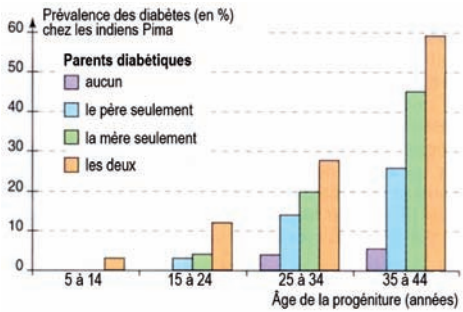
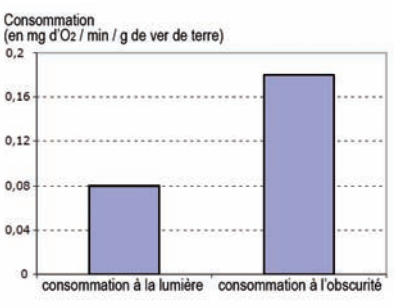
Utilité et utilisation des statistiques (de base) en biologie

Roland Douchet

Durant l'exercice de notre métier, nos élèves et nous mêmes sommes souvent confrontés à des données chiffrées issues d'un échantillonnage, qu'il s'agit d'interpréter dans le cadre d'une démarche scientifique (fig. 1). L'objectif ici est de montrer qu'une interprétation correcte ne peut se faire qu'en utilisant des statistiques.



A - Graphique extrait d'un livre de 1e S (édition Bordas)
Pour obtenir le graphique ci-dessus, toute la population des Indiens Pima n'a pas été sondée : seuls un certain nombre d'individus ont été étudiés (on dit qu'on a fait un échantillonnage de n individus) et les résultats concernant l'échantillon ont été traduits sur le graphique.



B - Graphique issu d'une expérience de TIPE
Pour obtenir ce graphique, le groupe de TIPE n'a pas étudié tous les vers de terre du monde entier, mais seulement quelques uns (échantillon de 5 vers pour les mesures à l'obscurité, de 7 vers pour les mesures à la lumière). Les résultats concernant ces 5 + 7 vers ont été traduits sur le graphique.

1. Exemples de données chiffrées issues d'un échantillonnage

Tout au long de cet article, nous montrerons, à travers des exemples concrets d'analyse de documents de lycée ou de résultats de TPE (Travaux personnels encadrés) ou TIPE (Travaux d'Initiative Personnelle Encadrés), la nécessité d'utiliser les statistiques pour pouvoir exploiter correctement des résultats. Nous présenterons de même les tests statistiques les plus souvent utilisés lors des TIPE (ils peuvent aussi être utilisés en TPE en fonction du

➤ **Mots clés** : test statistique, population, échantillon, estimation par intervalle, intervalle de confiance, barre d'erreur, écart type, comparaison de moyennes, comparaison de pourcentages, TPE, TIPE

■ **Roland Douchet** : professeur de SVT en BCPST1 à Angers, titulaire d'un master 2 de mathématiques appliquées à la biologie

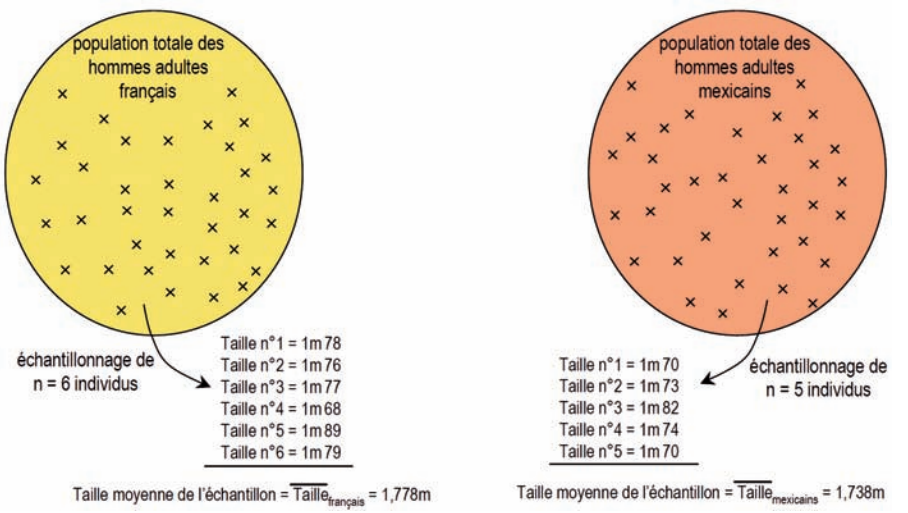
niveau des élèves et de leur curiosité mathématique).
*Il est à noter – afin de ne pas rebuter le plus grand nombre – que l'approche des statistiques choisie pour cet exposé est avant tout intuitive, et que le discours s'affranchit le plus possible du jargon des mathématiciens, sans pour autant perdre en rigueur. Les biologistes recherchant des précisions pourront lire avec profit les deux ouvrages suivants : « Statistique » (*Wonnacott and Wonnacott*) et « Biostatistique : une approche intuitive » (*Motulsky*).*

Introduction

La nécessité de l'utilisation des statistiques

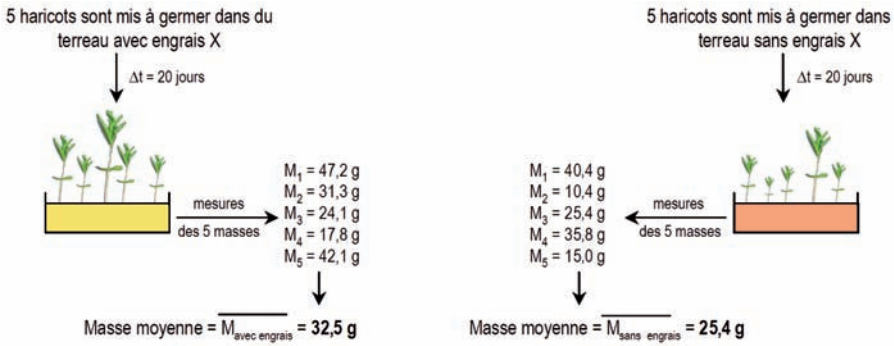
Prenons quelques exemples concrets montrant tout ce qu'on ne peut pas faire en absence de statistiques.

Exemple 1 : on cherche à déterminer si, en moyenne, les hommes français adultes sont plus grands que les hommes mexicains adultes. La figure 2 montre un protocole et son résultat associé. On peut dire que pour l'échantillon considéré, la taille moyenne des Français est effectivement supérieure à celle des Mexicains (1,778 > 1,738). Cependant, on ne peut décemment pas extrapoler ce résultat aux 2 populations (intuitivement, on sent bien que de nouveaux échantillonnages aboutiraient à d'autres moyennes. Il se pourrait même que pour certains échantillons, ce soit les Mexicains qui sont plus grands que les Français). Bref, on ne peut rien tirer d'intéressant d'un tel résultat…



2. Les hommes adultes français sont-ils en moyenne plus grands que les hommes adultes mexicains ? Exemple de protocole et résultats associés

Exemple 2 (tiré d'un TIPE de BCPST) : un groupe d'étudiants cherche à démontrer l'effet éventuel d'un engrais X sur la croissance de plants de haricot. La figure 3 montre le protocole et ses résultats associés. Ici encore, on constate que la moyenne des masses des 5 haricots avec engrais est supérieure à la moyenne des masses des 5 haricots sans engrais (32,1 > 25,4 g). Comme précédemment, on sent bien qu'en raison des fluctuations d'échantillonnage, les valeurs auraient pu être différentes si elles avaient été mesurées sur d'autres haricots. On ne peut ainsi pas du tout généraliser aux populations et dire par exemple que « l'engrais favorise la croissance des haricots » en général.



