Incertitudes de mesure en instrumentation - Etalonnage

1 L'incertitude de mesure

Soit une variable x dont la valeur réelle (mais pas forcément connue) est x_R,

On procède à n mesures, ou évaluations, de x appelées x_i ($1 \leq i \leq n$).

Ces mesures sont généralement différentes de x_R, cette différence est l'erreur, dont on distingue 2 composantes:

1.1 L'erreur systématique: e_B

La valeur moyenne de l'erreur systématique est non nulle: Le mesurage donne une valeur qui s'écarte systématiquement de la valeur vraie.

L'erreur systématique **n'est pas une variable aléatoire** et peut être difficile à déceler : c'est par exemple une erreur de parallaxe dans la lecture d'une indication (aiguille, bille d'un débitmètre…).

L'absence de contrôle et de correction des **facteurs d'influence** -température, pression, humidité, interférents…- amène une erreur systématique.

L'erreur systématique intervient dans la notion de **justesse** : une méthode d'analyse est juste quand on a pu éliminer l'erreur systématique. Cette élimination se fait le plus souvent à l'aide **d'étalons** qui ne doivent pas amener eux-mêmes une erreur .

Une erreur systématique peut être introduite par un **calcul** : une linéarité obtenue à partir d'une régression linéaire sur des points expérimentaux qui obéissent à une loi présentant une certaine "courbure" physique, comme la loi de Beer-Lambert à forte concentration, va introduire une erreur systématique, sauf en deux points de la régression.

D'une façon générale, on peut considérer que l'erreur systématique n'est finalement jamais évaluée car elle est:

- soit inconnue,
- soit connue et alors corrigée (par exemple par comparaison avec un étalon), auquel cas on l'annule.

1.2 L'erreur accidentelle ou aléatoire (e_A)

C'est une variable **aléatoire**. Lorsqu'on répète plusieurs fois le mesurage d'une grandeur physique ou chimique constante, on obtient généralement différentes valeurs plus ou moins dispersées (et qui sont souvent distribuées suivant une loi normale, voir plus loin) : à partir de cette "population" (résultats des mesurages), on va pouvoir estimer la qualité du mesurage et faire un certain nombre de tests.

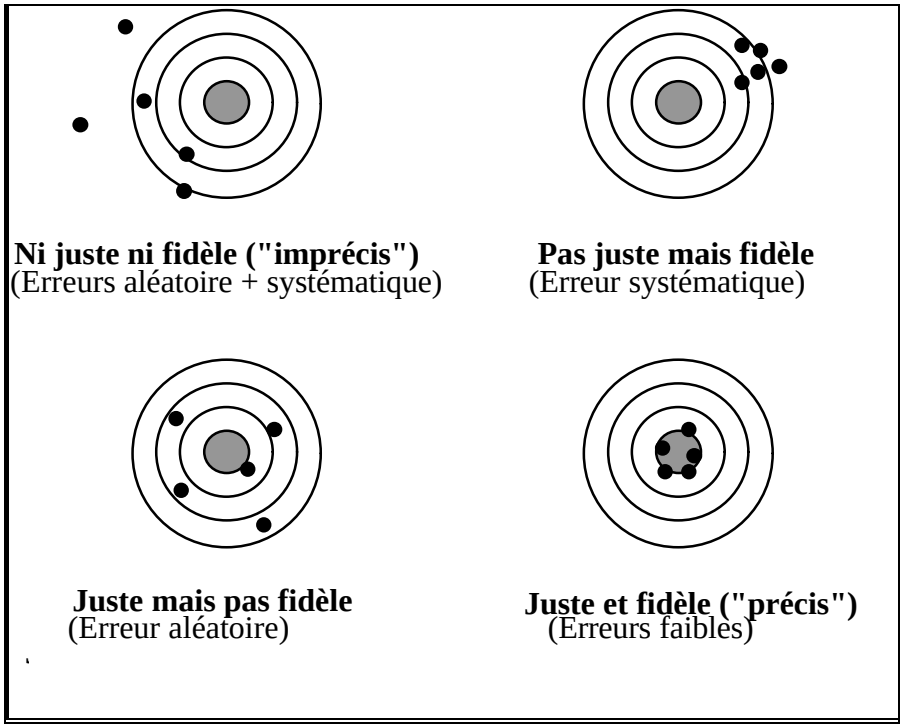
Cette dispersion est parfois masquée par un effet de discrétisation (ex: multimètre digital n'ayant pas assez de résolution), nous verrons plus loin que l'on doit alors introduire une erreur " de discrétisation"

L'erreur accidentelle permet d'introduire les notions de :

- **répétabilité** qui est définie comme l'étroitesse de l'accord entre les résultats de mesurages successifs d'une même grandeur, effectués avec la même méthode, par le même opérateur, avec les mêmes instruments de mesure, dans le même laboratoire, et à des intervalles de temps assez courts.

- **reproductibilité** qui est définie comme l'étroitesse de l'accord entre les résultats de mesurages successifs d'une même grandeur, dans le cas où les mesurages individuels sont effectués : suivant différentes méthodes, au moyen de différents instruments de mesure, par différents opérateurs dans différents laboratoires.

