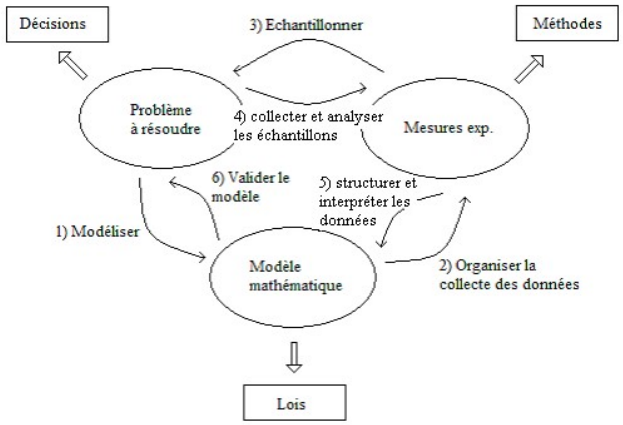
# UE XLP5CE041 : Chimiométrie

Chimiométrie = utilisation d'un ensemble d'outils mathématiques

pour analyser des données expérimentales de chimie

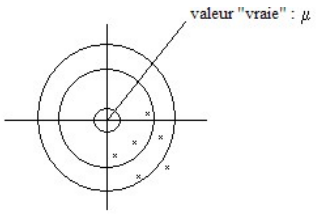
M. Le Guennec-Abguéguen
Département de Chimie, Nantes Université

Méthodologie générale :

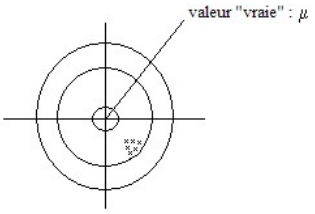


Qq caractéristiques importantes des méthodes d'analyses :

**la justesse** : c'est l'étroitesse de l'accord entre une mesure (ou la moyenne des mesures) et la valeur "vraie" de l'échantillon.




méthode ni juste, ni fidèle




méthode fidèle
mais pas juste

**la répétabilité** : il s'agit de la mesure de la fidélité si les opérateur, instrument(s), méthode, dans des délais courts sont les mêmes. Il est représenté par l'écart-type (s) si celui-ci est constant ou par le coefficient de variation s'il varie.

⇒ **la limite de répétabilité** (ou répétabilité), notée r, est l'écart maximal au niveau de confiance de 95% entre deux résultats obtenus sur un échantillon commun par une même méthode, un même opérateur travaillant sur le même appareil dans un délai court, c'est-à-dire selon des conditions de répétabilité.

**la reproductibilité** : il s'agit de la mesure de la fidélité lorsque n'importe quelle condition change (opérateur et/ou instrument(s) et/ou méthode et/ou délai d'exécution, etc.)

⇒ **la limite de reproductibilité** (ou reproductibilité), notée R, est l'écart maximal au niveau de confiance de 95% entre deux résultats obtenus sur un échantillon commun par deux opérateurs ou deux laboratoires différents, sur des appareils différents, parfois selon des méthodes d'analyse différentes, c'est-à-dire selon des conditions de reproductibilité.

Si l'on veut mesurer la reproductibilité d'une méthode, il sera nécessaire que tous les participants d'une analyse inter-laboratoire utilisent la même méthode.
Si l'on s'intéresse à la reproductibilité de l'analyte et non plus de la méthode, des méthodes différentes peuvent être utilisées

**la robustesse** : c'est l'importance des effets observés lorsque l'on fait subir de légères variations contrôlées aux conditions opératoires

## Le plan

**1re partie : analyses statistiques à une variable**

1. Notion de probabilité
2. Vérification du caractère aléatoire de collecte de données
3. Variable aléatoire, fonctions de distribution et de répartition
4. Représentations et valeurs caractéristiques d'une VA
5. Les lois de distribution
6. Vérification de la loi de probabilité d'une distribution
7. Intervalles de confiance
8. Comparaisons sur échantillons
9. Etude de cas

**2me partie : analyses statistiques à deux variables**

**=> les régressions linéaires simples**

1. Notion de corrélation et régression
2. La régression linéaire simple sans contrainte => RLS/MMCC
3. La régression linéaire simple avec contrainte => RLS/MMCF
4. Applications à la chimie analytique / Etude de cas
5. L'analyse des résidus

⇒ **2 évaluations : coefs 50-50**

# 1re partie :

# analyses statistiques

# à une variable

## 1. Notion de probabilité

définition : La **probabilité** est une évaluation du caractère **probable** d'un évènement. Lorsque le hasard intervient dans sa réalisation, celui-ci est dit aléatoire.

Soit une expérience, appelée épreuve, présentant n cas (résultats) possibles dont seulement $n_i$ sont favorables,

la probabilité de réalisation de l'évènement est : $prob(E_i) = \frac{n_i}{n} = f_i = p_i$

donc $0 \leq prob(Ei) \leq 1$

et $prob(\text{non } E_i) = prob(\overline{E_i}) = 1 - Pr(E_i) = q_i$

**=> exercice N°1**

## 2. Vérification du caractère aléatoire de collecte de données

Une donnée expérimentale correspond à $x_i = \mu + L_i + \varepsilon_i$

or $\mu$ n'est jamais connue, mais seulement estimée par m
et L et $\varepsilon$ sont décomposables en plusieurs sources d'erreurs :
évolution de l'échantillon dans le temps, erreurs dues aux solutions étalons, déréglage de l'appareil, erreurs de calcul, interférences, erreurs de préparation de l'échantillon, etc.

On introduit alors la notion d'erreurs
- **personnelles** (dues à un travail mal réalisé),
- **systématiques** (détérioration progressive d'une solution ou d'un échantillon, dérèglement progressif d'un appareil, ...)
- **aléatoires** (erreurs expérimentales simplement dues au hasard).

Le **test des signes** est l'une des méthodes permettant cette distinction des erreurs.

**étape 1** : Les N données expérimentales (une 20aine au minimum) étant rangées dans l'ordre **chronologique** de leur collecte, on commence par annoter une donnée d'un signe +, - ou 0 selon que la valeur considérée est respectivement plus grande, plus petite ou égale à la valeur qui a précède puis on comptabilise le nombre de groupes R de signes + et -

**étape 2** : On calcule la fonction discriminante $U = \frac{3R - 2N + 1}{\sqrt{1,6N - 2,9}}$

**étape 3** : Comme $prob(-1,96 < U < 1,96) = 0,95$
et $prob(-2,575 < U < 2,575) = 0,99$
les fluctuations expérimentales observées seront considérées comme aléatoire au niveau de confiance P si U est bien compris entre les valeurs limites au seuil de risque choisi

**=> exercice N°2**

## 3. Variable aléatoire, fonctions de distribution et de répartition

Soit la **variable** aléatoire étudiée, notée **X**, pouvant prendre les valeurs $x_i$ ($\in$ R) déterminées par le **résultat** d'une épreuve

la VA X peut être **discrète** ou **continue**

sa fonction de distribution est notée $f_X(x_i) = prob(X = x_i)$

sa fonction de répartition est notée $F_X(x_i) = prob(X \leq x_i)$

$$F_X(x_i) = \sum_{-\infty}^{x_i} f_X(x_i) \quad \text{si X est discrète}$$

$$F_X(x_i) = \int_{-\infty}^{x_i} f(x_i)dx \quad \text{si X est continue}$$

**=> exemples**

## 4. Représentations et valeurs caractéristiques d'une VA

⇒ **Représentations individualisées**

reprendre données exercice 2

⇒ **Représentations par classes**

reprendre données exercice 2

RQ. soient N données expérimentales,

nombre minimal de classes : $k \geq 1 + {}^{10}/_{3}\ \log N$ → **relation de Sturges**

amplitude maximale des classes = $(x_{max} - x_{min})/k$

→ par la suite, l'effectif de la classe est attribué au centre de la classe concernée

