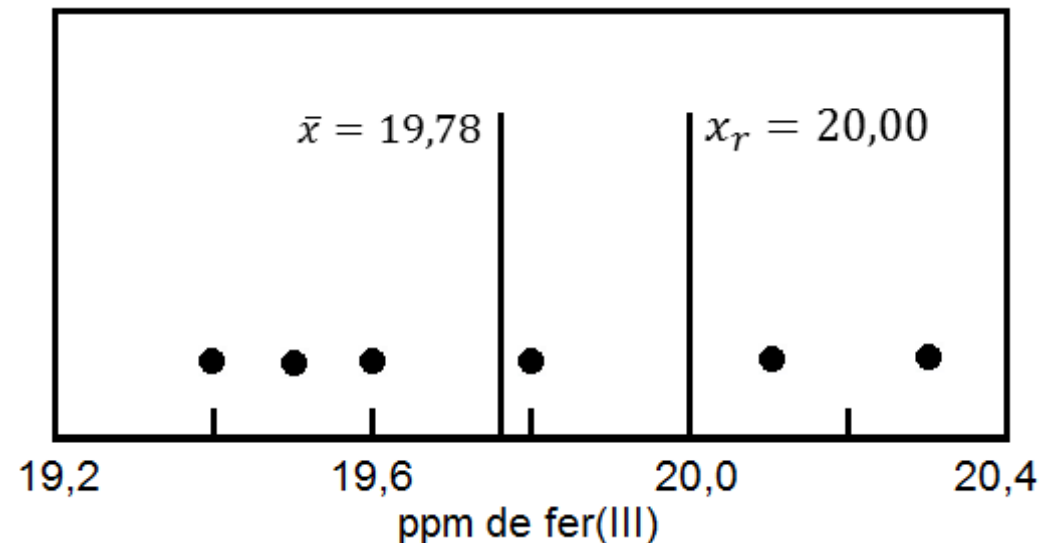


# Statistique appliquée

# Introduction

*Il est impossible d'effectuer une analyse chimique qui soit totalement exempte d'erreur ou d'incertitude. Tout ce qu'on peut espérer est de minimiser ces erreurs ou d'estimer leur grandeur de manière acceptable*

Exemple : résultat d'un dosage d'ion fer (III); 6 prises identiques d'une solution aqueuse qui contenait exactement 20,00 ppm de fer (III), analysés en suivant exactement le même protocole



# Introduction

- Les incertitudes de mesure ne peuvent jamais être complètement éliminées ; la valeur vraie d'une grandeur est toujours inconnue
- MAIS on peut souvent évaluer l'importance probable de l'erreur de mesure et déterminer les limites entre lesquelles la vraie valeur d'une grandeur mesurée a une certaine probabilité de se trouver
- Il est nécessaire d'estimer la *fiabilité* d'une analyse car des *données de qualité inconnue n'ont aucune valeur*.

# La fiabilité des données

La fiabilité des données recueillies peut être évaluée de différentes manières :

- Expériences révélant la présence d'erreur
- Analyse d'étalons de composition connue
- Consulter la littérature de chimie analytique...
- Étalonnage des appareils de mesure (amélioration des résultats de mesure)
- Tests statistiques appliqués aux résultats

La question à se poser est : quelle erreur maximale puis-je tolérer sur le résultat? *En règle générale, on ne peut se permettre de perdre du temps à produire des données plus fiables que ce qui est nécessaire*

# Définitions : précision

- La précision décrit la reproductibilité des mesures, c'est-à-dire la proximité de résultats qui ont été obtenus exactement de la même manière
- la précision d'un ensemble de données est donnée par :
  - l'écart-type,
  - la variance
  - le coefficient de variation.
- Tous ces termes dépendent de l'écart à la moyenne :

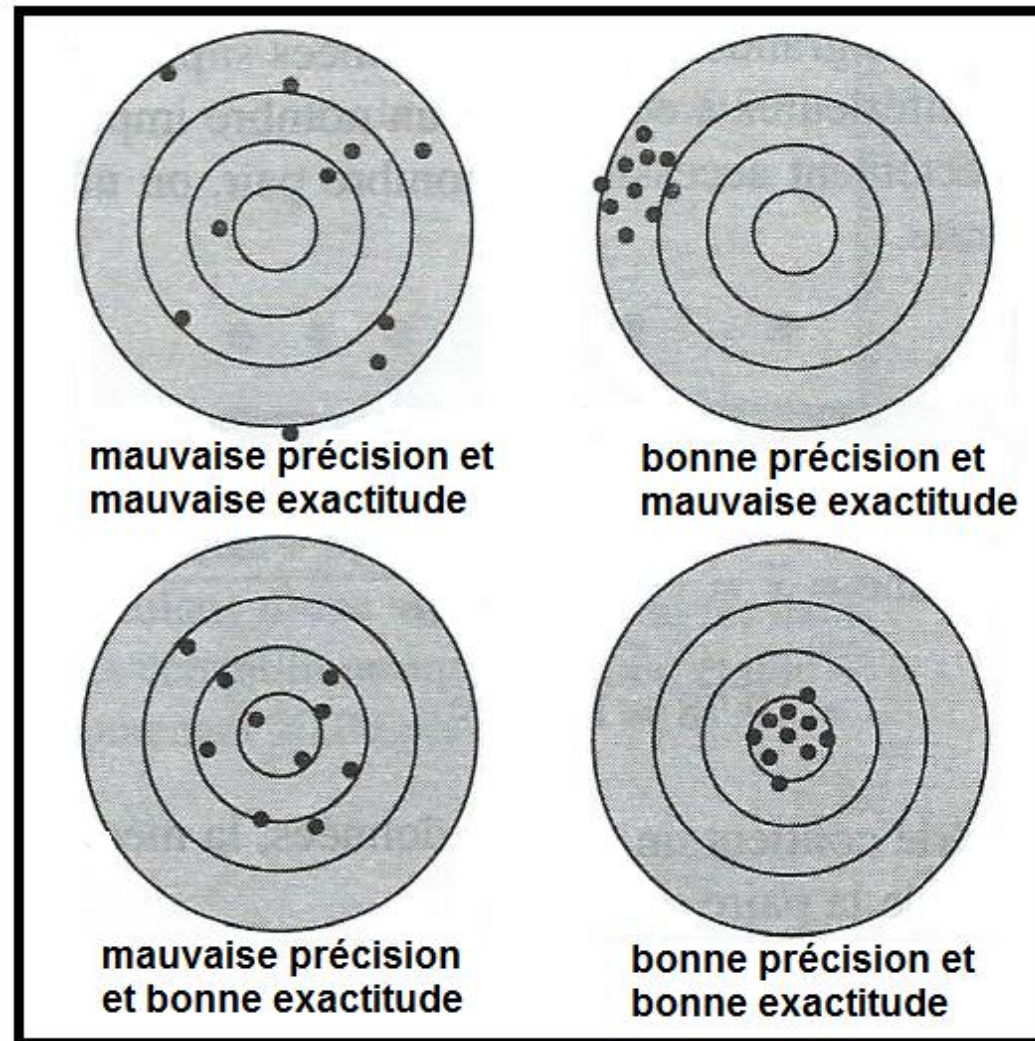
$$\text{écart à la moyenne} = d_i = |x_i - \bar{x}|$$