3. L'engrais X favorise-t-il la croissance (mesurée par la masse) des haricots? Exemple de protocole issu d'un TIPE de BCPST et résultats associés

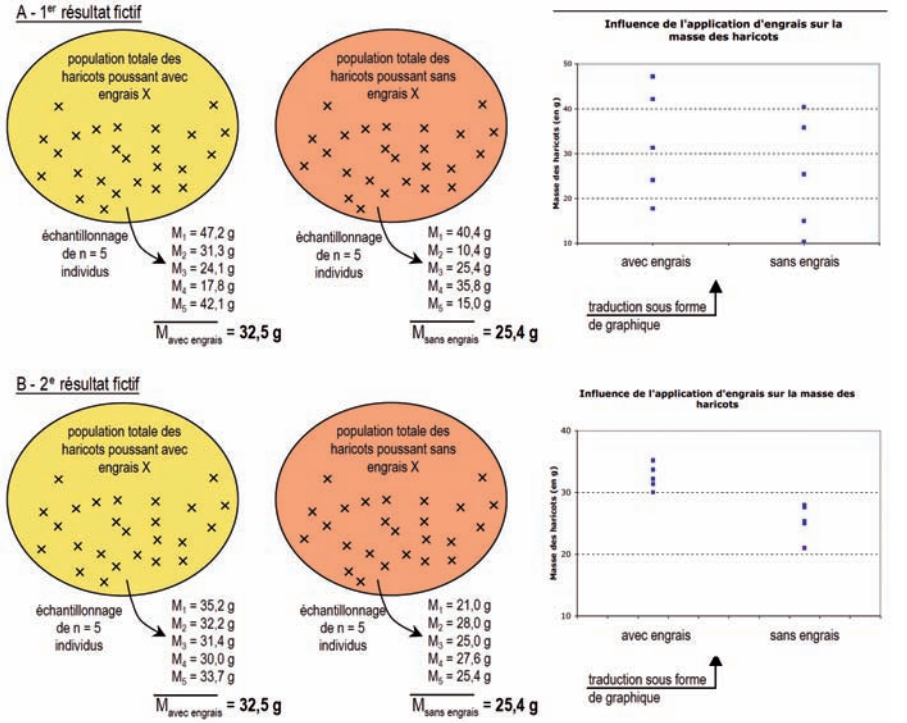
Exemples issus de la figure 1 : concernant la population des Indiens Pima, on ne peut par exemple pas dire a *priori* – pour les mêmes raisons que précédemment - que la prévalence des diabètes est plus forte quand les deux parents sont diabétiques par rapport au cas ou un seul ou aucun parent ne l'est. Concernant les vers de terre, on ne peut pas dire que la consommation en O_2 est plus forte à l'obscurité qu'à la lumière.

Se pose alors la question suivante : comment faire pour pouvoir « dire quelque chose », c'est à dire pour pouvoir **tirer des conclusions générales à partir d'un nombre limité de donnée**s (= pour **extrapoler** d'un échantillon à une population) ?

Principe de base de l'utilisation des statistiques pour la comparaison de résultats issus d'échantillonnages

Reprenons l'exemple précédent de la masse des plantules de haricot (*fig. 3 et 4A*). Intuitivement, on sent bien qu'il faut réaliser un échantillonnage de plusieurs individus (on dit « répétitions ») pour le paramètre mesuré, et calculer la moyenne résultante (ici, le nombre de répétitions est égal à 5 pour le cas avec et le cas sans engrais ; on dit n=5).

Cependant, nous avons vu que la simple comparaison des 2 moyennes ne donne aucun résultat généralisable aux populations (on peut traduire ceci de la façon sui-



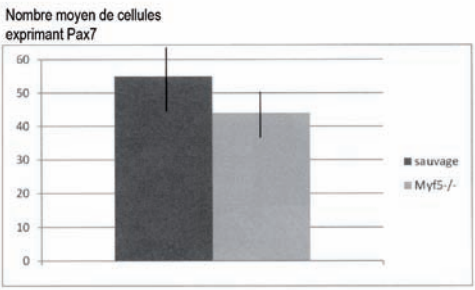
4. Présentation de deux résultats fictifs issus d'un protocole d'échantillonnage, visant à montrer qu'une simple comparaison de moyenne est une information insuffisante pour analyser des résultats.

vante : est-ce que 32,5 est vraiment statistiquement supérieur à 25,4 ?).

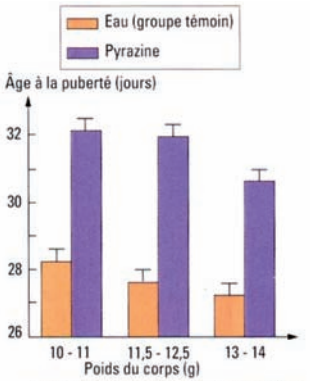
Sur la figure 4 sont représentés 2 résultats possibles pour cette expérience. Dans les 2 cas, la moyenne des haricots sans engrais est de 25,4 g, celle avec engrais de 32,5 g. Intuitivement cependant – dans le cas de la figure 4B - , on a l'impression – on a envie de dire qu'en général, les haricots poussent quand même mieux avec engrais, ce qui ne semble pas être le cas pour la figure 4A.

Ce qui est visible ici intuitivement est l'idée générale qu'il faut retenir : **l'analyse statistique se base sur des comparaisons de moyennes**. Cependant, la moyenne ne suffit pas lorsqu'on analyse et représente les résultats : il faut aussi avoir une idée de la **valeur des écarts des mesures par rapport à la moyenne**. Cette valeur est mesurée par la variance des mesures (ou sa racine carrée : l'écart type).

Sur les publications de chercheurs, les sujets de concours (CAPES, concours de BCPST…), et sur certains graphiques utilisés au lycée, l'information donnée est effectivement celle-ci : à chaque moyenne issue d'un échantillonnage de mesures est associé l'écart type des mesures, sous forme de barre centrée sur la moyenne (*fig. 5*).



On détermine par immunofluorescence le nombre moyen de cellules exprimant Pax7 par coupe de MAT de souris sauvages ou mutantes *Myf5-/-* âgées de 6 semaines. Les traits verticaux représentent l'écart-type.



Action de la pyrazine sur l'apparition de la puberté chez la souris

A - Graphique issu d'un sujet de concours (ENS 2009)
Ici, la signification des barres est clairement explicitée. Par exemple, chez les sauvages, le nombre moyen des cellules exprimant Pax7 dans l'échantillon est de 55, l'écart type est de 20 environ (la barre est centrée sur la moyenne et va de 45 (=55-20/2) à 65 (=55+20/2)).

B - Graphique issu d'un livre de TS (édition Belin)
Noter qu'ici, la signification des demi- barres n'est pas explicitée. Chaque demi-barre représente un demi écart-type (seules les demi barres supérieures sont représentées). Par rapport au cas A, elles sont de plus fermées par un trait horizontal. Le choix de ne représenter que des demi-barres et de les « fermer » est uniquement artistique.