⇒ **Valeurs caractéristiques :**

il existe deux catégories de valeurs caractéristiques : les paramètres de position et les paramètres de dispersion

- la **moyenne** : $\bar{X} = \sum_{i=1}^{n} x_i p_i$ si la VA est discrète

$$\bar{X} = \int_{-\infty}^{+\infty} x_i f_X(x_i) dx \quad \text{si la VA est continue}$$

**paramètres de position**

aussi notée m ou $\mu$

- le **mode**

- la **médiane**

- les quartiles, déciles, etc

⇒ **Détermination des quartiles $Q_1$, $Q_2$ (=Médiane), $Q_3$ :**

- le 1er quartile ($Q_1$) est la donnée de la série qui sépare les 25 % inférieurs des données
- le 2e quartile ($Q_2$) est la donnée de la série qui sépare les 50 % inférieurs des données
- le 3e quartile ($Q_3$) est la donnée de la série qui sépare les 75 % inférieurs des données

➢ **Représentations individualisées / Séries discrètes**

S'il y a *N* valeurs :

- le premier quartile est celui qui a le rang $\frac{N+3}{4}$
- le deuxième quartile (médiane) est celui qui a le rang $\frac{2N+2}{4}$ donc $\frac{N+1}{2}$
- le troisième quartile est celui qui a le rang $\frac{3N+1}{4}$

**Quand le rang d'un quartile n'est pas une valeur entière, on procède à une extrapolation linéaire :**

- Si le rang se termine par 0,25, alors le quartile est la moyenne entre $R_{inf}$ affecté du coefficient 3 et de $R_{sup}$ affecté du coefficient 1
- Si le rang se termine par 0,5, alors le quartile est la moyenne entre $R_{inf}$ et $R_{sup}$ (sans coefficient)
- Si le rang se termine par 0,75, alors le quartile est la moyenne entre $R_{inf}$ affecté du coefficient 1 et de $R_{sup}$ affecté du coefficient 3

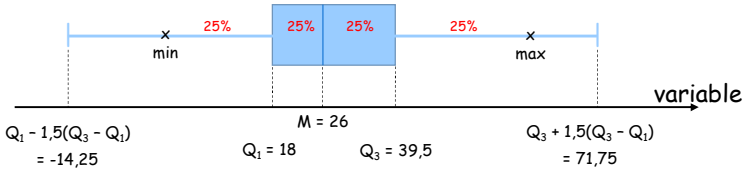
Exemple : soit la série 1, 11, 15, 19, 20, 24, 28, 34, 37, 47, 50, 61

$Q_1$ : rang $\frac{N+3}{4} = \frac{12+3}{4} = 3{,}75$ donc $Q_1 = \frac{15*1+19*3}{4} = 18$

$Q_2$ : rang $\frac{N+1}{2} = \frac{12+1}{2} = 6{,}5$ donc $Q_2 = \frac{24+28}{2} = 26$

$Q_3$ : rang $\frac{3N+1}{4} = \frac{3*12+1}{4} = 9{,}25$ donc $Q_3 = \frac{37*3+47*1}{4} = 39{,}5$

⇒ **Diagramme en boîte ou Boîte à moustaches :**



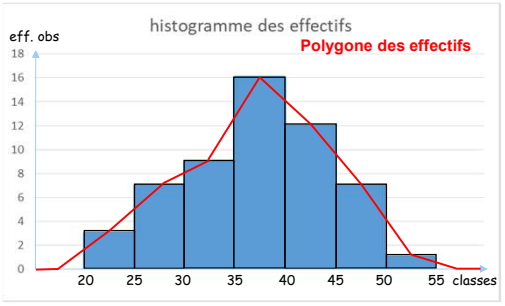

reprendre données exercice 2

## ➢ Représentations par classes

Réservées aux VA continues !

Exemple :

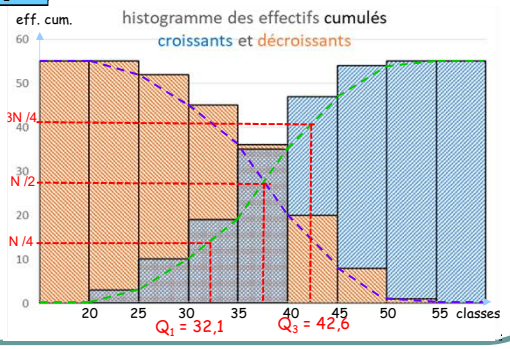
| classe | eff. obs. |
|---|---|
| ]20 ; 25] | 3 |
| ]25 ; 30] | 7 |
| ]30 ; 35] | 9 |
| ]35 ; 40] | 16 |
| ]40 ; 45] | 12 |
| ]45 ; 50] | 7 |
| ]50 ; 55] | 1 |



| classe | eff. obs. | eff. croiss. | eff. Décroiss. |
|---|---|---|---|
| ]20 ; 25] | 3 | 3 | 55 |
| ]25 ; 30] | 7 | 10 | 52 |
| ]30 ; 35] | 9 | 19 | 45 |
| ]35 ; 40] | 16 | 35 | 36 |
| ]40 ; 45] | 12 | 47 | 20 |
| ]45 ; 50] | 7 | 54 | 8 |
| ]50 ; 55] | 1 | 55 | 1 |

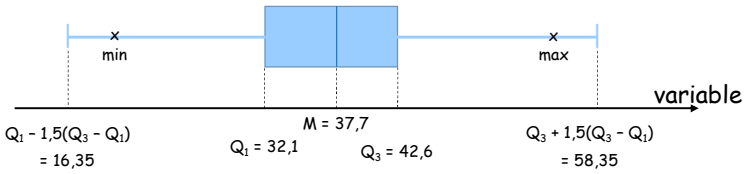
Polygone des effectifs cumulés décroissants

Polygone des effectifs cumulés croissants




RQ. Il n'y a pas de médiane ($Q_2$) si l'intersection des deux polygones est située en dessous de N /2

⇒ **Diagramme en boîte ou Boîte à moustaches :**




reprendre données exercice 2

puis comparer l'ensemble des résultats des deux types de représensations

**paramètres de dispersion**

- l'**écart-type** : $ET(X)=\sqrt{\sum_{i=1}^{n}(x_i-\bar{X})^2 p_i}$ si la VA est discrète

$ET(X)=\sqrt{\int_{-\infty}^{+\infty}(x_i-\bar{X})^2 f_X(x_i)dx}$ si la VA est continue

aussi noté s ou σ

RQ. la **variance** est notée var(X), $s^2$ ou $\sigma^2$

- l'**étendue**

- le **coefficient de variation** : $CV(\%)=100*\frac{s}{m}$

- l'**écart-type à la moyenne** : $s(m)=\frac{s}{\sqrt{N}}$

etc

reprendre exercice 2

## 5. Les lois de distribution

### 5.1 la loi Binomiale *B(N,p)*

=> loi de distribution la plus simple qui décrit un dénombrement de cas possibles.