En général, l'accord est moins bon quand il s'agit de reproductibilité. Ces deux types d'erreurs peuvent être illustrées par le tir à la cible:



Remarques

On parle de fidélité de l'instrument et de répétabilité des résultats. Attention à la confusion avec les termes anglo-saxons :

- justesse = accuracy

- répétabilité et reproductibilité = précision

Précision est un terme à éviter en français : on peut dire qu'un appareil "précis" est juste et fidèle.

Par définition, la **valeur moyenne** de l'erreur accidentelle, ou aléatoire, **est nulle**.

Dans la plupart des cas, les erreurs accidentelles ont une **distribution normale**. Cette hypothèse de distribution normale, valable pour 99% des cas, provient du fait que plusieurs sources indépendantes contribuent généralement à cette erreur. Or le *Théorème Central Limite* nous dit qu'une combinaison linéaire d'un nombre suffisamment grand de variables de distributions quelconques tend vers une distribution normale.

1.3 Caractérisation de la mesure et de son erreur:

On peut caractériser la mesure et son erreur aléatoire par:

- **une moyenne estimée** de la mesure qui est la moyenne arithmétique des n mesurages $x_1, x_2 \ldots x_i \ldots x_n$ faits pour caractériser une grandeur :

$$\overline{x} = \frac{\sum_1^n x_i}{n}$$

Si l'erreur systématique est nulle (ce que l'on suppose pour la suite), alors $\bar{x} \to x_R$ quand $n \to \infty$ (x_R = valeur **réelle**, à priori **inconnue**).

- **un écart-type** défini comme étant la racine carrée de la moyenne du carré de l'écart entre la mesure et la valeur réelle x_R:

$$s = \sqrt{\overline{\left((x_i - x_R)^2\right)}} = \sqrt{\frac{1}{n}\sum_1^n (x_i - x_R)^2}$$

Mais généralement, x_R est inconnu, on en a juste une estimation par la moyenne $\bar{x}$. On peut alors calculer une estimation de l'écart type notée s:

$$s = \sqrt{\frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2}\ .$$

Cette expression peut s'écrire aussi:

$$s = \sqrt{\frac{1}{n-1}\left[\sum_1^n x_i^2 - \frac{1}{n}\left(\sum_1^n x_i\right)^2\right]}$$

De même, l'écart type estimé tend vers l'écart type réel quand $n \to \infty$.

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

$$V = s^2$$

Remarque 1: En fait, le mot variance (et sa racine carrée, l'écart type) est souvent employé plus généralement pour caractériser la dispersion d'un jeu de n

valeurs d'une variable Z, même si cette variable n'est pas constante, et s'écrit alors:

$$Var(Z) = s^2(Z) = \frac{1}{n}\sum_1^n \left(Z_i - \bar{Z}\right)^2$$

On devrait donc, dans le cas précédent, parler de variance de l'erreur (ou écart type de l'erreur), mais ce n'est jamais le cas…

Si la distribution des erreurs est **gaussienne** (c'est généralement le cas), on peut alors calculer, à l'aide du tableau des **coefficients de Student**, la probabilité ***p*** pour qu'une mesure individuelle soit hors de l'intervalle $[\bar{x} - ts, \bar{x} + ts]$

Coefs de Student t

intervalle de conf.	90.0%	95.0%	98.0%	99.0%	99.9%
p	0.1	0.05	0.02	0.01	0.001
deg lib					
1	6.31	12.71	31.82	63.66	636.58
2	2.92	4.30	6.96	9.92	31.60
3	2.35	3.18	4.54	5.84	12.92
4	2.13	2.78	3.75	4.60	8.61
5	2.02	2.57	3.36	4.03	6.87
6	1.94	2.45	3.14	3.71	5.96
7	1.89	2.36	3.00	3.50	5.41
8	1.86	2.31	2.90	3.36	5.04
9	1.83	2.26	2.82	3.25	4.78
10	1.81	2.23	2.76	3.17	4.59
12	1.78	2.18	2.68	3.05	4.32
14	1.76	2.14	2.62	2.98	4.14
17	1.74	2.11	2.57	2.90	3.97
20	1.72	2.09	2.53	2.85	3.85
30	1.70	2.04	2.46	2.75	3.65
40	1.68	2.02	2.42	2.70	3.55
50	1.68	2.01	2.40	2.68	3.50
100	1.66	1.98	2.36	2.63	3.39
100000	1.64	1.96	2.33	2.58	3.29

t.s est appelé incertitude élargie et t est le coefficient de Student .

La derniere ligne (n grand) est calculée à partir de l'intégrale de la Gaussienne (fonction erf)

Dans le cas de l'estimation de l'écart-type, le nombre de degrés de liberté est n-1.

Ainsi, pour n suffisamment grand (n>20), 95% des mesures sont dans l'intervalle $[\bar{x} - 2s, \bar{x} + 2s]$, et 99.7% dans l'intervalle $[\bar{x} - 3s, \bar{x} + 3s]$

Si l'on prend maintenant pour mesure la moyenne de n mesures individuelles indépendantes, alors la valeur de l'écart type devient : $s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$

Bien que l'acquisition des données devienne plus longue, on a souvent intérêt à moyenner le plus grand nombre possible de mesures.

