■ Incertitudes expérimentales ■

1. Introduction et définitions

En sciences expérimentales, la grandeur que l'on souhaiter déterminer s'appelle le *mesurande* (valeur vraie X) qui est obtenu grâce à différentes *mesures* ou *mesurages* (valeur mesurée x). La mesure correspond à l'action de mesurer la variable aléatoire x dans l'espoir d'estimer la valeur vraie X qui est par essence inconnue puisqu'on la cherche !

Par exemple, quand on mesure la valeur du pH d'une solution, le mesurande est $X = pH$ et la mesure x est lue grâce au pHmètre.

La différence entre la valeur mesurée x et la valeur vraie X est *l'erreur de mesure* $E = x - X$. On distingue deux types d'erreur, les *erreurs aléatoires* et les *erreurs systématiques.*

1.1. Erreur aléatoire

Lorsque les conditions de répétabilité sont remplies c'est-à-dire lorsque les différentes mesures expérimentales sont réalisées rigoureusement dans les mêmes conditions opératoires, l'ensemble des mesures suit une distribution statistique de valeur moyenne $\bar{x}$, l'erreur aléatoire est par définition $E_{aléa} = x - \bar{x}$.

Par exemple : On considère des feuilles de papier sur lesquelles se trouvent un millier de points verts, rouges et bleus. On demande à un élève de compter le nombre de points verts sur chaque feuille. Comme les points sont petits et qu'il y en a beaucoup, les résultats des mesures x obtenues risquent d'être légèrement différents de la moyenne.

L'erreur aléatoire est généralement due aux conditions extérieures, variations de la température (entraînant une dilatation), de la pression, de l'humidité, ... L'erreur aléatoire est également liée à la *fidélité* de l'instrument de mesure (œil + cerveau dans l'exemple) c'est-à-dire son aptitude à donner des indications très voisines lors de l'application répétée du même protocole.

1.2. Erreur systématique

Par définition l'erreur systématique est la différence $E_{syst} = \bar{x} - X$ où $\bar{x}$ est la moyenne des différentes mesures. En théorie il faut une infinité de mesures pour obtenir $\bar{x}$, sinon il y a là encore une erreur sur l'estimation de la moyenne.

Par exemple, si l'expérience précédente est réalisée par un daltonien pour qui le vert et le rouge sont identiques, il commettra systématiquement la même erreur en plus de l'erreur aléatoire.

Les erreurs systématiques peuvent être dues à une erreur d'étalonnage (faire le zéro du spectrophotomètre avec une eau trouble), à l'oubli d'un paramètre (oublier que la période d'un pendule dépend en toute rigueur de l'angle initial), à une procédure erronée (ne pas tenir compte de la résistance du voltmètre qui crée un diviseur de tension), au vieillissement de l'appareil...

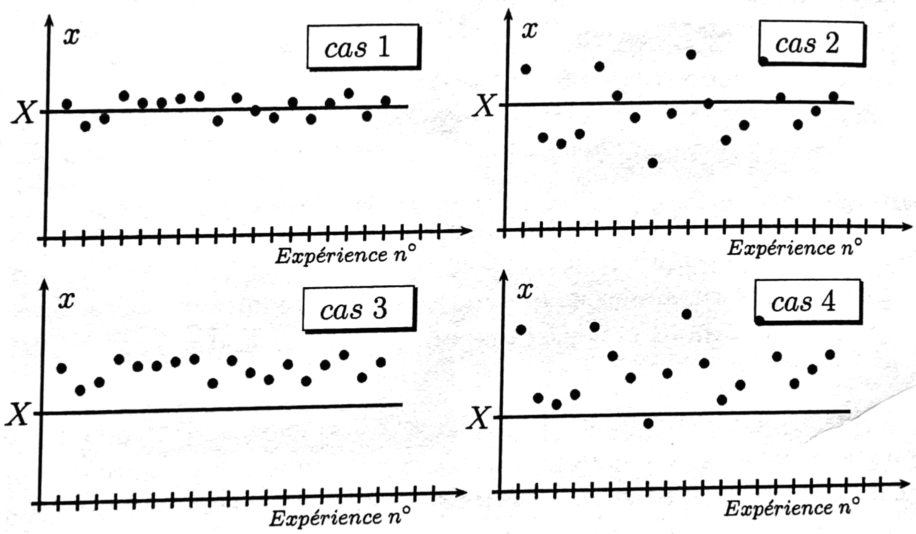
L'estimation de l'erreur systématique est appelée *biais de mesure* ou *erreur de justesse*. La *justesse* d'un instrument de mesure correspond à son aptitude à donner des indications exemptes d'erreur systématique.

