

## Determinazione di un segnale analitico in presenza di rumore distribuito in modo gaussiano

Ipotizzando che il responso  $y$  di un metodo analitico sia correlato linearmente alla concentrazione  $c$ , per una generica misura del responso si può scrivere la relazione:

$$y = \beta_0 + \beta_1 c + \varepsilon$$

dove  $\varepsilon$  è l'errore random associato alla misura.

Nell'ipotesi che la deviazione standard dei dati non sia dipendente dalla concentrazione e che l'errore sia distribuito in modo gaussiano, si ha:

$\varepsilon \sim N(0, \sigma_y)$ , e quindi si può scrivere anche:

$$y \sim N(\beta_0 + \beta_1 c, \sigma_y)$$

In definitiva, i valori del segnale  $Y$  ottenuti alle varie concentrazioni sono distribuiti in modo gaussiano intorno ad un valore che dipende linearmente dalla concentrazione ( $\beta_0 + \beta_1 c$ ) ed è caratterizzato da una deviazione standard pari a  $\sigma_y$ .

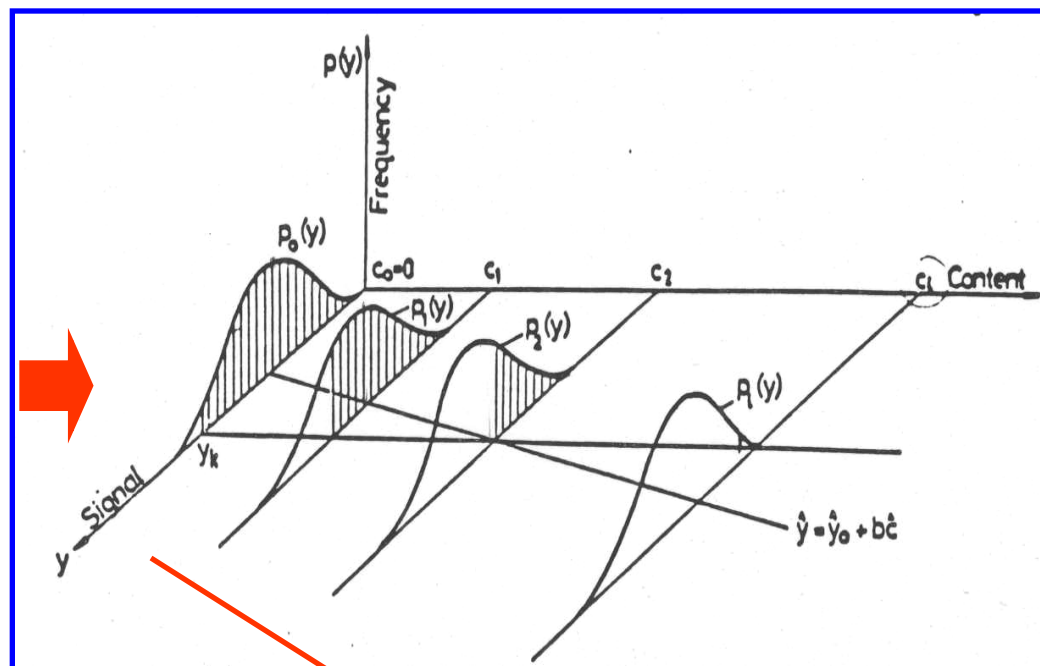
Quando si considera una soluzione in tutto simile a quelle standard impiegate per realizzare la retta di calibrazione "Y contro c" ma priva dell'analita ( $c = 0$ ), ossia la **soluzione di bianco**, il segnale registrato sarà:

$$y = \beta_0 + \varepsilon = y_0$$

Dunque anche il valore del segnale associato alla soluzione di bianco avrà una sua distribuzione.

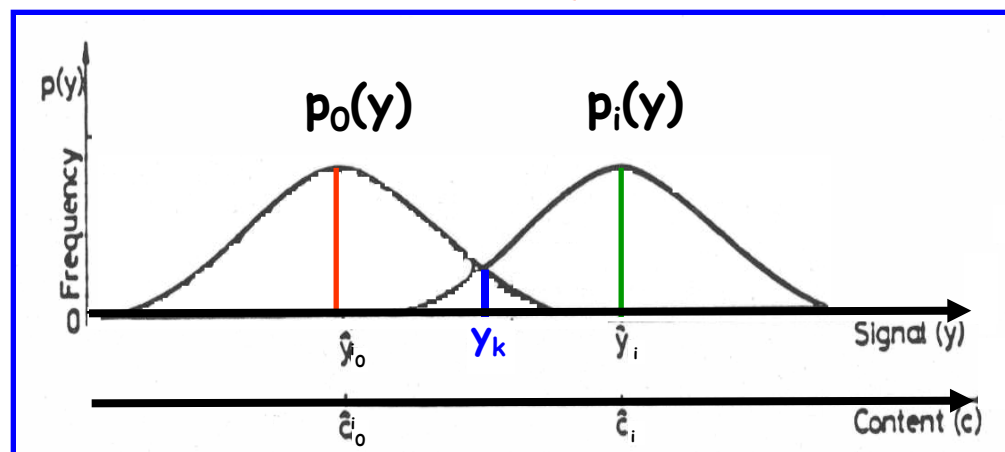
Nella terminologia della chimica analitica strumentale **la grandezza  $\varepsilon$  viene definita "rumore"**, ad indicare la variabilità del segnale che uno strumento fornisce anche in assenza di analita in grado di generarlo (quindi quello correlato al solvente puro o alla matrice priva dell'analita di interesse).