# Définition : exactitude

- L'exactitude exprime la proximité entre un résultat et sa valeur réelle ou présumée réelle et s'exprime par l'erreur (absolue ou relative)
- L'erreur absolue est donnée par la relation :  $E = x_i - x_r$
- L'erreur relative (plus souvent utilisée) est l'erreur absolue exprimée en pourcent

$$E_r = \frac{x_i - x_r}{x_r} \cdot 100\%$$

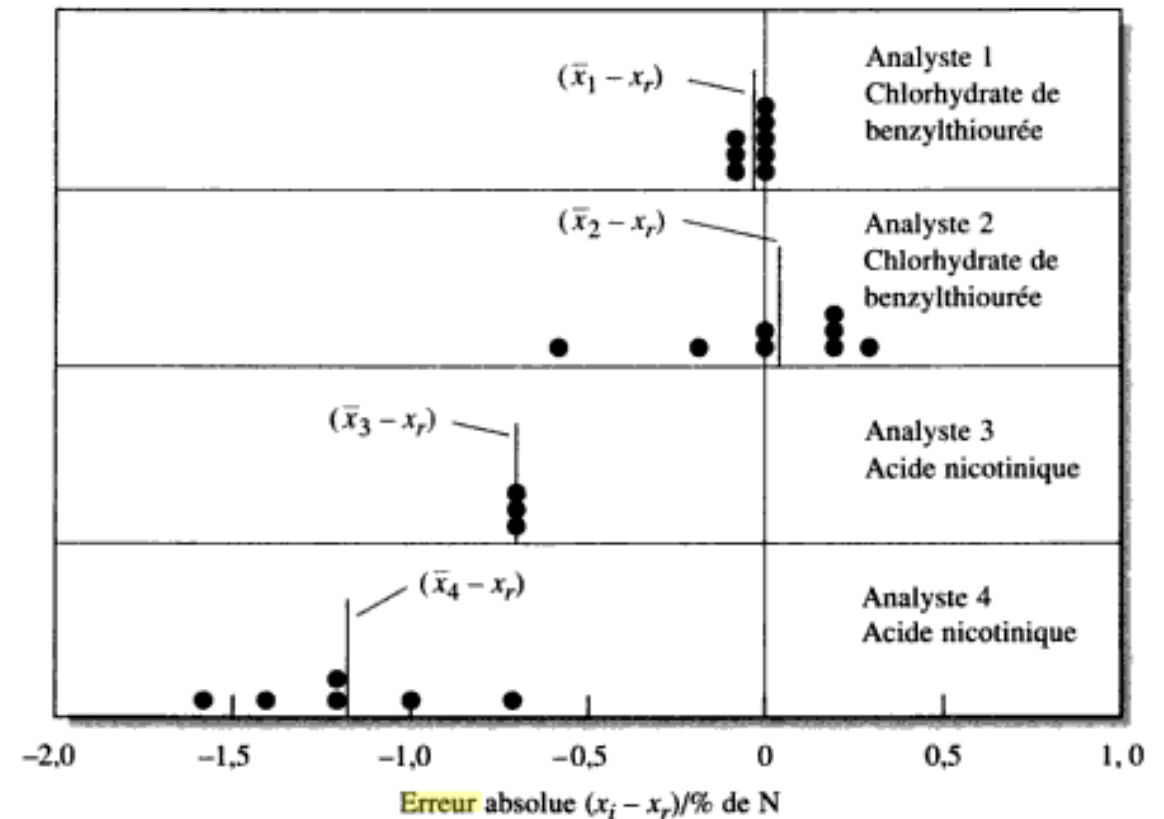
# Définitions : précision et exactitude



# Définition : les types d'erreurs dans les données expérimentales

- Si la précision d'une mesure se détermine facilement, l'exactitude est en revanche beaucoup plus difficile à déterminer (il faut connaître la valeur réelle)

Exemple : erreur absolue dans le microdosage Kjeldahl de l'azote dans deux composés : chlorhydrate de benzylthiourée et acide nicotinique



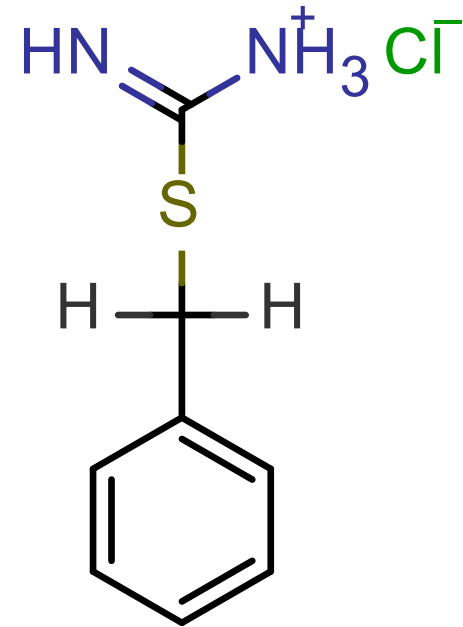


# Définition : les types d'erreurs dans les données expérimentales

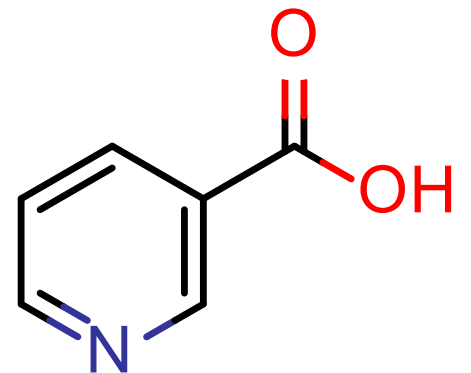
Explication : analyse de chlorhydrate de benzylthiourée et acide nicotinique par la méthode Kjeldahl

L'analyse consiste à décomposer les échantillons organiques dans l'acide sulfurique concentré et chaud, en présence de catalyseur, transformant l'azote de l'échantillon en sulfate d'ammonium. La quantité d'ammoniac dans le sulfate d'ammonium est finalement déterminée dans l'étape de mesure.