1.3. Erreur de mesure

D'après les définitions précédentes, l'erreur de mesure est la somme des deux précédentes :

$$E = \underbrace{(x - \bar{x})}_{E_{aléa}} + \underbrace{(\bar{x} - X)}_{E_{syst}} = x - X = E_{aléa} + E_{syst}$$

Exemple de mesure x pour le mesurande X :



Le cas 1 présente une faible erreur aléatoire (instrument fidèle) et une faible erreur systématique (instrument juste). Dans le cas 2 l'instrument de mesure est peu fidèle mais juste, dans le cas 3 il est fidèle mais peu juste et dans le cas 4 il est peu fidèle et peu juste. La détermination de l'incertitude se fait selon deux méthodes : type A (méthode statistique) ou type B (évaluation par données du constructeur ou connaissance du matériel).

1.4. Incertitudes

L'incertitude Δx estime l'importance de l'erreur aléatoire commise. En absence d'erreur systématique, elle définit un intervalle autour de la mesure qui inclut la valeur vraie avec une probabilité plus ou moins grande. L'évaluation de l'incertitude de la mesure due aux erreurs expérimentales est un domaine appelé <u>*métrologie*</u>.

Le mesurande se présente alors sous la forme $\boxed{X = \bar{x} \pm \Delta x}$

Δx est aussi appelée l'incertitude absolue et $\dfrac{\Delta x}{|\bar{x}|}$ l'incertitude relative.

Rq Dans le cas d'une évaluation de type B, s'il n'y a qu'une mesure elle est égale à la moyenne.

2. Évaluation de type A de l'incertitude

L'évaluation de type A de l'incertitude est réalisée par l'analyse statistique de la série de n mesures x_k indépendantes. La moyenne arithmétique est le meilleur estimateur du mesurande :

$$\boxed{\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k}$$

Il faut également étudier la dispersion de la distribution des résultats, on utilise alors l'écart-type :

$$\boxed{\sigma_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})^2}}$$

On peut également écrire :

$$\sum_{k=1}^{n}(x_k-\bar{x})^2 = \sum_{k=1}^{n}\left(x_k^2 - 2x_k\bar{x} + \bar{x}^2\right) = \sum_{k=1}^{n}x_k^2 - 2\sum_{k=1}^{n}x_k\bar{x} + \sum_{k=1}^{n}\bar{x}^2$$

$$= \sum_{k=1}^{n}x_k^2 - 2\bar{x}\underbrace{\sum_{k=1}^{n}x_k}_{n\bar{x}} + n\bar{x}^2 = \sum_{k=1}^{n}x_k^2 - n\bar{x}^2 = n\left(\overline{x^2} - \bar{x}^2\right)$$

$$\text{et donc } \sigma_x = \sqrt{\frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right)}$$

Rq On divise par $n-1$ et non par n, pour mémoire, on retiendra que lorsque $n=1$ (une seule mesure), l'écart-type n'est pas défini.

La variable $\bar{x}$ est elle-même une variable aléatoire qui dépend du nombre de mesures et qui possède une certaine distribution centrée sur la valeur du mesurande. Le théorème central limite stipule que l'écart-type de la moyenne $\sigma_{\bar{x}}$ vaut $\sigma_{\bar{x}} = \dfrac{\sigma_x}{\sqrt{n}}$ et si n tend vers l'infini alors la distribution tend vers une gaussienne centrée sur X.

L'estimation de l'incertitude est alors donnée par la formule :

$$\boxed{\Delta x = t_{68\,\%}\frac{\sigma_x}{\sqrt{n}}}$$

où t est le coefficient de Student qui dépend de la confiance accordée (ici 68 %) à l'expérience et de n. Pour une confiance de 68 %, c'est-à-dire que l'on a 68 chances sur 100 que X soit compris entre $\bar{x} - t\sigma_x/\sqrt{n}$ et $\bar{x} + t\sigma_x/\sqrt{n}$, le coefficient de Student prend les valeurs suivantes :

n	2	3	4	5	6	7	8	9	10	20	40	∞
$t_{68\,\%}$	1,84	1,32	1,20	1,14	1,11	1,09	1,08	1,07	1,06	1,03	1,01	1

En pratique $t \approx 1$ (avec plus de 7 mesures), les résultats se présentent alors sous la forme :

$$X = \bar{x} \pm \frac{\sigma_x}{\sqrt{n}} \quad \text{à 68 \% de confiance}$$

Pour une confiance de 95 %, le coefficient de Student prend les valeurs suivantes :

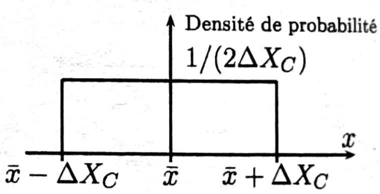
n	2	3	4	5	6	7	8	9	10	20	40	∞
$t_{95\,\%}$	12,7	4,3	3,18	2,78	2,57	2,45	2,37	2,31	2,26	2,09	2,02	1,96

3. Évaluation de type B de l'incertitude

Il s'agit d'évaluer l'incertitude par des méthodes autres que statistiques et principalement en utilisant les données du constructeur de l'appareil de mesure.

3.1. À partir de la classe de l'appareil

Le constructeur fournit la classe de l'appareil ΔX_C. On considère que l'indication donnée par le constructeur correspond à une distribution rectangulaire de largeur $2\Delta X_C$.



