n3} \cdot X_3 + \cdots + a_{np} \cdot X_p$$

- Poiché i componenti principali sono ottenuti ordinati secondo la varianza spiegata decrescente, generalmente i primi due consentono di rappresentare efficacemente lo spazio dei dati