Lors d'une expérience **répétée N fois** de façons identiques et indépendantes, elle ne peut donner **que 2 réponses possibles**.
La probabilité pour avoir **x succès** ($0 \le x \le N$) est

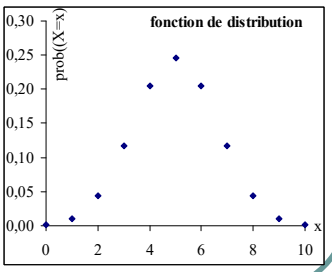
$f_X(x) = prob(X = x) = C_N^x\, p^x\, q^{(N-x)}$

$= \frac{N!}{x!(N-x)!} p^x\,(1-p)^{(N-x)}$

moyenne : $\mu = Np$

écart-type : $\sigma = \sqrt{Npq}$

=> loi discontinue, symétrique par rapport à x = μ et maximale à x = μ



RQ 1.

si N tend vers l'infini (> 50), p vers 0 (< 0,1) et $Np \leq 5$,
alors la loi *B(N, p)* tend vers une loi de Poisson de paramètre $\lambda = Np$

RQ 2.

si N tend vers l'infini (> 50), que p n'est pas trop petit (> 0,15) ni trop près de 1 et que $Np \geq 20$
alors la loi *B(N, p)* tend vers une loi de Normale de moyenne $\mu = Np$ et d'écart-type $\sigma = \sqrt{Npq}$

**=> exercice N°3**

## 5. Les lois de distribution

### 5.2 la loi de Poisson *P(λ)*

=> loi des évènements rares.

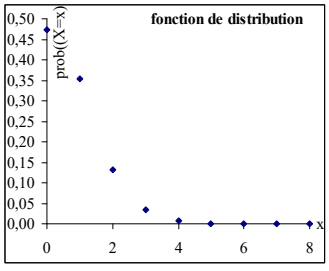
Soit une VA X qui peut prendre toutes les valeurs entières non négatives :

$f_X(x) = prob(X = x) = \frac{\lambda^x}{x!}\exp(-\lambda)$

moyenne : $\lambda$

écart-type : $\sigma = \sqrt{\lambda}$

=> loi discontinue, en forme de J inversé, maximale à x = 0 et tend vers 0 qd x augmente




RQ. si $\lambda$ augmente, alors la loi *P(λ)* tend vers une loi Normale de moyenne $\mu = \lambda$ et d'écart-type $\sigma = \sqrt{\lambda}$

**=> compl ex. N°3**

**=> exercice N°4**

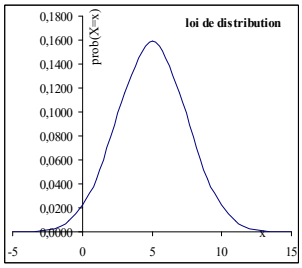
## 5. Les lois de distribution

### 5.3 la loi Normale $N(\mu, \sigma)$

=> aussi appelée loi de Gauss ou de Laplace-Gauss

Soit une VA X qui peut prendre toutes les valeurs réelles possibles :

$f_X(x) = prob(X = x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$




moyenne : $\mu$

écart-type : $\sigma$

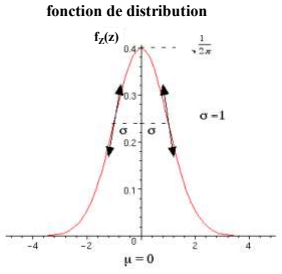
=> loi continue, en forme de cloche, symétrique par rapport à $x = \mu$ et maximale à $x = \mu$, tend vers zéro quand x tend vers $\pm\infty$

RQ. $\int_{-\infty}^{+\infty} f_X(x)dx = 1$

=> utilisation de la loi $N(\mu,\sigma)$ sous forme de $N(0,1)$ en posant $z = \frac{x-\mu}{\sigma}$

La VA Z est alors dite "centrée réduite" :

sa moyenne vaut zéro  son écart-type 1



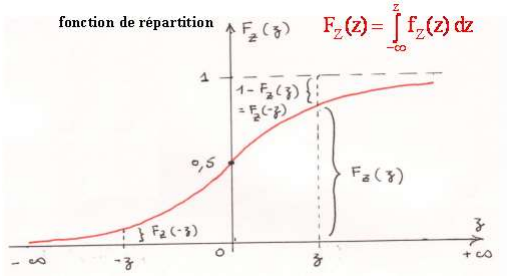

≈ 68% de l'effectif total appartient à l'intervalle $\mu \pm \sigma$,

≈ 95% de l'effectif total appartient à l'intervalle $\mu \pm 2\sigma$,

≈ 99% de l'effectif total appartient à l'intervalle $\mu \pm 2.5\sigma$,

≈ 99.8% de l'effectif total appartient à l'intervalle $\mu \pm 3\sigma$.

Rq: pour une loi normale moyenne = mode = médiane au seuil de risque $\alpha$




**⇒ table de la loi N(0,1)**

**=> exercice 5**

## 6. Vérification de la loi de probabilité d'une distribution

### 6.1. Test du Khi² (ou de Karl Pearson)

=> sert à comparer une distribution observée à une distribution théorique quelle que soit sa loi

**étape 1 :**
choisir une représentation individualisée si la loi de comparaison est discontinue, une représentation par classes si la loi de comparaison est continue

**étape 2 :**
ligne à ligne, on calcule la probabilité $p_i$ de la VA X à partir de la loi de distribution que l'on veut tester, d'où l'on va déterminer l'effectif théorique attendu $C_i = Np_i$

il faut que Ci > 5 !!!

**étape 3 :**

on calcule $\chi^2_{obs} = \sum_{i=1}^{k} \frac{(O_i - C_i)^2}{C_i} = \sum_{i=1}^{k} \frac{O_i^2}{C_i} - N$

que l'on compare à $\chi^2_{th}$ (table de Pearson) au seuil de risque $\alpha$
avec v = (k – p - 1) degrés de liberté

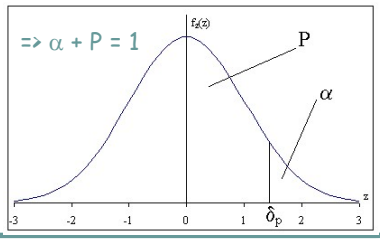
où p = 0 pour une loi quelconque,
= 1 pour une loi Binomiale ou de Poisson
= 2 pour une loi Normale

niveau de confiance P

$$prob(z \leq \delta_P) = P$$

seuil de risque $\alpha$

$$prob(z > \delta_P) = \alpha$$




**observation / conclusion :**

Si $\chi^2_{obs} < \chi^2_{th}$ alors la distribution observée est représentative de la loi de distribution étudiée au seuil de risque $\alpha$

**=> exercice 6**

**=> exercice 7**

**=> exercice 8**

**=> exercice 9**

## 6. Vérification de la loi de probabilité d'une distribution

### 6.2 Test de la droite de Henry

=> sert à comparer une distribution observée à une distribution de loi normale, et est basé sur la linéarisation de la fonction de répartition

**étape 1 :**

on calcule la moyenne et l'écart-type de l'échantillon que l'on prend pour estimateurs de la loi normale $N(\mu,\sigma)$

**étape 2 :**

pour chaque valeur $x_i$ de l'échantillon, on calcule le probit théorique

$(\text{prob th})_i = \frac{x_i - \mu}{\sigma} + 5$ (3 valeurs suffisent...)