Le résultat d'une mesure doit comporter **4 éléments :**

Ex : C_{NO} = 125.3 ppb ± 1.7 ppb (k=2)
1 2 3 4

1 : **Valeur numérique** avec un nombre correct de décimales

2 : **Unité**

3 : **Incertitude élargie** = t.σ (= intervalle de confiance x 2)

4 : **Le coefficient d'élargissement t utilisé** (généralement noté k : ex k=2)

Si la mesure est non nulle, on utilise souvent **l'incertitude** (élargie) **relative (à la mesure)**, calculée à partir de l'écart type relatif : $s_r = \frac{s}{|x|}$, souvent exprimée en pourcentage de la mesure.

1.4 Evaluation de l'incertitude :

On distingue 2 types de méthode :

Evaluation de type A : par analyse statistique de séries de mesures, à l'aide des formules du paragraphe précédent

Evaluation de typeB : Par tout autre moyen : généralement, on évalue l'effet sur l'incertitude finale des différentes sources d'incertitude, elles même évaluées :

- Par une méthode de type A,
- Par des données constructeur, d'étalonnage etc…

*Remarque: L'incertitude donnée par le constructeur peut être constante **pour un appareil donné**. Mais on ne connaît ni sa valeur, ni même son signe, elle sera donc considérée comme une erreur aléatoire.*

On utilise alors **les lois de propagation des écart-type** (ou des incertitudes, au facteur d'élargissement près) :

Si la mesure finale, y, est fonction de variables x_i, dont les $\sigma(x_i)$ sont connus:

$y=f(x_1 \ldots x_n)$, alors l'écart type de y s'écrit :

$$s^2(y) = \sum_1^n \left(\frac{\partial f}{\partial x_i} \right)^2 s^2(x_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} Cov(x_i, x_j)$$

Les termes $Cov(x_i,x_j)$ sont les covariances et sont donc nuls si les x_i sont indépendants.

(*rappel* : $Cov(X,Y) = \frac{1}{n}\sum_{1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$)

Exemples : *(variables indépendantes, donc pas de terme de covariance)*

1 : **Combinaison linéaire** $y = \sum_i a_i x_i$, alors $s^2(y) = \sum_i (a_i s(x_i))^2$

Ceci permet de retrouver la formule pour la **moyenne** (2 mesures successives ont des erreurs indépendantes car aléatoires) : $\bar{x} = \frac{1}{n}\sum_{1}^{n} x_i$ donc $s(\bar{x}) = \frac{1}{n}\sqrt{\sum_{1}^{n} s^2(x)} = \frac{s(x)}{\sqrt{n}}$

Ex: somme ou différence: $y = a + b - c$, alors $s(y) = \sqrt{s^2(a) + s^2(b) + s^2(c)}$

2 : **Produits et puissances** : $y = A\prod_i x_i^{a_i}$

Il est dans ce cas plus commode de considérer l'écart type relatif $\frac{s(y)}{|y|}$:

On a alors : $\left(\frac{s(y)}{y}\right)^2 = \sum_i \left(a_i \frac{s(x_i)}{x_i}\right)^2$

Ex : $y = K.\frac{a^2 b}{c}$ où K=constante, alors $\frac{s(y)}{|y|} = \sqrt{\left(\frac{2s(a)}{a}\right)^2 + \left(\frac{s(b)}{b}\right)^2 + \left(\frac{s(c)}{c}\right)^2}$

Ou encore: $s(y) = K\sqrt{\left(\frac{2ab}{c}s(a)\right)^2 + \left(\frac{a}{c}s(b)\right)^2 + \left(\frac{a^2 b}{c^2}s(c)\right)^2}$

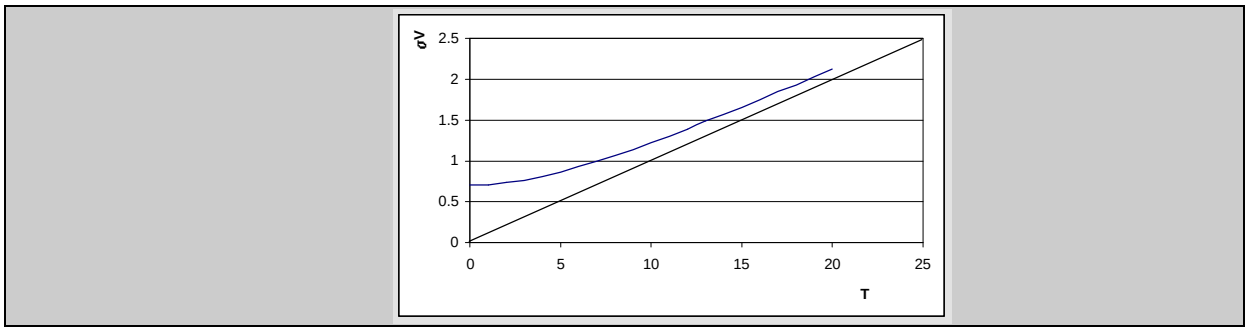
Exemple important: la droite de calibrage

Soit un capteur de température fournissant une tension V, ce capteur a été étalonné et on connaît des évaluations des coefficients a et b tels que: T=aV (cf chapitre étalonnage).

On a alors: $s_T = \sqrt{a^2 s_V^2 + V^2 s_a^2}$.

On voit alors que si V faible, $s_T = Cte = a s_V$

Et pour V suffisamment élevé, σ_T est proportionnel à V, $s_T = s_a.V$



Attention, ces lois ne seront qu'approximatives avec les estimations s des écart types.

