Traitement statistique des données pour le TIPE

1 Quelques rappels de statistiques

Définition 1.

Le **caractère** (ou la variable statistique) x désigne la propriété étudiée.
Une **série statistique** est l'ensemble des résultats d'une étude. On présente la série statistique essentiellement sous deux formes :
- En "vrac" : On liste toutes les valeurs obtenues pour le caractère $S = (x_1, ..., x_n)$
- Par "paquets" : On liste les valeurs du caractère et effectifs correspondants $S = ((a_1, n_1), ..., (a_r, n_r))$ avec $n = \sum_{k=1}^{r} n_k$.

Le terme x_i de la série statistique est la valeur prise par le caractère x pour l'individu numéro i de la population.
Les **modalités** $a_1, .. a_r$ sont les différentes valeurs prises par le caractère x et r est le nombre de modalités.
L'**effectif** n_j d'une modalité a_j représente le nombre d'individus de la population dont le caractère vaut a_j (i.e le nombre de fois que a_j apparaît parmi les x_i).
La **fréquence** f_j d'une modalité a_j est le quotient $f_j = \frac{n_j}{n}$ (i.e c'est la proportion de la population dont le caractère vaut a_j).

Définition 2.

On considère un caractère quantitatif x dont les modalités sont $a_1, ..., a_r$.
On considère alors la série statistique sur n individus : $((a_1, n_1), ..., (a_r, n_r))$.
On appelle **moyenne** de la série :

$$\bar{x} = \frac{x_1 + .. + x_n}{n} = \frac{1}{n} \sum_{k=1}^{r} n_k a_k$$

Définition 3.

La **variance** est la moyenne des carrés des écarts à la moyenne.

$$s_x^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \bar{x})^2 = \overline{x^2} - \bar{x}^2$$

où $x^2 = (x_1^2, ..., x_n^2)$
L'**écart type** est la racine carrée de la variance. $s_x = \sqrt{s_x^2}$.

Remarque 1 *La variance est aussi souvent notée σ_x^2 et l'écart-type σ_x.*

2 Des barres d'erreurs pour quoi faire ?

2.1 Pour mesurer l'imprécision expérimentale (incertitudes de type B)

C'est souvent ce que représentent les barres en physique.

2.2 Pour estimer une moyenne à partir d'un échantillon (incertitude de type A)

On cherche à estimer une moyenne sur une population à l'aide de la moyenne sur un échantillon. C'est souvent ce que vous cherchez à faire à l'issue de vos TIPE.

3 Quelles possibilités pour les barres d'erreurs ?

Les différentes options courantes concernant les barres d'erreurs sont les suivantes.

3.1 Pour mettre en avant la dispersion de la population étudiée

3.1.1 Maximum et minimum

Mettre pour extrémités des barres d'erreurs le minimum et le maximum des valeurs obtenues.

$$I = \left[\min_i x_i \ , \ \max_i x_i\right]$$

3.1.2 Ecart-type (ou déviation standard)

Mettre pour extrémités des barres d'erreurs la moyenne moins l'écart-type et la moyenne plus l'écart-type.

$$I = [\bar{x} - s_x, \bar{x} + s_x]$$

3.2 Pour mettre en avant la "confiance" que l'on peut avoir dans la moyenne calculée

3.2.1 Erreur standard de la moyenne

Mettre pour extrémités des barres d'erreurs la moyenne moins l'écart-type divisé par la racine carrée de la taille de l'échantillon et la moyenne plus l'écart-type divisé par la racine carrée de la taille de l'échantillon.

$$I = \left[\bar{x} - \frac{s_x}{\sqrt{n}}, \bar{x} + \frac{s_x}{\sqrt{n}}\right]$$

Vous pouvez ici également utiliser l'écart-type corrigé $s'_x = \sqrt{\frac{n}{n-1}} s_x$ à la place de l'écart-type. Faites attention, c'est souvent celui-ci que votre calculatrice vous donne.

Quel est son intérêt ?

Notons σ^2 la variance réelle du caractère dans la population. La variance s_x^2 de l'échantillon la sous-estime en moyenne d'un facteur $\frac{n-1}{n}$. En utilisant la variance corrigée $s_x'^2 = \frac{n}{n-1} s_x^2$, ce phénomène est corrigé en moyenne.

3.2.2 Intervalle de confiance de la moyenne

On travaille sur un échantillon statistique $x = (x_1, ..., x_n)$. On a $\overline{x} = \frac{x_1 + ... + x_n}{n}$ la moyenne de cette série. On cherche un intervalle centré en $\overline{x}$ qui a une grande probabilité de contenir la "vraie" moyenne (la moyenne sur l'ensemble de la population et pas simplement sur un échantillon).

Un intervalle de confiance prend donc la forme $[\overline{x} - a, \overline{x} + a]$ où il faut adapter a en fonction de la précision

attendue, de la "variabilité" de la série statistique et de la taille n de l'échantillon statistique.
Plus précisément, on donne un intervalle de confiance sous la forme

$$I = \left[\overline{x} - t\frac{s_x}{\sqrt{n}}, \overline{x} + t\frac{s_x}{\sqrt{n}}\right]$$

Ici on a donc pris en compte la "variabilité" avec l'écart type σ_x et la taille n de l'échantillon . Il reste à choisir t selon la précision attendue.
Il existe différentes méthodes pour déterminer ce t.