MAIS les composés contenant un cycle pyridinique comme l'acide nicotinique ne sont pas complètement décomposés dans  $\text{H}_2\text{SO}_4$  à moins de prendre des précautions particulières



chlorhydrate de  
benzylthiourée



acide nicotinique

# Définition : les types d'erreurs dans les données expérimentales

- Il existe trois grands types d'erreurs :
  - **L'erreur aléatoire** : causant la dispersion des résultats
  - **L'erreur systématique** : cause de la différence entre la moyenne des données et la moyenne présumée
  - **L'erreur grossière** : valeur excentrique par rapport à la moyenne, loin de toutes les autres valeurs obtenues

# Les erreurs systématiques

- Les erreurs systématiques sont caractérisées par un *biais*. Le biais affecte toutes les données d'un ensemble de mesure de la même manière et il porte un signe (positif ou négatif)
- Il y a trois types d'erreurs systématiques :
  - *Les erreurs instrumentales* (imperfection des dispositifs de mesure)
  - *Les erreurs dues à la méthode* : comportement non idéal des systèmes analytiques (lenteur de certaines réactions, réactions incomplètes, instabilité de certaines espèces, non spécificité de la plupart des réactifs...)
  - *Les erreurs personnelles* (manque de soin, inattention ou incompetence...)

# Effets des erreurs systématiques

- Constantes : d'autant plus gênantes si la grandeur mesurée est petite
  - L'excès de réactif nécessaire au virage de l'indicateur...
- Proportionnelles : fréquemment dues aux interférants
  - l'ion cuivre(II) réagit avec l'iodure de potassium pour donner de l'iode. La quantité d'iode mesurée est proportionnelle à la quantité d'ions cuivre(II). Mais le fer (III) réagit également avec l'iodure de potassium pour donner de l'iode... l'importance de l'erreur dépend donc de la proportion de la contamination par le fer(III)

# Les erreurs aléatoires dans les analyses

- Ces erreurs sont causées par des paramètres incontrôlables qui font partie de toute mesure physique ou chimique
- Elles ne peuvent être vraiment identifiées ou mesurées car elles sont trop petites pour être déterminées individuellement
- C'est l'effet cumulatif de ces erreurs qui cause *la dispersion des données d'une série de mesure autour de la moyenne*

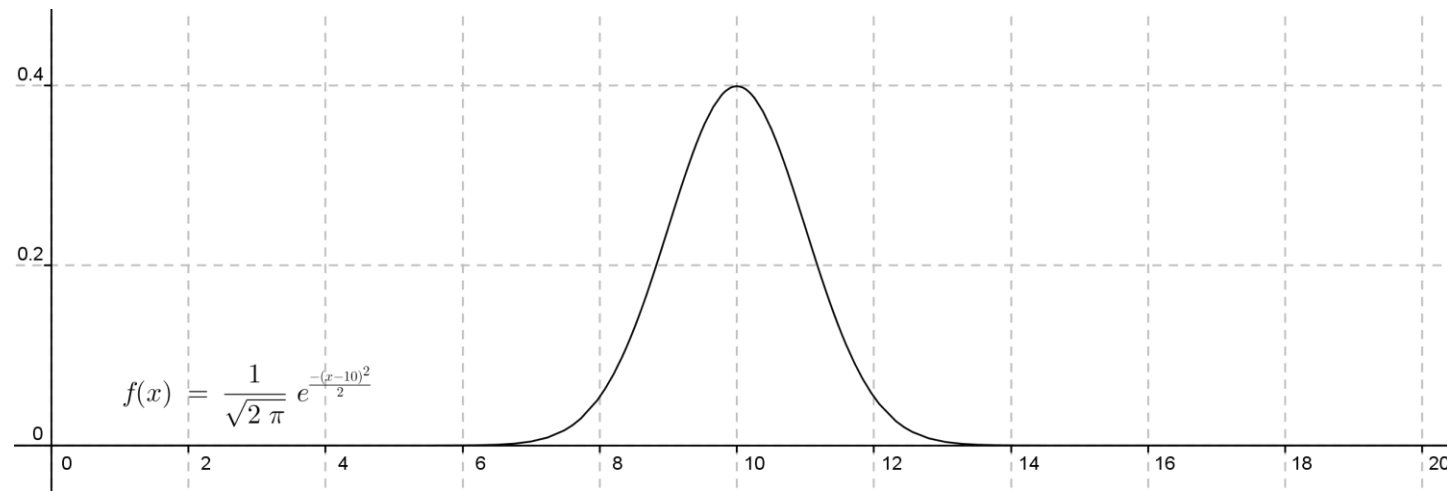
# Sources d'erreurs aléatoires

Exemple : Imaginons une situation où seulement 4 petites erreurs aléatoires se combinent pour donner une erreur globale. Chaque erreur a une probabilité égale de se produire et elle modifie le résultat final d'une quantité fixe  $\pm U$ .

<C:\Users\Bertrand\Documents\Bertrand\Cours\Institut Roger Lambion\Elements de chimie analytique\Tableau combinaisons possibles.xlsx>

# Distribution des données expérimentales

Dans la plupart des analyses quantitatives, on observe expérimentalement que la distribution des données est proche de la courbe de Gauss



# Traitement statistique de l'erreur aléatoire

- Hypothèse de départ : les erreurs aléatoires suivent une distribution gaussienne (ou normale)

## Échantillon et population

On distingue

- la moyenne de l'échantillon ( $\bar{x}$ ) : pour un échantillon limité tiré de la population de données
- la moyenne de la population ( $\mu$ ) : moyenne réelle pour cette population.