Indicando con  $p_0(y)$  la funzione densita' di probabilita' relativa ai valori del segnale del bianco e con  $p_i(y)$  quelle dei segnali corrispondenti alle varie concentrazioni  $c_i$ , si puo' usare una rappresentazione tri-dimensionale per visualizzare la situazione piu' generale (ossia quella in cui  $\beta_0 \neq 0$ ):



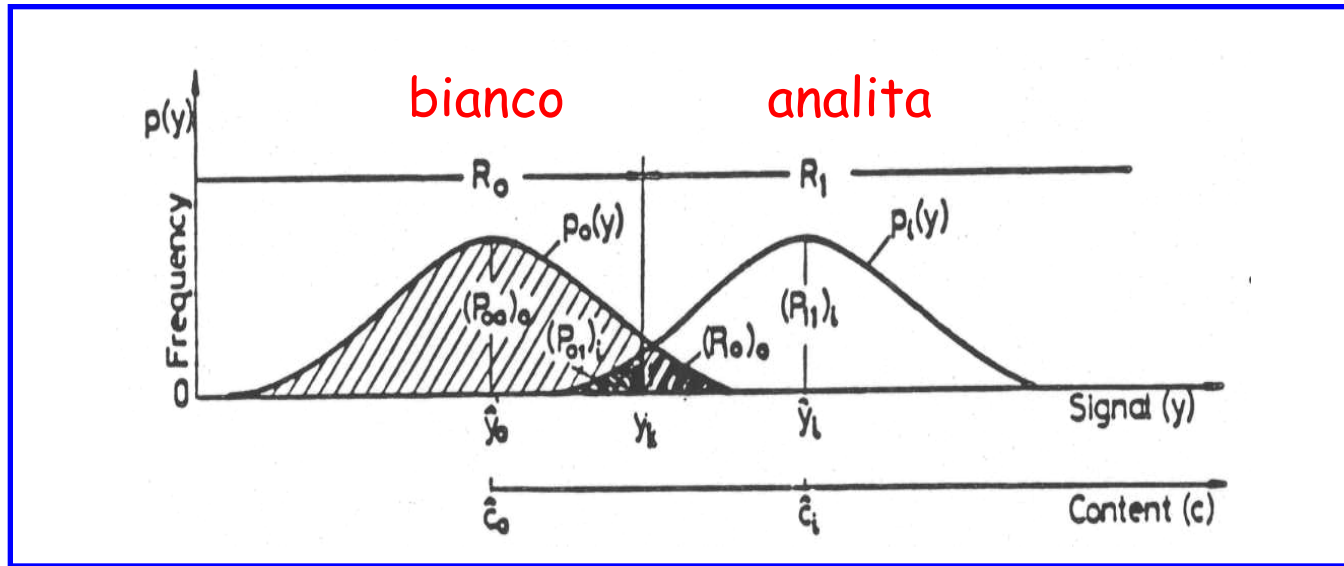
Se si guardasse la figura da sinistra a destra, in direzione perpendicolare all'asse che rappresenta il segnale  $y$ , si osserverebbe questa immagine:

più elevata è la concentrazione  $c_i$ , minore è la sovrapposizione fra le curve di densità di probabilità  $p_0(y)$  e  $p_i(y)$ .



## Soglia di decisione e rapporto segnale/rumore

La **decisione sull'attribuzione di un segnale all'analita o al bianco** sulla base delle rispettive densità di probabilità è una procedura analoga al **confronto fra due medie**:



Anche in questo caso si deve definire un **valore critico del segnale**, definito **soglia di decisione**,  $y_k$ , che suddivide l'intervallo dei valori di  $y$  in due regioni:

$R_0$  è la regione in cui il segnale si ritiene derivante solo dal bianco ( $y \leq y_k$ )

$R_1$  è la regione in cui si può ritenere il segnale derivante dall'analita ( $y > y_k$ )

Detta  $\sigma_y$  la deviazione standard caratteristica delle due distribuzioni, supposta uguale per entrambe, si definisce rapporto segnale/rumore la quantità:

$$r_{(S/N)} = (y_i - y_0) / \sigma_y$$

Tale rapporto, a parità di rumore, sarà tanto più elevato quanto maggiore sarà la concentrazione  $c_i$ , e quindi il segnale  $y_i$ .

Le probabilità corrispondenti alle diverse decisioni possibili possono essere calcolate mediante opportuni integrali delle due funzioni densità di probabilità disponibili.

Si usa il seguente simbolismo per indicare le diverse probabilità:

$(P_{xy})_z$  dove:

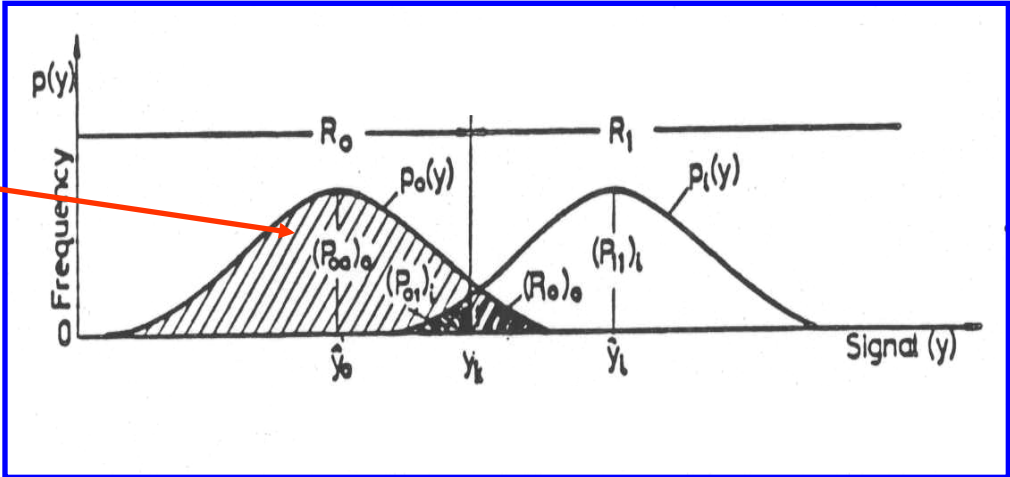
$x = 0/1$  rappresenta l'affermazione che l'analita non sia/sia presente;

$y = 0/1$  indica che realmente l'analita non sia/sia presente;

$z = 0/i$  indica che si usa la funzione  $p_0(y)$  o  $p_i(y)$  rispettivamente.