1. Le choix de $t = 1$ donne le cas de l'erreur standard à la moyenne. (Il correspond à un intervalle de confiance à environ 68%).
2. Le choix de $t = 2$ (plus précisément 1.96) est le choix classique (sur des échantillons suffisamment grands $n > 30$) pour avoir une précision à 95% (intervalle de confiance à environ 95%)
   On obtient dans ce cas :
   $$I = \left[\bar{x} - 2\frac{s_x}{\sqrt{n}}, \bar{x} + 2\frac{s_x}{\sqrt{n}}\right]$$
3. Le choix de $t = 3$ (sur des échantillons suffisamment grand $n \geq 30$) donne un intervalle de confiance à environ 99.7%.
4. Si les échantillons sont petits ($n \leq 30$), il faut "corriger" les valeurs de t ci-dessus pour prendre en compte de façon pertinente la petite taille de l'effectif (qui est nécessairement moins représentatif). Voir l'exemple ci-dessous.

Remarque 2 *Le sens d'intervalle à* 95% *doit être compris de la façon suivante : il y a* 95% *de chances que l'intervalle déterminé contienne la valeur moyenne.*

3.2.3 Intervalle de confiance de petits échantillons sur un exemple

Dans la très grande majorité des cas, les barres d'erreurs que vous devez placer représentent l'intervalle de confiance **sur une moyenne** ou **sur une proportion** (voir paragraphe suivant).
Dans le cas des barres d'erreur sur la moyenne, si votre échantillon est de petite taille ($n \leq 30$) et si le phénomène que vous considérez suit une distribution gaussienne, l'intervalle de confiance est donnée à l'aide de la **loi de Student**. Celle-ci permet de donner la probabilité que l'intervalle que vous obtenez contienne bien la moyenne "réelle", avec une précision que vous choisissez (90%, 95%, 99%, etc.).
Pour vérifier l'adéquation de vos données expérimentales avec une distribution gaussienne vous pouvez utiliser l'un des nombreux tests de normalité existants. Vous trouverez en annexe une méthode graphique basée sur le diagramme de Henry.
Par exemple, on souhaite mesurer l'angle que fait une goutte d'eau sur une feuille de capucine. On ne peut pas mesurer une infinité de gouttes d'eau ! On réalise donc des mesures sur 30 gouttes d'eau. L'ensemble des valeurs est donné dans le tableau ci-dessous.

135	132	127	133	125	130	131	127	132	133	135	130	125	137	129
120	152	130	139	135	137	134	133	134	131	130	125	136	123	135

Voyons les étapes pour obtenir un intervalle de confiance sur l'angle moyen "réel".

- On a 30 gouttes d'eau donc $n = 30$.
- On valide à l'aide du diagramme de Henry l'hypothèse d'une distribution normale (cf. Annexe 4).
- La moyenne de ces valeurs vaut $m = 131,8$.
- L'écart-type corrigé vaut $\sigma' = 5,90$.
- L'intervalle de confiance s'exprime sous la forme $[m - t\frac{\sigma'}{\sqrt{n}}, m + t\frac{\sigma'}{\sqrt{n}}]$ où t est à déterminer à l'aide de la loi de Student suivant le niveau de confiance souhaité.

Comment déterminer t ?

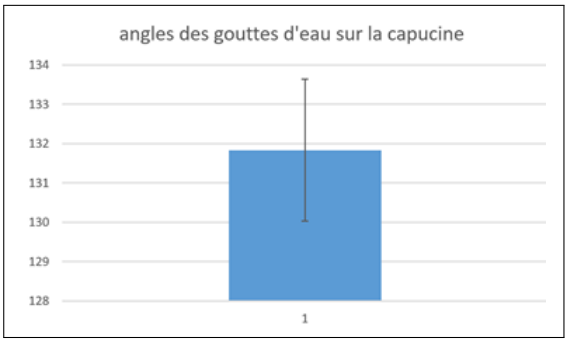
- On fixe tout d'abord le niveau de confiance souhaité, de la forme $1-\alpha$ (α correspond au niveau de risque). Plaçons-nous par exemple au niveau de confiance de 90%. On a alors $1-\alpha = 0,9$ soit $\alpha = 0,1$.
- On cherche alors le réel t positif tel que $\mathbb{P}(-t < T < t) = 1-\alpha$ où T est une variable aléatoire suivant la loi de Student à $k = n-1$ degrés de liberté. Cela revient, en utilisant les propriétés de symétrie de la loi de Student, à chercher t tel que $\mathbb{P}(T < t) = 1 - \frac{\alpha}{2}$.
  Dans notre exemple, on doit donc déterminer le réel t qui vérifie $\mathbb{P}(T < t) = 1-0,1/2 = 0,95$ avec T qui suit la loi de Student à 30-1 = 29 degrés de liberté.
- Il reste ne reste plus qu'à lire la table de la loi de Student (cf. Annexe 3) pour obtenir la valeur de t. On l'obtient en lisant la valeur à l'intersection de la ligne correspondant à $k = 29$ et de la colonne correspondant à $1-\alpha/2 = 95\%$, soit $t = 1,699$

Il ne reste plus qu'à calculer les bornes de l'intervalle de confiance :

- $m - t\frac{\sigma'}{\sqrt{n}} = 131,8 - 1,699 \times 5,90/\sqrt{30} = 130,0$
- $m + t\frac{\sigma'}{\sqrt{n}} = 131,8 + 1,699 \times 5,90/\sqrt{30} = 133,6$

On a donc 90% de chance que l'intervalle $[130,0; 133,6]$ contienne bien l'angle moyen fait par toutes les gouttes d'eau de l'Univers sur les feuilles de capucine !