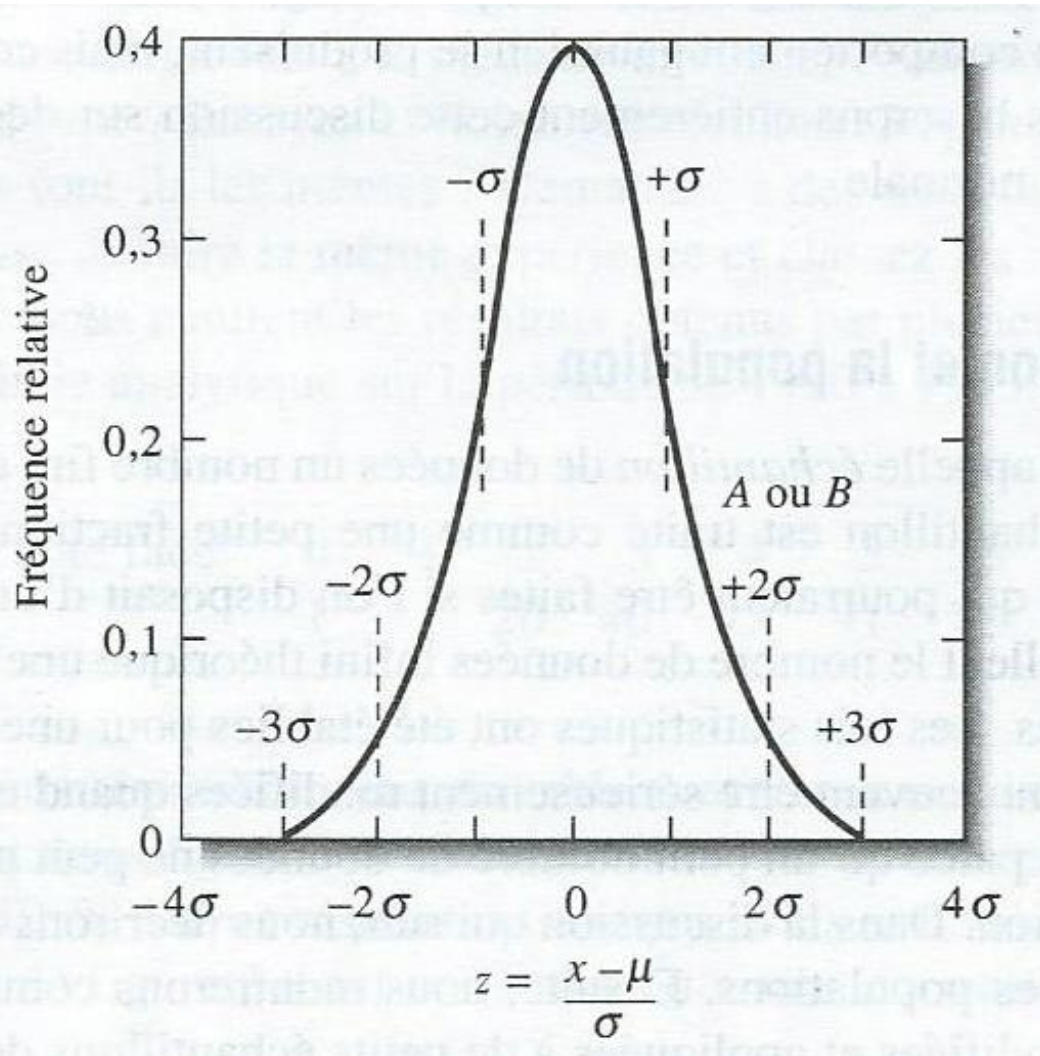
Lorsque  $N$  est petit,  $\bar{x}$  diffère de  $\mu$  car un petit échantillon ne représente pas exactement sa population. Lorsque  $N$  atteint 20 ou 30, cette différence devient négligeable.



## Écart-type de la population ( $\sigma$ )

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

La figure ci-contre montre une courbe d'erreur normale où l'abscisse est  $z$ , l'écart de la moyenne des données exprimées en unité d'écart-type.



- La moyenne est au point central de fréquence maximale
- Écarts positifs et négatifs répartis de manière symétrique
- La fréquence diminue exponentiellement lorsque la valeurs des écarts augmentent

# Traitement statistique de l'erreur aléatoire

**Mesure de la précision : écart-type de l'échantillon**

$$s = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}} \quad \text{ou} \quad s = \sqrt{\frac{\sum_{i=1}^N x_i^2 - \frac{(\sum_{i=1}^N x_i)^2}{N}}{N - 1}}$$

Où  $N$  = nombre d'échantillons

$x_i$  = valeur mesurée pour l'échantillon  $i$

# Traitement statistique de l'erreur aléatoire

Exercice : les résultats suivants ont été obtenus lors de l'analyse de la teneur en plomb d'un échantillon de sang : 0,752, 0,756, 0,751, 0,752 et 0,760 ppm de Pb. Calculez la moyenne et l'écart-type de cette série de données.

# Traitement statistique de l'erreur aléatoire

## Erreur type d'une moyenne

- Les valeurs calculées précédemment se rapportent à une seule mesure (un échantillon contenant N données);
- si une série d'échantillon contenant chacun N données est prise au hasard dans leur population, la moyenne de chaque série montrera de moins en moins de dispersion lorsque N augmente
- L'erreur type d'une moyenne est donné par  $s_m = \frac{s}{\sqrt{N}}$

# Traitement statistique de l'erreur aléatoire

## **Fiabilité de $s$ en tant que mesure de la précision**

- On améliore considérablement  $s$  si on augmente  $N$ . Pour des valeurs de  $N$  supérieures à environ 20 on se rapproche de la valeur théorique de l'écart-type d'une population.

Pour améliorer la fiabilité de  $s$  on peut **grouper des données**

- Dans ce cas les données de séries d'échantillons semblables, amassées au cours du temps, peuvent être groupées de manière à obtenir une estimation de  $s$  qui soit meilleure que la valeur de chaque sous ensemble isolé.

# Traitement statistique de l'erreur aléatoire

## **Grouper les données aux fins d'améliorer la fiabilité de s**

- Il faut admettre que les sources d'erreur aléatoire sont les mêmes pour toutes les mesures

→ hypothèse valable lorsque les échantillons ont des compositions voisines et s'il ont été analysés exactement de la même manière.

Calcul des écarts-types groupés :

$$s_{groupé} = \sqrt{\frac{\sum_{i=1}^{N_1} (x_i - \bar{x}_1)^2 + \sum_{i=1}^{N_2} (x_i - \bar{x}_2)^2 + \sum_{i=1}^{N_3} (x_i - \bar{x}_3)^2 + \dots}{N_1 + N_2 + N_3 + \dots - N_t}}$$

# Traitement statistique de l'erreur aléatoire

Spécimen	# d'échantillons analysés	Teneur en Hg (en ppm)	Moyenne (en ppm)	Somme des carrés des écarts à la moyenne
1	3	1,80, 1,58, 1,64		
2	4	0,96, 0,98, 1,02, 1,10		
3	2	3,13, 3,35		
4	6	2,06, 1,93, 2,12, 2,16, 1,89, 1,95		
5	4	0,57, 0,58, 0,64, 0,49		
6	5	2,35, 2,44, 2,70, 2,48, 2,44		
7	4	1,11, 1,15, 1,22, 1,04		

Calculez un écart-type groupé pour cette méthode, en vous basant sur les trois premières colonnes de données

# Autres termes pour exprimer la précision des échantillons de données

La **variance** est simplement égale au carré de l'écart-type :

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

L'unité de la variance est donc l'unité du résultat élevée au carré.

En général on utilise plutôt l'écart-type pour caractériser la précision, car il est plus facile de comparer une mesure et sa précision si elles sont dans les mêmes unités



# Autres termes pour exprimer la précision des échantillons de données

## **écart-type relatif et coefficient de variation**

- L'écart-type relatif se calcule en divisant l'écart-type par la moyenne de la série de données. Il est exprimé en ‰ ou en % suivant que l'on multiplie le résultat par 1000 ou par 100.

$$s_r = (s/\bar{x}) \cdot 1000\text{‰}$$

- lorsque l'écart-type relatif est multiplié par 100%, on l'appelle *coefficient de variation*

$$CV = (s/\bar{x}) \cdot 100\%$$

# Autres termes pour exprimer la précision des échantillons de données

**L'étendue ( $w$ )** ou amplitude d'une série de données

- Différence entre la plus grande et la plus petite des valeurs de la série.

