Chapitre[4] Analyse quantitative : méthodes statistiques

Samir Kenouche - Département des Sciences de la Matière - UMKB

Module : Spectroscopie atomique et moléculaire - Niveau Master 1

Version corrigée, améliorée et augmentée

Résumé

Toutes les mesures expérimentales sont oblitérées d'incertitudes. L'expérimentateur doit savoir quantifier ces erreurs et les réduire autant que possible. Il est d'usage d'utiliser indifféremment les terminologies erreur et incertitude. Ces deux termes ont des significations distinctes. L'erreur de mesure représente l'écart entre la valeur mesurée et sa valeur effective (ou parfaite), qui est inaccessible par l'expérience. D'un autre côté, l'incertitude de mesure est une estimation de l'intervalle dans lequel on trouve la valeur mesurée avec une certaine probabilité. A la lumière de ces définitions, on comprend que l'incertitude est une approximation de l'erreur de mesure, d'où l'utilisation d'outils statistiques pour l'estimer.

L'analyse statistique est à la base de toutes les approches visant à chiffrer les incertitudes expérimentales. Cette analyse implique un traitement mathématique rigoureux et complexe. Toutefois, l'expérimentateur n'a pas systématiquement besoin de maîtriser toutes les subtilités mathématiques afin de concrétiser cette démarche. C'est pourquoi, dans ce chapitre, l'accent est plutôt mis sur les aspects interprétation et signification physique des incertitudes expérimentales avec un minimum de mathématiques.

" En mathématiques, on ne comprend pas les choses ... on s'y habitue ..."

Cf. John Von Neumann, Mathématicien (1903 - 1957).

Table des matières

I. Quantification des erreurs

L'expérience montre que la répétition d'une mesure expérimentale génère systématiquement des résultats légèrement différents. Dans la plupart des cas, ces résultats sont plus au moins dispersés autour d'une certaine valeur centrale. En effet, on peut légitimement se poser la question : d'où provient cette dispersion ?. Cette question nous amène à définir dans un premier temps les diverses incertitudes affectant les mesures expérimentales.

S. Kenouche est docteur en Physique de l'Université de Montpellier et docteur en Chimie de l'Université de Béjaia.
Site web : voir http://www.sites.univ-biskra.dz/kenouche
Document fait le 01.12.2020.

i. Erreurs aléatoires

En répétant les mesures de l'absorbance (A) d'une substance chimique avec un spectrophomètre, on constate de façon permanente des résultats sensiblement différents. Ceci est vrai même si l'expérience est répétée dans des conditions expérimentales rigoureusement identiques. Cela est lié à l'incapacité des dispositifs expérimentaux à produire des mesures avec une précision infinie. Cette fluctuation des mesures est liée par exemple à l'incapacité du monochromateur à filtrer avec une précision infinie les longueurs d'onde (λ) émises. Il fournit toujours une bande passante $\lambda + \delta\lambda$. Dans ce cas, on qualifiera $\delta\lambda$ *d'erreur aléatoire*.

$$\lambda + \delta\lambda \Rightarrow A + \delta A \tag{1}$$

Le mot aléatoire, il faut le prendre ici au sens des probabilités. En effet, même avec des conditions expérimentales rigoureusement identiques, la valeur de $\delta\lambda$ change légèrement. Cette erreur aléatoire suit une loi statistique .

ii. Erreurs systématiques

Ces erreurs proviennent d'un mauvais réglage du dispositif de mesure, d'une erreur dans la démarche expérimentale et/ou d'une erreur dans la modélisation du phénomène étudié. Par exemple, en spectrophotométrie UV-Vis, on peut envisager la présence d'impuretés dans l'échantillon à analyser, ou bien une erreur dans le calcul de la concentration. On peut aussi envisager que l'expérimentateur n'ait pas pris le soin de vérifier la validité de la loi de *Beer-Lambert* ou l'appareil n'est pas installé sur une surface stable, car un petit mouvement du dispositif influera sur le résultat de la mesure. Notons Δ, le terme qui tient compte des erreurs systématiques. L'erreur sera donc la somme des deux contributions : *aléatoire* et *systématique*

$$\lambda + \delta\lambda + \Delta \Rightarrow A + \delta A + \Delta A \tag{2}$$

Les erreurs aléatoires sont facilement détectables. En revanche, les erreurs systématiques sont difficiles à éliminer si leurs causes exactes ne sont pas clairement identifiées. Ces erreurs produisent un biais (écart) constant pendant les mesures. Ainsi, en l'absence d'erreurs systématiques ($\Delta = 0$), la valeur mesurée est :

$$a = \bar{a} \pm \delta a \tag{3}$$

La valeur vraie de a se trouve dans l'intervalle $\bar{a} \pm \delta a$. Avec $\bar{a}$ est la meilleure estimation de la vraie valeur de a.