5. Représentation de la variabilité des mesures sur les graphiques issus de documents de lycée, sujets de concours, travaux de chercheurs…

Rappel de quelques notions de mathématiques utilisées en statistiques

Finalement, pour pouvoir interpréter les résultats issus de documents ou d'expériences de TPE, TIPE, il va être nécessaire de savoir calculer les moyennes et variances (ou écarts type) de plusieurs mesures.

– Soit un paramètre X (ex : taille de Français, masse des haricots, consommation en O_2, etc…)

– Soit un échantillonnage de n mesures de ce paramètre : $\{x_1 ; x_2 ; x_3 \ldots ; x_n\}$

La moyenne de ces *n* mesures est : $\overline{X} = E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$

La variance de ces *n* mesures est : $V(X) = S^2(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))^2.$

Elle mesure l'écart à la moyenne des différentes mesures.

L'écart type de ces *n* mesures est : $S(X) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))^2}$

Application

Pour la figure 3, la moyenne de masse des haricots sans engrais est :

$$\overline{X} = \frac{1}{5}(47{,}2 + 31{,}3 + 24{,}1 + 17{,}8 + 42{,}1) = 32{,}5$$

La variance est :

$$Var(X) = \frac{1}{5}\left((47{,}2 - 32{,}5)^2 + (31{,}3 - 32{,}5)^2 + (24{,}1 - 32{,}5)^2 + (17{,}8 - 32{,}5)^2 + (42{,}1 - 32{,}5)^2\right) = 119{,}27$$

L'écart type est :

$$S(X) = \sqrt{Var(X)} = \sqrt{119{,}27} = 10{,}9$$

Voyons maintenant comment utiliser concrètement les statistiques afin de pouvoir exploiter différents résultats. Le but n'étant pas de présenter tous les tests statistiques, nous nous focaliserons ici uniquement sur les tests visant à **estimer des moyennes de populations à partir d'un échantillonnage**, et à comparer **2 valeurs issues d'un échantillonnage entre elles**, car c'est effectivement ce qui est le plus utilisé en TIPE/TPE. Les premiers tests présentés seront l'occasion d'aborder le principe de leur construction.

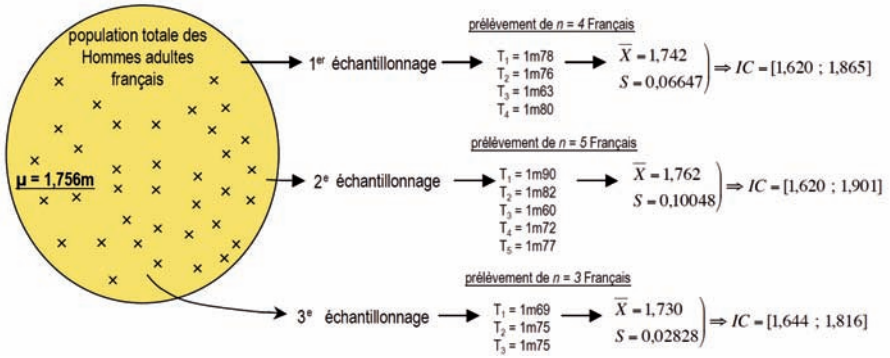
Comment estimer la moyenne d'une population μ à partir de la moyenne obtenue par un échantillonnage de cette population ($\overline{X}$) ?

Les principes généraux de l'estimation par intervalle

Considérons l'exemple concret suivant : on veut connaître la moyenne de la taille des hommes adultes français (la vraie moyenne de la population totale est appelée μ et cette moyenne est de 1,756 m d'après l'INSEE et encore, cette moyenne est elle-même estimée par statistiques !). Pour cela, on échantillonne *n* Français, ce qui donne une moyenne de taille pour l'échantillon appelé X .

La question est la suivante : **à partir de $\overline{X}$, peut-on avoir une idée de** μ ?

La figure 6 (début) montre 3 échantillonnages ayant abouti chacun à une valeur différente du fait des fluctuations d'échantillonnage. La vraie valeur de μ est impossible à avoir par cette méthode (il faudrait en effet échantillonner toute la population !). En revanche, grâce aux statistiques, on peut estimer un intervalle dans lequel la vraie valeur de μ a de bonnes chances de se trouver. On dit que l'on procède à une estimation par intervalle. L'intervalle obtenu est appelé **intervalle de confiance (IC)**. Cet intervalle dépend du niveau de confiance exigé. Par exemple, si on veut être sûr à 100 % que le μ réel se trouve dans l'intervalle de confiance, alors il sera très large (par exemple dans l'exemple de la taille des Français, on peut dire que μ est compris dans l'intervalle [0m – 3m] : on est sûr d'avoir juste, mais l'information est en contrepartie très peu intéressante. Si au contraire, on veut un intervalle restreint, alors le niveau de confiance baissera.



6. Estimation de la moyenne d'une population par intervalle de confiance calculé à partir de plusieurs échantillons

Par conséquent, quand on calcule un intervalle de confiance, il faut toujours indiquer le niveau de confiance souhaité. En biologie, l'intervalle de confiance est conventionnellement calculé pour un niveau de confiance de 95% (on dit aussi « au seuil de risque de 5% »). Ceci signifie que q**uand on donne un intervalle de confiance à 95%, le μ réel a 95% de chance d'être dans l'intervalle de confiance calculé**.

Sur la figure 6 (fin) sont présentés les différents IC (calculés par la formule (1) présentée après) dans l'exemple de la taille des Français. On voit que pour les 3 échantillons considérés, la vraie moyenne μ est bien comprise dans les différents intervalles proposés. Si on avait considéré 100 échantillons (et non 3), on aurait trouvé à peu près 95 intervalles contenant μ, et 5 ne la contenant pas.

Finalement, comment calcule-t-on l'intervalle de confiance ? Intuitivement, on sent que :

- l'intervalle de confiance est centré sur la moyenne $\overline{X}$;

- plus l'écart type de l'échantillon S est petit, plus l'intervalle va être petit (si les mesures issues de l'échantillonnage sont proches, alors il est logique de penser que la moyenne $\overline{X}$ de l'échantillon estime assez bien la vraie moyenne μ de la population) ;

- plus le nombre de mesures n est élevé, plus $\overline{X}$ tend vers μ (dans le cas extrême où n est égal au nombre d'individus de la population : $\overline{X} = \mu$).

La formule de calcul de l'intervalle de confiance peut alors être compréhensible :

$$\text{L'intervalle de confiance à 95 \% est égal à } \left[\overline{X} - t_{(n-1)}\frac{S}{\sqrt{n-1}} ; \overline{X} + t_{(n-1)}\frac{S}{\sqrt{n-1}}\right] \quad (1)$$

La signification du paramètre appelé t(n-1) ne sera pas détaillée ici. t(n-1) est une valeur donnée par la table figure 7.

n-1	t	n-1	t
1	12,706	12	2,179
2	4,303	13	2,160
3	3,182	14	2,145
4	2,776	15	2,131
5	2,571	20	2,086
6	2,447	25	2,060
7	2,365	30	2,042
8	2,306	40	2,021
9	2,262	60	2,000
10	2,228	120	1,980
11	2,201	∞	1,960

7. Valeur de t à 95 % de confiance (modifié d'après Motulsky)
Exemple : pour un effectif d'échantillonage de n=5, la valeur de t à considérer dans la formule (1) est 2,776.