Exercice : pour la série de données suivante déterminez (a) la variance, (b) l'écart-type relatif en parties pour mille, (c) le coefficient de variation et (d) l'étendue

Teneur en plomb d'un échantillon de sang : 0,752, 0,756, 0,752, 0,751 et 0,760 ppm de Pb.

# Exercices

Bouteille	% (m/v) en sucre résiduel	Nombre d'observations	Écarts à la moyenne des résultats individuels
1	0,94	3	0,050, 0,10, 0,08
2	1,08	4	0,060, 0,050, 0,090, 0,060
3	1,20	5	0,05, 0,12, 0,07, 0,00, 0,08
4	0,67	4	0,05, 0,10, 0,06, 0,09
5	0,83	3	0,07, 0,09, 0,10
6	0,76	4	0,06, 0,12, 0,04, 0,03

- Évaluer l'écart-type pour chaque série de données
- Grouper les données pour établir un écart-type absolu pour la méthode

# Application de la statistique au traitement et à l'évaluation des données

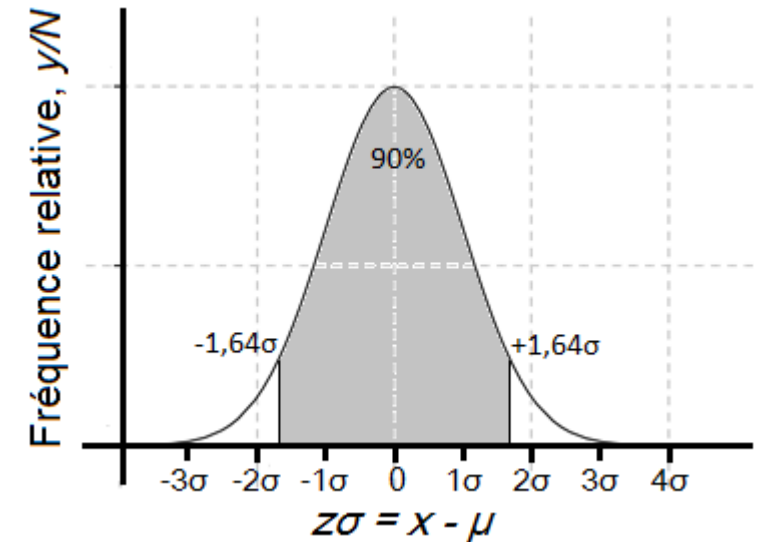
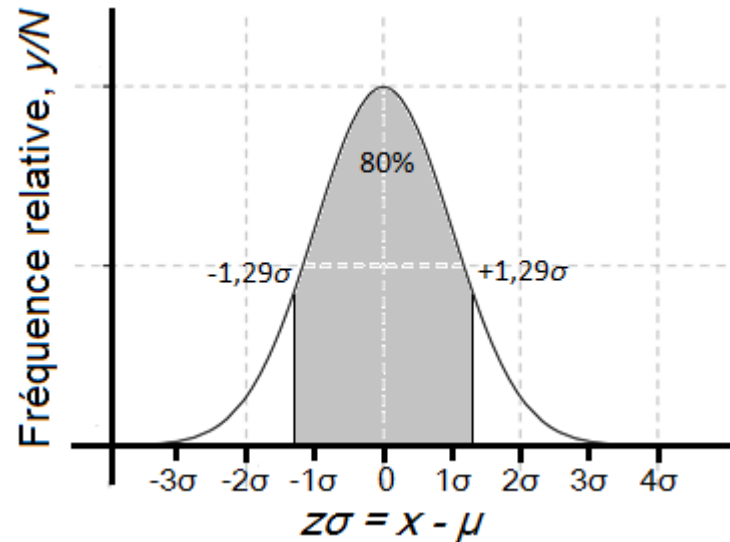
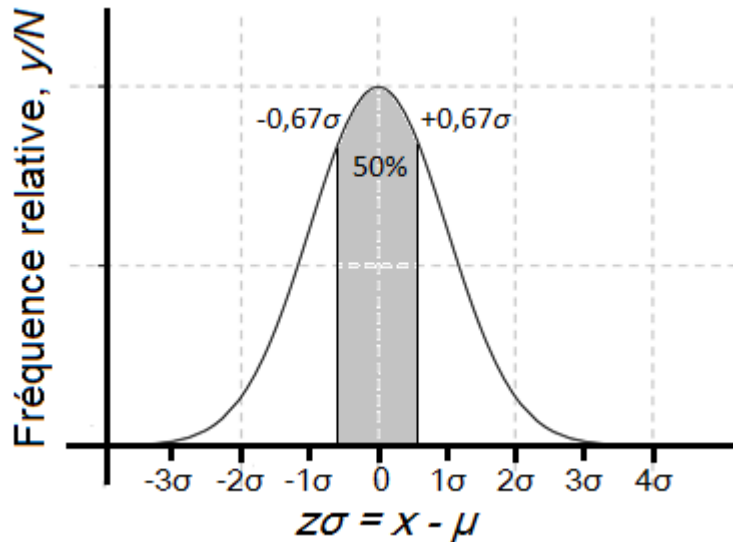
# Limites de confiance

La valeur exacte de la moyenne  $\mu$  (moyenne d'une population) n'est jamais connue exactement MAIS on peut fixer des limites autour d'une moyenne  $\bar{x}$  entre lesquelles  $\mu$  se situe avec une certaine probabilité.

**On appelle ces limites *limites de confiance* entre lesquelles se définit un *intervalle de confiance***

La largeur de l'intervalle de confiance dépend de l'exactitude avec laquelle on connaît  $s$ , c'est-à-dire de combien  $s$  diffère de  $\sigma$  (écart-type de la population)

# Limites de confiance lorsque $s$ est une bonne approximation de $\sigma$



Les relations ci-dessus permettent de définir autour d'une mesure un domaine dans lequel la moyenne réelle a une certaine probabilité de se trouver.

Dans 90% des cas, la moyenne réelle est à moins de  $1,64\sigma$  de toute mesure effectuée. Le degré de confiance est de 90% et l'intervalle de confiance vaut  $\pm z\sigma = \pm 1,64\sigma$

# Limites de confiance lorsque $s$ est une bonne approximation de $\sigma$

Expression générale des limites de confiance (LC) pour une seule mesure :

$$\text{LC\% pour } \mu = x \pm z\sigma$$

où  $z$  est l'écart de la moyenne exprimée en unité d'écart-type

Pour la moyenne de  $N$  mesures, on emploie l'erreur type de la moyenne :

$$\text{LC\% pour } \mu = \bar{x} \pm \frac{z\sigma}{\sqrt{N}}$$

Et donc

$$IC = \pm \frac{z\sigma}{\sqrt{N}}$$

# Limites de confiance lorsque $s$ est une bonne approximation de $\sigma$

Degré de confiance /%	$z$
50	0,67
68	1,00
80	1,29
90	1,64
95	1,96
96	2,00
99	2,58
99,7	3,00
99,9	3,29

Le tableau ci-contre donne le degré de confiance pour différentes valeurs de  $z$

Exercice 1 : calculez les limites de confiances à 80% et 95% pour (a) une mesure à 1,80 ppm de Hg et (b) la valeur moyenne 1,67 ppm de Hg obtenues avec 3 mesures. On suppose que  $s \rightarrow \sigma = 0,10$



# Limites de confiance lorsque $s$ est une bonne approximation de $\sigma$

Degré de confiance /%	$z$
50	0,67
68	1,00
80	1,29
90	1,64
95	1,96
96	2,00
99	2,58
99,7	3,00
99,9	3,29

Exercice 2: pour  $\bar{x} = 1,67$  ppm de Hg et  $\sigma = 0,10$ , calculez le nombre de mesures nécessaires pour ramener à  $\pm 0,07$  l'intervalle de confiance à 95%.