**étape 3 :**

pour chaque valeur $x_i$ de l'échantillon, on cherche le probit observé à partir des fréquences relatives cumulées

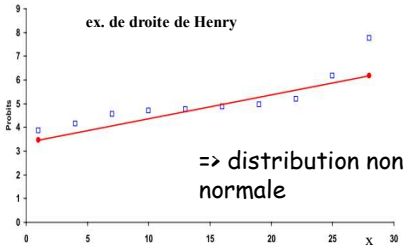
$F_i^* = \frac{\sum_{j=1}^{x} n_j - 0{,}5}{N}$ et de la table des probits

**étape 4 :**

on trace le graphique probits = $f(x_i)$

**observation / conclusion :**

Si les probits observés s'alignent à proximité de la droite des probits théoriques de façon aléatoire alors la distribution observée suit une loi $N(\mu,\sigma)$




=> distribution non normale

**=> exercice 10**

## 6. Vérification de la loi de probabilité d'une distribution

### 6.3 Test de Shapiro-Wilk

=> sert à comparer une distribution observée à une distribution de loi normale (il faut un nombre pair de données expérimentales)

**étape 1 :**

classer les données expérimentales dans l'ordre croissant puis en mettre la 1re moitié en colonne 1 de haut en bas et dans la colonne 2 de bas en haut

**étape 2 :**

ligne à ligne, la colonne 3 est la différence entre la colonne 2 et celle qui précède

**étape 3 :**

recopier en colonne 4 depuis la 1re table de Shapiro-Wilk les coefficients appropriés

**étape 4 :**

en colonne 5, multiplier ligne à ligne les deux colonnes qui précèdent puis tout additionner et mettre au carré $\Rightarrow$ on obtient $b^2$

**étape 5 :**

calculer $\Delta^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{N}$ puis $W_{obs} = \frac{b^2}{\Delta^2}$

**observation / conclusion :**

Si $W_{obs} > W_{th}$ lu dans la 2me table de Shapiro-Wilk alors la distribution observée appartient à une loi $N(\mu,\sigma)$ au seuil de risque $\alpha$

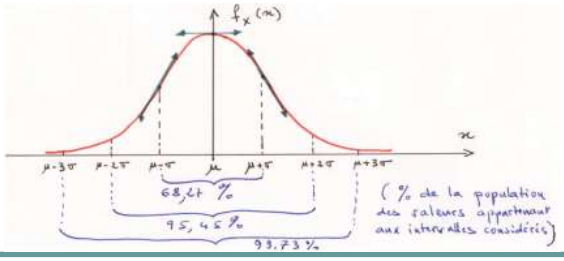
**=> exercice 11**

## 7. Intervalles de confiance

sont de deux types et ne s'appliquent que si la distribution étudiée suit une loi normale

Intervalle de confiance sur les valeurs expérimentales (autour de $\mu$)

l'intervalle de confiance au niveau de probabilité P est l'intervalle qui contient P% des valeurs expérimentales observées : $x = \mu \pm \delta_P$




Soit une VA X qui suit une loi $N(\mu,\sigma)$,

on veut chercher l'$IC_P$ tel que $prob(\mu - \delta_P < x \leq \mu + \delta_P) = P$

si on pose $z = \frac{x-\mu}{\sigma}$ la variable Z suit une loi $N(0,1)$

on a donc $$prob\left\{\frac{-\delta_P}{\sigma} < \frac{x-\mu}{\sigma} \leq \frac{\delta_P}{\sigma}\right\} = P$$

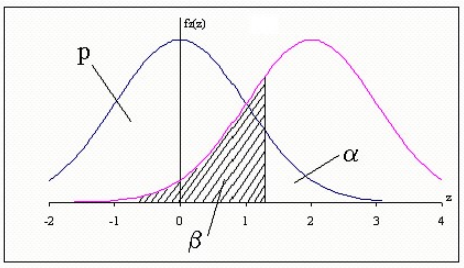
$$= prob\left(-z < \frac{x-\mu}{\sigma} \leq z\right)$$
$$= F_Z(z) - F_Z(-z)$$
$$= 2F_Z(z) - 1$$

on cherche donc dans la **table N(0,1) z tel que $F_Z(z) = (P+1)/2$**

et on aura pour $IC_P$ : $x = \mu \pm \delta_P = \mu \pm z\sigma \approx m \pm zs$

Toutes les valeurs expérimentales **qui sortent de cet $IC_p$** sont dites **aberrantes** au niveau de confiance P

les rejeter consiste à prendre un risque de 1re espèce ($\alpha$), par opposition au risque de 2me espèce ($\beta$) qui consiste à conserver une valeur incluse dans l'$IC_p$ mais qui appartiendrait à une autre population




**=> exercice 12**

Test de Dixon

=> autre méthode pour repérer des valeurs expérimentales aberrantes

**étape 1 :**
classer les N données expérimentales dans l'ordre croissant

**étape 2 :**
calculer $R_{obs}$ pour les valeurs expérimentales extrêmes selon

si $N \leq 7$, $R_{obs} = \frac{x_2 - x_1}{x_N - x_1}$ et $\frac{x_N - x_{N-1}}{x_N - x_1}$

si $8 \leq N \leq 12$, $R_{obs} = \frac{x_2 - x_1}{x_{N-1} - x_1}$ et $\frac{x_N - x_{N-1}}{x_N - x_2}$

si $N \geq 13$, $R_{obs} = \frac{x_3 - x_1}{x_{N-2} - x_1}$ et $\frac{x_N - x_{N-2}}{x_N - x_3}$

**observation / conclusion :**

Si $R_{obs} < R_{th}$ à 5% alors la valeur extrême peut être conservée avec un risque de 2me espèce ($\beta$) acceptable

Si $R_{obs} > R_{th}$ à 1% alors la valeur extrême est aberrante et doit être rejetée

**=> exercice 13**

Intervalle de confiance sur la valeur vraie de $\mu$ : $\mu = m \pm \delta_P$

Soit une VA X qui suit une loi $N(\mu,\sigma)$,

on cherche ici l'$IC_P$ tel que $prob(m - \delta_P < \mu \leq m + \delta_P) = P$

puisque l'on part d'un **échantillon** de taille N (rarement possible d'étudier une population complète) la table $N(0,1)$ n'est plus utilisable

⇒ **table de Student-Fisher**

et $\delta_P = t\frac{s}{\sqrt{N}}$ au seuil de risque $\alpha$ bilatéral et avec v = (N-1) ddl

$\frac{s}{\sqrt{N}}$ : écart-type à la moyenne

**=> exercice 14**

RQ 1. si N suffisamment grand, t est remplaçable par z tel que

$$F_Z(z) = (P+1)/2$$

RQ 2. si s ≈ cste quel que soit N, $\frac{s}{\sqrt{N}}$ diminue si N augmente

RQ 3. si s varie, on lui préfère le coefficient de variation :

$$CV\ (\%) = 100\ s/m$$

RQ 4. la valeur $p = 100\delta_P/m$ est appelée précision, on connait alors ***μ* à p% près au seuil de risque α**

**=> exercice 15**

**=> exercice 16**

**=> exercice 17**

**=> exercice 18**

**=> exercice 19**

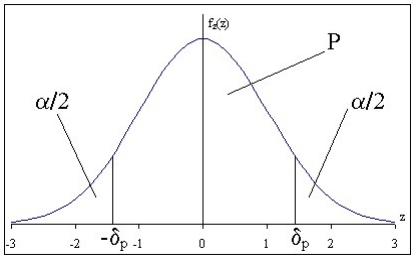
## 8. Comparaisons sur échantillons

### 8.1 Les hypothèses

Un test statistique commence toujours par une hypothèse que l'on cherche à vérifier à un seuil de risque $\alpha$ / au niveau de confiance P ($\alpha$ + P = 100% = 1)

=> sont de 2 types :

Hypothèse "nulle" $H_0$ : $\lambda = \lambda_0$ = **hypothèse bilatérale** (symétrique le plus souvent)

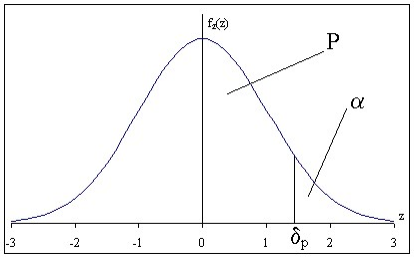



$$prob\left(-\delta_P \le z \le \delta_P\right) = P$$

$$prob\left(z \le -\delta_P \text{ ou } z \ge \delta_P\right) = 2\frac{\alpha}{2} = \alpha$$

Hypothèses "alternatives" $H_1$ : $\lambda > \lambda_0$ = **hypothèse unilatérale** à G.