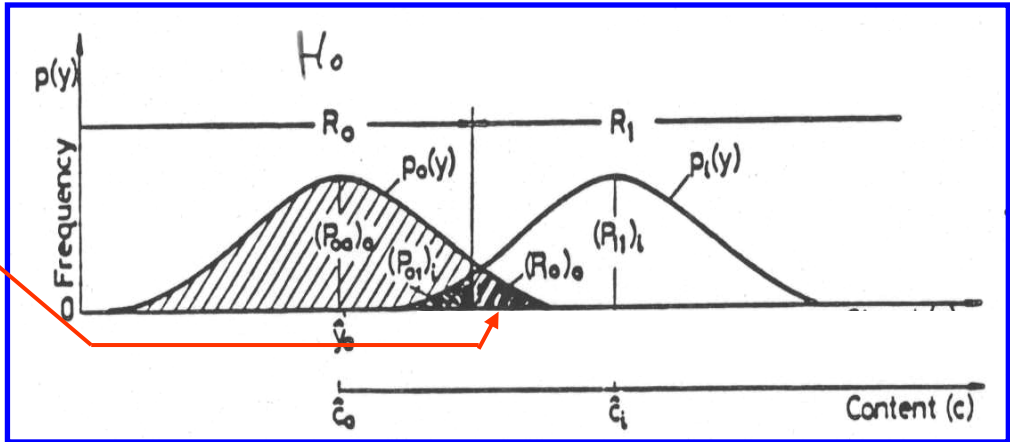
$(P_{00})_0$ : probabilità di affermare che l'analita sia assente quando ciò è vero

$$(P_{00})_0 = \int_{-\infty}^{y_k} p_0(y) dy$$



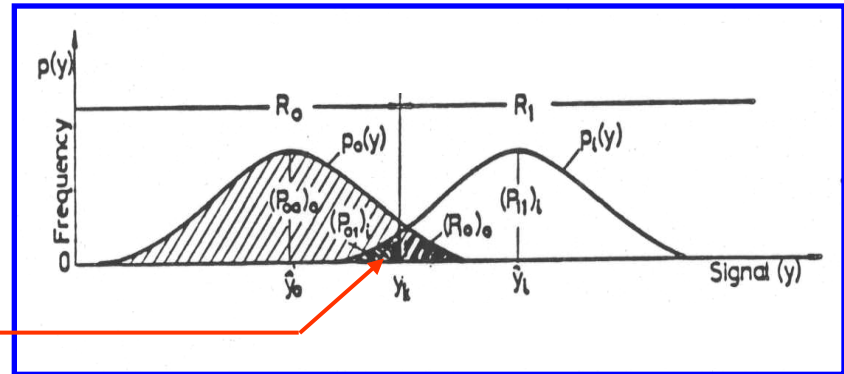
$(P_{10})_0$ : probabilità di affermare che l'analita sia presente quando ciò NON è vero (ossia il segnale deriva comunque dal bianco)

$$(P_{10})_0 = \int_{y_k}^{+\infty} p_0(y) dy$$



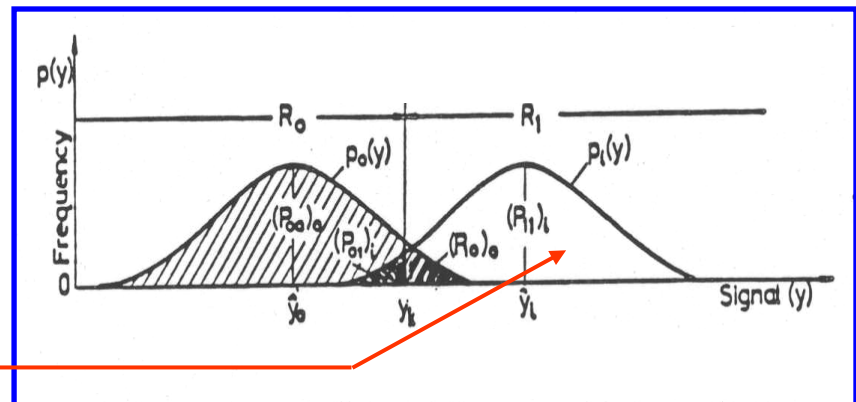
$(P_{01})_i$ : probabilità di affermare che l'analita NON sia presente quando invece lo è

$$(P_{01})_i = \int_{-\infty}^{y_k} p_i(y) dy$$



$(P_{11})_i$ : probabilità di affermare che l'analita sia presente quando ciò è vero

$$(P_{11})_i = \int_{y_k}^{+\infty} p_i(y) dy$$



## Criterio di Neyman-Pearson per la soglia di decisione

Il **criterio di Neyman-Pearson** consiste nel determinare la soglia di decisione in base alla **probabilità di falsa rivelazione**, ossia la probabilità di stabilire che l'analita sia presente quando ciò non è vero, corrispondente alla grandezza  $(P_{10})_0$ .

Poiché si è supposto che le funzioni di densità di probabilità  $p(y)$  siano gaussiane, la funzione  $p_0(y)$  avrà la forma:

$$p_0(y) = \frac{1}{\sigma_y \sqrt{2\pi}} \exp \left[ -\frac{(y - y_0)^2}{2\sigma_y^2} \right]$$

e quindi:

$$(P_{10})_0 = \frac{1}{\sigma_y \sqrt{2\pi}} \int_{y_k}^{+\infty} \exp \left[ -\frac{(y - y_0)^2}{2\sigma_y^2} \right] dy$$

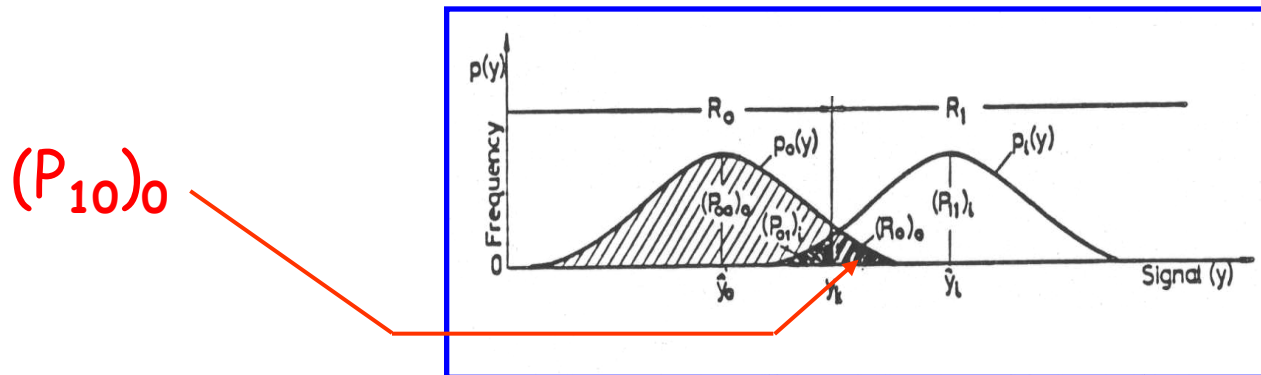


Se si introduce la **variabile normale standardizzata**  $z = (y-y_0)/\sigma_y$  si può scrivere:

$$(P_{10})_0 = \frac{1}{\sqrt{2\pi}} \int_{z_k}^{+\infty} \exp\left[-\frac{z^2}{2}\right] dz$$

dove  $z_k = (y_k - y_0)/\sigma_y$  rappresenta il valore standardizzato corrispondente alla soglia di decisione ed è, di fatto, l'incognita dell'equazione ora scritta, in quanto l'integrale, se sviluppato, è funzione di  $z_k$ .