Réponse : 8

# Limites de confiances lorsque $\sigma$ est inconnu

Par manque de temps ou parce que la quantité d'échantillon est insuffisante, il n'est pas possible d'estimer correctement  $\sigma$ .

Il faut donc que la série de mesure permette de déterminer non-seulement la moyenne mais aussi d'en déterminer la précision.

Comme  $s$  est calculée avec un petit nombre de mesures, elle peut être incertaine  $\Rightarrow$  les limites de confiances vont être plus étendues que lorsqu'on connaît  $\sigma$ ...

# Limites de confiances lorsque $\sigma$ est inconnu

Pour tenir compte de la valeur de  $s$ , on utilise le paramètre statistique  $t$  équivalent de  $z$  dans les équations précédentes :

$$t = \frac{x - \mu}{s}$$

$t$  dépend du degré de confiance désiré mais également du nombre de degré de liberté impliqué dans le calcul de  $s$ . Du coup,  $t \rightarrow z$  lorsque le nombre de degré de liberté tend vers l'infini.

Les limites de confiances sont données par la relation suivante :

$$\text{LC\% pour } \mu = \bar{x} \pm \frac{ts}{\sqrt{N}}$$

# Limites de confiance lorsque $\sigma$ est inconnu

Exemple : un chimiste a obtenu les données suivantes pour la teneur en alcool d'un échantillon de sang : %  $\text{C}_2\text{H}_5\text{OH}$  = 0,084, 0,089 et 0,079. calculez les limites de confiance à 95 % pour la moyenne en supposant

- (a) qu'on n'a aucune autre information sur la précision de la méthode
- (b) Que, sur base de mesures antérieures, on sait que  $s \rightarrow \sigma = 0,005\%$  de  $\text{C}_2\text{H}_5\text{OH}$

D'après les tables de  $t$  (*dia suivante*) :  $t = 4,30$  pour 2 degrés de liberté et 95%

<i>Valeurs de t pour divers degrés de probabilité</i>					
	Facteurs pour les intervalles de confiance				
degrés de liberté	80%	90%	95%	99%	99,90%
1	3,078	6,314	12,71	63,66	636,6
2	1,886	2,92	4,303	9,925	31,6
3	1,638	2,353	3,182	5,841	12,92
4	1,533	2,132	2,776	4,604	8,61
5	1,476	2,015	2,571	4,032	6,869
6	1,44	1,943	2,447	3,707	5,959
7	1,415	1,895	2,365	3,499	5,408
8	1,397	1,86	2,306	3,355	5,041
9	1,383	1,833	2,262	3,25	4,781
10	1,372	1,812	2,228	3,169	4,587
11	1,363	1,796	2,201	3,106	4,437
12	1,356	1,782	2,179	3,055	4,318
13	1,35	1,771	2,16	3,012	4,221
14	1,345	1,761	2,145	2,977	4,14
∞	1,282	1,645	1,96	2,576	3,291

# Aide statistique à la vérification d'hypothèses

Le développement scientifique repose en grande partie sur la vérification d'hypothèses.

Les résultats expérimentaux étant rarement en parfait accord avec les prédictions d'un modèle théorique, il faut « apprécier » le désaccord observé : est-ce une conséquence des erreurs aléatoires ou est-ce une erreur systématique dans la procédure de mesure?

⇒ certains tests statistiques sont utilisés pour porter de tels jugements

On utilise le concept d'hypothèse nulle qui postule que deux grandeurs observées (une théorique et une expérimentale) sont les mêmes

Exemple : si la différence observée entre valeur théorique et expérimentale est égale ou supérieure à la différence qui se produira 5 fois sur 100 (degré de probabilité de 5%), on peut considérer que l'hypothèse nulle n'est pas valable et que la différence est significative pour un degré de confiance de 95%.

Les tests consistent à comparer :

- Une moyenne expérimentale avec une valeur vraie
- Deux moyennes expérimentales
- Les écart-types de deux séries de mesures

# Comparaison de la précision des mesures

**Utilisation du test  $F$**  : les données ne doivent pas nécessairement provenir du même échantillon ; cependant il faut que les causes d'erreur aléatoire soient les mêmes.

**Hypothèse nulle** : les précisions des séries de données sont identiques.

⇒ On compare la valeur  $F$ , rapport des variances des deux mesures, aux valeurs critiques consignées dans les tables.

$$F = \frac{s_x^2}{s_y^2}$$

*Remarque : on place toujours la variance la plus élevée au numérateur.*



Valeurs critiques pour $F$ au degré de probabilité 5 %								
	Degrés de liberté (Numérateur)							
Degrés de liberté (dénominateur)	2	3	4	5	6	12	20	$\infty$
2	19,00	19,16	19,25	19,30	19,33	19,41	19,45	19,50
3	9,55	9,28	9,12	9,01	8,94	8,74	8,66	8,53
4	6,94	6,59	6,39	6,26	6,16	5,91	5,80	5,63
5	5,79	5,41	5,19	5,05	4,95	4,68	4,56	4,36
6	5,14	4,76	4,53	4,39	4,28	4,00	3,87	3,67
12	3,89	3,49	3,26	3,11	3,00	2,69	2,54	2,30
20	3,49	3,10	2,87	2,71	2,60	2,28	2,12	1,84
$\infty$	3,00	2,60	2,37	2,21	2,10	1,75	1,57	1,00

Au degré de probabilité 5 %, si les écarts-types de 2 mesures sont les mêmes, les valeurs de  $F$  ne seront dépassées que dans 5 % des cas...

# Comparaison de la précision des mesures

Le test  $F$  peut être utilisé pour :

- 1) Décider si une méthode A est plus précise qu'une méthode B
- 2) Décider si il y a une différence entre les précisions de deux méthodes

Si le test  $F$  permet de conclure que l'hypothèse nulle est vérifiée, on peut **grouper les 2 écarts-types**

# Comparaison de la précision des mesures

*Exemple* : l'écart-type (+ de 100 mesures) d'une méthode de mesure classique de dosage du CO dans les mélanges gazeux vaut 0,21 ppm de CO. Une modification de la méthode donne une valeur de  $s = 0,15$  ppm de CO avec 12 degrés de liberté. Une deuxième modification, donne  $s = 0,12$  ppm de CO, aussi avec 12 degrés de liberté.