Quand vous représentez votre résultat, veillez bien à préciser ce que représente votre barre d'erreur en mentionnant la précision (ici, 90%) et la méthode employée.
Exemple de présentation :



La barre d'erreur représente l'intervalle de confiance de la moyenne à 90%. Elle a été obtenue à l'aide de la loi de Student.

Attention!
NE JAMAIS UTILISER LES BARRES D'ERREURS AUTOMATIQUE D'EXCEL.
Vous ne savez pas du tout quel traitement statistique est appliqué derrière ! La plupart du temps, elles sont complètement fausses... et ça se voit. **Deux cas de figure doivent vous alerter :**

- **Les barres d'erreur sont les mêmes pour toutes les moyennes :** alors ça, quelle coïncidence troublante ! Etes-vous sûr d'avoir obtenu pile poil le même écart-type pour toutes vos conditions ?
- **Les barres d'erreur sont proportionnelles aux moyennes :** pourquoi les petits angles varieraient-ils moins que les grands ?

Le jury sait que ces deux erreurs sont "classiques" et va donc les chercher prioritairement dans les rapports de TIPE.

Exercice 1 (A vous de jouer !) On a mesuré aussi l'angle fait par des gouttes d'eau sur une peinture hydrophobe. Cette fois, seules 20 mesures ont été réalisées. L'ensemble des valeurs est donné par le tableau ci-dessous.

124	129	115	132	129	130	124	125	126	135
131	129	136	135	121	129	127	130	123	135

Calculer la moyenne et l'écart-type corrigé de cette série statistique puis déterminer l'intervalle de confiance à 90% en utilisant la loi de Student. La différence avec la moyenne obtenue pour la feuille de capucine est-elle "statistiquement significative" ici ?

Remarque 3 *Vous réalisez ici que la notion de "statistiquement significative" est* ***relative*** *puisqu'elle dépend de la précision choisie pour l'intervalle de confiance. Deux intervalles de confiance peuvent ne pas se recouper lorsqu'ils sont déterminés avec une précision de 90%, mais se recouper lorsqu'ils sont déterminés avec une précision de 99%. En effet, pour une série statistique donnée, la longueur de l'intervalle de confiance augmente avec le niveau de confiance.*

3.2.4 Intervalle de confiance sur une proportion

Nous avons à notre disposition un échantillon aléatoire simple de n individus et nous souhaitons inférer à partir de ses seules valeurs la **proportion** p d'un caractère donné **au sein de la population dont il est issu**. Les exemples sont nombreux : proportion d'une essence donnée dans une forêt, proportion de grains d'un diamètre inférieur à 2 mm dans un sable, proportion de micas dans un granite, etc.

Notons f la proportion du caractère dans l'échantillon. Si l'échantillon est bien aléatoire et **vérifie** $nf \geq 5$ **et** $n(1-f) \geq 5$ **ou** $n \geq 30$, **l'intervalle de confiance de** p **à 95%** est donné par

$$\left[f - 1,96\sqrt{\frac{f(1-f)}{n}}; f + 1,96\sqrt{\frac{f(1-f)}{n}} \right].$$

Cela signifie que cet intervalle a 95% de chance de contenir la proportion p.
Vous verrez dans le cours de mathématiques de 2ème année d'où provient cette expression de l'intervalle de confiance.

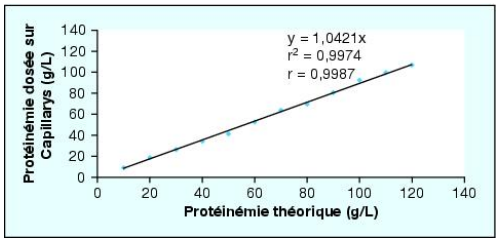
Exemple 1 *Disposant de la proportion f de filles en BCPST2 au lycée Lakanal, on cherche à estimer la proportion p de filles inscrites au concours Agro au niveau national. En 2016, il y avait 62 filles sur les 87 étudiants des deux classes de deuxième année, soit $f = \frac{62}{87} = 0,713$.*
Comme $n \geq 30$, on peut utiliser l'intervalle de confiance donné plus haut.
On a $1,96\sqrt{\frac{f(1-f)}{n}} = 1,96\sqrt{\frac{0,713(1-0,713)}{87}} = 0,095$.
L'intervalle de confiance de p au niveau 95% est donc $[0,713 - 0,095; 0,713 + 0,095] = [0,618; 0,808]$.
Il y a 95% de chances que l'intervalle de confiance contienne la proportion réelle p. Les statistiques nationales pour l'année 2016 assurent que 2189 filles étaient inscrites au concours Agro sur les 3140 candidats, soit $p = 0,697$.

Exercice 2 (A vous de jouer !) En analysant un échantillon aléatoire de 1236 grains issus d'une surface de sédimentation, on a obtenu que 12% d'entre eux avait son diamètre supérieur à 2 mm. Notons p la proportion réelle de grains de diamètre supérieur à 2 mm pour l'ensemble de la surface.
Après avoir vérifié que vous pouvez utiliser l'expression de l'intervalle de confiance sur une proportion donné ci-dessus, vous calculerez l'intervalle de confiance de p au niveau de confiance 95% obtenu à partir de l'échantillon.