Per valori di  $(P_{10})_0$  via via più piccoli  $z_k$  aumenta e, automaticamente, anche  $y_k$  aumenta, dunque si sposta verso destra nel seguente grafico:



Fissato un valore per la probabilità  $(P_{10})_0$  si determina automaticamente il valore di soglia standardizzato  $z_k$  e la corrispondente **soglia di decisione**:

$$y_k = y_0 + z_k \sigma_y$$

La **regola su cui si basa la decisione** è dunque:

- ✓ se  $y \leq y_0 + z_k \sigma_y$  si stabilisce che **l'analita non è presente** nel campione;
- ✓ se  $y > y_0 + z_k \sigma_y$  si stabilisce che **l'analita è presente** nel campione.

**Se si replicano N misure sul campione** sottoposto al test le condizioni del criterio di Neyman-Pearson diventano:

se  $\bar{y} \leq y_0 + z_k \sigma_y / \sqrt{N}$   $\Rightarrow$  **l'ipotesi che l'analita sia assente può essere accettata**

se  $\bar{y} > y_0 + z_k \sigma_y / \sqrt{N}$   $\Rightarrow$  **l'ipotesi che l'analita sia presente può essere accettata**

Il criterio tiene conto dunque della **media dei valori** ottenuti dalle N misure del segnale sul campione sottoposto al test.

## Calcolo della soglia di decisione

Supponiamo di voler stabilire la soglia di decisione  $y_k$  corrispondente ad una probabilità di falsa rivelazione dell'1%, ossia a  $(P_{10})_0 = 0.01$ .

Tale probabilità corrisponde a considerare sulla curva di densità di probabilità normale standardizzata un valore di  $z$  corrispondente ad un'area sottesa pari al 99%, ossia  $z = 2.33$ .

Essendo  $z_k = 2.33$  la soglia di decisione, in termini di  $y_k$ , sarà:

$$y_k = y_0 + 2.33 \sigma_y$$

pertanto la regola di decisione sarà:

se:  $y \leq y_0 + 2.33 \sigma_y$  l'analita è assente

se:  $y > y_0 + 2.33 \sigma_y$  l'analita è presente

Nel caso di **N misure replicate** la soglia di decisione sarà invece:

$$y_k = y_0 + 2.33 \sigma_y / \sqrt{N}$$

e quindi la regola di decisione sarà:

se  $\bar{y} \leq y_0 + 2.33 \sigma_y / \sqrt{N}$  l'analita è assente

se  $\bar{y} > y_0 + 2.33 \sigma_y / \sqrt{N}$  l'analita è presente

## Limite di rivelabilità in chimica analitica

Per **limite di rivelabilità (LOD)** in chimica analitica si intende:

*"la minima concentrazione di analita che può essere rivelata ad un certo livello di fiducia (o ad un certo rapporto segnale/rumore, S/N) con un particolare metodo analitico strumentale".*

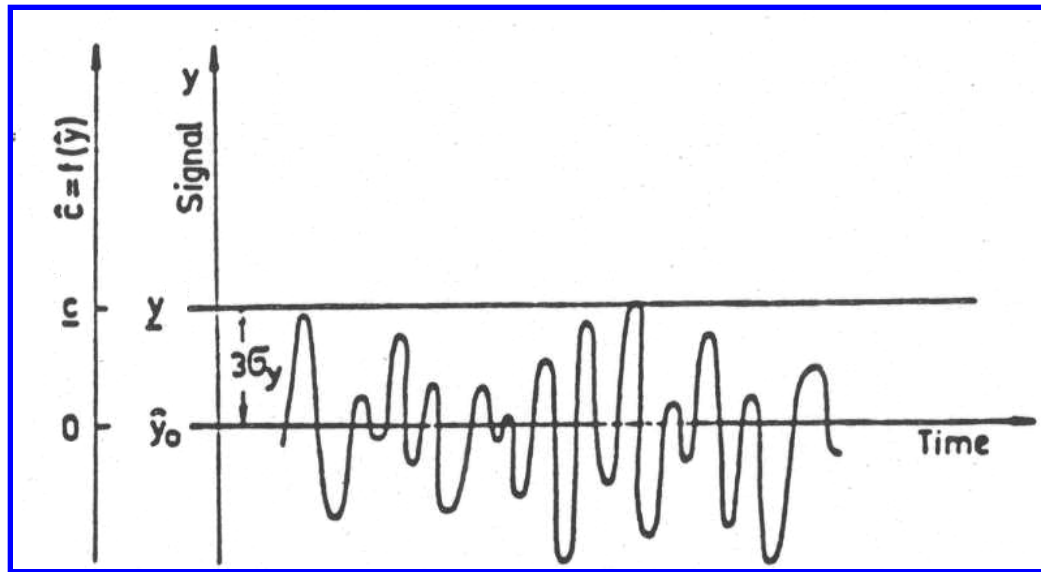
Questo parametro, fondamentale per valutare quanto sia sensibile un metodo analitico, è strettamente legato alla **soglia di decisione**, quindi presuppone una determinazione su base statistica.

### Criterio di Kaiser per il limite di rivelabilità

Il criterio di Kaiser stabilisce che il limite di rivelabilità sia dato dalla concentrazione di analita a cui corrisponde un segnale pari a:

$$y = y_0 + 3 \sigma_y$$

dove  $y_0$  è il valore medio del segnale ottenuto dal bianco mentre  $\sigma_y$  è la deviazione standard di tale segnale.