On montre alors que

$$\boxed{\Delta x = k\frac{\Delta X_C}{\sqrt{3}}}$$

$k = 1$ pour une confiance à 68 % et $k = 2$ pour une confiance à 95 %

Justification :
On appelle distribution de probabilité d'une grandeur aléatoire x à valeurs continues la fonction $x \mapsto f(x)$ telle que la probabilité pour que x soit compris entre a et b s'écrive :

$$P(a < x < b) = \int_a^b f(x)\mathrm{d}x$$

$f(x)$ est donc une densité de probabilité : $f(x)\mathrm{d}x$ est la probabilité que la variable aléatoire x prenne une valeur comprise entre x et $x + \mathrm{d}x$. On a alors : $f(x) > 0$ et $1 = \displaystyle\int_{-\infty}^{+\infty} f(x)\mathrm{d}x$.

La valeur moyenne (ou espérance) $\bar{x}$ d'une grandeur aléatoire x à valeur continue est donnée par :

$$\bar{x} = \int_{-\infty}^{+\infty} x f(x) \mathrm{d}x$$

L'écart-type σ de la distribution des valeurs de x est défini par :

$$\sigma^2 = \int_{-\infty}^{+\infty} (x - \bar{x})^2 f(x) \mathrm{d}x$$

C'est une mesure de la dispersion des valeurs de x autour de sa valeur moyenne. Plus les valeurs de x se concentrent autour de la moyenne, plus l'écart-type est faible. σ^2 est appelé variance de la distribution.

Pour une distribution rectangulaire de largeur $2\Delta X_C$, on a nécessairement une hauteur $1/(2\Delta X_C)$ afin d'avoir une aire unitaire. Le calcul donne :

$$\sigma^2 = \int_{\bar{x}-\Delta X_C}^{\bar{x}+\Delta X_C} (x-\bar{x})^2 \frac{1}{2\Delta X_C} \mathrm{d}x = \frac{1}{2\Delta X_C} \left[\frac{(x-\bar{x})^3}{3} \right]_{\bar{x}-\Delta X_C}^{\bar{x}+\Delta X_C} = \frac{\Delta X_C^2}{3}$$

$$\text{puis } \sigma = \frac{\Delta X_C}{\sqrt{3}} = \Delta x$$

3.2. Mesures analogiques

Pour un appareil de mesure analogique (appareil à cadran, lecture d'un réglet...), l'incertitude de lecture est estimée à partir de la valeur d'une graduation (précision de l'appareil). On considère que la graduation correspond à une distribution rectangulaire de largeur *graduation* (ou *division*) alors :

$$\boxed{\Delta x = k \frac{graduation}{2\sqrt{3}} = k \frac{graduation}{\sqrt{12}}}$$

$k = 1$ pour une confiance à 68 % et $k = 2$ pour une confiance à 95 %

3.3. Appareils numériques

Pour les appareils numériques, le constructeur indique pour la précision un pourcentage p de la valeur lue et un nombre N de digit (un digit correspond au dernier chiffre affiché). p correspond à une erreur de calibrage (variable d'un appareil à l'autre), donc systématique. Elle peut devenir aléatoire si on répète la mesure avec des appareils différents de la même série, et être traitée par une évaluation de type A. N correspond à une erreur aléatoire de type bruit de fond.

$$\boxed{\Delta x = k \left(\frac{p \times \text{valeur lue}}{\sqrt{3}} + \frac{N \,\text{digit}}{\sqrt{3}} \right)}$$

$k = 1$ pour une confiance à 68 % et $k = 2$ pour une confiance à 95 %

L'incertitude ne comporte qu'un seul chiffre significatif et la valeur numérique de la mesure doit être cohérente avec cette incertitude quant au dernier chiffre donné.

3.4. Présentation des résultats

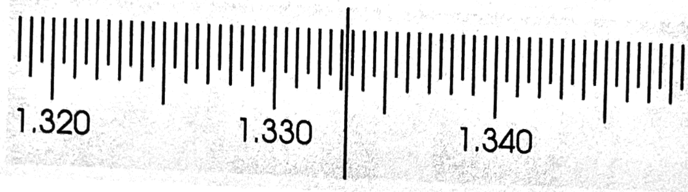
Les résultats sont présentés avec 1 seul chiffre significatif pour l'incertitude et autant de décimales pour la mesure et l'incertitude. Il faut également préciser la confiance.

Question 1 Des élèves mesurent l'indice de réfraction absolue de l'eau à l'aide d'un réfractomètre d'Abbe et une lampe à vapeur de sodium. On peut lire sur les copies :

- $n = 1,333 \pm 0,0005$ à 68 % de confiance
- $n = 1,333 \pm 0,0018$ à 68 % de confiance
- $n = 1,3329 \pm 0,005$ à 68 % de confiance
- $n = 1,3330 \pm 0,0005$ à 68 % de confiance

1) Quel est le résultat présenté correctement ? Proposer une correction pour les autres.