Noter ici que l'intervalle de confiance s'écrit sous la forme d'un intervalle $[\overline{X} - a ; \overline{X} + a]$ et non sous la forme $\overline{X} \pm a$ (sous cette dernière forme, le a correspond classiquement au demi écart-type).

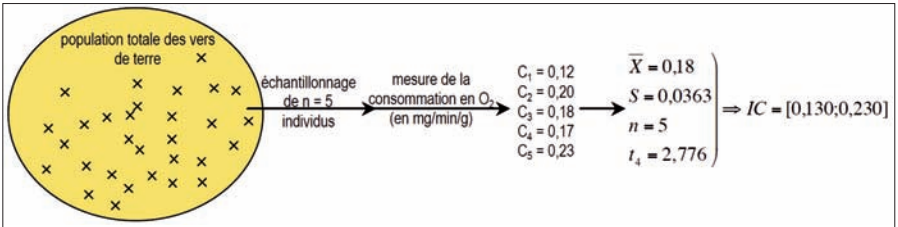
Application de l'estimation par intervalle à partir d'un exemple concret issu de TIPE

Dans cette application, des étudiants cherchent à déterminer, dans des conditions expérimentales données, la consommation en O_2 de vers de terre à l'obscurité. N'ayant pas accès à la population mondiale des vers de terre, les étudiants réalisent un échantillonnage de 5 vers de terre pour lesquels ils mesurent la consommation en O_2. La figure 8A montre l'analyse statistique issue de cette expérience.

Remarques concernant l'estimation par intervalle

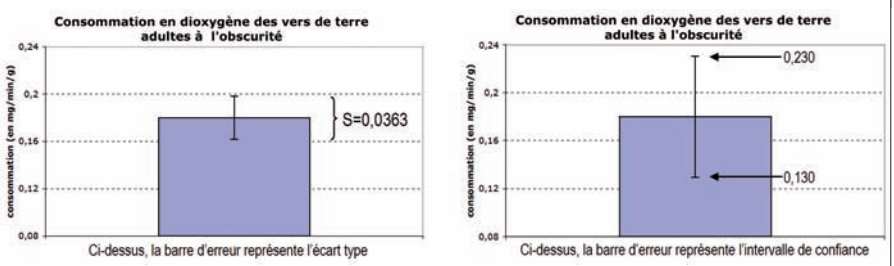
L'intervalle de confiance fournit un intervalle dans lequel la vraie moyenne de la population a de fortes chances de se trouver. Il est donc bien plus informatif que la moyenne de l'échantillon. Il est à noter – du fait du niveau de confiance à 95 % – qu'il se peut que la vraie moyenne soit hors de l'intervalle (dans l'exemple précédent, il se pourrait peut-être qu'en réalité, la consommation moyenne des vers de terre soit de 0,10 mg/min/g par exemple…). Ceci ne peut pas être vérifié. On touche là la limite fondamentale des statistiques.

Certaines publications de chercheurs présentent les résultats sous forme de graphiques exprimés en moyenne associée à des barres représentant les intervalles de confiance (voir *fig. 8B*). Ceci a l'intérêt de matérialiser la zone dans laquelle se trouve la vraie moyenne (à l'erreur près de 5 %). Concernant les barres d'erreur, il est à noter qu'il n'y a pas de représentation universelle : nous avions vu par exemple dans la figure 5 que les barres d'erreur pouvaient aussi représenter l'écart type.



A. Traitement statistique des résultats

Conclusion : la consommation moyenne des vers de terre (sous-entendu : de la population générale = ce qui nous intéresse) a 95 % de chance de se trouver entre 0,130 et 0,230 mg/min/g. Concrètement, on s'affranchit de cette phrase à rallonge en concluant : « les vers de terre consomment statistiquement entre 0,130 et 0,230 mg/min/g ».



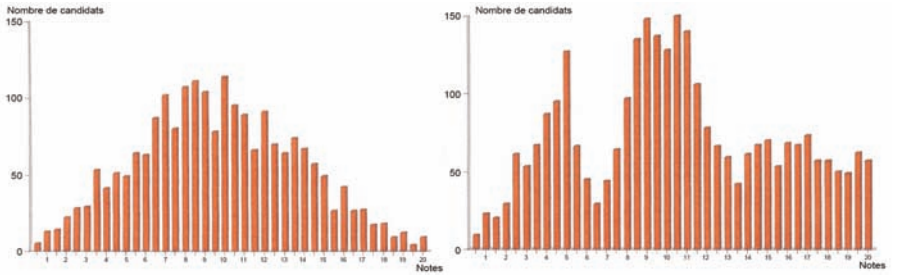
B - Deux représentations graphiques possibles des résultats ci-dessus

Remarque : certains chercheurs représentent quant à eux l'erreur standard à la moyenne ($ESM = S/\sqrt{n-1}$). Celle-ci correspond à l'intervalle de confiance divisé par la valeur tn-1. Elle matérialise donc - comme l'intervalle de confiance - la précision avec laquelle on connaît la moyenne de la population. L'ESM est assez souvent représentée, ce qui peut s'expliquer par une raison simple mais peu avouable : c'est elle qui donne les plus petites barres d'erreur, et donc qui fait la meilleure impression…

8. Application concrète : estimation de la consommation moyenne en dioxygène de vers de terre à l'obscurité à partir d'un échantillonnage

D'autres chercheurs représentent quant à eux l'erreur standard à la moyenne ($ESM = S/\sqrt{n-1}$) (voir figure 8B)… En règle générale, les barres d'erreur ne peuvent donc être interprétées correctement que si nous savons ce qu'elles représentent : par exemple, représenter un écart type (qui matérialise la variabilité des mesures) ou un intervalle de confiance (qui matérialise l'intervalle dans lequel la vraie moyenne a 95 % de chance de se trouver) n'est pas du tout la même chose. Le mieux est donc d'indiquer clairement - noir sur blanc – la signification des barres dans la légende de chaque graphique. Ceci n'est malheureusement pas toujours le cas.

Ce qu'on vient de dire ici ne s'applique en fait que quand la variable étudiée suit une loi normale. Sans rentrer dans les détails, ceci signifie que la variable mesurée (taille des Français, consommation en O_2…) a pour fonction de densité une courbe en forme de cloche (appelée courbe de Gauss) (*fig. 9*). On admettra ici que c'est effectivement le cas pour l'essentiel des variables utilisées en biologie. Toutefois, dans le cas d'une variable non normale, ce qu'on vient de dire s'applique tout de même quand $n \geq 20$.



9. Exemple de distributions quasi normale ou non normale
Histogrammes des notes aux épreuves écrites de chimie (à gauche) et de biologie-A (à droite) du concours Agro-Véto 2010 (population de 2918 étudiants)
Dans la population « notes en chimie », la distribution est « grossièrement » normale. Ceci signifie que l'on peut estimer un intervalle de confiance pour la moyenne de chimie (μ) à partir d'un échantillon de n notes prises aléatoirement, en appliquant la formule (1).
Dans la population « notes en biologie », la distribution n'est pas du tout normale. Ceci signifie que l'on ne peut pas estimer un intervalle de confiance pour la moyenne de biologie (μ) à partir d'un échantillon de n notes prises aléatoirement, en appliquant la formule (1) (sauf si n est suffisamment grand).