Parmi les effet contribuant à l'incertitude, ont doit souvent considérer l'incertitude due à la discrétisation :

1.5 l'incertitude due à la discrétisation :

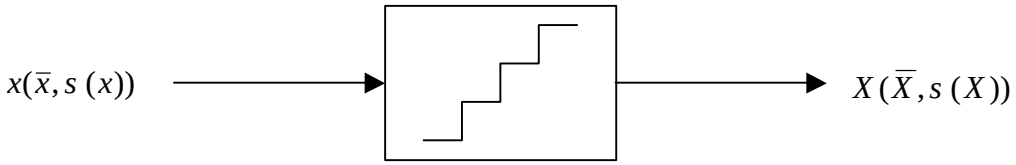
La discrétisation intervient lorsque les valeurs finales constituent un ensemble discret (et généralement fini). On peut alors définir un pas de mesure (ou parfois digit, ou LSB, Least Significant Byte) qui est le plus petit écart mesurable.

On parle aussi de résolution du système d'acquisition, à ne pas confondre avec l'incertitude.

Ex: - multimètre digital affichant 000.00 V, le digit ici est de 10 mV

- Convertisseur analogique digital: composant électronique permettant de convertir une tension en un entier. Exemple: convertisseur 0-5V, 12 bits: résultat = entier de 12 bits, donc le LSB est $5/(2^{12}) = 1.22$ mV

- règle graduée: pas de mesure de 1 mm, on suppose que l'on n'essaie pas de lire "entre les graduations"



Lorsqu'on lit une valeur X, alors cette valeur peut correspondre à une valeur réelle x comprise entre X-δ/2 et X+δ/2 avec une distribution uniforme. On introduit donc une erreur supplémentaire comprise entre -δ/2 et +δ/2

On peut montrer que l'écart type lié à la discrétisation est :

$$s_d = \sqrt{\frac{1}{d}\int_{-\frac{d}{2}}^{\frac{d}{2}} x^2 dx} = \frac{d}{2\sqrt{3}} = 0.29d$$

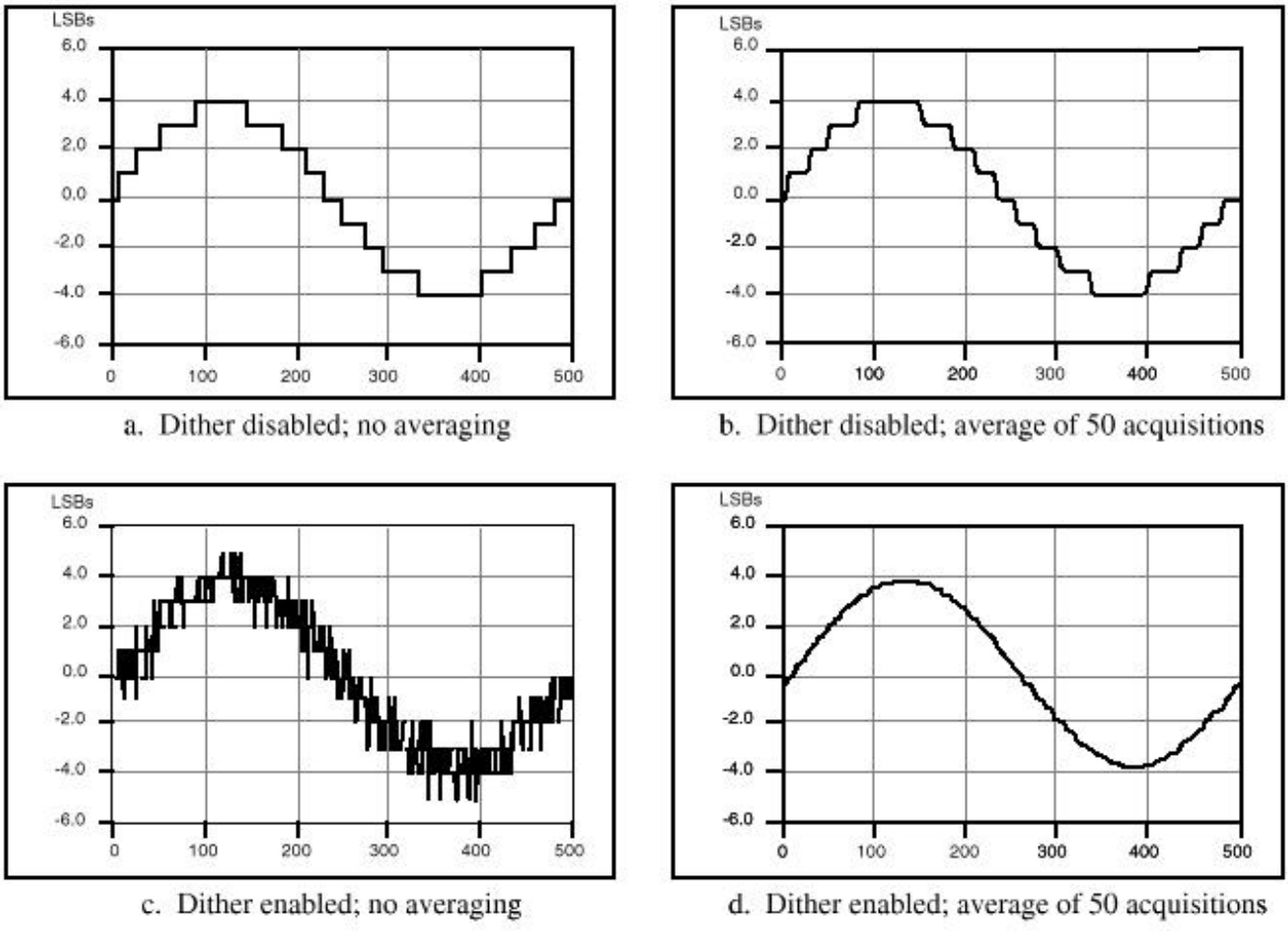
L'écart type global sera donc: $s(X) = \sqrt{s^2(x) + s_d^2}$