2) Sachant qu'une mesure est la suivante, comment présenter le résultat avec une confiance à 68 % ?



Réponse 1

- "$n = 1,333 \pm 0,0005$ à 68 % de confiance" est mal écrit, cette présentation peut se corriger selon : $n = 1,3330 \pm 0,0005$ ou $n = 1,333 \pm 0,001$.
- "$n = 1,333 \pm 0,0018$ à 68 % de confiance" présente une incertitude à 2 chiffres significatifs, il faut corriger selon : $n = 1,333 \pm 0,002$.
- "$n = 1,3329 \pm 0,005$ à 68 % de confiance" présente plus de chiffres après la virgule que l'incertitude, il faut écrire : $n = 1,333 \pm 0,005$.
- "$n = 1,3330 \pm 0,0005$ à 68 % de confiance" est une présentation convenable.

La ligne de lecture indique un indice centré entre 1,3330 et 1,3335. Pour une mesure analogique, l'incertitude est $\Delta n = \dfrac{graduation}{\sqrt{12}}$ avec une graduation de 0,0005, il vient $\Delta n = 0,00014$ et on peut présenter le résultat ainsi :

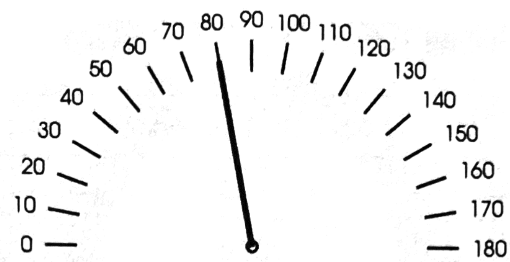
$$n = 1,3333 \pm 0,0002 \text{ à } 68\,\% \text{ de confiance}$$

3.5. Exemples

Question 2 Représenter les grandeurs physiques sous la forme $X \pm \Delta X$ avec l'incertitude à 95 % de confiance.

1) Une pipette indique 25 mL ± 0,05 mL.

2) Le compteur de vitesse affiche :



3) Une résistance affiche $R = 200\ \Omega \pm 5\,\%$.
4) Un ampèremètre affiche $5,21$ mA, la précision est de $(1\,\% \pm 2\,\text{digit})$:

Réponse 2

1) Pour une pipette indique 25 mL ± 0,05 mL (classe de l'appareil), l'incertitude sur la mesure du volume est $\Delta V = 2 \times 0,05/\sqrt{3} \approx 0,0577$ mL (à 95 % de confiance) donc le volume prélevé vaut :

$$\boxed{V = 25,00 \pm 0,06\ \text{mL}}\ \text{pour une confiance à } 95\,\%.$$

2) Le compteur de vitesse (analogique) affiche 80 km/h et une graduation correspond à 10 km/h donc $\Delta v = 2 \times 10/\sqrt{12} \approx 5,8$ km/h d'où :

$$\boxed{v = 80 \pm 6\ \text{km/h}}\ \text{pour une confiance à } 95\,\%.$$

3) Une résistance affiche $R = 200\ \Omega \pm 5\,\%$, $\Delta R_C = 200 \times 5\% = 10\ \Omega$ (classe de l'appareil) et l'incertitude vaut $\Delta R = 2 \times 10/\sqrt{3} \approx 11,5$ (à 95 %).
La résistance est donc :

$$\boxed{R = 0,20 \pm 0,02\ \text{k}\Omega}\ \text{pour une confiance à } 95\,\%.$$

4) Un ampèremètre affiche $5,21$ mA, la précision est de $(1\% \pm 2\ \text{digit})$: ce qui correspond donc à $\Delta I = 2\dfrac{5,21 \times 0,01 + 0,02}{\sqrt{3}} = 0,083$ A, l'intensité mesurée est :

$$\boxed{I = 5,21 \pm 0,09\ \text{A}}\ \text{au niveau de confiance de } 95\%.$$

4. Propagation des incertitudes

L'incertitude Δy d'une mesure y lorsqu'elle est obtenue à partir d'autres grandeurs $x_{k\in[1,n]}$ indépendantes (ou non corrélées) telle que $y = f(x_1, x_2, ...)$ s'écrit :

$$\Delta y = \sqrt{\sum_{k=1}^{n} \left(\frac{\partial f}{\partial x_k} \Delta x_k\right)^2}$$

On retient les cas particuliers suivants :

$z = x + y$ ou $z = x - y$	$\Rightarrow$	$\Delta z = \sqrt{\Delta x^2 + \Delta y^2}$
$z = \lambda x$ avec $\lambda = \text{cste}$	$\Rightarrow$	$\Delta z = \sqrt{\lambda^2 \Delta x^2} = \lvert\lambda\rvert\, \Delta x$
$z = x \times y$ ou $z = x/y$	$\Rightarrow$	$\Delta z = \sqrt{y^2\Delta x^2 + x^2\Delta y^2}$ ou $\dfrac{\Delta z}{z} = \sqrt{\left(\dfrac{\Delta x}{x}\right)^2 + \left(\dfrac{\Delta y}{y}\right)^2}$
$z = \lambda x^a \times y^b$	$\Rightarrow$	$\dfrac{\Delta z}{z} = \sqrt{\left(a\dfrac{\Delta x}{x}\right)^2 + \left(b\dfrac{\Delta y}{y}\right)^2}$