Des tableurs grand public comme Excel ou Numbers peuvent calculer pour nous les intervalles de confiance. Il est à noter cependant que la formule utilisée est une formule approchée (notamment au niveau de la valeur t) si bien que les résultats donnés sous estiment l'intervalle de confiance, sauf quand l'effectif (= n) est grand, ce qui n'est très souvent pas le cas en TIPE ou TPE…Les logiciels de statistiques (Statel…) utilisent en revanche la même formule que celle présentée ci-dessus.

Comment comparer entre elles deux moyennes obtenues par l'expérience ?

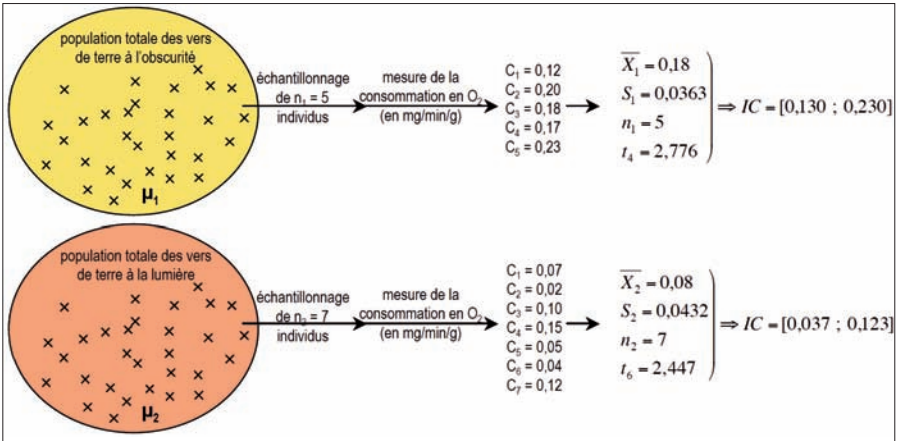
Une première possibilité : la comparaison des intervalles de confiance des moyennes des 2 populations

Reprenons l'exemple concret des vers de terre : les étudiants du groupe de TIPE veulent déterminer si la lumière influence la consommation moyenne en dioxygène des vers de terre adultes.

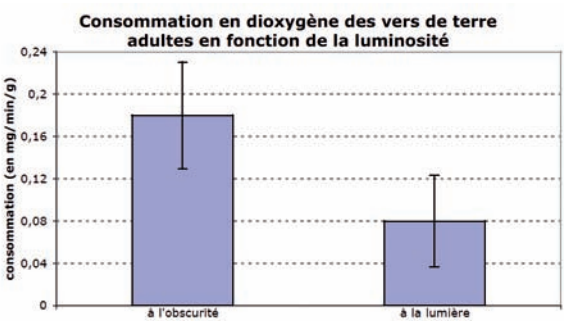
N'ayant pas accès aux populations mondiales des vers de terre cultivés à la lumière et à l'obscurité, des mesures sont effectuées sur 7 vers de terre à la lumière et 5 vers de terre à l'obscurité (échantillonnage). L'analyse statistique est présentée figure 10A.

Pratiquement, un intervalle de confiance est tracé pour chacune des 2 moyennes (*fig. 10B*). On considèrera que si les intervalles ne se recoupent pas, alors les 2 moyennes réelles des 2 populations sont statistiquement différentes.

Dans cet exemple, les 2 intervalles ne se recoupent pas : on peut donc affirmer que la consommation à l'obscurité est statistiquement supérieure à celle à la lumière.



A - Echantillonnage et résultats obtenus



B - Graphique obtenu (les barres représentent les intervalles de confiance)
La vraie moyenne µ1 est située entre 0,130 et 0,230 (au seuil de 95 %). La vraie moyenne µ2 est située entre 0,037 et 0,123. Les 2 intervalles ne se recoupent pas : on peut dire que, statistiquement, la consommation des vers de terre à l'obscurité est plus élevée qu'à la lumière.

10. Exemple concret : comparaison de deux moyennes issues d'une expérience de TIPE

Dans le cas où les intervalles se recoupent, il est important de noter ici que ceci ne signifie pas que les consommations sont identiques : ceci signifie juste qu'on ne peut pas voir de différence significative par cette approche visuelle, ce qui n'est pas la même chose (on dit que cette méthode est peu puissante). Ceci est très fréquent en TIPE et laisse les étudiants déçus de leur résultat (surtout quand la simple lecture des résultats bruts semblait pourtant montrer l'existence d'une différence, que les statistiques n'ont pu par cette méthode mettre en évidence). Comment faire alors pour réussir à prouver cette différence (si vraiment elle existe) ?

La formule de l'intervalle de confiance donnée en (1) est très intéressante : elle montre clairement deux façons de réduire l'intervalle : faire baisser S ou augmenter *n*. Concrètement pour les étudiants, il s'agit de manipuler le plus précisément possible (pour faire baisser S (même si l'essentiel de la variabilité est biologique, c'est - à - dire le fait que les personnes, les animaux, les cellules soient différents les uns des autres, ne peut pas être diminuée) et d'essayer de travailler avec le plus grand effectif possible (pour faire augmenter *n*).

Une deuxième possibilité : calculer un intervalle de confiance pour la différence des moyennes des deux populations

Cette méthode permet de voir des différences statistiques non mises en évidence par la comparaison visuelle des 2 intervalles de confiance unitaires de la méthode précédente. C'est elle qui est utilisée en techerche (parfois sous le nom « test de Student »). Elle a le mérite d'être facilement compréhensible. La méthode est basée sur ce qui a été fait précédemment lors de la construction de l'intervalle de confiance de la formule (1). Par analogie, on peut estimer un intervalle de confiance pour la différence de 2 moyennes de 2 populations (appelées μ_1 et μ_2) à partir des 2 moyennes issues d'un échantillonnage des 2 populations (appelées $\overline{X_1}$ et $\overline{X_2}$).
La formule de calcul de l'intervalle de confiance de $\mu_1 - \mu_2$ est alors la suivante :

L'intervalle de confiance à 95 % est égal à

$$\left[\overline{X}_1 - \overline{X}_2 - t_{(n1+n2-2)}\sqrt{\frac{S^2}{n_1}+\frac{S^2}{n_2}};\overline{X}_1 - \overline{X}_2 + t_{(n1+n2-2)}\sqrt{\frac{S^2}{n_1}+\frac{S^2}{n_2}}\right] \quad \textbf{(2)}$$

S^2 est la variance commune des 2 échantillons et vaut : $S^2 = \dfrac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$

S_1 et S_2 sont les écart-types correspondant respectivement aux 2 échantillons issus des populations 1 et 2. La valeur *t* est donnée par la table figure 7.

La figure 11 montre l'application concrète de cette formule avec le cas des vers de terre.