Dithering:

Si l'on fait une moyenne $\overline{X}$ de n mesures x_i ET si $\sigma(x)$ est suffisamment grand (au moins du même ordre de grandeur que σ_d), alors $s(\overline{X}) \rightarrow \frac{s(x)}{\sqrt{n}}$ et la distribution de l'erreur tend à devenir gaussienne (d'après le Théorème central limite). Cette méthode (dite "dithering") est utilisée afin d'améliorer "artificiellement" le "pas de mesure" de certains systèmes d'acquisition. Mais cela se fait au détriment de la vitesse (n mesures nécessaires).

Paradoxe: l'ajout de bruit peut (dans ce cas très précis) améliorer la mesure!

Si $\sigma(x) << \sigma_d$, alors X=constante et $\overline{X} = X$, dithering impossible! Certains dispositifs peuvent *optionnellement* ajouter à x un signal de *bruit blanc* pour augmenter $\sigma(x)$ (cartes National Instrument).



Dithering (ajout d'un bruit de 0.5 LSB RMS) sur un signal faible sinusoï dal (-6 à +6 LSB) (National Instruments)

1.6 La limite de détection (lower detection limit LDL):

Exemple: analyseur chimique (mesure de concentration)

Dans certains cas, il est important de pouvoir prendre une décision sur la présence ou l'absence d'un composé dans un échantillon : toxicité, qualité d'usage liées à ce composé à une concentration déterminée, il faudra alors une méthode de mesure adaptée à cette concentration.

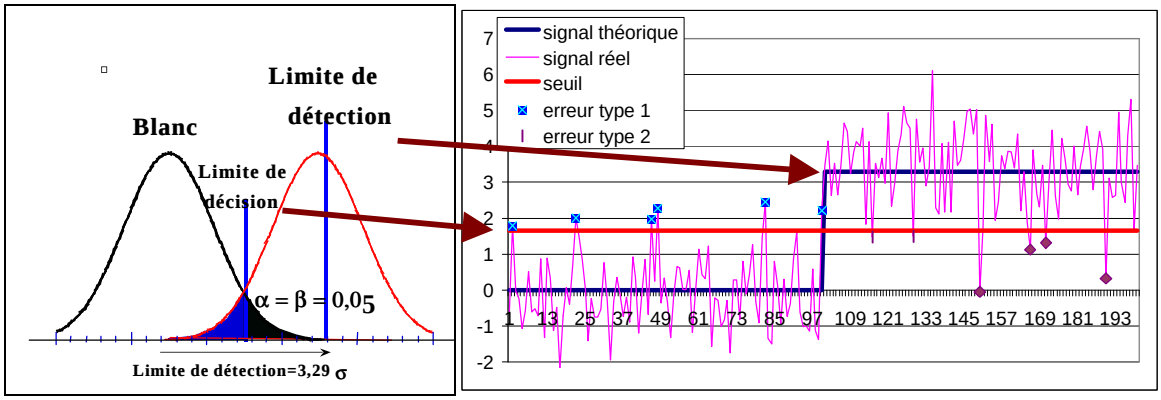
Il faudra donc pouvoir annoncer, avec un certain risque, la présence ou l'absence de ce composé.

- déclarer que le composé est présent alors qu'il n'y est pas, c'est le risque α de 1ère espèce

- déclarer que le composé est absent alors qu'il est présent, c'est le risque β de 2ème espèce.

Si on considère que la distribution est normale avec un écart-type σ, on peut utiliser les tests d'hypothèse, ces tests étant unilatéraux puisqu'il n'existe pas de concentrations négatives. La probabilité de 5% va être « bloquée à droite », on montre que cette limite de 5% correspond à une valeur de $1{,}64\sigma$ (et non pas $1{,}96\ \sigma$ qui s'applique aux tests bilatéraux).

On appelle cette valeur **seuil (ou limite) de décision** (on compare la mesure avec elle pour décider si le composé est présent ou pas).



Lorsqu'on trouve une valeur expérimentale (ou une moyenne) égale au seuil de décision (1.64σ), et que le composé est absent, la probabilité d'erreur de première espèce (α) sera de 5%. Par contre, si le composé est présent avec une concentration égale au seuil de décision, la probabilité d'erreur de seconde espèce (β) sera de 50%, en effet, dans 50% des cas, on annoncera une concentration nulle alors qu'elle ne l'est pas!

On ne pourra égaler les 2 risques α et β, que si la mesure est égale à 2 fois le seuil de décision. Si c'est 5 % on doit donc se placer à 2 x 1,64 σ (et non pas 1,96 σ car risque unilatéral !) donc 3,28 σ, dans ces conditions $\alpha = \beta = 5$ %.

(Le risque total sera toujours 5% et pas 10 % puisque on ne peut pas avoir les 2 types d'erreur en même temps).