Question 3 — La masse du Soleil La troisième loi de Kepler relie la période T et le demi-grand axe a de l'orbite d'une planète autour de son étoile : $\dfrac{T^2}{a^3} = \dfrac{4\pi^2}{\mathcal{G}M}$ avec $\mathcal{G}$ la constante de la gravitation universelle et M la masse de l'étoile.
Déterminer la masse M_S du Soleil et l'incertitude ΔM_S sur cette masse.

Données :

- $\mathcal{G} = (6{,}674\,08 \pm 0{,}015\,\%) \cdot 10^{-11}\ \text{m}^3{\cdot}\text{kg}^{-1}{\cdot}\text{s}^{-2}$
- $T_{Terre} = (31\,556\,925 \pm 1)$ s
- $a_{Terre} = (1{,}495\,982\,620 \pm 0{,}001\,\%) \cdot 10^{11}$ m

Réponse 3 La masse du Soleil s'obtient facilement selon

$$M_S = \frac{4\pi^2 a^3}{\mathcal{G}T^2} = 1{,}988\,65 \cdot 10^{30}\ \text{kg}$$

On garde une précision relative d'environ 10^{-5} (comparable à celle de $\mathcal{G}$ et a).
La propagation des incertitudes donne

$$\frac{\Delta M_S}{M_S} = \sqrt{\left(3\frac{\Delta a}{a}\right)^2 + \left(2\frac{\Delta T}{T}\right)^2 + \left(\frac{\Delta \mathcal{G}}{\mathcal{G}}\right)^2}$$

avec pour 68 % de confiance

$$\frac{\Delta \mathcal{G}}{\mathcal{G}} = \frac{0{,}000\,15}{\sqrt{3}}\ ; \quad \frac{\Delta a}{a} = \frac{0{,}000\,01}{\sqrt{3}}\ ; \quad \frac{\Delta T}{T} = \frac{1}{\sqrt{3}} \cdot \frac{1}{31\,556\,925}$$

Un calcul donne

$$\frac{\Delta M_S}{M_S} = 8{,}8 \cdot 10^{-5} \quad \text{soit} \quad \Delta M_S = 1{,}8 \cdot 10^{26}\ \text{kg}$$

D'où le résultat

$$\boxed{M_S = (1{,}988\,7 \pm 0{,}000\,2) \cdot 10^{30}\ \text{kg à 68 \% de confiance}}$$

Question 4 — Incertitude d'un dosage

On considère le dosage d'une solution de soude de concentration C_B par une solution d'acide oxalique de concentration C_A dont un volume $V_A = 20$ mL est placé dans le bécher ($H_2C_2O_4$: $pK_A = 1,2$ et 4,2). L'équivalence du dosage est repérée par le virage de la phénophtaléine pour un volume $V_e = 16,0$ mL.

La solution d'acide a été préparée en pesant $m = 500,2$ mg d'acide oxalique dihydraté solide ($\mathcal{M} = 126,07$ g) à l'aide d'une balance de précision 0,1 mg. L'acide est dissous dans une fiole jaugée de volume $V_0 = 100$ mL de classe A (± 0,10 mL). On introduit ensuite les 20 mL à l'aide d'une pipette jaugée de 20 mL de classe A (± 0,020 mL). La burette utilisée est de classe A, graduée tous les 0,1 mL, elle délivre jusqu'à 25 mL avec une tolérance de ± 0,05 mL à 20 °C.

Déterminer la concentration de la soude et son incertitude à 68 % de confiance.

Réponse 4 — Incertitude d'un dosage

On détermine d'abord la concentration de l'acide et son incertitude :

$$C_A = \frac{m/\mathcal{M}}{V_0} = 0,0397 \text{ mol}\cdot\text{L}^{-1}$$

- $\Delta m = 0,1/\sqrt{3} = 0,058$ mg donc $m = 500,20 \pm 0,06$ mg,
- $\Delta V_0 = 0,10/\sqrt{3} = 0,058$ mL donc $V_0 = 100,00 \pm 0,06$ mL.

La formule de propagation donne

$$\frac{\Delta C_A}{C_A} = \sqrt{\left(\frac{\Delta m}{m}\right)^2 + \left(\frac{\Delta V_0}{V_0}\right)^2} = 6,1\cdot 10^{-4}$$

Au virage $V_e = 16$ mL, les deux acidités ont été dosées car le $pK_I \approx 9$ de l'indicateur est largement supérieur au $2^{\text{ème}}$ pK_A de l'acide alors :

$$n_{OH^- ajouté} = 2n_{H_2C_2O_4\,init} \text{ d'où } \boxed{C_B = 2C_A\frac{V_A}{V_e}}$$

Numériquement : $\boxed{C_B = 0{,}099\,25 \text{ mol}\cdot\text{L}^{-1}}$

Pour le volume d'acide, $\Delta V_A = 0,02/\sqrt{3} = 0,012$ mL donc $V_A = 20,00 \pm 0,02$ mL.