4 Régression linéaire et incertitudes

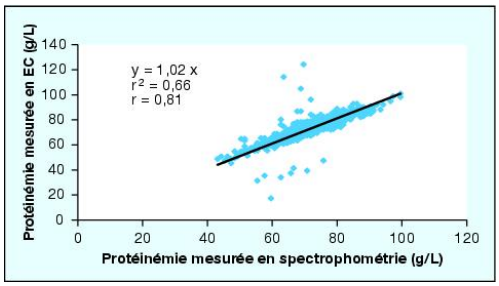
Vous serez souvent amenés à déterminer l'adéquation de données expérimentales avec une loi linéaire du type $y = ax + b$, à obtenir une estimation des paramètres a et b ainsi que la précision de cette estimation.

On suppose que l'on dispose de deux séries de n valeurs $x_1, \cdots, x_n$ et $y_1, \cdots, y_n$. **On supposera que les incertitudes sur les x_i sont négligeables devant celles sur les y_i.** On note $u(y_i)$ l'incertitude-type sur y_i.
En général, les valeurs x_i correspondent à différentes valeurs, que vous avez la possibilité de choisir, d'une grandeur physique X et les y_i aux valeurs d'une autre grandeur Y que vous observez pour la valeur x_i choisie (ou plutôt, si le travail est bien fait, à la moyenne des valeurs observées lors des différentes réalisations de l'expérience).
Par exemple, en spectrophotométrie, on se place à une certaine longueur d'onde et on mesure l'absorbance d'une solution colorée en fonction de la concentration d'une gamme étalon. D'après la loi de Beer-Lambert, ces deux grandeurs sont reliées linéairement. Une fois le coefficient multiplicatif estimé, on pourra mesurer l'absorbance d'une solution mystère et en déduire sa concentration.
Voici ci-dessous deux exemples d'utilisation de régression linéaire issus d'un article de recherche.



Comparaison des valeurs de protéinémies mesurées sur Capillarys 2 et celles théoriques attendues, sur des dilutions successives du pool de sérum normal (PSN) avec une solution aqueuse de NaCl à 0,15 mol/L.

Ce premier exemple montre une régression linéaire sur des points "isolés". Le suivant concerne un nuage de points assez dense.



Comparaison des valeurs de protéinémies trouvées en électrophorèse capillaire (EC) sur le Capillarys 2 (Sebia®) et en spectrophotométrie sur Modular (Roche®) pour les 859 sérums sélectionnés.

4.1 Cas où les y_i ont la même incertitude-type

Si toutes les mesures y_i présentent la même incertitude-type, notée $u(y)$, on utilise la méthode des moindres carrés pour déterminer la droite qui approche le mieux les points $(x_1, y_1), \cdots, (x_n, y_n)$. Pour cela, on cherche les valeurs de a et b qui minimisent la quantité $\sum_{i=1}^{n}(y_i - (ax_i + b))^2$. On détermine analytiquement l'unique solution de ce problème :

$$a = \frac{s_{xy}}{s_x^2} \quad \text{et} \quad b = \overline{y} - \frac{s_{xy}}{s_x^2}\overline{x}$$

où $s_{xy} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} y_j\right) = \overline{xy} - \overline{x}\,\overline{y}$ désigne la covariance de la série double $(x_1, y_1), \cdots, (x_n, y_n)$.

Les incertitudes-types sur les coefficients a et b sont alors données par les formules suivantes :

$$u(a) = \frac{u(y)}{s_x\sqrt{n}} \quad \text{et} \quad u(b) = u(y)\frac{\sqrt{\overline{x^2}}}{s_x\sqrt{n}}$$

Remarque 4 *Si l'incertitude $u(y)$ n'est pas connue, on peut la remplacer par $u_{stat}(y)$, qui est l'estimation statistique de $u(y)$:*

$$u_{stat}(y) = \sqrt{\frac{1}{n-2}}\sum_{i=1}^{n}(y_i - (ax_i + b))^2$$

Remarque 5 *Si vous n'êtes pas intéressés par les incertitudes sur les coefficients a et b ou si les incertitudes sur les y_i sont du même ordre, vous pouvez vous contenter de la régression linéaire simple décrite ci-dessus et vous pouvez lire directement la section **Interprétation**. Sinon, ce qui suit est pour vous.*

4.2 Pour aller plus loin : cas où les y_i ont des incertitudes-types différentes

On utilise une méthode des moindres carrées pondérée de façon à privilégier les mesures présentant les incertitudes-types les plus faibles. On cherche alors les valeurs de a et b qui minimisent la quantité $\sum_{i=1}^{n} w_i(y_i - (ax_i + b))^2$, avec $w_i = 1/u(y_i)^2$. On trouve comme valeurs de a et b :

$$a = \frac{\overline{xy}^w - \overline{x}^w\overline{y}^w}{(s_x^w)^2} \quad \text{et} \quad b = \overline{y}^w - \frac{s_{xy}^w}{(s_x^w)^2}\overline{x}^w$$

avec

$$\overline{xy}^w = \frac{\sum_{i=1}^{n} w_i x_i y_i}{\sum_{i=1}^{n} w_i}, \quad \overline{x}^w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \overline{y}^w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \quad \overline{x^2}^w = \frac{\sum_{i=1}^{n} w_i x_i^2}{\sum_{i=1}^{n} w_i}$$

$$(s_x^w)^2 = \overline{x^2}^w - (\overline{x}^w)^2, \quad (s_{xy}^w)^2 = \overline{xy}^w - \overline{x}^w\overline{y}^w$$