Le **seuil (ou limite) de détection** est donc défini comme étant la plus petite quantité ***détectable*** en égalisant les probabilités des 2 types de risque à 5%. Il est donc égal à **2 fois le seuil de décision** et à environ 3 fois (3.28) l'écart type de la mesure proche de 0 (car l'écart type n'est pas forcément constant sur toute la gamme).

De même, pour une probabilité d'erreur de 1%, il sera égal à 4.66σ, et pour 0.1%, à 6.2σ.

Plus on exigera une détection fiable (probabilité d'erreur faible), plus le seuil de détection devra être élevé (pour un écart type de signal donné). On aura donc aussi intérêt à diminuer le niveau de bruit de l'analyseur, notamment en effectuant un moyennage des mesures.

On peut aussi définir un seuil de détection de variation de la mesure: on est alors obligé de soustraire 2 mesures ("faire un blanc"), or $s(a-b) = \sqrt{s^2(a) + s^2(b)}$

$$\sigma = \sqrt{\sigma_B^2 + \sigma_B^2} = \sigma_B \sqrt{2}$$

Si on fait la détermination en soustrayant le blanc, il faut alors faire intervenir le facteur $\sqrt{2}$, et le seuil de détection est: $S = 3.28 * \sqrt{2} * s(x)$, ($\sigma$ évalué au voisinage du point de fonctionnement).

2 L'étalonnage

L'étalonnage (calibration in English) d'un instrument de mesure ou d'un capteur, consiste à modéliser le signal de sortie du capteur (appelé Xvariable) en fonction de la variable mesurée (appelée Yvariable). Pendant l'étalonnage, les Yvariables, considérées comme étalons, sont évaluées, ainsi que les Xvariables associées.

Le modèle est alors une fonction F: $\hat{Y} = F(X)$

L'utilisation normal du modèle, ou prédiction, permet d'évaluer les $\hat{Y}$ à l'aide des seuls X et du modèle F.

Ex: capteur de pression: L'étalonnage nécessite de générer plusieurs valeurs de pressions $Y_1..Y_n$ représentatives de la gamme du capteur. Ces pressions sont connues soit à l'aide d'un autre capteur, soit parce que l'on utilise des grandeurs étalon. Pour chaque valeur de pression, une mesure de tension en sortie du capteur X_i est réalisée. On peut alors évaluer la fonction f qui est généralement une droite de régression, on fait alors l'hypothèse de linéarité.

En prédiction (utilisation normale du capteur), la droite de régression permet de calculer la valeur de la pression à partir de la seule valeur de tension du capteur.

2.1 Méthode générale

La méthode générale consiste à:

- Faire une hypothèse sur le type de loi mathématique décrite par la fonction F (ex: linéaire)

- Déterminer, par une méthode adéquate, les paramètres de la fonction rendant celle-ci la "plus performante possible" en terme de modélisation: On cherche généralement à minimiser la somme S des carrés des erreurs de prédiction:

$$S = \sum \left\| Y_i - \hat{Y}_i \right\| \quad \text{où } \hat{Y}_i \text{ est la valeur calculée par le modèle.}$$

2.2 performance de la prédiction

L'écart type estimé de l'erreur de prédiction comprend 2 composantes: $s^2 = s_m^2 + s_a^2$ qui sont:

2.2.1 L'erreur de modélisation

σ_m représente l'erreur de modélisation qui peut avoir 2 sources:

- Les données (X et Y) utilisées lors de l'étalonnage contiennent une composante d'erreur aléatoire. Le modèle ne peut donc être parfait si le nombre de "points" d'étalonnage est fini (mais peut tendre à l'être si ce nombre est grand…)

- L'hypothèse sur la forme de la fonction F peut (et même est généralement) fausse: il s'agit donc ici d'erreur systématique qui est, comme cela a déjà été dit plus haut, soit inconnue soit corrigée. On distingue 3 cas:

    1. Le modèle est trop simple: on parle de sous-modélisation, c'est par exemple le choix d'une droite de régression alors qu'il existe une courbure.

    2. Le modèle est trop complexe, on parle alors de sur-modélisation. Ex: choix d'un polynôme de degré 5 alors qu'il y a peu de points expérimentaux (au moins 5 quand-même!). Le danger est ici de **modéliser des particularités des échantillons**, dont leurs erreurs aléatoires, alors que le but est de **modéliser l'information commune**. Ce danger tend à s'effacer si le nombre de points est suffisamment grand.

    3. Hypothèse fausse, sans commentaire…

2.2.2 L'erreur aléatoire "liée à la mesure"

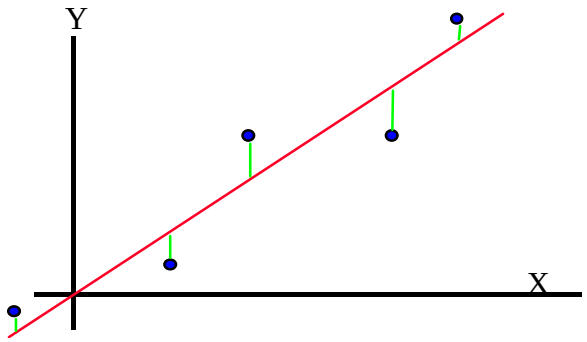
La mesure des Xvariables est entachée d'une erreur aléatoire caractérisée par l'écart type estimé σ_X. L'erreur aléatoire σ_a de mesure est la composante due à cette erreur, propagée à travers le modèle selon les lois citées plus haut.