Le volume équivalent s'obtient grâce à la différence de deux lectures (le 0 de la burette et le 16,0 mL de l'équivalence) : $V_e = V_{fin} - V_{zéro}$.

L'incertitude vaut alors $\Delta V_e = \sqrt{\Delta V_{fin}^2 + \Delta V_{zéro}^2}$.

- $V_{zéro}$ présente une incertitude due à la lecture $\Delta V_{lecture} = \dfrac{graduation}{\sqrt{12}}$. Comme la graduation est usuellement de 0,1 mL, il vient $\Delta V_{lecture} = 0,03$ mL.
- V_{fin} présente une incertitude qui prend en compte :
  - la lecture $\Delta V_{lecture}$ (comme pour le zéro),

- la classe de la burette, pour cette manipulation (classe A) $\Delta V_C = 0,05$ mL donc $\Delta V_{burette} = \Delta V_C/\sqrt{3} = 0,03$ mL,
- le volume propre d'une goutte (qui n'est pas réductible) soit 0,05 mL et donc $\Delta V_{goutte} = 0,05/\sqrt{3} = 0,03$ mL,
- l'erreur de titrage due à la méthode (saut de pH, virage de l'indicateur,...), par défaut on peut considérer que cette erreur est nulle si le virage se fait à la goutte près.

On a alors $\Delta V_{fin} = \Delta V_{titrage} + \Delta V_{goutte} + \Delta V_{burette} + \Delta V_{lecture} = 0,09$ mL.
Finalement $\Delta V_e = \sqrt{\Delta V_{fin}^2 + \Delta V_{zéro}^2} = 0,1$ mL et $V_e = 16,0 \pm 0,1$mL.

Comme $C_B = 2C_A\dfrac{V_A}{V_e}$, la formule de propagation donne :

$$\frac{\Delta C_B}{C_B} = \sqrt{\left(\frac{\Delta C_A}{C_A}\right)^2 + \left(\frac{\Delta V_e}{V_e}\right)^2 + \left(\frac{\Delta V_A}{V_A}\right)^2} = 6,4 \cdot 10^{-3}$$

$$\text{d'où} \quad \Delta C_B = 6,3 \cdot 10^{-4}\ \text{mol·L}^{-1}$$

En conclusion :

$\boxed{C_B = 99,3 \pm 0,7\ \text{mmol·L}^{-1}}$ pour une confiance de 68 %

5. Les régressions linéaires

5.1. Présentation

Des relevés expérimentaux donnent N mesures fournissant un ensemble de couples (x_i, y_i) pour $i \in [\![1, N]\!]$. Le problème consiste à chercher une relation linéaire entre ces valeurs selon le modèle $y(x) = ax + b$.

L'estimateur des moindres carrés ordinaires consiste à minimiser la fonction ϕ, somme des carrés des écarts entre les valeurs prédites et les valeurs observées par rapport aux deux paramètres a (coefficient directeur) et b (ordonnée à l'origine). Dans le cas d'une approximation linéaire, il est facile de donner une expression analytique des estimateurs $\hat{a}$ et $\hat{b}$ de a et b.

$$\begin{array}{rcl} \phi : \quad \mathbb{R}^2 & \to & \mathbb{R} \\ (a,b) & \mapsto & \displaystyle\sum_{i=1}^{N} (y_i - (ax_i + b))^2 = \sum_{i=1}^{N} \left(y_i^2 + a^2x_i^2 + b^2 - 2ax_iy_i - 2by_i + 2abx_i\right) \end{array}$$

$$\text{Au minimum} \quad \left(\frac{\partial \phi}{\partial a}\right)_b = 0 \quad \text{et} \quad \left(\frac{\partial \phi}{\partial b}\right)_a = 0$$

$$\text{soit} \begin{cases} \left(\dfrac{\partial \phi}{\partial a}\right)_b \left(\hat{a}, \hat{b}\right) = 0 = \displaystyle\sum_{i=1}^{N} 2\hat{a}x_i^2 - 2x_iy_i + 2\hat{b}x_i \\ \left(\dfrac{\partial \phi}{\partial b}\right)_a \left(\hat{a}, \hat{b}\right) = 0 = \displaystyle\sum_{i=1}^{N} 2\hat{b} - 2y_i + 2\hat{a}x_i \end{cases}$$