Ces deux modifications sont-elles significativement plus précises que la méthode originale?

Pour la 1<sup>ère</sup> modification : 
$$F_1 = \frac{s_{\text{réf}}^2}{s_1^2} = \frac{(0,21)^2}{(0,15)^2} = 1,96$$

Pour la 2<sup>e</sup> modification : 
$$F_2 = \frac{s_{\text{réf}}^2}{s_2^2} = \frac{(0,21)^2}{(0,12)^2} = 3,06$$

# Comparaison de la précision des mesures

Pour la méthode de dosage classique,  $s_{\text{réf}} \rightarrow \sigma$  ; le nombre de degré de liberté est  $\infty$ .  
Valeur critique de  $F = 2,30$ .

- la valeur de  $F$  pour la 1<sup>ère</sup> modification est inférieure à 2,30 l'hypothèse nulle est vérifiée avec une certitude de 95 %.
- la valeur de  $F$  pour la 2<sup>e</sup> modification est supérieure à 2,30 l'hypothèse nulle est infirmée avec une certitude de 95 %.

⇒ la 2<sup>e</sup> modification est donc significativement plus précise que l'originale.

# Exercice

Une méthode titrimétrique pour le dosage du calcium dans le calcaire a été testée par l'analyse d'un calcaire du NIST qui contient 30,15 % de CaO. Le résultat moyen de quatre analyses a été 30,26 % de CaO, avec un écart-type de 0,085 %. En groupant les données de plusieurs analyses, on a établi que  $s \rightarrow \sigma = 0,094$  % de CaO.

- (a) Les données indiquent-elles la présence d'une erreur systématique au degré de confiance 95 %?
- (b) Les données indiquent-elles la présence d'une erreur systématique au degré de confiance 95 % si l'on ne disposait d'aucune valeur groupée de  $\sigma$ ?

Réponse : (a) oui, (b) non

# Détection d'erreurs grossières

Si un résultat d'une série de donnée est excentrique il faut prendre la décision de la garder ou de le rejeter.

Cette décision sera fonction des normes (indulgentes ou exigeantes) qui sont imposées.

## Le test $Q$

Le test  $Q$  est un test statistique simple qui consiste à calculer la valeur expérimentale  $Q_{exp}$  en divisant par l'étendue  $w$  la valeur absolue de la différence entre le résultat suspect  $x_q$  et son plus proche voisin  $x_n$  :

$$Q_{exp} = |x_q - x_n| / w$$

On compare la valeur  $Q_{exp}$  aux valeurs critiques de rejet  $Q_{crit}$  selon le tableau le tableau suivant

<b>TABLEAU 4-4</b> Valeurs critiques pour le quotient de rejet $Q^*$			
Nombre d'observations	$Q_{crit}$ (Rejeter si $Q_{exp} > Q_{crit}$ )		
	90% de confiance	95% de confiance	99% de confiance
3	0,941	0,970	0,994
4	0,765	0,829	0,926
5	0,642	0,710	0,821
6	0,560	0,625	0,740
7	0,507	0,568	0,680
8	0,468	0,526	0,634
9	0,437	0,493	0,598
10	0,412	0,466	0,568

# Exemple

L'analyse d'un échantillon de calcite fournit des fractions massiques en CaO de 55,95, 56,00, 56,04, 56,08 et 56,23. La dernière valeur semble anormale ; faut-il la conserver ou la rejeter?

La différence entre 56,23 et 56,08 vaut 0,15, l'étendue (56,23 – 55,95) est égale à 0,28. Dès lors,

$$Q_{exp} = \frac{0,15}{0,28} = 0,54$$

Pour cinq mesures,  $Q_{crit}$  vaut 0,642 pour un degré de confiance de 90%. Si l'on se satisfait de ce degré de confiance, on conservera la valeur excentrique puisque  $0,54 < 0,642$ .

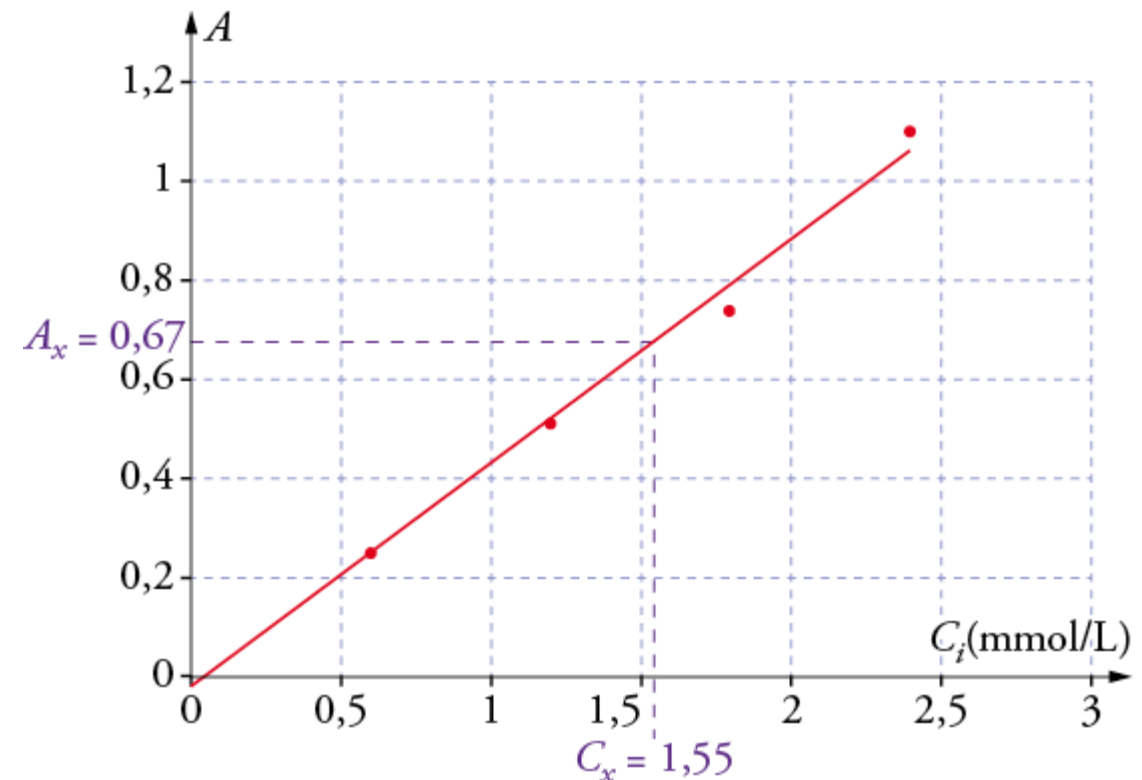


# Régression linéaire

Si on peut mettre en évidence une relation linéaire significative entre deux caractères quantitatifs continus, on peut chercher à formaliser la relation moyenne qui unit ces deux variables.