$\lambda \le \lambda_0$ = **hypothèse unilatérale** à D.




ex. test unilatéral à droite

$$prob\left(z \le \delta_P\right) = P$$

$$prob\left(z > \delta_P\right) = \alpha$$

**Quel que soit le test, tant que la valeur observée est inférieure à la valeur théorique donnée par les tables, les différences observées entre les valeurs comparées sont dues au hasard. En conséquence, on acceptera l'hypothèse $H_0$ et on rejettera l'hypothèse $H_1$**

Rejeter $H_0$, alors qu'elle est peut-être vraie, ou accepter $H_1$ alors qu'elle est peut-être fausse, constitue un risque de première espèce : risque $\alpha$.

Accepter $H_0$, alors qu'elle est peut-être fausse, ou rejeter $H_1$, alors qu'elle est peut-être vraie, constitue un risque de deuxième espèce : risque $\beta$.

## 8. Comparaisons sur échantillons

### 8.2 Le test de Student-Fisher

Comparaison d'une moyenne expérimentale m à une valeur de référence $m_0$

**étape 1 :**
on pose l'hypothèse en fonction de la question que l'on se pose

**étape 2 :**
on détermine la valeur observée de la fonction discriminante, ici :

$$t_{obs} = \frac{|m - m_0|}{s/\sqrt{N}}$$

**étape 3 :**
on compare $t_{obs}$ à une valeur limite $t_{th}$ lue dans la table de Student au seuil de risque $\alpha$, uni- ou bilatéral selon l'hypothèse posée, et avec v = (N - 1) ddl

**étape 4 :**
si $t_{obs} < t_{th}$, alors on accepte $H_0$ ou on rejette $H_1$, selon l'hypothèse posée

=> exercice 20
=> exercice 21

## Comparaison de deux moyennes expérimentales $m_1$ et $m_2$

**étape 1 :**
on pose l'hypothèse en fonction de la question que l'on se pose

**étape 2 :**
on détermine la valeur observée de la fonction discriminante,
ici :

$$t_{obs} = \frac{|m_1 - m_2|}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \qquad \text{avec} \quad s = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}}$$

écart-type estimé commun

**étape 3 :**
on compare $t_{obs}$ à une valeur limite $t_{th}$ lue dans la table de Student au seuil de risque $\alpha$, uni- ou bilatéral selon l'hypothèse posée, et avec $v = (N_1 + N_2 - 2)$ ddl

**étape 4 :**
si $t_{obs} < t_{th}$, alors on accepte $H_0$ ou on rejette $H_1$, selon l'hypothèse posée

**=> exercice 22**

**=> exercice 23**

## Comparaison d'une fréquence expérimentale f à une valeur de référence p

**étape 1 :**
on pose l'hypothèse en fonction de la question que l'on se pose

**étape 2 :**
on détermine la valeur observée de la fonction discriminante,
ici :

$$t_{obs} = \frac{|f - p|}{\sqrt{\frac{p(1-p)}{N}}}$$

**étape 3 :**
on compare $t_{obs}$ à une valeur limite $t_{th}$ lue dans la table de Student au seuil de risque $\alpha$, uni- ou bilatéral selon l'hypothèse posée, et avec $v = (N - 1)$ ddl

**étape 4 :**
si $t_{obs} < t_{th}$, alors on accepte $H_0$ ou on rejette $H_1$, selon l'hypothèse posée

**=> exercice 24**

Comparaison de deux fréquences expérimentales $f_1$ et $f_2$

**étape 1 :**
on pose l'hypothèse en fonction de la question que l'on se pose

**étape 2 :**
on détermine la valeur observée de la fonction discriminante, ici :

$$t_{obs} = \frac{|f_1 - f_2|}{\sqrt{p(1-p)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \quad \text{avec} \quad p = \frac{N_1 f_1 + N_2 f_2}{N_1 + N_2 - 2}$$

probabilité estimée commune

**étape 3 :**
on compare $t_{obs}$ à une valeur limite $t_{th}$ lue dans la table de Student au seuil de risque $\alpha$, uni- ou bilatéral selon l'hypothèse posée, et avec $v = (N_1 + N_2 - 2)$ ddl

**étape 4 :**
si $t_{obs} < t_{th}$, alors on accepte $H_0$ ou on rejette $H_1$, selon l'hypothèse posée

**=> exercice 25**

## 8. Comparaisons sur échantillons

### 8.3 Le test de Pearson (Khi²)

Comparaison d'une variance expérimentale $s^2$ à une valeur de référence $\sigma_0^2$

**étape 1 :**
on pose l'hypothèse en fonction de la question que l'on se pose

**étape 2 :**
on détermine la valeur observée de la fonction discriminante, ici :

$$\chi^2_{obs} = v\frac{s^2}{\sigma_0^2}$$

avec v = N si $\mu$ connue sinon (N-1) ddl

**étape 3 :**
on compare $\chi_{obs}$ à une valeur limite $\chi_{th}$ lue dans la table de Pearson au seuil de risque $\alpha$ avec v ddl

**étape 4 :**
il faudra alors accepter l'hypothèse

$s^2 < \sigma_0^2$ si $\chi^2_{obs} < \chi^2_{th,1-\alpha}$

$s^2 = \sigma_0^2$ si $\chi^2_{th,1-\alpha/2} < \chi^2_{obs} < \chi^2_{th,\alpha/2}$

$s^2 > \sigma_0^2$ si $\chi^2_{obs} > \chi^2_{th,\alpha}$

**=> exercice 26**

**=> exercice 27**

## 8. Comparaisons sur échantillons

### 8.4 Le test de Fisher-Snedecor

Comparaison de deux variances expérimentales $s_1^2$ et $s_2^2$

**étape 1 :**

seule hypothèse possible : $H_1$ $s_1^2 > s_2^2$

**Le test est unilatéral par défaut**

**étape 2 :**

on détermine la valeur observée de la fonction discriminante, ici :

$$F_{obs} = \frac{s_1^2}{s_2^2}$$ tjs > 1 !

**étape 3 :**

on compare $F_{obs}$ à une valeur limite $F_{th}$ lue dans la table de Fisher au seuil de risque $\alpha$ avec $v_1$ et $v_2$ = N et/ou (N-1) ddl selon si $\mu_1$ et/ou $\mu_2$ sont ou non connues

**étape 4 :**

si $F_{obs} < F_{th}$, on rejette $H_1$

**=> exercice 28**

**=> exercice 29**

## 9. Etude de cas

# 2me partie :

# analyse statistique

# à deux variables

# => les régressions linéaires simples

## 1. Notions de corrélation et régression

Quand peut-on dire qu'il y a corrélation ?