A. Distribution des erreurs aléatoires

L'expérience montre qu'en répétant un grand nombre de fois une mesure expérimentale, on obtient presque toujours pour les résultats de la mesure, une distribution de probabilité *Gaussienne*. Ceci est d'autant plus vrai que les erreurs systématiques soient négligeables et il y a de nombreuses sources d'erreurs aléatoires indépendantes. Cette distribution est tellement fréquente en sciences expérimentales (chimie, physique ...) qu'on l'appelle dans la littérature, la distribution *normale*. Ce phénomène provient du *théorème Central Limite* dont l'énoncé est :

Soit X, une variable aléatoire d'espérance mathématique μ, de variance σ_x^2 et dont la loi de probabilité est quelconque[1]*. Soit $\bar{X}_n$, une nouvelle variable aléatoire définie comme la moyenne calculée sur n mesures. Si la variance de X est fini, alors la distribution de $\bar{X}_n$ tend vers une loi normale pour n grand. De plus, $\bar{X}_n$ aura comme espérance μ et variance σ_x^2/n.*

TABLE I: Un cas pratique

$1^{ère}$ mesure	$2^{ème}$ mesure	$3^{ème}$ mesure	$4^{ème}$ mesure	...	effectif	$\bar{X}_n$
1.32	0	1	0	...	2	$\bar{X}_1 = \frac{2\,1.32}{4}$
1.28	1	1	1	...	4	$\bar{X}_2 = \frac{4\,1.28}{4}$
1.27	2	0	2	...	5	$\bar{X}_3 = \frac{5\,1.27}{4}$
1.29	3	2	2	...	8	$\bar{X}_4 = \frac{8\,1.29}{4}$
1.30	1	2	1	...	5	$\bar{X}_5 = \frac{5\,1.30}{4}$
1.31	0	1	1	...	3	$\bar{X}_6 = \frac{3\,1.31}{4}$
1.26	0	0	0	...	1	$\bar{X}_7 = \frac{1\,1.26}{4}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮

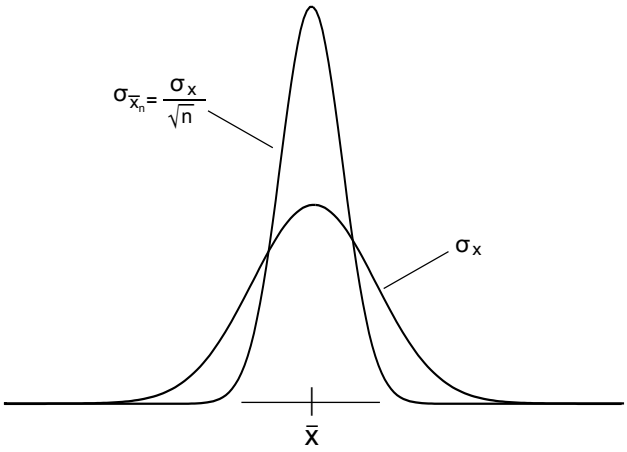

FIGURE 1: Gain de précision pour $n \rightarrow$ grand

La figure ci-dessus illustre clairement que la détermination de $\bar{X}$ [2] est $\sqrt{n}$ fois plus précise que celle obtenue à partir d'une seule mesure. Ainsi, plus n est grand plus la distribution se rétrécit et donc l'intervalle de confiance contenant $\bar{X}$ se rétrécit également, d'où le gain en précision.

B. *Distribution Gaussienne*

Une loi de distribution *Gaussienne* ayant une valeur moyenne $\bar{x}$ et un écart-type σ a la forme mathématique :

$$\rho(x) = \frac{1}{\sqrt{2\,\pi}\sigma} \times exp\left(-\frac{(x-\bar{x})^2}{2\sigma^2}\right) \tag{4}$$

1. La puissance de ce théorème tient au fait qu'aucune hypothèse n'est exigée sur la densité de probabilité de X. La variable aléatoire X peut avoir n'importe quelle distribution. Pour n élevé cette distribution tendra presque toujours vers la loi normale.
2. $\bar{X}$ est la meilleur estimation de X

La distribution de *Gauss* est symétrique autour de la moyenne et les points d'inflexion sont situés à une distance σ de l'axe de symétrie $x = \bar{x}$. Il arrive très souvent qu'on soit confronté à comparer deux ou plusieurs grandeurs ne s'exprimant pas dans les mêmes unités et/ou ayant des ordres de grandeur différents. Pour cela, il s'avère judicieux de *centrer* (mêmes ordres de grandeur) et de *réduire* (sans unités) les variables d'origines :

$$z = \frac{x - \bar{x}}{\sigma} \quad \Rightarrow \quad \rho(z) = \frac{1}{\sqrt{2\,\pi}} \times exp\left(-\frac{z^2}{2}\right) \tag{5}$$