Dans l'exemple de la figure 10, on a :

$$\begin{cases} n_1 = 5 \text{ et } n_2 = 7 \text{ (effectifs des 2 échantillons)} \\ \overline{X}_1 - \overline{X}_2 = 0{,}08 \text{ (différence des 2 moyennes obtenues par échantillonnage)} \\ t_{n1+n2-2} = t_{10} = 2{,}228 \\ S^2 = \dfrac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} = \dfrac{5*0{,}0363^2 + 7*0{,}04324^2}{10} = 0{,}00197 \end{cases}$$

L'application de la formule (2) donne l'intervalle de confiance pour la différence de moyenne $\mu_1 - \mu_2$:

$$\left[\overline{X}_1 - \overline{X}_2 - t_{(n1+n2-2)}\sqrt{\frac{S^2}{n_1}+\frac{S^2}{n_2}};\overline{X}_1 - \overline{X}_2 + t_{(n1+n2-2)}\sqrt{\frac{S^2}{n_1}+\frac{S^2}{n_2}}\right]$$

soit

$$\left[0{,}10 - 2{,}228\sqrt{\frac{0{,}00197}{5}+\frac{0{,}00197}{7}};0{,}10 + 2{,}228\sqrt{\frac{0{,}00197}{5}+\frac{0{,}00197}{7}}\right]$$

soit

$$[0{,}10 - 0{,}058; 0{,}10 + 0{,}058]$$

soit

$$[0{,}042\ ;\ 0{,}158]$$

Ceci signifie qu'il y a 95 % de chance que la différence de consommation réelle (sous-entendu : dans la population) entre les vers de terre à l'obscurité et à la lumière soit comprise entre 0,042 et 0,158.
Le « 0 » n'étant pas dans l'intervalle, on peut donc dire que **en général, les vers de terre consomment en moyenne plus de dioxygène à l'obscurité qu'à la lumière.**
L'intervalle de confiance a aussi l'intérêt de nous estimer (à l'aide d'un intervalle) la différence de consommation.

11. Calcul de l'intervalle de confiance de la différence de deux moyennes à partir de l'exemple de la figure 10

Conclusion et remarques concernant la comparaison de deux moyennes issues de l'expérience

Les deux méthodes qui viennent d'être abordées permettent de pouvoir prouver

statistiquement une différence (ou non) entre deux moyennes issues d'un échantillonnage. Elles sont pour cela très utiles dans l'interprétation de nombreuses expériences.

Comme dans la première partie, ce que l'on vient de dire ne s'applique rigoureusement que dans des cas précis : il faut que les 2 variables à comparer suivent une loi normale et aient une variance assez proche (des logiciels comme Statel testent l'égalité des variances). Si ce n'est pas le cas, mais que les effectifs n_1 et n_2 sont supérieurs à 20, ce que l'on vient de dire s'applique aussi.

Dans les cas où les conditions d'application ne sont pas réunies, il faut alors utiliser un autre test, appelé test des rangs de Wilcoxon. Ce dernier est intuitif et assez simple d'utilisation. La figure 12 montre son principe à travers deux exemples concrets. L'idée est que le calcul ne porte pas sur les valeurs numériques des mesures issues des échantillons représentatifs des populations, mais sur leurs rangs attribués suite au classement des valeurs par ordre croissant. Ceci permet de s'affranchir des conditions de normalité et d'homogénéité des variances indispensables à la fiabilité des tests basés sur les intervalles de confiance. En contre partie, dans le cas de données chiffrées, l'information sur les différences numériques entre les mesures n'est pas prise en compte (ce qui explique pourquoi ce test est moins puissant que les tests sur valeurs chiffrées).

Comment estimer le pourcentage d'une population π à partir du pourcentage obtenu par un échantillonnage de cette population (P) ?

Tout ce qui a été abordé jusqu'à présent concerne des données chiffrées résultant de mesures moyennées. Dans le cas où des résultats sont exprimés sous forme de proportion, on peut facilement étendre ce qui vient d'être vu, en retouchant légèrement les différentes formules.

Dans le cas présent, on va montrer comment estimer un pourcentage d'une population à partir du pourcentage obtenu par échantillonnage de cette population.

Considérons l'exemple concret suivant : des étudiants cherchent à comparer la qualité gustative de deux sels que l'on notera A et B pour ne pas froisser les habitants de Guérande ou de Camargue. Pour cela, ils font goûter les deux sels (sur une tartine beurrée) à de nombreux individus cobayes et ceux-ci doivent indiquer laquelle des 2 tartines salées ils préfèrent.

Dans le cas où *n* est assez grand (concrètement : $n \geq 20$), l'intervalle de confiance se calcule selon le même principe que ce qui a été vu dans la première partie (la variance étant ici égale à P(1-P), et la valeur t étant remplacée par 1,96) :

$$\text{L'intervalle de confiance à 95 \% est égal à } \left[P - 1{,}96\sqrt{\frac{P(1-P)}{n}} ; P + 1{,}96\sqrt{\frac{P(1-P)}{n}} \right] \quad (4)$$

Exemple 1 : On cherche à déterminer si la présence de mouron sur une parcelle de poivrons influence le nombre de graines par fruit de poivron

Tableau des résultats (chaque chiffre représente le nombre de graines trouvées dans un fruit échantillonné)

Nombre de graines/fruit en absence de mouron	Nombre de graines/fruit en présence de mouron
60 ; 71 ; 48 ; 56 ; 72 ; 44 ; 22 ; 57 ; 68 ; 47	32 ; 27 ; 50 ; 19 ; 18 ; 40 ; 7 ; 13 ; 18 ; 45 ; 52
M = 10 fruits testés	

N = 21 fruits testés en tout

1re étape : classement des données et affectation d'un rang à chaque donnée

Nombre de graines/ fruit en absence de mouron	Nombre de graines/ fruit en présence de mouron
	7
	13
	18
	18
	19
22	
	27
	32
	40
44	
	45
47	
48	
	50
	52
56	
57	
60	
68	
71	
72	

traduction en « rang »

Rangs en absence de mouron	Rangs en présence de mouron
	1
	2
	3
	4
	5
6	
	7
	8
	9
10	
	11
12	
13	
	14
	15
16	
17	
18	
19	
20	
21	

2e étape : calcul de la somme des Rangs (=SR) d'une des 2 colonnes

Par exemple ici, pour la colonne de gauche: $SR_{absence\ de\ mouron} = 6 + 10 + 12 + 13 + 16 + 17 + 18 + 19 + 20 + 21 = 152$

Intuitivement, on sent que s'il n'y a pas d'« effet mouron », alors les rangs 1 à 21 sont dispersés dans les 2 colonnes. Si au contraire le mouron augmente le nombre de graines par fruit, alors les rangs faibles sont plutôt dans la colonne de gauche (donc SR calculé est faible). Enfin, si le mouron diminue le nombre de graines par fruit, alors les rangs élevés sont plutôt dans la colonne de gauche (donc SR calculé est fort).