Les incertitudes-types sur les coefficients a et b sont alors données par les formules suivantes :

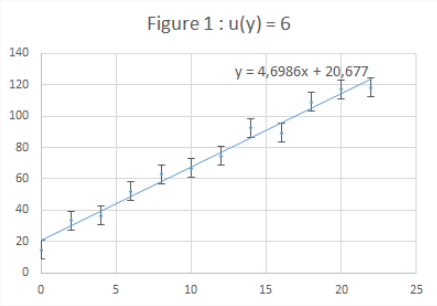
$$u(a) = \frac{1}{s_x^w\sqrt{\sum_{i=1}^{n} w_i}} \quad \text{et} \quad u(b) = \frac{\sqrt{\overline{x^2}^w}}{s_x^w\sqrt{\sum_{i=1}^{n} w_i}}$$

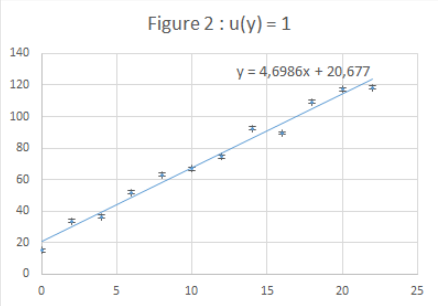
4.3 Interprétation

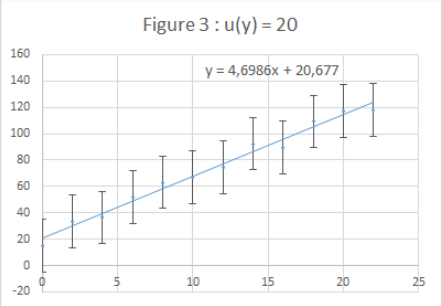
Le coefficient de corrélation linéaire $r_{xy} = s_{xy}/s_x s_y$ (noté r en l'absence d'ambiguité) donne une indication sur la qualité de la régression mais **ne suffit pas** à valider ou invalider un modèle linéaire. On pourra juger graphiquement de la pertinence du modèle en représentant sur un même graphe le nuage de points, la droite de régression linéaire ainsi que les barres d'erreurs pour chaque mesure y_i, correspondant à l'intervalle $[y_i - u(y_i), y_i + u(y_i)]$.
Il convient d'éliminer les points aberrants ou de refaire les mesures avant de réaliser la régression linéaire.

Nous allons réaliser à trois reprises une régression linéaire simple sur les séries de données ci-dessous à l'aide du logiciel Excel. A chaque fois, nous imposons une incertitude-type $u(y)$ constante sur la deuxième série de données et successivement égale à 6 (Fig. 1), à 1 (Fig. 2) et à 20 (Fig. 3).

x	0	2	4	6	8	10	12	14	16	18	20	22
y	14,79	33,52	36,50	51,88	63,11	66,94	74,58	92,46	89,50	109,29	117,40	118,37







Sur la figure 1, l'écart entre les points expérimentaux et la droite est du même ordre de grandeur que les incertitudes. On peut valider le modèle linéaire.
Sur la figure 2, les incertitudes sont petites par rapport aux écarts à la droite. La loi linéaire n'est pas validée. Soit la loi n'est pas linéaire, soit les incertitudes ont été sous-estimées.
Sur la figure 3, les incertitudes sont grandes par rapport aux écarts des points à la droite. Le modèle linéaire n'est pas a priori rejeté mais les imprécisions sont grandes : de nombreuses droites peuvent intercepter l'ensemble des barres d'erreur.

Remarque 6 *Vous remarquerez que dans les trois cas, l'équation de la droite des moindres carrés est la même. Les incertitudes sur la série y n'interviennent pas* a priori. *Ce n'est en réalité pas le cas dès lors que l'on précise l'incertitude sur les coefficients a et b (voir ci-dessous).*

	A	B	C
1	x	y	x^2
2	0	14,79	0
3	2	33,52	4
4	4	36,5	16
5	6	51,88	36
6	8	63,11	64
7	10	66,94	100
8	12	74,58	144
9	14	92,46	196
10	16	89,5	256
11	18	109,29	324
12	20	117,4	400
13	22	118,37	484
14			
15	Ecart-type série x		Moment d'ordre 2 empirique de la série x
16	6,90410506		168,6666667
17			
18	u(y)	u(a)	u(b)
19	1	0,0418121	0,543020989
20	6	0,2508726	3,258125936
21	10	0,41812101	5,430209893

Voici une copie écran du logiciel Excel montrant le calcul de $u(a)$ et $u(b)$ pour les différentes valeurs de $u(y)$. Voici les différentes étapes suivies.