On dit qu'il y a **corrélation** lorsqu'il y a possibilité de relier deux variables (ou plus) par un lien mathématique

linéaire simple : $y = y_0 + bx$

linéaire multiple : $y = y_0 + b_1x_1 + b_2x_2 + \ldots$

polynomial simple : $y = y_0 + b_1x + b_2x^2 + \ldots$

parabolique, logarithmique, exponentielle, etc

Ce lien est quantifié par le **coefficient de corrélation**, noté r, ou R selon le cas.

Dans la pratique : **$-1 \leq r$ (ou R) $\leq 1$**

et l'on a affaire à des nuages de points orientés

**=> exemples**

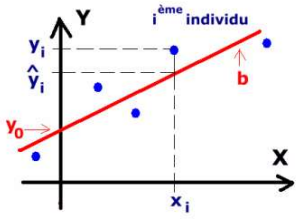
La **régression** est un ensemble de méthodes statistiques utilisées pour analyser la relation d'une variable par rapport à une ou plusieurs autres.

Il en existe deux sortes couramment utilisées :

- la régression par la **Méthode des Moindres Carrés Conventionnels** (MMCC) => $y_0$ déterminé avec b
- la régression par la **Méthode des Moindres Carrés Forcés** (MMCF) => $y_0$ connu et fixe, seul b est déterminé

Par la Méthode des Moindres Carrés, on cherche à minimiser au maximum les écarts entre la valeur expérimentale $y_i$ de la variable expliquée Y et sa valeur calculée $\hat{y}_i$ par le modèle mathématique

=> obtention de la courbe la plus ajustée possible



Le modèle de régression le plus connu est le modèle de régression linéaire $y = y_0 + bx$

De nombreux modèles apparemment non linéaires peuvent être transformés en modèles linéaires, exemple :

| | | |
|---|---|---|
| $y = a\ b^x$ | => $Lny = Lna + x\ Lnb$ | soit $y' = y'_0 + b'x$ |
| $y = a\ x^b$ | => $Lny = Lna + b\ Lnx$ | soit $y' = y'_0 + bx'$ |
| $y = a\ exp(bx)$ | => $Lny = Lna + bx$ | soit $y' = y'_0 + bx$ |
| $y = \frac{x}{ax+b}$ | => $1/y = a + b/x$ | soit $y' = y_0 + bx'$ |

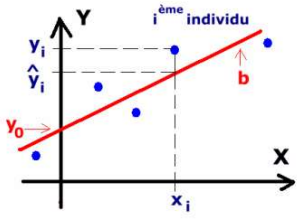
etc

## 2. La régression linéaire simple sans contrainte => RLS/MMCC

=> exercice 30

On dispose de N points expérimentaux dont les variations de la variable Y (expliquée) sont expliquées par les variations de la variable X (explicative) selon la relation $\hat{y} = y_0 + bx$

La MMCC permet d'obtenir la "meilleure" droite de régression, c'est à dire la droite la plus ajustée possible aux données expérimentales, en minimisant la Somme des Carrés des Résidus :



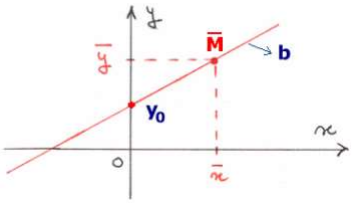

$$SCR = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}(y_i - y_0 - bx_i)^2$$
$$= \sum y_i^2 - y_0 \sum y_i + 2b \sum x_i y_i + N y_0 + 2b y_0 \sum x_i + b^2 \sum x_i^2$$

Les coefficients de sensibilité $y_0$ et b sont alors déterminés par les dérivées simultanées

$$\frac{\partial SCR}{\partial y_0}=0 \quad \text{et} \quad \frac{\partial SCR}{\partial b}=0$$

dont la résolution conduit à $y_0=\frac{\sum y_i - b\sum x_i}{N}=\bar{y}-b\bar{x}$

$$b=\frac{N\sum x_i y_i-\sum x_i * \sum y_i}{N\sum x_i^2-\left(\sum x_i\right)^2}$$




=> la droite de régression passe par le point moyen $\bar{M}(\bar{x},\bar{y})$



1. Le coefficient de corrélation linéaire r

r quantifie le lien entre la variable expliquée et la variable explicative

$$|r|=\sqrt{\frac{\text{variation expliquée}}{\text{variation totale}}}=\sqrt{1-\frac{\text{variation résiduelle}}{\text{variation totale}}}$$
$$=\sqrt{1-\frac{SCR_{\min}}{(N-1)\,s^2(y)}}$$

la variation totale de y est $\sum(y_i-\bar{y})^2=\sum y_i^2-\frac{\left(\sum y_i\right)^2}{N}=(N-1)\,s^2(y)$

la variation expliquée par le modèle de régression est $\sum(\hat{y}_i-\bar{y})^2$

la variation résiduelle est $\sum(y_i-\hat{y}_i)^2=SCR_{\min}=\sum y_i^2-y_0\sum y_i-b\sum x_i y_i$

telles que

**variation totale = variation expliquée + variation résiduelle**

on peut montrer que l'on a aussi r = b $\frac{s(x)}{s(y)}$
donc r et b sont de même signe

r = 0 => pas de corrélation, |r| = 1 => corrélation "parfaite"

$r^2$ = coefficient de détermination et indique le pourcentage de variation expérimentale expliquée par le modèle de régression

2. Vérification de l'existence significative d'une corrélation

=> test de Student de r contre 0

hypothèse $H_0$ : r = 0, absence de corrélation

puis on calcule $t_{obs} = \frac{|r-0|}{s(r)}$ avec $s(r) = \sqrt{\frac{1-r^2}{v}}$



que l'on compare à $t_{th}$ au seuil de risque $\alpha$ bilatéral avec v = (N-2) ddl

si $t_{obs} < t_{th}$, $H_0$ est acceptée et il n'y a pas de corrélation

si $t_{obs} > t_{th}$, $H_0$ est rejetée et il y a une corrélation significative au seuil de risque $\alpha$

3. Vérification de la significativité des coefficients de sensibilité

**étape 1** : détermination des écart-types s(y/x), $s(y_0)$ et s(b)

écart-type lié : $s(y/x) = \sqrt{\frac{\text{variation résiduelle}}{\text{nbre de degrés de liberté}}} = \sqrt{\frac{SCR_{min}}{v}} = s_{lié}$

Rq 1. s(y/x) pondère la SCR par le nbre de ddl, ainsi un petit échantillon donnera un écart-type lié important même si $SCR_{min}$ est faible

Rq 2. une bonne régression aura s(y/x) le plus faible possible

écart-type sur la pente : s(b) = $\frac{s(y/x)}{\sqrt{\sum(x_i-\bar{x})^2}} = \frac{s_{lié}}{\sqrt{(N-1)\,s^2(x)}}$

écart-type sur l'ordonnée à l'origine : $s(y_0) = s(b)\sqrt{\frac{\sum x_i^2}{N}}$