3e étape : réalisation concrète du test des rangs de Wilcoxon

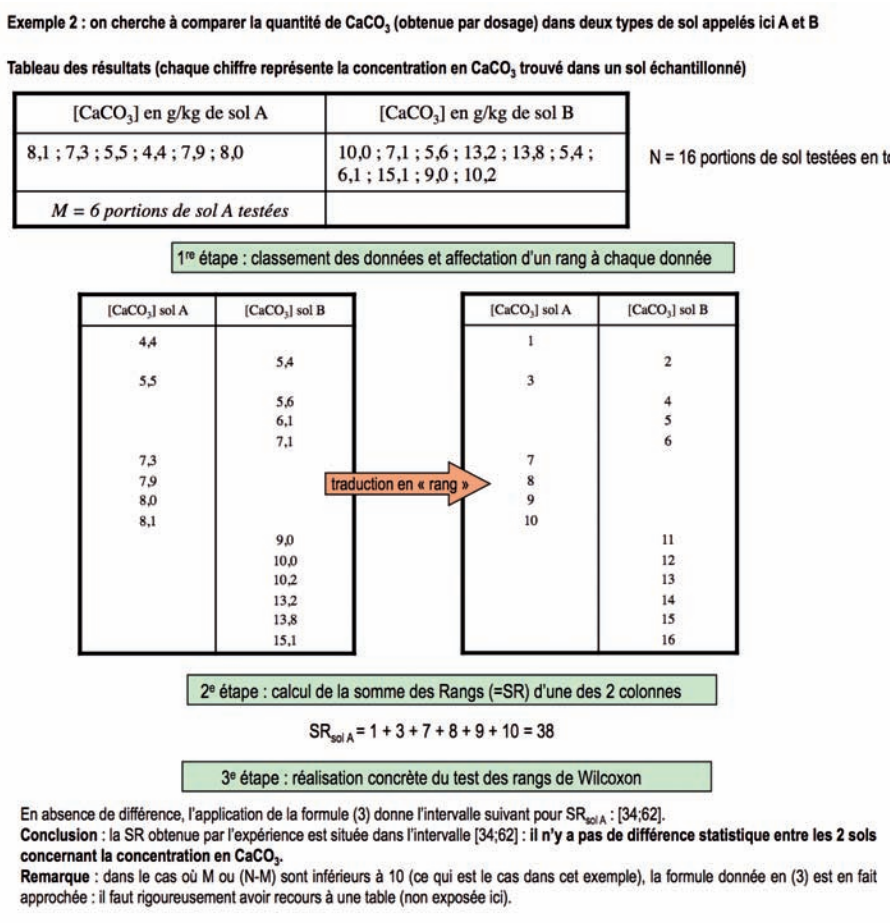
On admet que s'il n'y a pas d'effet du paramètre testé (ici : présence de mouron), alors le SR obtenu par l'échantillonnage a 95 % de chance de se trouver compris dans l'intervalle suivant :

$$\left[\frac{M(N+1)}{2} - 1{,}96\sqrt{\frac{M(N+1)(N-M)}{12}}; \frac{M(N+1)}{2} + 1{,}96\sqrt{\frac{M(N+1)(N-M)}{12}}\right] \quad (3)$$

Si SR est en dehors de l'intervalle, alors on considère qu'il y a un effet au seuil habituel des 5%.

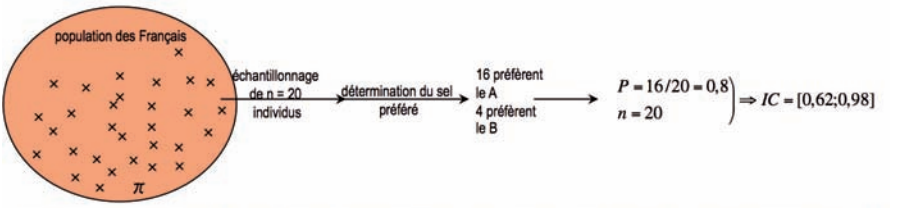
Application au cas ci-dessus : en prenant M=10 et N = 21, on trouve l'intervalle suivant : [82;138].
Conclusion : la SR trouvée est plus élevée que la SR attendue dans le cas d'absence d'effet (puisque 152 > 138). On peut donc dire que **la présence de mouron diminue le nombre de graines dans les fruits de poivron.**
Remarque : nous nous sommes intéressés ici à la colonne de gauche et avons classé par ordre croissant. Ceci est sans conséquence (on peut prendre la colonne de droite (dans ce cas M = 11) ou classer par ordre décroissant (dans ce cas le SR obtenu aurait été inférieur à 82).

12. Principe du test des rangs de Wilcoxon à travers 2 exemples concrets : 1er exemple



12. Principe du test des rangs de Wilcoxon à travers 2 exemples concrets : 2e exemple

La figure 13 présente l'analyse statistique des résultats. Dans cet exemple, on peut donc dire qu'il y a 95 % de chance que le vrai pourcentage soit entre 62 et 98 %. 50 % étant hors intervalle, on conclue donc à une différence significative entre les deux sels. Notons ici encore que l'on peut se tromper, mais on ne le saura jamais : par analogie, la probabilité d'obtenir 16 piles en lançant 20 fois une pièce équilibrée n'est clairement pas nulle (bien qu'inférieure à 5%).



Conclusion : le pourcentage réel de la population a 95 % de chance de se trouver entre 62 et 98. Au seuil des 5 %, on peut donc dire que **62 à 98% de la population préfère le sel A.**

13. Application concrète : estimation du pourcentage de français qui préfèrent le sel A à partir d'un échantillonnage de cette population

Remarque : la formule ci-dessus n'est valable que pour de grands échantillons ($n \geq 20$). Dans le cas d'échantillons plus petits, on ne peut pas faire d'estimation statistique comme jusqu'à présent. On peut à la place réaliser un test d'hypothèse en utilisant la loi binomiale. Les tests d'hypothèse ne pouvant être abordés ici, on ne traitera pas ce cas.

Comment comparer entre eux deux pourcentages obtenus par l'expérience ?

Comme précédemment avec la comparaison de deux moyennes, on peut ici tracer un intervalle de confiance pour chacun des deux pourcentages et considérer que si les deux intervalles ne se recoupent pas, alors les deux pourcentages réels des deux populations sont différents.

Comme vu dans la 2e partie, une deuxième possibilité – plus puissante mais moins visuelle – est de calculer un intervalle de confiance pour la différence des pourcentages des deux populations ($\pi_1 - \pi_2$), à partir des pourcentages obtenus par les deux échantillons (P1 et P2).

La formule de l'intervalle de confiance de $\pi_1 - \pi_2$ est alors la suivante :

L'IC à 95 % est

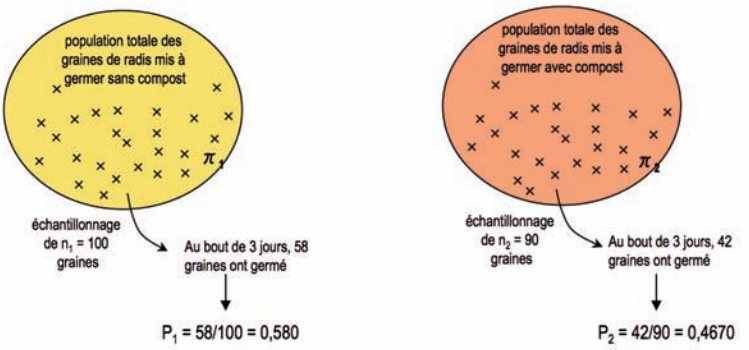
$$\left[P_1 - P_2 - 1{,}96\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}} ; P_1 - P_2 + 1{,}96\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}} \right] \quad \textbf{(5)}$$

La figure 14 montre une application de cette formule, dans laquelle des étudiants cherchent à déterminer si la présence de compost dans un sol a une influence sur la capacité de germination de graines de radis.

Remarque :

Comme dit dans la 3e partie, ce test ne marche que dans le cas de grands échantillons ($n \geq 20$).

La question à laquelle les étudiants veulent répondre est la suivante : le compost influence-t-il la germination des graines de radis sur 3 jours? On cherche donc à savoir si le pourcentage de germination sans compost (π_1) est différent du pourcentage avec compost (π_2).
Dans l'expérience présentée ci-dessous, les étudiants disposent du pourcentage de germination d'un échantillon de graines mis à germer avec ou sans compost (P_1 et P_2).