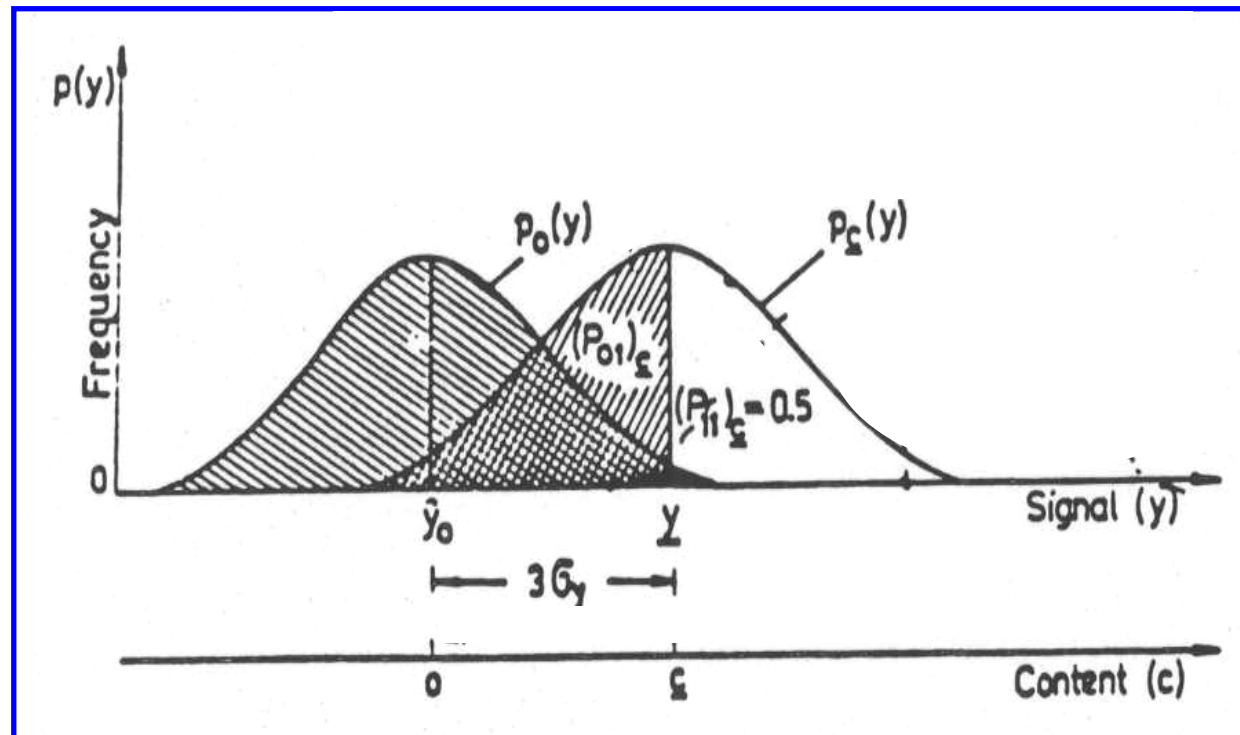
La curva riportata in figura rappresenta la congiunzione dei punti corrispondenti a centinaia di valori di responsi del bianco misurati nel tempo ed è di fatto una rappresentazione del rumore associato alla misura.

Il criterio di Kaiser equivale a considerare per la soglia di decisione un valore  $z_k = 3$ , ossia fissare una probabilità di falsa rivelazione  $(P_{10})_0 = 0.0013$  (0.13 %).

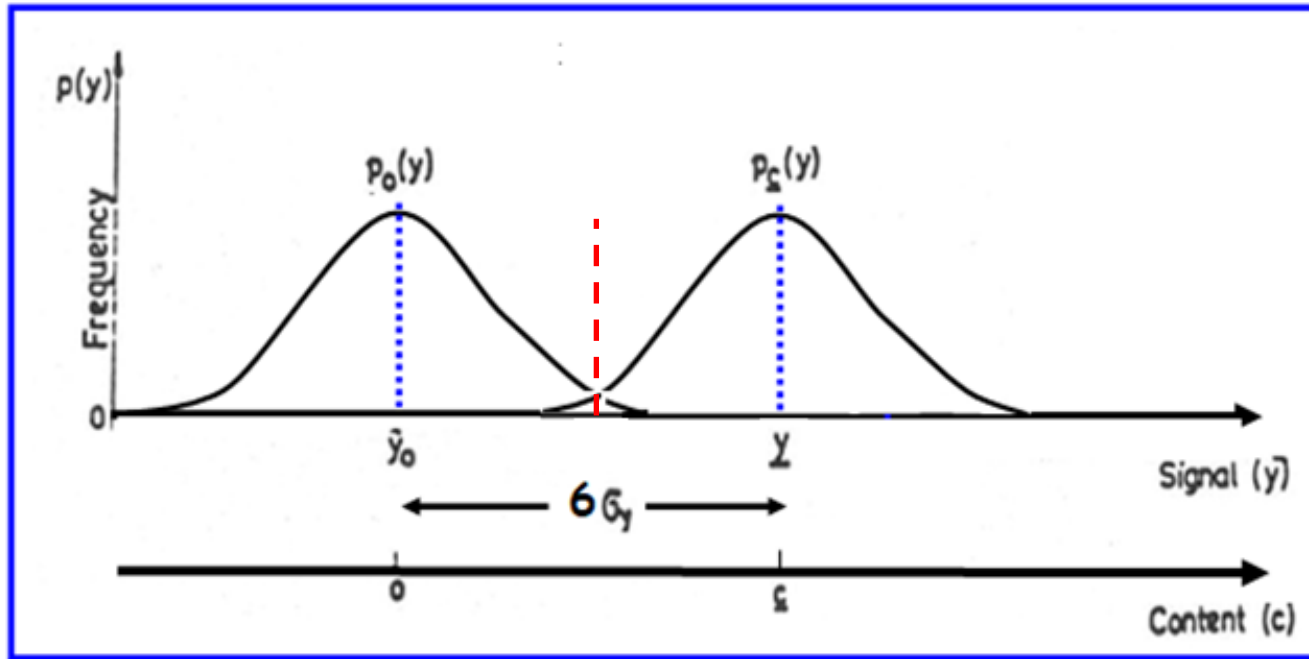
Dal valore di  $\underline{y}$  così stabilito si ricava poi il corrispondente valore di concentrazione  $\underline{c}$ , ossia il LOD associato ad un rapporto  $S/N = 3$ .