**étape 2** : test de $y_0$ contre 0

=> test de Student

hypothèse $H_0$ : $y_0 = 0$ (RLS/MMCF(0,0))

puis on calcule $t_{obs} = \frac{|y_0 - 0|}{s(y_0)}$

que l'on compare à $t_{th}$ au seuil de risque $\alpha$ bilatéral avec v = (N-2) ddl

si $t_{obs} < t_{th}$, $H_0$ est acceptée et on devra basculer en MMCF(0,0)

si $t_{obs} > t_{th}$, $H_0$ est rejetée et on reste en MMCC

au seuil de risque $\alpha$

**étape 3** : test de b contre 0

=> test de Student

hypothèse $H_0$ : b = 0 (absence de proportionnalité)

puis on calcule $t_{obs} = \frac{|b-0|}{s(b)}$

que l'on compare à $t_{th}$ au seuil de risque $\alpha$ bilatéral avec v = (N-2) ddl

si $t_{obs} < t_{th}$, $H_0$ est acceptée et il n'existe pas de proportionnalité significative ente X et Y => $y = y_0 = \overline{y}$

si $t_{obs} > t_{th}$, $H_0$ est rejetée et la proportionnalité entre X et Y est significative => la relation $y = y_0 + bx$ a un sens

au seuil de risque $\alpha$

Rq. b peut être testé contre n'importe quelle valeur théorique :

si $b = b_{th}$, la pente $exp^{tale}$ ne diffère pas de façon significative de la pente théorique

si $b = b_{ref}$, deux méthodes de dosages auront la même sensibilité

si $b < b_{ref}$, ou $b > b_{ref}$, alors la méthode de dosage étudiée sera moins, ou plus, sensible que la méthode de référence

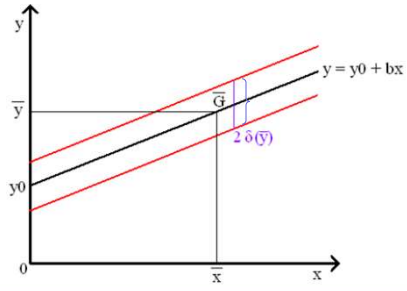


## 4. Droites d'incertitude - Points aberrants

on détermine les coordonnées du point moyen $\overline{G}$ de coordonnées $(\overline{x}, \overline{y})$

et l'incertitude sur $\overline{y}$ : $\delta(\bar{y}) = t \frac{s_{lié}}{\sqrt{N}}$

puis on trace les droites parallèles à la droite de régression passant par $(0, y_0 \pm \delta(\overline{y}))$ et par $(\overline{x}, \overline{y} \pm \delta(\overline{y}))$




tout point expérimental qui sort de la bande d'incertitude est considéré aberrant à $\alpha$%

## 5. Analyse de variance - tableau d'ANOVA

| source de variation | somme des carrés des écarts | ddl | carrés moyens | $F_{exp}$ |
|---|---|---|---|---|
| modèle de régression | $\sum(\hat{y}_i-\bar{y})^2$ | 1 | $\sum(\hat{y}_i-\bar{y})^2$ /1 = CMR | $\frac{CMR}{s^2(y/x)}$ |
| variation résiduelle | $\sum(y_i-\hat{y}_i)^2=SCR_{min}$ | N-2 | $SCR_{min}$ /(N-2) = $s^2(y/x)$ | |
| variation totale | $\sum(y_i-\bar{y})^2=(N-1)s^2(y)$ | | | |

si $F_{exp} > F_{th}(\alpha, v_1=1, v_2=(N-2)$ ddl) lue dans la table de Fisher

alors le modèle explique la variation expérimentale observée

avec un taux de $r^2$ % de variation expliquée

Rq 1. la régression sera d'autant meilleure que $F_{exp}$ sera élevé

Rq 2. en RLS/MMCC, on peut montrer que $F_{exp} = t^2_{obs}$ du test de Student de r contre 0



## 6. Valeurs prédites et intervalles de confiance

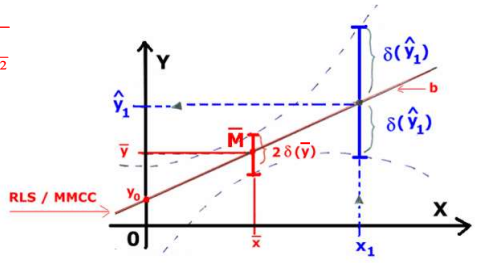
=> sont de 2 types selon que l'on considère une valeur $y_i$ isolée ou répétée

### Cas d'une mesure isolée

la valeur prédite pour une valeur donnée $x_i$ est $\hat{y}_i = y_0 + bx_i$

son incertitude, au seuil de risque $\alpha$% bilatéral, avec v = (N-2) ddl est donnée par

$$\delta(\hat{y}_i) = t\, s_{lié} \sqrt{1+\frac{1}{N}+\frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$

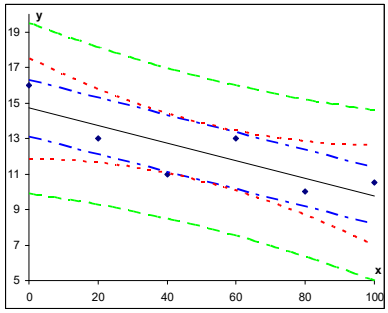


Cas d'une mesure répétée

on montre que la moyenne des valeurs prédites pour une valeur donnée $x_i$ est également $\overline{\hat{y}_i} = \hat{y}_i = y_0 + bx_i$

son incertitude, au seuil de risque $\alpha$% bilatéral, avec v = (N-2) ddl est donnée par

$$\overline{\delta(\hat{y}_i)} = t\, s_{lié} \sqrt{\frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$




droite de régression
droites d'incertitude
courbes de prédiction de mesures isolées
courbes de prédiction de mesures répétées

## 3. La régression linéaire simple avec contrainte => RLS/MMCF

=> exercice 31

Dans certains cas, on s'attend à une valeur bien établie de $y_0$ :

ex. loi de Beer-Lambert $A = \varepsilon\, c\, l$ (spectroscopie d'absorption)

loi de Henry $P = k\, c$ (chromatographie "head space")

loi de la cryoscopie $T = T_{fus} - K^{cr}\, c$ (t° de congélation d'une solution)

etc

**Après contrôle,**

**on doit trouver que $y_0$ ne diffère pas de façon significative de la valeur fixée et appliquer la méthode des Moindres Carrés Forcés (MMCF)**

Cas où $y_0 = 0$ : $y = b_f x$

La Somme des Carrés de Résidus devient :

$$SCR_f = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}(y_i - b_f x_i)^2$$
$$= \sum y_i^2 - 2b\sum x_i y_i + b^2 \sum x_i^2$$

et sa minimisation par $\frac{\partial SCR_f}{\partial b_f} = 0$ conduit à $b_f = \frac{\sum x_i y_i}{\sum x_i^2}$

et $SCR_{f\ min} = \sum y_i^2 - b_f \sum x_i y_i$



$$\Rightarrow |r_f| = \sqrt{1 - \frac{\text{variation résiduelle}}{\text{variation totale}}} = \sqrt{1 - \frac{SCR_{\min}}{(N-1)\,s^2(y)}}$$

$s(r_f) = \sqrt{\frac{1-r_f^2}{v}}$ avec v = (N-1) ddl

écart-type lié : $s_{lié} = \sqrt{\frac{\text{variation résiduelle}}{\text{nbre de degrés de liberté}}} = \sqrt{\frac{SCR_{\min}}{v}}$