La résolution du système donne alors :

$$\begin{cases} 0 = \sum_{i=1}^{N} x_i \left(\hat{a}x_i + y_i - \hat{b}\right) \\ 0 = \sum_{i=1}^{N} \hat{b} - y_i + \hat{a}x_i \end{cases} \Leftrightarrow \begin{cases} 0 = \sum_{i=1}^{N} x_i \left(\hat{a}x_i + y_i - \hat{b}\right) \\ N\hat{b} = \sum_{i=1}^{N} y_i - \hat{a}x_i \end{cases}$$

$$\Leftrightarrow \begin{cases} 0 = \sum_{i=1}^{N} x_i \left(\hat{a}x_i + y_i - \hat{b}\right) \\ \hat{b} = \bar{y} - \hat{a}\bar{x} \end{cases}$$

En posant : $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ et $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ les moyennes des y_i et x_i.

$$\Leftrightarrow \begin{cases} 0 = \sum_{i=1}^{N} x_i \left(\hat{a}\left(x_i - \bar{x}\right) + \left(y_i - \bar{y}\right)\right) \\ 0 = \sum_{i=1}^{N} \left(\hat{a}\left(x_i - \bar{x}\right) + \left(y_i - \bar{y}\right)\right) \end{cases}$$

En multipliant la 2^e^ égalité par $\bar{x}$, la soustraction des deux égalités donne :

$0 = \sum_{i=1}^{N} \left(x_i - \bar{x}\right)\left(\hat{a}\left(x_i - \bar{x}\right) + \left(y_i - \bar{y}\right)\right)$ finalement,

$$\boxed{\hat{a} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2} = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\cdot\bar{y}}{\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2} = \frac{\operatorname{cov}\left(X,Y\right)}{\operatorname{var}\left(X\right)}} \quad \text{et} \quad \boxed{\hat{b} = \bar{y} - \hat{a}\bar{x}}$$

$\operatorname{var}(X)$ est la variance de la variable aléatoire X prenant les valeurs $(x_i)_i$ et $\operatorname{cov}(X,Y)$ est la covariance des variables aléatoires X et Y.

On obtient alors un estimateur du coefficient directeur et de l'ordonnée à l'origine, mais quel crédit accorder à ces résultats ?

Par définition le coefficient de détermination est :

$$R^2 = r^2 = \frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)\operatorname{var}(Y)}$$

Il exprime le rapport entre la variance expliquée par le modèle et la variance totale (r est appelé coefficient de corrélation). Il mesure le pourcentage d'explication du modèle par la régression linéaire et peut aussi s'écrire :

$$R^2 = \frac{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2} \quad \text{où} \quad \hat{y}_i = \hat{a}x_i + \hat{b}$$

Ce coefficient présente un intérêt uniquement si l'hypothèse même de la relation linéaire a été validée. Il est alors nécessaire de faire un test d'hypothèse pour confirmer ou infirmer le modèle.

En fait R^2 sert à infirmer l'existence d'une relation linéaire entre X et Y, en effet s'il est proche de zéro alors $\text{cov}(X,Y) \approx 0$ indique que les variables aléatoires X et Y sont indépendantes. Si R^2 est proche de 1, il ne faut pas conclure immédiatement à la dépendance linéaire des variables, il est préférable d'effectuer un test.

5.2. L'incertitude sur les estimateurs

On introduit les résidus calculés ou estimés selon :

$$e_i = y_i - \hat{y}_i$$

La variance σ^2 est estimée par la variation résiduelle : $\sigma^2 = \dfrac{1}{N}\sum_{i=1}^{N} e_i^2$.

Une analyse statistique permet de donner les intervalles de confiance pour les estimateurs de la pente et de l'ordonnée à l'origine, il peut être montré que :

$$\boxed{\Delta\hat{a} = t \cdot \frac{\sigma}{\sigma_x} \cdot \sqrt{\frac{1}{N}}} \quad \text{et} \quad \boxed{\Delta\hat{b} = \Delta\hat{a}\sqrt{\sigma_x^2 + \bar{x}^2}}$$

où t est le coefficient de Student introduit précédemment

5.3. Test de significativité globale du F de Fisher

Il s'agit d'un test d'analyse statistique qui compare le F théorique de Fisher et le F_{calc} calculé à partir de la régression. Le F théorique de Fisher est tabulé selon le risque α de se tromper sur l'hypothèse et le nombre N de valeurs : on le note $F_{\alpha,1,N}$ (le 1 rappelle que l'on effectue une régression simple et non multiple).

Dans la suite, on considère un risque de se tromper sur la pertinence de la régression de $\alpha = 1\,\%$. Le F_{calc} s'obtient selon la formule $F_{calc} = \dfrac{(N-2)\,r^2}{1-r^2}$. Le test affirme que la régression linéaire est significative (avec un risque de se tromper de α) si $F_{calc} = \dfrac{(N-2)\,r^2}{1-r^2} > F$ soit $r > \sqrt{\dfrac{F}{N-2+F}}$.

Par exemple, dans le cas où $N = 25$, $F_{\alpha=0.01,1,N=25} = 7,77$, par conséquent la régression est valable si $r > \sqrt{\dfrac{F}{N-2+F}} = 0,502$!