Il criterio di Kaiser non si preoccupa della probabilità di affermare che non ci sia analita nel campione quando invece esso è presente,  $(P_{01})_{\underline{c}}$ .

Tuttavia, se si fissa a  $3\sigma_y$  la differenza fra  $\underline{y}$  e  $y_0$  tale probabilità può essere pari anche al 50 %, quindi molto elevata, come evidenziato dal seguente grafico:



Si può dimostrare che soltanto se la differenza fra  $\bar{y}$  e  $y_0$  fosse pari a  $6\sigma_y$ , mantenendo la soglia di decisione a  $3\sigma_y$ , tale probabilità sarebbe praticamente trascurabile:



In queste condizioni il valore di  $\underline{c}$  sarebbe la minima concentrazione di analita rivelabile con buona precisione da una singola misura.

Talvolta si usa proprio tale valore, ossia il LOD a  $S/N = 6$ , come limite di rivelabilità di un metodo analitico strumentale, tuttavia il **criterio di Kaiser** resta il più diffuso nella definizione del valore del LOD.



## Relazione fra limite di rivelabilità e rapporto segnale/rumore

A prescindere dal valore numerico di  $z_k$  adottato per il limite di rivelabilità, si può notare che esso corrisponde alla quantità:

$$(\underline{y} - y_0) / \sigma_y$$

che, per definizione, rappresenta il **rapporto segnale/rumore (S/N)** in corrispondenza del segnale  $\underline{y}$ .

Nell'ipotesi che il segnale  $y$  dipenda linearmente dalla concentrazione si può scrivere:

$$y = y_0 + b_1 c \quad \text{ossia} \quad \underline{y} = y_0 + b_1 \underline{c}$$

dove  $b_1$  è la pendenza della retta che rappresenta la dipendenza del segnale dalla concentrazione.

Il limite di rivelabilità,  $\underline{c}$ , sarà dato dunque da:

$$\underline{c} = (\underline{y} - y_0) / b_1$$

Detto  $\underline{r}_{(S/N)}$  il rapporto segnale su rumore in corrispondenza del segnale  $\underline{y}$ , si può ricavare la relazione:

$$\underline{c} = \underline{r}_{(S/N)} \sigma_y / b_1$$

Fissato un certo rapporto segnale/rumore il limite di rivelabilità sarà tanto più basso (e quindi migliore):

- ✓ quanto maggiore è la pendenza  $b_1$ , ossia la sensibilità del metodo
- ✓ quanto più precisa è la misura di  $y$ , ossia quanto minore è  $\sigma_y$ .

## Stima del limite di rivelabilità direttamente dal rapporto segnale/rumore

La relazione:

$$\underline{c} = \underline{r}_{(S/N)} \sigma_y / b_1$$

si puo' sfruttare per una determinazione del limite di rivelabilita' a partire da una **misura diretta del rapporto segnale/rumore**. Le diverse fasi della procedura sono:

### 1) Determinazione di $\underline{r}_{(S/N)}$ a diverse concentrazioni $c$ dell'analita

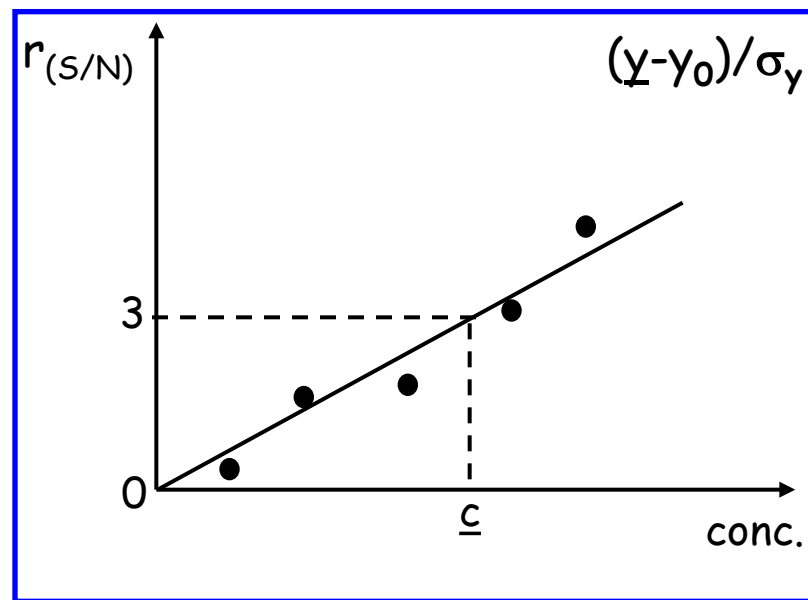
La determinazione prevede:

- ❖ la misura del **segnale del bianco** su **un certo numero di replicati**, con valutazione della sua deviazione standard, che rappresenta il rumore;
- ❖ la misura del **segnale analitico** in una serie di soluzioni standard dell'analita effettuando piu' replicati per ciascuna soluzione;

❖ il calcolo del valore medio e della deviazione standard di  $r_{(S/N)}$  alle varie concentrazioni.

2) Interpolazione dei dati di  $r_{(S/N)}$  in funzione di  $c$  con il metodo dei minimi quadrati.

I dati di  $r_{(S/N)}$  ottenuti in funzione di  $c$  vengono trattati con la regressione lineare convenzionale o pesata, a seconda dei casi.



3) Determinazione del limite di rivelabilita'

Il LOD si ricava scegliendo il valore di  $r_{(S/N)}$  che si desidera adottare, ad esempio 3, e ricavando il corrispondente valore di  $c$  dalla retta di regressione (esattamente come si fa per una concentrazione a partire dal segnale).

## Stima del limite di rivelabilità dalla regressione lineare

Siano date  $n$  coppie di valori  $(c_i, y_i)$ , che rappresentano il set di dati ottenuto per effettuare la calibrazione di un metodo analitico.