D'après la formule (5), l'intervalle de confiance à 95% de la différence des pourcentages $\pi_1 - \pi_2$ est :

$$\left[P_1 - P_2 - 1{,}96\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}\,;\, P_1 - P_2 + 1{,}96\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}\right]$$ c'est à dire [-0,028 ; 0,255]

Conclusion : ceci signifie qu'il y a 95 % de chance que la différence entre les pourcentages réels de germination avec et sans compost soit comprise entre -0,028 et 0,255.
Le « 0 » étant dans l'intervalle, on ne peut donc pas dire que le compost en général favorise ou inhibe la germination des graines de radis les 3 premiers jours. On dit **le compost n'a statistiquement pas d'influence sur la germination des graines de radis les 3 premiers jours.**

14. Application concrète: comparaison de deux pourcentages issus d'une expérience de TIPE

Il y a une autre façon de réaliser ce test, en utilisant la statistique du Chi2. Celle-ci est un test d'hypothèse basé sur la comparaison des effectifs réels et des effectifs théoriques attendus en absence de différence. Le test du Chi2 est aussi beaucoup utilisé (notamment en génétique). Il n'est pas abordé ici mais donne exactement les mêmes résultats que le test de l'intervalle de confiance de la différence des pourcentages (les 2 tests sont en effet équivalents).

Conclusion

L'objectif de ce texte a été de montrer l'apport des statistiques dans l'exploitation de résultats issus d'expériences en biologie : elles permettent de **tirer des conclusions générales à partir d'un nombre limité de données** (c'est à dire de pouvoir extrapoler à une population ce que l'on voit à partir d'un échantillon).

En TIPE ou TPE, une initiation aux concepts de base des statistiques (notamment le fait qu'une simple comparaison de 2 moyennes n'est pas suffisante) permet aux élèves de réfléchir en amont sur le nombre de végétaux – animaux (*n*) à incorporer dans tel ou tel protocole et donc de pouvoir bâtir des expérimentations dont les résultats puissent ensuite être analysés.

Nous nous sommes ici focalisés uniquement sur les méthodes permettant l'**estimation statistique**, c'est à dire comment estimer une moyenne ou un pourcentage à partir d'un échantillonnage. Cette estimation varie selon la taille de l'échantillon et la dispersion. Ces deux notions sont combinées en statistiques pour donner un intervalle de confiance de la moyenne ou du pourcentage de la population. La figure 15 récapitule sous forme de guide les intervalles de confiance abordés dans le texte.

CE QU'ON VEUT FAIRE	CE QU'ON UTILISE	LES CONTRAINTES D'UTILISATION
Estimer la moyenne d'une population	Construction d'un IC à 95% (partie I formule (1)) : $\mathbf{IC} = \left[\overline{X} - t_{(n-1)}\frac{S}{\sqrt{n-1}}; \overline{X} + t_{(n-1)}\frac{S}{\sqrt{n-1}}\right]$	La variable suit une loi normale ou $n \geq 20$
Comparer 2 moyennes	1ère possibilité : Construction d'un IC à 95% pour les 2 moyennes et observation de leur éventuel recoupement (peu puissant)	La variable suit une loi normale ou $n \geq 20$
	2ème possibilité : Construction d'un IC à 95% pour la différence des moyennes (partie II formule (2)) : $\mathbf{IC} = \left[\overline{X}_1 - \overline{X}_2 - t_{(n1+n2-2)}\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}; \overline{X}_1 - \overline{X}_2 + t_{(n1+n2-2)}\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}\right]$	La variable suit une loi normale et les variances sont proches ou $n \geq 20$
	3ème possibilité : test des rangs de Wilcoxon (partie II formule (3)) : Calcul de l'intervalle de SR en absence de différence et comparaison avec le SR obtenu par l'expérience	Aucune
Estimer le pourcentage d'une population	Construction d'un IC à 95% (partie III formule (4)) : $\mathbf{IC} = \left[P - 1{,}96\sqrt{\frac{P(1-P)}{n}}; P + 1{,}96\sqrt{\frac{P(1-P)}{n}}\right]$	$n \geq 20$
Comparer 2 pourcentages	1ère possibilité : Construction d'un IC à 95% pour les 2 pourcentages et observation de leur éventuel recoupement (peu puissant)	$n \geq 20$
	2ème possibilité : Construction d'un IC à 95% pour la différence des pourcentages (partie IV formule (5)) : $\mathbf{IC} = \left[P_1 - P_2 - 1{,}96\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}; P_1 - P_2 + 1{,}96\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}\right]$	$n \geq 20$

15. Guide pratique des tests statistiques pouvant être utilisés en TPE et TIPE, et permettant une estimation par intervalle

Comme pour les jeux de plateau, de nombreuses extensions des tests présentés ici sont disponibles (selon si les échantillons sont appariés ou non, selon si on veut faire des sommes de rang en prenant en compte les valeurs numériques des mesures,..etc…). De même, des tests permettant de comparer plus de 2 valeurs entre elles existent (les fameuses ANOVA (ANalysis Of VAriance) et leurs pendants non paramétriques ; le test du Chi2 d'indépendance…). Les lecteurs intéressés pourront en trouver des descriptions dans les ouvrages Statistiques (Wonacott and Wonacott) et Biostatistiques (Motulsky).

Un autre domaine des statistiques concerne les **tests d'hypothèse**, qui sont aussi très fortement utilisés dans les articles de recherche : ils aident à décider s'il est vraisemblable qu'une différence observée soit due au hasard. L'idée sous-jacente est la suivante : s'il n'y a pas de différences entre deux populations, quelle est la probabilité de sélectionner aléatoirement des échantillons entre lesquels on trouve une

aussi grande ou plus grande différence que celle observée. La réponse est donnée sous forme de probabilité, appelée P-value. Si la P-value est petite (souvent : P<0,05), alors on considérera que la différence est significative. Cet aspect n'a pas pu être abordé ici.

Les statistiques ne sont pas « magiques » et ne donnent pas de conclusions solides (*cf* l'incertitude à 5 %). L'expression « statistiquement significatif » est séduisante mais ne veut pas dire « c'est différent ». Nous n'avons pas vu ici « ce qu'il y a derrière » les différentes formules admises. A moins d'étudier clairement les statistiques sous l'angle mathématique, il faut accepter en confiance une grande partie des statistiques : on peut très bien apprendre à utiliser et interpréter les différents tests sans en comprendre entièrement les mécanismes. C'est une situation fréquente en sciences (on peut interpréter les résultats d'une sonde à dioxygène même si on ne comprend pas à fond son mécanisme). Il faut juste en savoir assez sur le fonctionnement de l'outil pour savoir à quoi il sert et éviter de l'utiliser dans des situations non appropriées.

Notons enfin pour finir que si les statistiques nous sont d'une grande aide, elles ne font pas tout et ne garantissent malheureusement pas un travail de qualité : elles ne peuvent pas nous aider à surmonter les problèmes classiques de validité de l'échantillonnage par exemple (dans l'exemple du test de goût de la figure 13, l'échantillonnage a été fait sur des lycéens d'un lycée particulier et la généralisation a été faite sur la population française : l'échantillon est souvent moins diversifié que la population ciblée, ou la représente mal, ce qui met à mal la généralisation effectuée…).

Remerciements : Florence Manero, Michel Borrat-Michaud et Franck Lepage pour leur relecture du manuscrit.

■