La variable centrée réduite z suit dans ce cas, la *loi normale centrée et réduite* $\mathcal{N}(0;1)$. Comme il a été souligné précédemment, une grande majorité de grandeurs expérimentales se décrivent, au moins en première approximation, par cette distribution. Ceci explique son importance en sciences expérimentales. La loi normale est caractérisée par deux paramètres : la valeur moyenne $\bar{x}$ associée à la "vraie" valeur de la grandeur et la largeur à mi-hauteur σ associée à l'incertitude expérimentale. C'est la raison pour laquelle le résultat d'une expérience s'écrit sous la forme : $\bar{x} \pm k\,\sigma$. Le facteur d'élargissement k dépend du seuil de signification choisi pour cet intervalle de confiance.

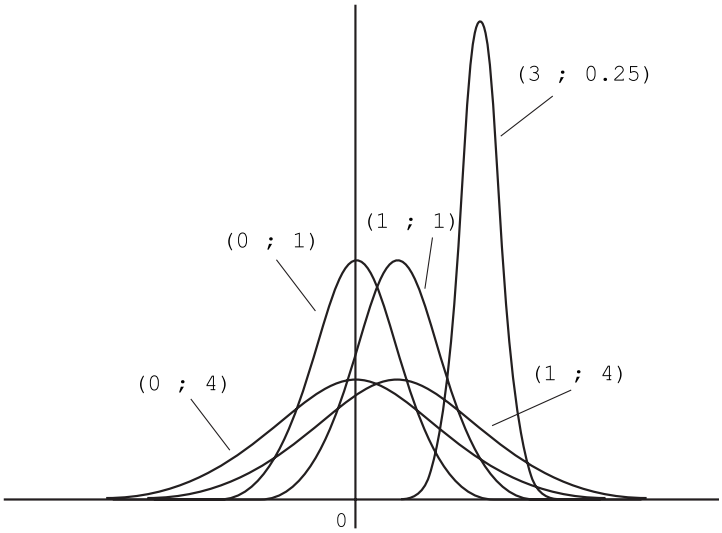

FIGURE 2: Allure de la loi normale en fonction des paramètres ($\bar{x}$, σ)

A partir de la figure ci-dessus, on voit clairement que plus la valeur de σ est grande, plus la distribution en question est large. La probabilité [3] d'avoir une mesure comprise dans l'intervalle $\bar{x} \pm k\,\sigma$ est :

$$p(\bar{x} - k\sigma < x < \bar{x} + k\,\sigma) = \int_{\bar{x}-k\sigma}^{\bar{x}+k\sigma} \rho(x)\,dx \tag{6}$$

La densité de probabilité $p(x)$ a les propriétés suivantes :

$$\rho(x) \geq 0 \tag{7}$$

3. Cette intégrale n'a pas de primitive explicite. Cela signifie qu'il est impossible de l'exprimer algébriquement à partir des fonctions usuelles (polynômes, exponentielle, logarithme ....). Les calculatrices et les tableurs permettent de calculer des valeurs approchées de cette intégrale. Il existe des tables donnant directement une valeur approchée selon le seuil de signification fixé.

$$p(x) = \int_{-\infty}^{+\infty} \rho(x)\, dx = 1 \tag{8}$$

Pour un niveau de confiance de 95% $(k = 2)$ [4], il vient :

$$p(\bar{x} - 2\sigma < x < \bar{x} + 2\,\sigma) = \int_{\bar{x}-2\sigma}^{\bar{x}+2\sigma} \rho(x)\, dx \Rightarrow p(-2 < z < +2) = \int_{-2}^{+2} \rho(z)\, dz = 0.95$$

Pour $k = 2$, nous obtenons :

$$p(\bar{x} - 2\sigma < x < \bar{x} + 2\,\sigma) \iff p(\bar{x} - 2\sigma - \bar{x} < x - \bar{x} < \bar{x} + 2\,\sigma - \bar{x})$$

$$\iff p(-2\sigma < x - \bar{x} < +2\,\sigma) \iff p(-2\frac{\sigma}{\sigma} < \frac{x - \bar{x}}{\sigma} < +2\,\frac{\sigma}{\sigma})$$

$$\Rightarrow p(-2 < z < +2) \simeq 0.95$$

Ainsi, 95% des réalisations de la variable aléatoire se trouvent dans l'intervalle $\bar{x} \pm 2\,\sigma$. La moyenne de cet intervalle étant la valeur la plus probable de la variable aléatoire.

a. Exercice d'application

Une entreprise pharmaceutique produit en grande quantité des comprimés analgésiques. Soit X la variable aléatoire qui, à chaque comprimé prélevé au hasard dans la production, associe la masse du principe actif, en milligrammes. On suppose que la variable aléatoire X suit la loi normale de moyenne $505.5\,mg$ et d'écart-type $18\,mg$.