D'une façon générale, l'erreur relative $\frac{s(y)}{|y|}$ sera d'autant plus amplifiée que le modèle sera complexe (ex: polynôme de degré élevé avec de forts coefficients en valeur absolue). *(Il s'agit plus exactement de l'erreur relative à la variance des variables).*

2.3 La régression linéaire

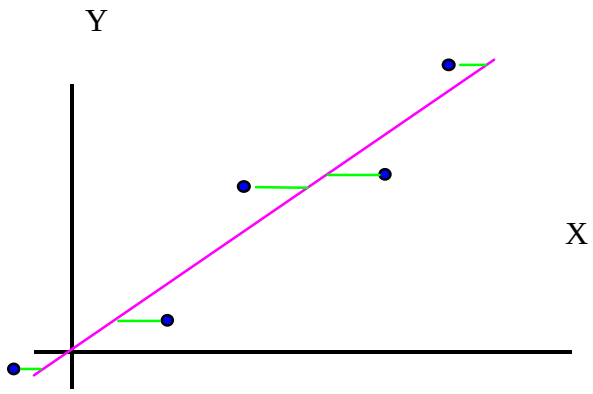
C'est le cas le plus simple qui sera seul traité ici. Il s'agit de l'hypothèse linéaire dans le cas monovariable (Xvariable et Yvariable = scalaire).

Il y a en fait 2 cas:



Minimisations des erreurs sur Y

- **Regression de y/x:** On cherche les coefficients a et b tels que $\hat{y} = ax + b + e_y$ où e_y constitue l'erreur sur les yvariables qui sera minimisée. On suppose donc ici que l'erreur est apportée par les Yvariables.



Minimisation des erreurs sur X

- **Regression de x/y:** On cherche les coefficients a' et b' tels que $\hat{x} = a'y + b' + e_x$ où e_x constitue l'erreur sur les xvariables qui sera minimisée. On suppose donc ici que l'erreur est apportée par les Xvariables. Pour la prédiction, on utilise: $\hat{y} = \frac{1}{a'}x - \frac{b'}{a'}$ Cette méthode est plus appropriée lorsque les Yvariables sont des étalons de qualité.

Les résultats sont différents ($a' \neq \frac{1}{a}$ et $b' \neq \frac{-b}{a}$) mais généralement proches. Toutefois, une erreur aléatoire sur les Xvariables dans le premier cas ou sur les Yvariables dans le second cas entraîne une erreur systématique, ou biais, sur a et b (par rapport à leurs valeurs REELLES). Mais:

1: Cette erreur est souvent négligeable.

2: Elle n'est dommageable que si l'on cherche à déterminer la valeur réelle de a et b (mesure d'une grandeur physique). Dans le cas de l'étalonnage, où l'on veut déterminer des grandeurs inconnues à partir d'une droite, on montre que la méthode reste valable.

En fait, dans la plupart des cas, on a une erreur aléatoire sur les Xvariables ET sur les Yvariables…

2.3.1 Caractéristiques de la régression y/x:

On suppose donc que σ_x=0 et aussi que σ_y ne dépend pas de x.

On peut montrer que les valeurs "optimales" de a et b sont:

$$a = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)}$$

$b = \bar{y} - a\bar{x}$ (la droite de régression passe par $\bar{x}, \bar{y}$).

Estimation de "l'écart type des résidus": $s_{y/x} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}}$ où $\hat{y}_i = ax_i + b$

($y_i - \hat{y}_i$ est appelé résidu de la $i^{ème}$ mesure = composante non modélisée)

Ecart type estimé de la pente: $s_a = \frac{s_{y/x}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$

Ecart type estimé de l'ordonnée à l'origine: $s_b = s_{y/x}\sqrt{\frac{\sum_i x_i^2}{n\sum_i (x_i - \bar{x})^2}}$

2.3.2 Estimation de l'écart type de l'erreur de prédiction:

Soit $\bar{X}$ la mesure obtenue à partir de la moyenne de n mesures individuelles. Par hypothèse (régression de y sur x), seule la mesure de y (donc pendant l'étalonnage) comporte une erreur caractérisée par $s_{y/x}$.

Soient a et b la pente et ordonnée à l'origine estimés lors d'une regression linéaire de y sur x portant sur m points x_i, $1 \le i \le m$, mesurés dans les mêmes conditions.

Une estimation de Y sera: $\hat{Y} = a\bar{X} + b$

Alors l'estimation de l'écart type de prédiction s'écrit: $s(\hat{\bar{Y}}) = s_{y/x}\sqrt{\frac{1}{n} + \frac{1}{m} + \frac{(\bar{X} - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$

(où x_i et $\bar{x}$ sont respectivement la i^{eme} mesure d'étalonnage et la moyenne des mesures d'étalonnage)

Et la mesure s'écrit: $Y = a\bar{X} + b \pm ts(Y)$ (k=…)

Où t est le coefficient de Student correspondant à un nombre de degrés de liberté égal à m-2.