Le ipotesi fondamentali per la stima del limite di rivelabilità a partire dalla regressione lineare sono:

- ✓ le concentrazioni  $c_i$  si possono ritenere **affette da un'incertezza trascurabile**;
- ✓ i segnali  $y_i$  sono affetti **da un'incertezza distribuita in modo gaussiano**.

## Procedura

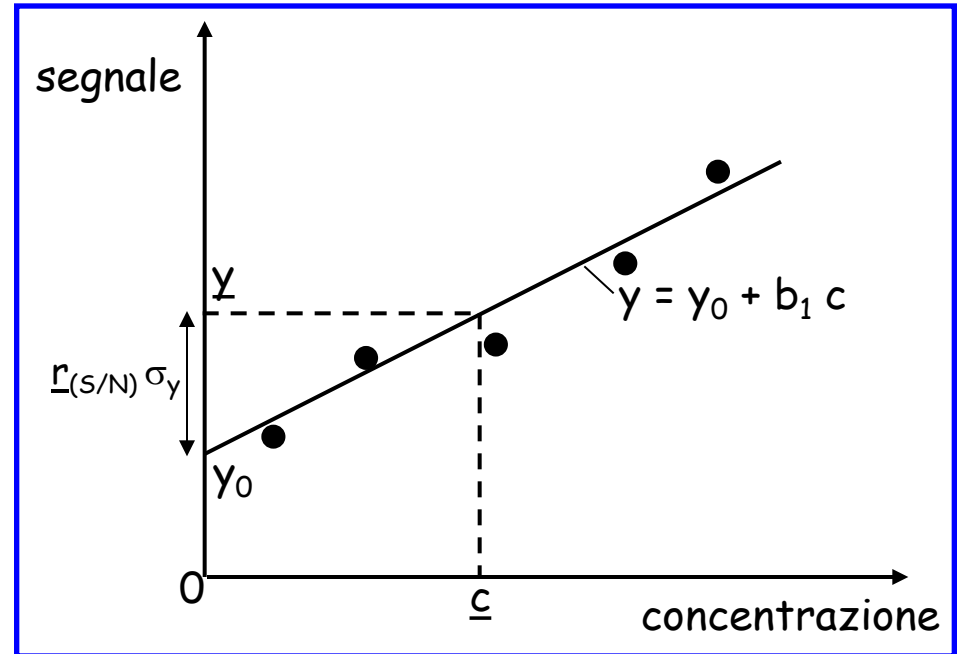
Si valuta inizialmente il valore del segnale,  $\underline{y}$ , a cui corrisponde il rapporto segnale/rumore,  $\underline{r}_{(S/N)}$ , prescelto per il limite di rivelabilità, tipicamente 3.

Essendo:

$$(\underline{y} - y_0)/\sigma_y = \underline{r}_{(S/N)}$$

risulta:

$$\underline{y} = y_0 + \underline{r}_{(S/N)} \sigma_y$$



Se l'equazione della retta di regressione è  $y = y_0 + b_1 c$ , si ricava facilmente che il **limite di rivelabilità** e' dato da:

$$\underline{c} = (\underline{y} - y_0)/b_1 = \underline{r}_{(S/N)} \sigma_y / b_1$$

In generale si adotta il criterio di Kaiser, per cui:  $r_{(S/N)} = 3$ , mentre al posto di  $\sigma_y$  si possono usare:

✓ il valore della deviazione standard sui residui della regressione lineare:

$$s_{y/x} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}}$$

oppure:

✓ il valore della deviazione standard sull'intercetta, che nella nuova notazione è:

$$s_{y_0} = s_{y/x} \sqrt{\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}}$$

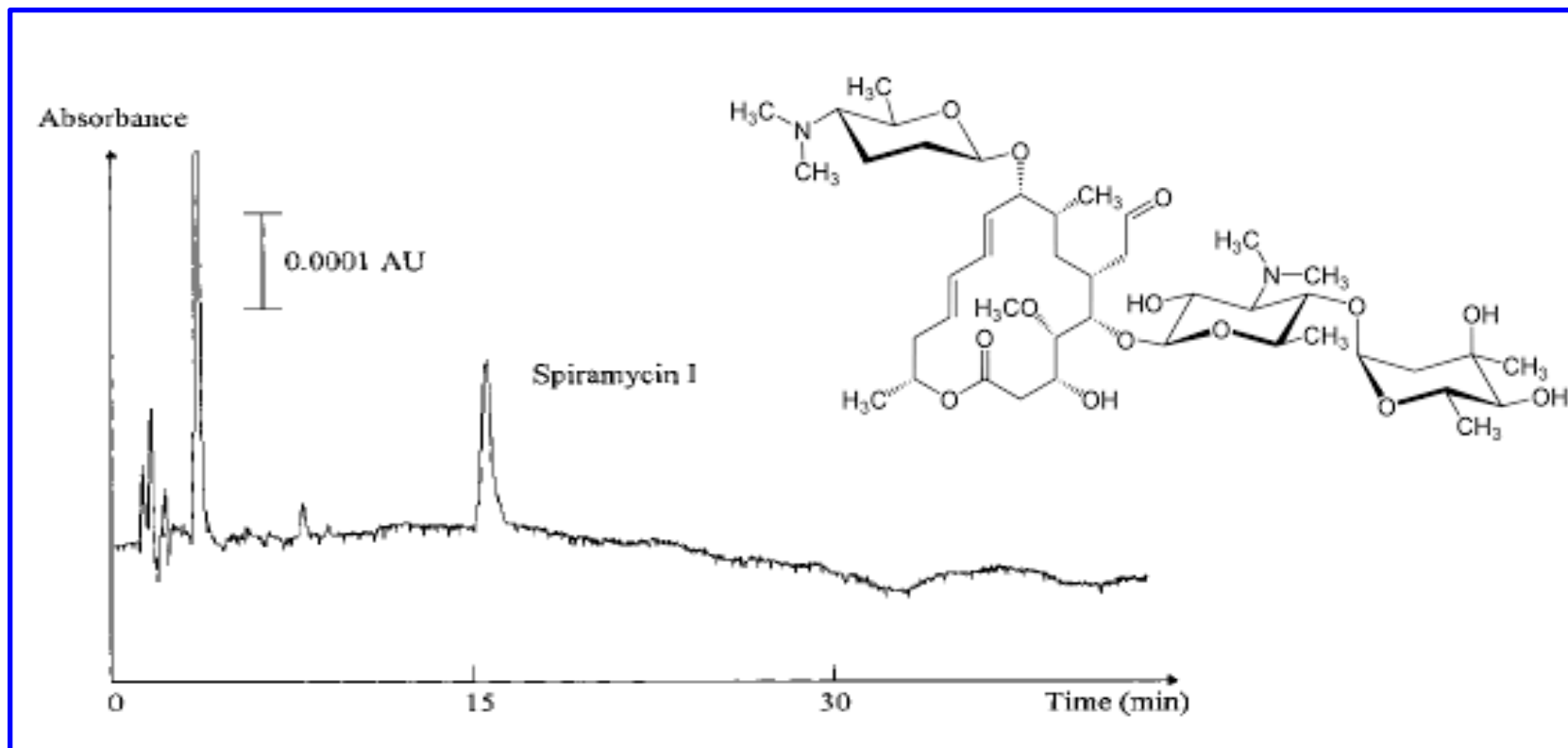
Dalle equazioni riportate si deduce che il valore indicato per **il limite di rivelabilita'** dipende:

- ✓ **dal rapporto segnale/rumore scelto** (a sua volta legato alla probabilità di falsa rivelazione adottata)
- ✓ **dallo stimatore adottato per  $\sigma_y$ , ossia  $s_{y/x}$  o  $s_{y_0}$** , che a loro volta cambieranno a seconda che si usi una regressione lineare convenzionale o pesata.



Confronto fra diversi metodi per la determinazione del LOD:  
un esempio sperimentale (*Analytical Chemistry*, 1999, 71, 2672)

Un esempio concreto di confronto fra diverse stime del LOD è stato riportato a proposito della determinazione dell'antibiotico spiramicina mediante HPLC di ripartizione in fase inversa su colonna C8 con rivelazione UV (232 nm).

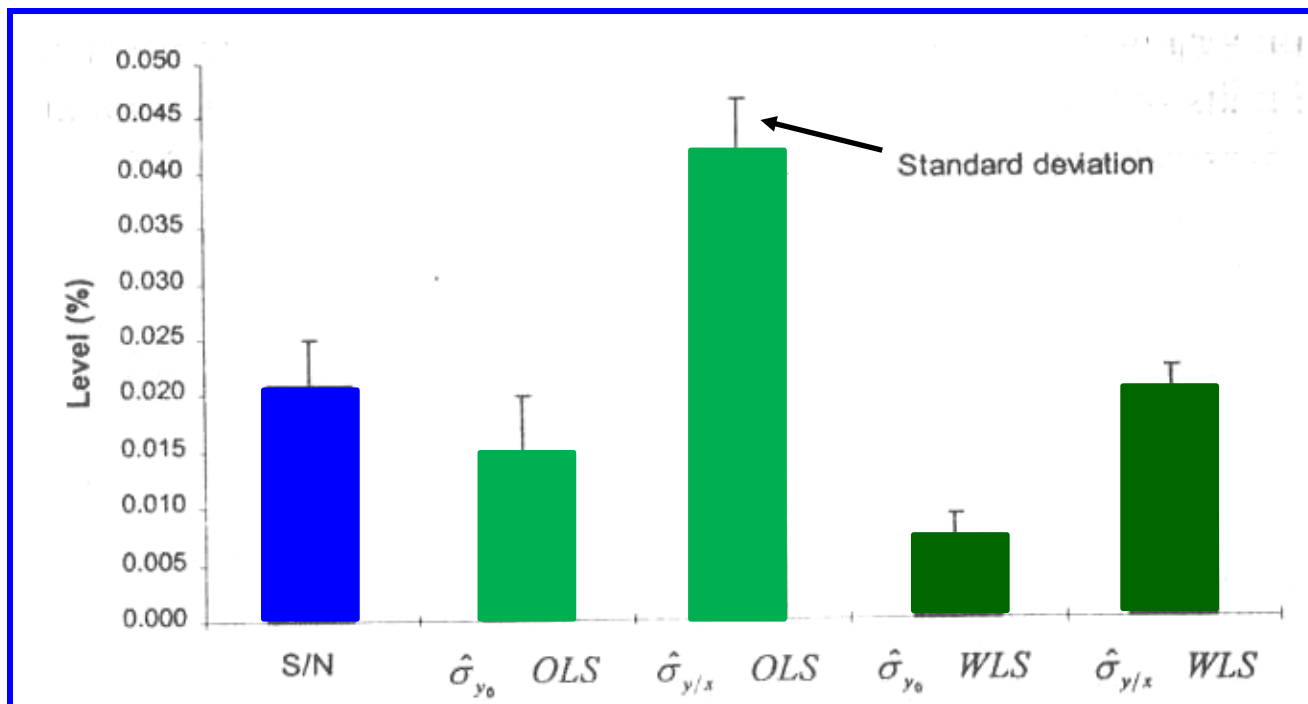


Sono state effettuate 6 misure replicate su 8 soluzioni di spiramicina a concentrazione variabile da 0 a 2.5 ppm.

Il LOD del metodo e' stato valutato scegliendo in tutti i casi il criterio di Kaiser, ossia ad un  $r_{(S/N)} = 3$ , ma usando 5 approcci diversi:

- ✓ determinazione diretta del  $r_{(S/N)}$  in funzione della concentrazione di spiramicina ( $S/N$ );
- ✓ determinazione con i minimi quadrati convenzionali (Ordinary Least Squares, OLS) usando  $s_{y/x}$ ;
- ✓ determinazione con i minimi quadrati convenzionali (Ordinary Least Squares, OLS) usando  $s_{y0}$ ;
- ✓ determinazione con i minimi quadrati pesati (Weighted Least Squares, WLS) usando  $s_{y/x}$ ;
- ✓ determinazione con i minimi quadrati pesati (Weighted Least Squares, WLS) usando  $s_{y0}$ .

Come si può notare, sono stati ottenuti valori di LOD (espressi come percentuale della massima concentrazione adottata nella calibrazione, ossia di 2.5 ppm) diversi a seconda dei casi.



In ogni caso essi individuano l'ordine di grandezza della sensibilità del metodo (le diverse stime del LOD sono comprese fra 25 e 75 ppb).