1) Déterminer la probabilité qu'un comprimé, tiré au hasard, ait une masse du principe actif comprise entre $492.4\,mg$ et $510.3\,mg$
2) Un comprimé est déclaré défectueux si la masse du principe actif est, soit inférieure à $485.7\,mg$, soit supérieure à $520.7\,mg$. Calculer la probabilité qu'un comprimé tiré au hasard soit défectueux.

Solution :

$$1)\ \text{On cherche} \quad p(492.40 \leq m \leq 510.30) = p(492.40 < m < 510.30) = ?$$

On calcule d'abord les variables centrées réduites correspondantes :

$$z_1 = \frac{492.4 - 505.5}{18} = -0.73 \qquad \text{et} \qquad z_2 = \frac{510.30 - 505.5}{18} = 0.27$$

$$z = \frac{m - 505.5}{18} \quad \Rightarrow \quad p(z_1 \leq z \leq z_2) = p(z \leq z_2) - p(z \leq z_1) = p(z \leq 0.27) - p(z \leq -0.73)$$

$$= p(z \leq 0.27) - \underbrace{p(z > 0.73)}_{1-p(z\leq 0.73)} = p(z \leq 0.27) - (1 - p(z \leq 0.73)) = 0.60642 - (1 - 0.76730)$$

$$\Rightarrow p(z_1 \leq z \leq z_2) = 0.37$$

4. De façon analogue pour k = 1 $\Rightarrow p(z) \simeq 0.68$ et $k = 3 \Rightarrow p(z) \simeq 0.99$. Le choix du niveau de confiance dépend de la précision qu'on souhaite donner au résultat. En sciences expérimentales on prend très souvent $k = 2$.

Ou bien en terme de pourcentage $p(z_1 \leq z \leq z_2) = 37\% \Rightarrow$ nous avons ainsi 37 chances sur 100 d'avoir un comprimé ayant une masse du principe actif comprise entre 492.40 et 510.30 mg.

2) D'après les données, on comprend que pour avoir un comprimé qui n'est pas défectueux il faut calculer la probabilité :

$$p_1(485.70 \leq m \leq 520.70) = ?$$

Ensuite on détermine le complément de cette probabilité :

$$p_2 = 1 - p_1$$

Où p_2 représente la probabilité qu'un comprimé tiré au hasard soit défectueux. Avec une démarche similaire que la première question il vient :

$$p_1 = p(z_1 \leq z \leq z_2) = p(z \leq 0.84) - p(z \leq -1.10) = p(z \leq 0.84) - (1 - p(z \leq 1.10))$$
$$= 0.79955 - (1 - 0.86433) = 0.67 \;\Rightarrow\; p_2 = 1 - 0.67 = 0.33$$

Les probabilités sont calculées [5] à partir de la table de la loi Normale centrée et réduite

C. *Propagation des incertitudes*

Dans cette section, on commencera d'abord par examiner le cas où les variables aléatoires $X_1, X_2, ..., X_n$ sont indépendantes. Cela signifie que la valeur d'une erreur sur une mesure donnée ne dépend pas de celle sur une autre mesure [6]. Soit Y une réponse, physique ou chimique, fonction des grandeurs $X_1, X_2, ..., X_n$:

$$Y = f(X_1, X_2, ..., X_n) \tag{9}$$

Les incertitudes ΔX_i commises sur chaque mesure X_i se propagent et affectent la grandeur Y suivant la relation [7] :

$$\Delta Y = \sqrt{\sum_{i=1}^{n} \left(\frac{\delta f}{\delta X_i}\right)^2 \times (\Delta X_i)^2} \tag{10}$$

Avec, ΔY est l'erreur absolue commise sur la grandeur Y et ΔX_i est l'erreur absolue commise sur la mesure X_i. Si la condition, d'erreurs aléatoires indépendantes, n'est pas vérifiée, la relation (10) de combinaison quadratique ne s'applique pas. En outre, la formule (10) est appliquée sans distinction afin d'évaluer les incertitudes de type A et celles de type B. Si les variables X_i sont corrélées deux à deux, l'équation (10) prend la forme :

$$\Delta Y^2 = \sum_{i=1}^{n} \left(\frac{\delta f}{\delta X_i}\right)^2 \times (\Delta X_i)^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{\delta f}{\delta X_i} \frac{\delta f}{\delta X_j} \Delta(X_i, X_j) \tag{11}$$

5. Les valeurs des probabilités sont déterminées à partir de la table de la loi normale centrée et réduite. Il faut systématiquement faire en sorte de transformer la probabilité en question selon l'écriture $p(Z \leq z)$ afin d'utiliser cette table.