Les méthodes analytiques sont basées sur une courbe d'étalonnage déterminée expérimentalement. Cette courbe exprime une quantité mesurée  $y$  en fonction d'une quantité connue

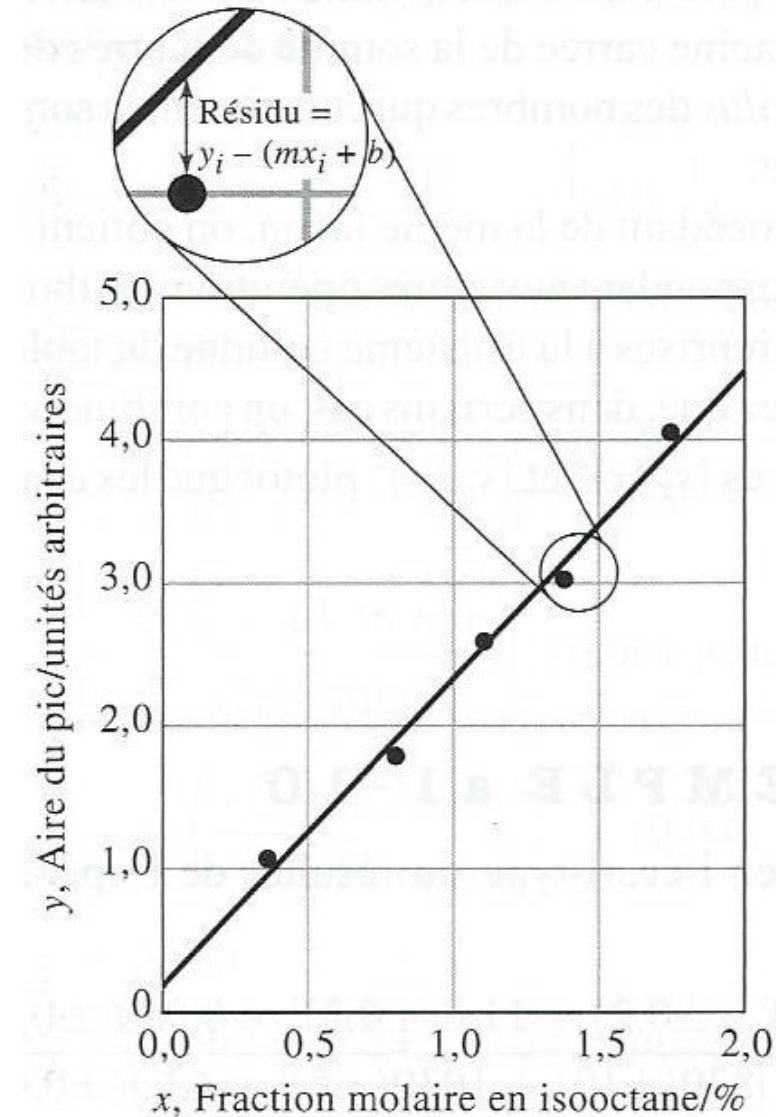
Ci-contre : courbe d'étalonnage de la concentration en glucose-DNS pour une méthode spectrophotométrique de dosage du glucose dans le vin.



# Régression linéaire

En raison des erreurs aléatoires, l'alignement des points n'est pas parfait ; pour trouver la « **meilleure** » droite compatible avec les points, on peut utiliser la méthode des moindres carrés (régression linéaire)

L'axe vertical de chaque point à la droite est appelé *résidu*. La droite des moindres carrés est celle qui **minimise la somme des carrés des résidus de tous les points**. La méthode donne également les écarts-types pour  $m$  et  $b$



# Établissement de la droite de régression par la méthode des moindres carrés

## Hypothèses

1. Il existe vraiment une **relation linéaire** entre la variable mesurée  $y$  et la concentration de l'analyte  $x$  :

$$y = mx + b$$

Où  $b$  = ordonnée à l'origine  
 $m$  = pente de la droite

2. **Tout écart** de mesure avec la droite résulte d'**erreurs de mesure** (pas d'erreur sur la valeur de  $x$ )

⇒ On va rechercher la droite des moindres carrés

# Établissement de la droite de régression par la méthode des moindres carrés

On définit 3 grandeurs :

$$S_{xx} = \sum (x_i - \bar{x})^2 \quad \text{Somme des écarts à la moyenne pour les valeurs individuelles de } x$$

$$S_{yy} = \sum (y_i - \bar{y})^2 \quad \text{Somme des écarts à la moyenne pour les valeurs individuelles de } y$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

Ces trois grandeurs vont être utilisées pour caractériser la droite de régression

# Caractérisation de la droite de régression

1.  $m$  la pente de la droite

$$m = S_{xy}/S_{xx}$$

2.  $b$  l'ordonnée à l'origine

$$b = \bar{y} - m\bar{x}$$

3.  $s_r$  l'écart-type des résidus

$$s_r = \sqrt{\frac{S_{yy} - m^2 S_{xx}}{N - 2}}$$

4.  $s_m$  l'écart-type de la pente

$$s_m = \sqrt{\frac{s_r^2}{S_{xx}}}$$

5.  $s_b$  l'écart-type de l'ordonnée à l'origine

$$s_b = s_r \sqrt{\frac{\sum x_i^2}{N \sum x_i^2 - (\sum x_i)^2}}$$

# Caractérisation de la droite de régression

6.  $s_c$ , l'écart-type pour les résultats obtenus à l'aide de la courbe d'étalonnage

$$s_c = \frac{s_r}{m} \sqrt{\frac{1}{M} + \frac{1}{N} + \frac{(\bar{y}_c - \bar{y})^2}{m^2 S_{xx}}}$$

Dans cette dernière équation :  $\bar{y}_c$  est la moyenne d'une série de  $M$  analyses lorsqu'on utilise une courbe d'étalonnage qui contient  $N$  points

# Établissement des courbes d'étalonnage par la méthode des moindres carrés

*Exemple 1* : appliquez la méthode des moindres carrés aux résultats expérimentaux donnés par le tableau ci-dessous.

Données d'étalonnage pour une méthode chromatographique de dosage de l'isooctane dans un mélange d'hydrocarbures	
Fraction molaire en isooctane, $x_i/\%$	Aire du pic $y_i$
0,352	1,09
0,803	1,78
1,08	2,60
1,38	3,03
1,75	4,01

# Établissement des courbes d'étalonnage par la méthode des moindres carrés

*Exemple 2* : la courbe d'étalonnage précédente a été utilisée pour le dosage chromatographique de l'isooctane dans un mélange d'hydrocarbures. On a obtenu une aire de pic de 2,65. Calculez la fraction molaire en isooctane et l'écart-type du résultat si l'aire est

- (a) le résultat d'une seule mesure
- (b) la moyenne de 4 mesures