1. *On calcule en colonne C (cellules C2 à C13) les carrés des valeurs de la série x.*
2. *On calcule l'écart-type s_x de la série x en cellule C16. Pour cela on utilise la commande* `=ECARTYPE.PEARSON(A2:A13)`. ***Attention****, la fonction* `=ECARTYPE.STANDARD(A2:A13)` *renvoie la valeur $\sqrt{n/(n-1)}s_x$.*
3. *On calcule $\overline{x^2}$ la moyenne de la série des carrés en cellule C16 à l'aide de l'instruction* `=MOYENNE(C2:C13)`.
4. *On calcule $u(a)$ et $u(b)$ en utilisant les formules vues plus haut. Par exemple, pour $u(y) = 1$, on calcule $u(a)$ en cellule B19 à l'aide de l'instruction* `=A19/(A$16*RACINE(12))` *et $u(b)$ en cellule C19 à l'aide de l'instruction* `=A19*RACINE(C$16)/(A$16*RACINE(12))`. *La valeur 12 correspond à la taille n de l'échantillon.*

5 Conclusion

La connaissance de ces méthodes d'obtention des barres d'erreur doit orienter le choix de vos expériences ou mesures de TIPE. En effet, les expériences en biologie et géologie nécessitent un traitement statistique pour avoir une quelconque valeur scientifique. Il est donc vivement conseillé d'avoir en tête la ligne directrice suivante : "je mesure tel paramètre ou telle proportion sur un grand nombre d'échantillons, afin d'estimer un intervalle de confiance". **Une expérience sans quantification validée statistiquement ne vaut pas grand chose.**

Voici en résumé ce qu'il faut retenir :

Notations

	proportion	moyenne	variance	écart-type	variance corrigée	écart-type corrigé
population	p	m	σ^2	σ		
échantillon	f_n	$\bar{x}$	s^2	s	s'^2	s'

Intervalle de confiance

paramètre	conditions	intervalle de confiance à 95%
proportion	$nf_n \geq 5$ et $n(1-f_n) \geq 5$ ou $n \geq 30$	$\left[f - 1,96\sqrt{\frac{f(1-f)}{n}}; f + 1,96\sqrt{\frac{f(1-f)}{n}}\right]$
moyenne	$n \geq 30$	$\left[\bar{x} - 1,96\frac{s_x}{\sqrt{n}}, \bar{x} + 1,96\frac{s_x}{\sqrt{n}}\right]$
moyenne	$n \leq 30$, ditribution normale	$\left[\bar{x} - t\frac{s'_x}{\sqrt{n}}, \bar{x} + t\frac{s'_x}{\sqrt{n}}\right]$ voir section 3.2.4 pour la détermination de t

Distribution choisie

Petit échantillon ($n \leq 30$) : loi de Student
Grand échantillon ($n > 30$) : loi normale

Au final, quelles barres d'erreur choisir ?

Cela dépend de l'objectif que vous poursuivez et de la série statistique dont vous disposez.

• Si vous souhaitez montrer une différence entre les moyennes de deux échantillons, mieux vaut opter pour l'erreur standard de la moyenne ou l'intervalle de confiance à 95%. Mieux vaut alors disposer d'un grand nombre de mesures :

- Erreur standard de la moyenne : voir section 3.2.1
- Intervalle de confiance à 95% pour un grand nombre de mesures : voir section 3.2.2
- Cas particulier d'une proportion : voir section 3.2.4
- Intervalle de confiance à 95% pour un petit nombre de mesures dans le cas d'une distribution normale : voir section 3.2.3

Vous pourrez alors conclure, sans faire une trop grosse erreur, que si les barres d'erreur ne se chevauchent pas, alors les moyennes sont significativement différentes.
• Si vous voulez mettre en évidence une grande variabilité à l'intérieur de chaque échantillon, des barres d'erreur correspondant au minimum-maximum (voir section 3.1.1) ou à l'écart-type (voir section 3.1.2) feront l'affaire.

Annexe 1 - Barres d'erreurs sous Excel

Le plus simple probablement pour faire des barres d'erreur correctes sous Excel c'est de rentrer soi-même la précision voulue.

Si on reprend l'exemple ci-dessus, on a obtenu une moyenne à 131.3, ce qu'on peut représenter par exemple par un diagramme en barre de hauteur 131.3.

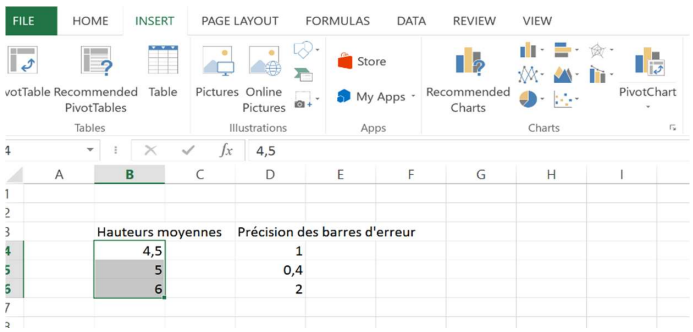
On choisit alors l'option barre d'erreur d'Excel et dans la partie marge d'erreur, on choisit précision et on rentre le résultat du calcul $t\frac{s_x}{\sqrt{n}}$ (dans le cas présent ce calcul donne $2.045 \times \frac{6.11}{\sqrt{30}} \simeq 2.28$).