Cette condition est évidemment trop peu restrictive, si elle peut avoir un sens en économie ou analyse de populations, elle ne présente pas beaucoup d'intérêt en physique.

En conclusion, il va falloir faire preuve de bon sens plutôt que d'utiliser des tests statistiques.

Rq Le F de Fisher est obtenu avec Python selon :

```
In [1]:  import scipy.stats as ss

In [2]:  ss.f.isf(0.01,1,25)
```

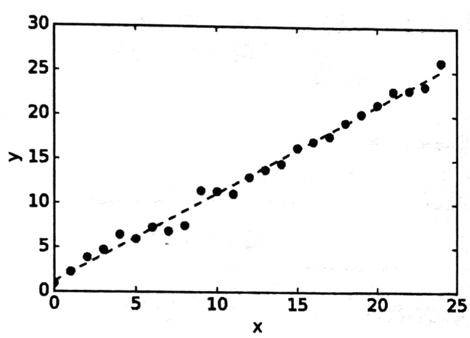
5.4. Exemple

On considère 2 exemples volontairement biaisés de $N = 25$ mesures (artificielles).

- Pour l'expérience numérique 1 : la figure 1 a été réalisée selon la loi $y_i = 1 + x_i + \varepsilon_i$ avec ε_i une variable aléatoire répartie uniformément dans l'intervalle $[-1,5\,;\,1,5]$ pour $x_i = i, \forall i \in [\![1, N]\!]$.
- Pour l'expérience numérique 2 : la figure 2 a été obtenue selon la loi $y_i = (1 + x_i)^{1,3}$ pour $x_i = i, \forall i \in [\![1, N]\!]$.

Dans les deux cas le coefficient de Student utilisé pour le calcul des incertitudes est donné à 99 % ce qui entraîne une incertitude élevée sur les estimateurs.
D'après sa réalisation l'expérience 1 correspond à une relation linéaire entre les variables x et y du type $y = ax + b$ avec $a = 1$ et $b = 1$ alors que l'expérience 2 simule une loi de puissance. Si on ne considère que le coefficient de corrélation, on peut être tenté de dire que l'expérience 2 est meilleure (puisque son coefficient est plus proche de 1), ce qui est évidemment faux.

Expérience 1

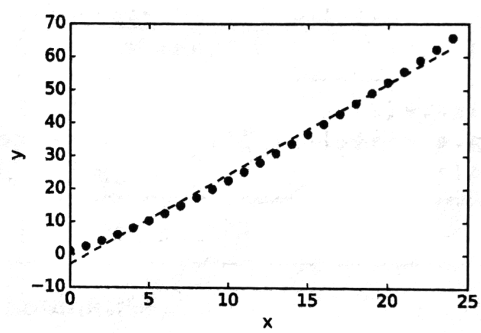


r = 0.99256481196

a = 9.98e-01 +/- 4.30e-02 à 99%

b = 1.24e+00 +/- 5.61e-01 à 99%

Expérience 2



r = 0.996359342223

a = 2.74e+00 +/- 2.21e-01 à 99%

b = -3.00e+00 +/- 3.09e+00 à 99%

En revanche, l'expérience 1 présente une incertitude (pour une confiance à 99 %) relative de 4,3 % sur a et environ 50 % sur b, alors que, l'expérience 2 indique une incertitude relative de environ 10 % sur a et supérieure à 100 % sur b.
Au regard des incertitudes sur les estimateurs c'est bien l'expérience 1 qui semble la plus fidèle.

En conclusion, il faut garder un esprit critique sur ses résultats et analyser l'ensemble de ses analyses statistiques : coefficient de corrélation mais surtout incertitudes sur les estimateurs.

On donne le script Python :

```
import numpy as np
import matplotlib.pyplot as plt
import random as rd
import scipy.stats as ss
def reglin(x,y):
    N = len(x)
```

```
    r = np.corrcoef(x,y)[0][1]  # coeff de corrélation
    [a,b] = np.polyfit(x,y,1) # estimateurs
    yth = (a*x+b)
    erreur = y-yth
    s = np.sum(erreur**2)/N
    student = ss.t.interval(.99,N)[1]
    Da = student*s*(1/np.sqrt(N*np.var(x)))
    Db = student*s*np.sqrt(1/N+np.mean(x)**2/(N*np.var(x)))

    print('r  =   '+str(r))
    print('a = {:.2e} +/- {:.2e} à 99%'.format(a,Da))
    print('b = {:.2e} +/- {:.2e} à 99%'.format(b,Db))
    plt.xlabel('x',fontsize=14)
    plt.ylabel('y',fontsize=14)
    plt.plot(x,y,'ok')
    plt.plot(x,a*x+b,'--k',linewidth = 1)

N = 25
x = np.array([i for i in range(N)])

plt.figure(1)
y = np.array([1+x[i]+(rd.random()-.5)*3 for i in range(N)])
reglin(x,y)

plt.figure(2)
y = np.array([1+x[i] for i in range(N)])**1.3
reglin(x,y)
```