écart-type sur la pente : $s(b_f) = \frac{s_{lié}}{\sqrt{\sum x_i^2}}$

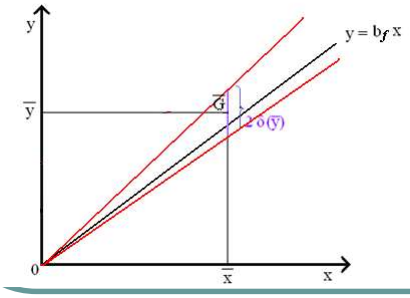
Rq. $s_f(\hat{y}_0) = s(y/x)\sqrt{\frac{1}{N} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$

Droites d'incertitude - Points aberrants

on détermine les coordonnées du point moyen $\overline{G}$ de coordonnées $(\overline{x}, \overline{y})$

et l'incertitude sur $\overline{y}$ : $\delta(\overline{y}) = t\frac{s_{lié}}{\sqrt{N}}$

puis on trace les droites passant par (0,0) et par $(\overline{x}, \overline{y} \pm \delta(\overline{y}))$




tout point expérimental qui sort du cône d'incertitude est considéré aberrant à $\alpha$%

Cas où $y_0$ connu et fixe mais $\neq 0$ : $y - y_0 = b_f x$

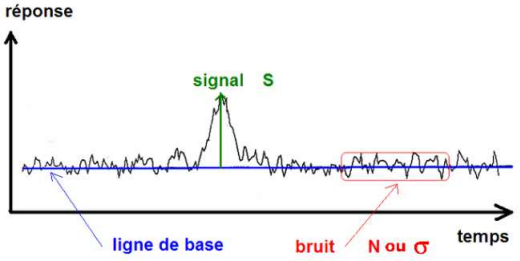
mêmes formules mais en remplaçant $y_i$ par $(y_i - y_0)$

=> changement de variable

**4. Applications à la chimie analytique**

**=> exercice 32**

1. Bruit de fond - Limites de détection et de quantification



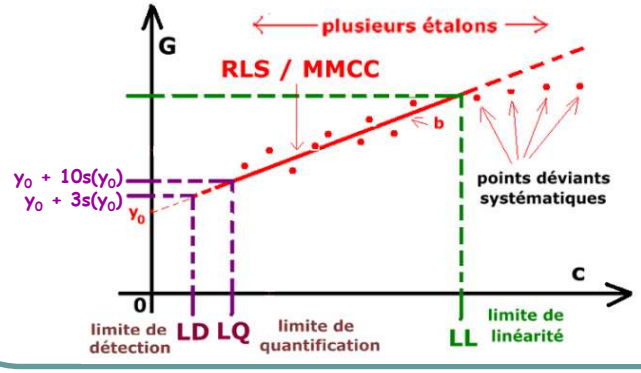
La **limite de détection** (LD) est la plus petite concentration de l'analyte pouvant être détectée avec une incertitude acceptable, mais non quantifiée dans les conditions de l'expérience

On considère que ceci est réalisé lorsque le signal dépasse $3s(y_0)$
ceci correspond à un seuil de risque de 0,13%

La **limite de quantification** (LQ) est la plus petite concentration de l'analyte pouvant être quantifiée avec une incertitude acceptable dans les conditions de l'expérience

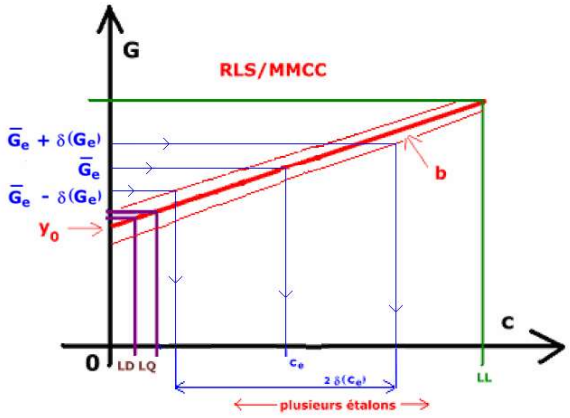
On considère que ceci est réalisé lorsque le signal dépasse $10s(y_0)$
ceci correspond à un seuil de risque de 5%

On ne peut conclure à la présence de l'analyte que si sa concentration est supérieure à LD.
On ne peut la doser qu'entre des concentrations comprises entre LQ et LL (limite de linéarité)

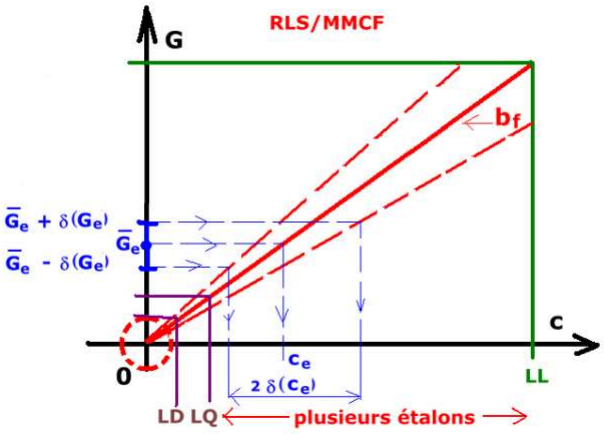


## 2. Détermination de la concentration d'une solution inconnue

Cas de la RLS/MMCC




Cas de la RLS/MMCF



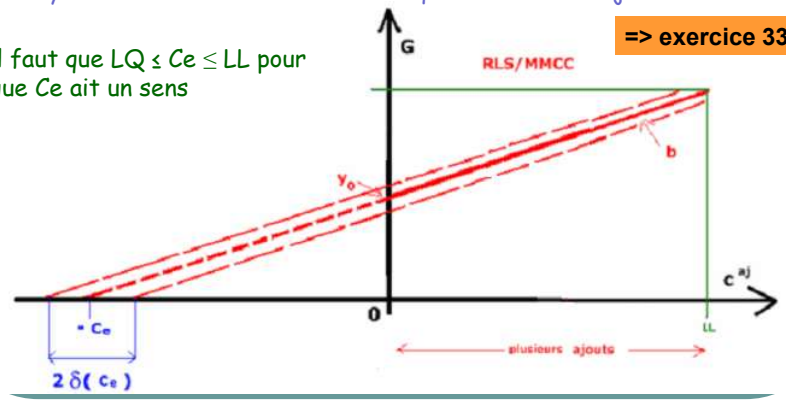

**=> finir exercices 31 et 32**

## 3. Méthode des ajouts dosés

Cet étalonnage est réalisé en ajoutant une quantité connue $c_i^{aj}$ de l'analyte à doser l'échantillon inconnu qui en contient déjà

il faut que LQ ≤ Ce ≤ LL pour que Ce ait un sens

**=> exercice 33**




## 4. Etude de cas

## 5. L'analyse des résidus

Soit un modèle mathématique quelconque $y = f(x_1, x_2, ...)$

→ régression polynomiale simple

→ régression linéaire multiple

→ régression logarithmique

→ expressions qcq, etc

Le modèle est arrêté suite aux analyses statistiques vues précédemment (coef de corrélation global et partiels, coefs de sensibilité, $s_{lié}$, $F_{exp}$, ...)

➢ La recherche des points aberrant passe par l'analyse des résidus standardisés :

$$RS = \frac{\hat{y} - y}{s_{lié}}$$

il faut |RS| < 2 à P = 95%
2,5 à 99%
3 à env 100%

L'analyse des résidus termine l'analyse d'un modèle mathématique et permet de le valider définitivement :

1. Ils doivent suivre une loi normale centrée et réduite

2. Ils doivent être indépendant des variables étudiées => graphiques RS = f(y) et RS = f($x_i$)

Exemple d'analyse complète :

⇒ TP