Surface	Feuille capucine fraîche		Feuille capucine sèche		Tuile + peinture		Bois + peinture		Tuile sans peinture*		Bois sans peinture*		
													*: surfaces témoins
Volume	50µl	100µl	50µl	100µl	50µl	100µl	50µl	100µl	50µl	100µl	50µl	100µl	
	135	117	124	130	118	118	123	133	38	41	41	45	
	132	118	129	115	129	125	131	129	42	52	39	52	
	127	121	115	132	126	136	127	125	36	45	61	55	
	138	128	132	130	131	130	126	127	35	48	55	47	
	125	117	129	123	128	129	135	119	39	44	38	46	
	130	125	130	117	134	132	121	130					
	131	118	124	129	130	125	127	123					
	127	112	125	120	135	134	134	128	Moyenne :	Moyenne :	Moyenne :	Moyenne :	
	132	113	126	131	122	126	123	129	38	46	46.8	49	
	133	122	135	126	133	135	130	122					
	135	135	131	118	124	128	125	129	Ecart-type :	Ecart-type :	Ecart-type :	Ecart-type :	
	132	132	129	123	135	127	127	126	2.738612788	4.183300133	10.49761878	4.301162634	
	125	128	136	127	126	141	132	135					
	137	119	135	129	131	129	135	132					
	129	135	127	121	125	130	133	127					
	120	125			126	128	134	133					
	152	122			133	126	135	128					
	130	120	Moyenne :	Moyenne :	124	130	124	125					
	136	126	128.4666667	124.7333333	135	123	129	130					
	135	124			129	130	127	126					
	123	119	Ecart-type :	Ecart-type :									
	126	127	5.38339599	5.560918047									
	136	128			Moyenne :	Moyenne :	Moyenne :	Moyenne :					
	125	124			128.7	129.1	128.9	127.8					
	130	130											
	131	126			Ecart-type :	Ecart-type :	Ecart-type :	Ecart-type :					
	124	120			4.791439735	5.025199656	4.529435889	3.941680112					
	133	127											
	134	119											
	137	123											
	Moyenne :	Moyenne :											
	131.3333333	123.3333333											

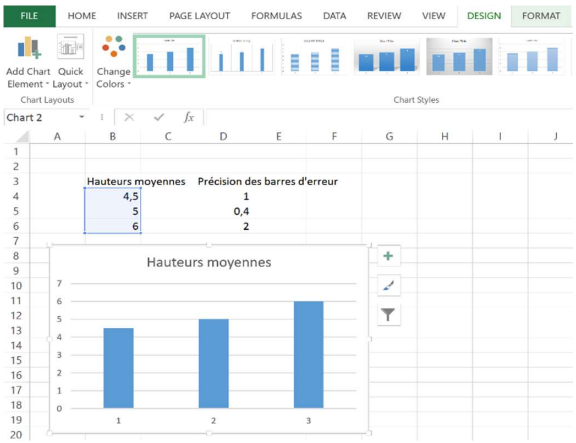
Annexe 2

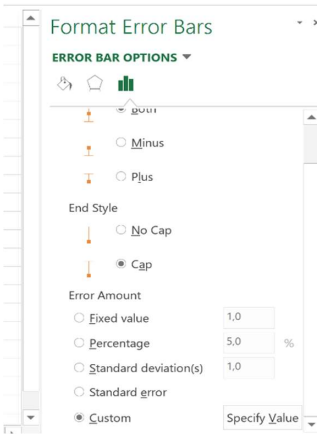
Modalités pratiques : comment entrer ses barres d'erreur sur Excel 2013 (version anglaise ici)

1) Vous entrez dans deux colonnes différentes vos moyennes et les valeurs que vous avez envie d'imposer aux barres d'erreur. Ne surtout pas choisir les barres d'erreur automatiques. Vous devez savoir à quoi correspondent vos barres !
2) Vous sélectionnez vos valeurs de moyennes. Vous cliquez sur « insert » (insertion) puis « chart » (graphique) puis le type de graphique que vous voulez : ex ici un histogramme.

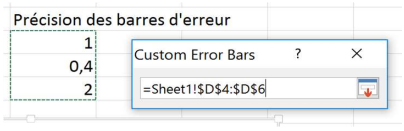


3) L'histogramme est maintenant là. Vous voulez lui ajouter les barres d'erreur. Allez dans l'onglet « design », « add chart element » (tout à gauche), « error bars », « more error bar options ». Une fenêtre "format error bars" apparaît sur le côté gauche.

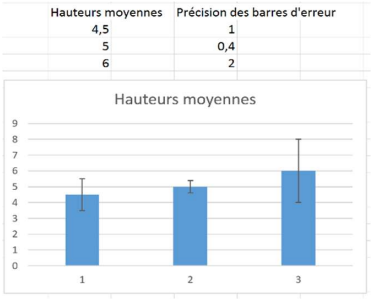




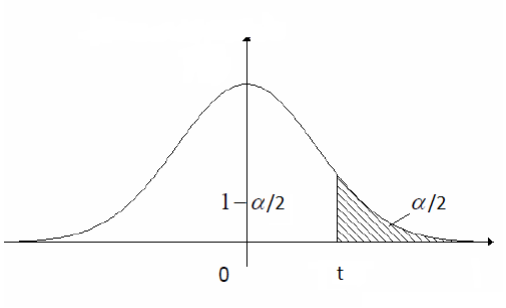
4) Cliquez sur "Custom" (tout en bas) puis "Specify value". Il vous demande alors quelles valeurs vous voulez mettre de part et d'autre de votre moyenne. Vous devez, pour chacune (« positive value » et « negative value »), cliquer sur la flèche rouge et sélectionner les cases correspondant à vos valeurs de barres d'erreur.



5) Les barres d'erreur sont donc personnalisées pour chaque moyenne.



Annexe 3 - Loi de Student



Si T suit une loi de Student à k degrés de liberté, la table donne le réel t tel que $\mathbb{P}(T \leq t) = 1 - \frac{\alpha}{2}$.