6. Concrètement si $Y = f(x_1; \Delta x_1, x_2; \Delta x_2) \Rightarrow$ il existe aucune relation mathématique entre les variables aléatoires x_1 et x_2. Une variation pour l'une des variables n'entraîne pas une variation prévisible pour l'autre

7. Cette relation dérive d'un développement de Taylor. Il est supposé que les variables aléatoires X_i suivent une distribution de Gauss et que les valeurs $\Delta X_i / Y << 1$ restent suffisamment faibles.

Avec $\Delta(X_i, X_j) = \Delta(X_j, X_i)$ est la covariance associée à X_i et X_j. D'un autre côté le coefficient de corrélation vaut :

$$r(X_i, X_j) = \frac{\Delta(X_i, X_j)}{\sigma_{X_i}\,\sigma_{X_j}} \tag{12}$$

En combinant les relations (11) et (12) on obtient :

$$\Delta Y^2 = \sum_{i=1}^{n}\left(\frac{\delta f}{\delta X_i}\right)^2 \times (\Delta X_i)^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{\delta f}{\delta X_i}\frac{\delta f}{\delta X_j}\,\sigma_{X_j}\,\sigma_{X_j}\,r(X_i, X_j) \tag{13}$$

Les coefficients de corrélation sont plus aisément interprétables que les covariances. En effet, pour des variables décorrélées $\Rightarrow r(X_i, X_j) = 0$, on retrouve la relation (10).

b. Exercice d'application

Comme application de la relation de propagation des erreurs, considérons la formule du dosage

$$C_x = \frac{C_0 \times V_e}{V}$$

Dans ce cas la concentration inconnue C_x est une fonction s'écrivant comme :

$$C_x = f(C_0, V_e, V)$$

il vient :

$$dC_x = \sqrt{\left(\frac{\delta C_x}{\delta C_0}\right)^2 \times (dC_0)^2 + \left(\frac{\delta C_x}{\delta V_e}\right)^2 \times (dV_e)^2 + \left(\frac{\delta C_x}{\delta V}\right)^2 \times (dV)^2}$$

$$dC_x = \sqrt{\left(\frac{V_e}{V}\right)^2 \times (dC_0)^2 + \left(\frac{C_0}{V}\right)^2 \times (dV_e)^2 + \left(-\frac{C_0\,V_e}{V^2}\right)^2 \times (dV)^2}$$

Ensuite on passe aux variations finies, cela se traduit par le remplacement des éléments différentiels par les incertitudes sur les grandeurs correspondantes ($d \Rightarrow \Delta$). De plus, les incertitudes s'ajoutent, donc tous les signes négatifs redeviennent des signes positifs.

$$\Delta C_x = \sqrt{\left(\frac{V_e}{V}\right)^2 \times (\Delta C_0)^2 + \left(\frac{C_0}{V}\right)^2 \times (\Delta V_e)^2 + \left(\frac{C_0\,V_e}{V^2}\right)^2 \times (\Delta V)^2}$$

Où ΔC_0 est l'erreur commise sur la concentration de la solution titrante, ΔV_e est l'erreur commise sur le volume équivalent et ΔV est l'erreur commise sur le volume de la solution titrée. Notons que ΔV_e est estimée par :

$$\Delta V_e = \sqrt{(\Delta V_{lec})^2 + (\Delta V_{lec} + \Delta V_{bur} + \Delta V_g)^2}$$

Avec, ΔV_{lec} est l'erreur de lecture, estimée à partir de la moitié de la graduation (0.025 mL). ΔV_{bur} est l'incertitude liée à la burette (0.02 mL) et $\Delta V_g = 0.05$ mL (incertitude liée à la goutte). Ainsi, la concentration C_x est déterminée avec une précision $C_x \pm \Delta C_x$.

c. Exercice supplémentaire

En utilisant la formule de propagation des erreurs, calculer l'incertitude commise sur la grandeur f :

$$f = a \times x \quad ; \quad f = x^a \times y^b \quad ; \quad f = \frac{a \times x}{b \times y} \quad ; \quad f = a \times x + b \times y \quad ; \quad f = \frac{a}{x^2} + \frac{b}{y^2}$$

Commenter les résultats obtenus.