Dans cette expression, le terme $\frac{1}{n}$ représente l'erreur liée à la mesure de Y (ecart type estimé: $\frac{s_{y/x}}{\sqrt{n}}$)

Le terme restant correspond à l'erreur de modélisation, d'écart type estimé: $s_{y/x}^2\left(\frac{1}{m}+\frac{\left(\overline{X}-\overline{x}\right)^2}{\sum\left(x_i-\overline{x}\right)^2}\right)$. Ce terme passe par un minimum pour $\overline{X}=\overline{x}$, moyenne des x_i d'étalonnage. Le résultat sera donc "meilleur" en milieu de gamme qu'aux extrémités.

Bien que cela ne soit pas forcément rigoureux, on peut extrapoler au cas où l'erreur vient de la Xvariable (écart type estimé s_x):

$$s(\hat{\overline{Y}})=\sqrt{(\frac{1}{n}+\frac{1}{m})a^2s_x^2+\frac{\left(\overline{X}-\overline{x}\right)^2 s_{y/x}}{\sum\left(x_i-\overline{x}\right)^2}}\approx s_{y/x}\sqrt{\frac{1}{n}+\frac{1}{m}+\frac{\left(\overline{X}-\overline{x}\right)^2}{\sum\left(x_i-\overline{x}\right)^2}}$$

2.3.3 Mesure de la qualité d'ajustement d'une régression linéaire:

On peut considérer que la quantité d'information totale contenue dans une variable Y est proportionnelle à sa variance $Var(Y)=\frac{1}{n}\sum_1^n\left(Y_i-\overline{Y}\right)^2$

Remarque 2 (rappel): Il s'agit ici de l'estimation de la variation de différentes valeurs d'une variable, à ne pas confondre avec $s^2(Y)$, estimation de l'écart de différentes mesures par rapport à une même valeur à priori inconnue de la variable, et donc de l'erreur; d'où le terme en 1/n au lieu de 1/(n-1) (cf remarque 1).

On peut ensuite décomposer cette quantité d'information totale (variance de Y) en deux quantités complémentaires : celle qui peut être reconstituée à partir de la connaissance de la variable X (variance des valeurs estimées de Y, notées $\hat{Y}$) et celle qui **ne peut pas** être reconstituée à partir de la connaissance de X (variance des résidus de la régression ε=Y-$\hat{Y}$). Au total, on peut définir la relation suivante :

Var(Y)	**=**	**Var($\hat{Y}$=aX+b)**	**+**	**Var(ε)**
information totale	=	*information modélisée*	+	*information résiduelle*

La qualité de l'ajustement correspond donc au rapport entre l'information totale sur Y et l'information effectivement reconstituée à partir de la connaissance procurée par la variable X. Cette qualité d'ajustement varie entre 0% (X n'apporte aucun élément de prévision sur Y) et 100% (la connaissance des valeurs de X permet de prévoir intégralement les valeurs de Y) et dépend de l'intensité de la corrélation entre X et Y.

Ce rapport est le coefficient de détermination:

Qualité d'ajustement = $\frac{Var(\hat{Y})}{Var(Y)}=\frac{\sum\left(\hat{Y}-\hat{\overline{Y}}\right)^2}{\sum(Y-\overline{Y})^2}=\frac{a^2Var(X)}{Var(Y)}$ **= coeff. de détermination R_d**

On rencontre aussi souvent le coefficient de corrélation: $R_c=\frac{Cov(X,Y)}{s(X)s(Y)}$

On montre que $R_d=R_c^2$

2.4 Autres méthodes de modélisation: (voir biblio)

La modélisation Multivariable (ou analyse Multivariable): Les variables X et Y ne sont plus des scalaires mais des vecteurs, le but est alors de trouver une fonction vectorielle F telle que Y=F(X).

Si cette fonction est linéaire, elle est alors une matrice. De nombreuses méthodes linéaires existent: Classic Least Squares, Partial Components Regression, Partial Least Square. Ces 2 dernières sont basées sur la décomposition en composantes principales.

Dans le cas non linéaire, les réseaux de neurones (Neural Network) sont de plus en plus utilisés.

Voir document "Analyse Multivariables" de P. Breuil Ecole des Mines St Etienne : http://www.emse.fr/ECOLE/FRENCH/SPIN/instrum/pb/anamu/index.html

3 Bibliographie:

- "Statistics for analytical Chemistry" 3 rd ed. J.C. Miller and J.N. Miller John Wiley & Sons, 1998
- "Multivariate Statistical Methods, A Primer" B.F.J. Monley, Chapman & Hall 1986
- *"Multivariate Calibration"* **Harald Martens, Tormod Naes** ed: John Wiley & sons Chichester
- "*Practical Guide to Chemometrics"* **Stephen John Haswell** ed: Marcel Dekker, Inc New York
- CETAMA "Statistique appliquée à l'exploitation des mesures" 2e éd. Masson 1986

Sites WWW:

http://www.emse.fr/fr/transfert/spin/formation/axes/inst/capmes/ : Ce document + exercices sous Excel

http://www.galactic.com/Algorithms/ : algorithmes de calibrage, plutôt multivariable.

http://physics.nist.gov/cuu/Uncertainty/index.html: excellent site simple et succinct, ce document s'en est beaucoup inspiré.

http://rfv.insa-lyon.fr/~jolion/STAT/poly.html : cours complet et touffu de probas & stats.

http://www-sv.cict.fr/lsp/Besse/Hyper/modlinhtml/modlinhtml.html : idem pour la modélisation statistique (la régression linéaire dans tous ses états).

http://math.uc.edu/~brycw/classes/147/blue/tools.htm : collection de liens, books online et applets.