1-$\alpha/2$ / k	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.445	0.703	0.883	1.230	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.443	0.697	0.876	1.214	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.442	0.695	0.873	1.209	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.441	0.694	0.870	1.204	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.440	0.692	0.868	1.200	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.439	0.690	0.865	1.194	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.438	0.689	0.863	1.191	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.438	0.688	0.862	1.189	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.438	0.688	0.861	1.187	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.437	0.687	0.860	1.185	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.437	0.686	0.859	1.183	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.437	0.686	0.858	1.182	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.436	0.685	0.858	1.180	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.436	0.685	0.857	1.179	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.436	0.684	0.856	1.177	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.435	0.684	0.855	1.176	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385
35	0.255	0.434	0.682	0.852	1.170	1.306	1.690	2.030	2.438	2.724	3.340
40	0.255	0.434	0.681	0.851	1.167	1.303	1.684	2.021	2.423	2.704	3.307

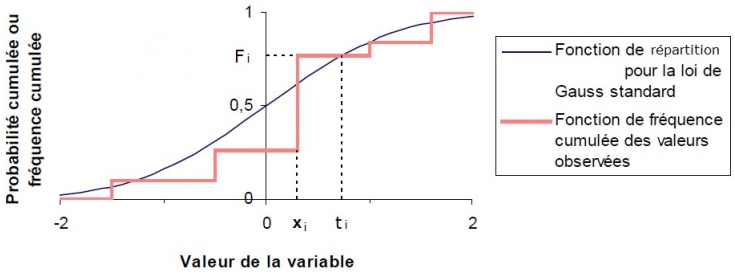
Annexe 4 - Diagramme de Henry

Considérons n réalisations $x_1, \cdots, x_n$ d'une variable aléatoire X. Connaissant $\mathbb{P}(X \leq x_i)$, on peut trouver le réel t_i tel que $\phi(t_i) = \mathbb{P}(X \leq x_i)$, où ϕ est la fonction de répartition de la loi normale centrée réduite. Si X est une variable gaussienne de moyenne m et de variance σ^2, les points de coordonnées (x_i, t_i) appartiennent à la droite d'équation $t = \dfrac{x-m}{\sigma}$. En effet, la variable centrée réduite associée à X, $X^* = \dfrac{X-m}{\sigma}$ suit la loi normale centrée réduite et

$$\mathbb{P}(X \leq x_i) = \mathbb{P}(X^* \leq \frac{x_i - m}{\sigma}) = \phi\left(\frac{x_i - m}{\sigma}\right)$$

La fonction ϕ réalisant une bijection de $\mathbb{R}$ dans $]0,1[$, on a bien $t_i = \dfrac{x_i - m}{\sigma}$.
Cette constatation va être le fondement du test de normalité présenté ici, qui va consister à afficher le nuage de points (x_i, t_i) et à vérifier de visu son alignement.
Dans la pratique, on ne connait pas la loi de la variable X, on a seulement accès à un n échantillon de mesures, correspondant à n réalisations de la variable X, dont les valeurs peuvent se répéter.
Voilà comment construire le diagramme de Henry à partir de ces mesures :

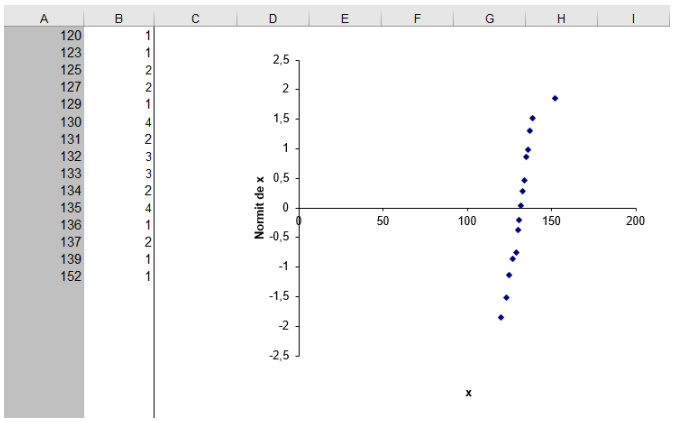
1. Pour chaque valeur x_i, on calcule la fréquence cumulée F_i, c'est-à-dire le nombre de valeurs (éventuellement non disctintes) du n échantillon de données, inférieures ou égales à x_i, divisé par n.
2. On détermine à l'aide de la table de la loi normale centrée réduite le réel t_i vérifiant $\phi(t_i) = F_i$. Autrement dit, t_i est le quantile d'ordre F_i de la loi normale centrée réduite.



3. On représente les couples (x_i, t_i) par un nuage de points.

Interprétation : Si le nuage de points obtenu est proche d'une droite, on peut considérer que X suit une loi normale.

Sur la page internet de la revue Modulad (`http://modulad.fr/excel.htm`) vous pouvez télécharger un document Excel (zippé) contenant des exemples mais aussi une feuille de calcul permettant de tester vos données. Utilisée pour les mesures de l'angle d'une goutte d'eau sur une feuille de capucine, on obtient le résultat ci-dessous.



La colonne A contient les valeurs observées rangées par ordre croissant. En colonne B est indiqué combien de fois chaque valeur est observée.
Le diagramme de Henry obtenu montre que les données recueillies sont compatibles avec une loi normale, le nuage de points étant proche d'une droite.

Référence : CHAPELAIN, Kathy et GRENIER, Emmanuel. Comment construire un diagramme de Henry avec Excel et comment l'interpréter. Revue MODULAD, 2006, vol. 1, no 35.