II. Analyse de la variance

Le principe de l'analyse de la variance consiste à décomposer la variance totale en une somme de variances inter- et intra-groupes :

$$\underbrace{y_i - \bar{y}}_{\text{err. Totale}} = \underbrace{\bar{y}_i - \bar{y}}_{\text{err. factorielle}} + \underbrace{y_i - \bar{y}_i}_{\text{err. expérimentale}} \tag{14}$$

Pour la somme des carrés des écarts, on aura :

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{\text{var. totale}} = \underbrace{\sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2}_{\text{var. factorielle}} + \underbrace{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}_{\text{var. expérimentale}} \tag{15}$$

Le travail de l'expérimentateur consiste à chercher à minimiser le plus possible l'erreur factorielle. En d'autres mots, il faudra vérifier que la somme des carrés des écarts des erreurs factorielles soit très proche de la somme des carrés des écarts des erreurs expérimentales. Si l'on souhaite aller plus loin dans la comparaison des deux variances (expérimentale et factorielle), on peut utiliser la statistique de Fisher, notée F_{obs}.

La statistique (ou test) de Fisher consiste à comparer le rapport entre la variance factorielle et la variance expérimentale. Cette statistique (rapport des deux variances suivant leur degrés de liberté respectifs) permet de quantifier la probabilité que la variance factorielle soit négligeable devant la variance expérimentale. Plus la valeur de F_{obs} est faible plus la variance factorielle est négligeable devant l'erreur expérimentale. La relation (15) est à la base de l'analyse de la variance. Pour J échantillons, la généralisation de l'équation (15) conduit à :

$$\frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - \bar{y})^2 = \frac{1}{I} \sum_{i=1}^{I} (\bar{y}_i - \bar{y})^2 + \frac{1}{I \times J} \sum_{i=1}^{I} \left(\sum_{j=1}^{J} (y_{ij} - \bar{y}_i)^2 \right) \tag{16}$$

Avec, I et J ($I \times J = n$, est le nombre total de données) sont respectivement les nombres de lignes et de colonnes. En multipliant (16) par n, on obtient :

$$\sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - \bar{y})^2 = J \sum_{i=1}^{I} (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{I} \left(\sum_{j=1}^{J} (y_{ij} - \bar{y}_i)^2 \right) \tag{17}$$

L'équation (17) se lit :

$$\underbrace{SC_{\mathrm{T}}}_{\text{Totale}} = \underbrace{SC_{\mathrm{F}}}_{\text{Factorielle}} + \underbrace{SC_{\mathrm{R}}}_{\text{Résiduelle}} \tag{18}$$

Avec,

- $SC_{\mathbf{T}}$: Dispersion des données autour de la moyenne générale.
- $SC_{\mathbf{F}}$: Dispersion des moyennes de chaque groupe autour de la moyenne générale (variance inter-groupe).
- $SC_{\mathbf{R}}$: Dispersion des données à l'intérieur de chaque groupe autour de sa moyenne (variance intra-groupe).

La statistique de Fisher est donnée par :

$$F_{\text{obs}} = \frac{\frac{SC_{\text{F}}}{df_1}}{\frac{SC_{\text{R}}}{df_2}} \Longrightarrow F_{\text{obs}} = \frac{SC_{\text{F}}}{SC_{\text{R}}} \times \frac{df_2}{df_1} \tag{19}$$

Avec df (The degrees of freedom), df_1 est le nombre de degré de liberté de SC_{F} et df_2 est le nombre de degré de liberté de SC_{R}.

$$df_1 = J - 1 \quad \text{et} \quad df_2 = \text{nombre total de données - J} \tag{20}$$

Si $F_{\text{obs}} < F_\alpha(df_1, df_2)$ alors la variance SC_{F} n'est pas significativement supérieure à la variance SC_{R}. Sinon, la variance SC_{F} est significativement supérieure à la variance SC_{R}. La quantité statistique $F_\alpha(df_1, df_2)$ représente le quantile d'ordre α de la loi de Fisher à df_1 et df_2 degrés de liberté.

Dans les logiciels de statistique, les résultats de l'analyse de la variance sont rassemblés sous forme d'un tableau, appelé *table de l'analyse de la variance* (One-way analysis of variance). Dans le cas présent (anova à un facteur), cette table se présente sous la forme suivante :

TABLE II: Table de l'analyse de la variance

Source	SS	df	MS	F	probability
Inter	CM_F	df_1	CM_F/df_1	F_{obs}	p-value
Intra	CM_R	df_2	CM_R/df_2		
Total	CM_T	$df_1 + df_2$			

d. Exemple d'application

TABLE III: Résultats par appareil

Essai	Appareil A	Appareil B	Moyenne
1	36.7000	27.9500	32.3250
2	34.5000	32.3500	33.4250
3	34.1000	36.9000	35.5000
4	35.5000	37.8000	36.6500
5	36.0500	28.8500	32.4500
6	37.3000	33.6000	35.4500

Solution

Il sera question de l'analyse de variance à un facteur (une seule influence). Dans cet exercice, le facteur est la méthode d'analyse et la réponse est la masse de la substance. A cet effet, on utilisera la relation :

$$\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\bar{y})^2}_{SC_T}=\underbrace{J\sum_{i=1}^{I}(\bar{y}_i-\bar{y})^2}_{SC_F}+\underbrace{\sum_{i=1}^{I}\left(\sum_{j=1}^{J}(y_{ij}-\bar{y}_i)^2\right)}_{SC_R} \tag{21}$$

Le terme SC_T compare toutes les mesures du tableau à la moyenne générale $\bar{y}$. Le terme SC_F compare les moyennes des méthodes d'analyse A et B ($\bar{y}_i$) à la moyenne générale. Le terme SC_R compare la moyenne d'une méthode d'analyse donnée à la moyenne obtenue par la méthode elle-même. Ce dernier terme traduit l'influence des incertitudes expérimentales. Le nombre de données expérimentales n = 12. Le nombre de ligne I = 6 et le nombre de colonne J = 2.

$$\bar{y}=34.3000 \qquad \bar{y}_A=35.6917 \qquad \bar{y}_B=32.9083 \tag{22}$$

$$\sum_{i=1}^{I=6}(\bar{y}_i-\bar{y})^2=(32.3250-34.30)^2+(33.4250-34.30)^2+(35.5000-34.30)^2$$
$$+(36.65000-34.30)^2+(32.4500-34.30)^2+(35.4500-34.30)^2=16.3737$$
$$\Rightarrow J\times\sum_{i=1}^{I=6}(\bar{y}_i-\bar{y})^2=2\times 16.3737=32.7474$$

Le deuxième terme :

$$\sum_{i=1}^{I=6}\left(\sum_{j=1}^{J=2}(y_{ij}-\bar{y}_i)^2\right)=(36.7000-32.3250)^2+(27.9500-32.3250)^2$$
$$+(34.5000-33.4250)^2+(32.3500-33.4250)^2$$
$$+(34.1000-35.5000)^2+(36.9000-35.5000)^2$$
$$+(35.5000-36.6500)^2+(37.8000-36.6500)^2$$
$$+(36.0500-32.4500)^2+(28.2500-32.4500)^2$$
$$+(37.3000-35.4500)^2+(33.6000-35.4500)^2$$
$$\Rightarrow\sum_{i=1}^{I=6}\left(\sum_{j=1}^{J=2}(y_{ij}-\bar{y}_i)^2\right)=79.9225$$
$$F_{obs}=\frac{32.7474}{79.9225}\times\frac{10}{1}\Rightarrow F_{obs}=4.0973 \tag{23}$$

Or d'après la table des fractiles de la loi de Fisher on lit $F_{0.95}(1,10)=4.9600$. Nous avons bien $F_{obs}<F_{0.95}(1,10)$, on en déduit que le type de méthode d'analyse n'a pas une influence significative sur le résultat de la masse de la substance considérée.

On peut également quantifier cette analyse statistique en calculant la p-value du test de Fisher selon la démarche suivante :

$$\begin{cases} \mathcal{H}_0 \Longrightarrow SC_F = 0 \\ \mathcal{H}_1 \Longrightarrow SC_F \neq 0 \\ \text{p-value} = \mathbb{P}(F > F_{obs}) \end{cases}$$

Si p-value $= \mathbb{P}_{\mathcal{H}_0}(F > F_{obs}) < \alpha \Longrightarrow$ l'hypothèse $\mathcal{H}_0$ est rejetée au risque 5%.
Sous Matlab, la probabilité $\mathbb{P}(F > F_{obs})$ est calculée selon :

$$\text{p-value} = \mathbb{P}(F > F_{obs}) = 1 - \mathbb{P}(F < F_{obs}) = 1 - \Phi(F_{obs}) \tag{24}$$

Avec,

$$\Phi(F_{obs}) = \mathbb{P}(F < F_{obs}) = \int_{-\infty}^{F_{obs}} \rho(x|df_1, df_2)\, dx \tag{25}$$

La fonction $\rho(x|df_1, df_2)$ est la distribution de Fisher (Fisher's probability density function) de la variable x pour les degrés de liberté df_1 et df_2. Avec, $\Phi(F_{obs})$ est la fonction de répartition de la loi de Fisher ou encore la distribution cumulative de la loi de Fisher (Fisher's cumulative distribution function). L'hypothèse H_0 est acceptée dans le cas où :

$$1 - \Phi(F_{obs}) \geq 0.05 \tag{26}$$

et elle est rejetée si

$$1 - \Phi(F_{obs}) < 0.05 \tag{27}$$

```
clc ; clear all ;
% Samir Kenouche le 20/02/2019
% Test de Fisher, par defaut alpha = 0.05
fobs = 4.0973 ; df1 = 1 ; df2 = 10 ;
pvalue = 1 - fcdf(fobs, df1, df2)
pvalue = 0.0705
```

La p-value calculée est supérieure au seuil 0.05 (5%), dans ce cas l'hypothèse nulle $\mathcal{H}_0$ est acceptée et l'hypothèse alternative $\mathcal{H}_1$ est rejetée. On retrouve ainsi le résultat précédent. La p-value $=$ 0.0705 signifie qu'il y a 07 chances sur 100 d'avoir une valeur de SC_F égale à zéro. En effet, cette variance a donc une forte chance d'être égale à zéro par rapport au seuil fixé $\alpha = 0.05$. Elle est donc significativement différente de zéro. On comprend ainsi que l'interprétation des résultats dépendra du seuil de signification fixé. Pour un seuil $\alpha = 0.10$ c'est l'hypothèse $\mathcal{H}_1$ qui sera plutôt acceptée. Du coup, se pose la problématique du choix de seuil de signification notamment pour les p-values proches du seuil fixé.

III. Comparaison de deux moyennes

Un expérimentateur peut se trouver dans une situation où il devra comparer les moyennes de deux séries de mesures. Imaginons par exemple qu'un dosage d'une substance donnée soit réalisé par deux chimistes, ou par deux méthodes d'analyse différentes ou bien pendant deux saisons différentes (par exemple en hiver et au printemps). Le test de signification est construit en calculant la variable de *Student* :

$$t = \frac{|\bar{a}_1 - \bar{a}_2|}{\sqrt{V_{a_1} + V_{a_2}}} \tag{28}$$

Avec, les variances sur les mesures $a_{1,i}$ et $a_{2,i}$ sont données par :

$$V_{a_1} = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (a_{1,i} - \bar{a}_1)^2 \quad \text{et} \quad V_{a_2} = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (a_{2,i} - \bar{a}_2)^2 \tag{29}$$

Avec n_1 et n_2 sont les nombres de répétitions effectuées au même point expérimental respectivement pour $a_{1,i}$ et $a_{2,i}$. Le rapport (28) suit une loi de Student à $\nu = (n_1 - 1) + (n_2 - 1)$ degrés de liberté.

e. Exercice d'application :

Considérons une étude portant sur la mesure de la masse (mg) d'un principe actif, pour un comprimé donné, menée par deux laboratoires pharmaceutiques différents. Les résultats obtenus sont rassemblés dans le tableau ci-dessous.

TABLE IV: Résultats par laboratoire

Essai	Laboratoire A	Laboratoire B
1	500.1	509.5
2	480.5	500.0
3	490.3	502.2
4	488.2	489.8
5	500.0	500.6
6	502.0	499.2
7	500.4	489.2
8	505.8	500.9
9	485.6	497.1
10	498.5	487.9

Solution

$$\bar{a}_1 = 495.1400 \quad \bar{a}_2 = 497.6400 \quad V_{a_1} = 69.4227 \quad V_{a_2} = 46.3004$$

$$t_{obs} = \frac{|495.1500 - 497.6400|}{\sqrt{69.4227 + 46.3004}} \Rightarrow t_{obs} = 0.2324$$

Nous avons $t_{(n_1+n_2-2),\alpha} = t_{(18,0.05)} = 2.10$, voir la table de *Student*. Pour les étudiants (es) qui ont du mal à se familiariser avec le calcul des probabilités, peuvent comparer directement le quantile

$t_{(n_1+n_2-2),\alpha}$ à celui calculé à partir de la statistique de *Student*.

Le test impose : $\mathbb{P}\left(T > t_{(n_1+n_2-2),\alpha}\right) = \alpha$

Si $|t_{obs}| > t_{(n_1+n_2-2),\alpha} \Rightarrow \mathbb{P}\left(T > t_{obs}\right) < \mathbb{P}\left(T > t_{(n_1+n_2-2),\alpha}\right) \Rightarrow \mathbb{P}\left(T > t_{obs}\right) < \alpha$

L'hypothèse $\mathcal{H}_0$ est rejetée au risque α

Si $|t_{obs}| < t_{(n_1+n_2-2),\alpha} \Rightarrow \mathbb{P}\left(T > t_{obs}\right) > \mathbb{P}\left(T > t_{(n_1+n_2-2),\alpha}\right) \Rightarrow \mathbb{P}\left(T > t_{obs}\right) > \alpha$

L'hypothèse $\mathcal{H}_0$ est acceptée au risque α

$t_{obs} < 2.10 \Rightarrow \mathcal{H}_0$ est acceptée au risque 5%

On en déduit que les deux moyennes ne sont pas significativement différentes et par conséquent les essais menés par les deux laboratoires produisent les mêmes résultats.