Les statistiques en biologie expérimentale : pourquoi ? Comment ? (deuxième partie)

Dans le dernier numéro de «Regard sur la biochimie», nous avons vu quel était l'intérêt du *t-test* en biologie expérimentale. Il existe plusieurs variantes de ce test, qu'il faut savoir choisir en fonction de la situation ; il faut aussi savoir que le *t-test* n'est pas applicable dans toutes les circonstances – il convient, alors, de savoir s'en rendre compte, et de connaître les alternatives à privilégier. C'est l'objet de ce deuxième et dernier article.

Les variantes du *t-test* :

Lorsqu'il cherche à utiliser le *t-test*, quel que soit le programme informatique utilisé (R, Excel, …), l'expérimentateur est amené à choisir parmi différentes options du test. Voici comment choisir les plus appropriées :

1. «variances égales» ou «variances différentes» ? La variance mesure la dispersion des données de chaque série de valeurs (sur la figure 1 de notre article de la semaine dernière, à consulter sur http://www.sfbbm.fr/regards.php, la variance du jeu de données n°1 était plus grande que celle du jeu de données n°2 : les données étaient plus étalées de part et d'autre de leur moyenne). Dans la version initiale du *t-test* (écrite au début du XX^{ème} siècle ; elle est appelée «*Student's t-test*»), il fallait que les variances des deux jeux de données soient égales pour que le test soit applicable (en réalité, elles ne sont jamais parfaitement égales, comme toujours : les deux nombres ne seront jamais égaux, avec tous leurs chiffres après la virgule ; il faut donc qu'elle soit similaires – le terme consacré est : «homogènes»). Si elles sont très différentes, il ne faut pas utiliser le *t-test* sous sa version originale, mais une version ultérieure (appelée «*Welch's t-test*»), qui tolère les hétérogénéités de variances. Par contre, si elles sont similaires («homogènes»), il est préférable d'utiliser la version initiale du *t-test* («*t-test* à variances égales» ou «*t-test* à variances homogènes» ou «*Student's t-test*»), qui est plus précise que le *Welch's t-test*.
2. «données appariées» ou non ? On utilise le *t-test* à données appariées lorsqu'il existe un lien logique évident entre chaque réplicat d'une série de valeurs, et un réplicat unique de l'autre série de valeurs, et qui est susceptible d'affecter le résultat de la mesure. Ce lien logique peut être : ces deux mesures ont été réalisées sur le même objet (par exemple : si l'expérience consiste à peser des patients avant et après un traitement amaigrissant, il existe un lien logique évident entre chaque réplicat du premier jeu de données, et un réplicat du deuxième : la mesure a été faite sur le même individu). Apparier les données donne au test plus de précision : dans notre exemple, elle permet au test de s'affranchir de l'hétérogénéité des poids des différents patients (plutôt que de comparer, en bloc, les deux séries de valeurs, le test comparera le poids de chaque patient avant traitement, à son poids après traitement). Il est donc préférable d'apparier les données chaque fois qu'elles peuvent l'être : en éliminant l'hétérogénéité entre les réplicats de la mesure, on donne au test les moyens de détecter des différences plus faibles.
3. «*one-tailed*» ou «*two-tailed*» ? Tout ce que nous avons vu jusqu'ici concernait le «*two-tailed t-test*», qui cherche à déterminer si deux jeux de données sont significativement différents. Le «*one-tailed t-test*» teste si l'un des deux jeux de données est supérieur à l'autre (il teste donc un unique sens de variation). Il ne faut utiliser cette variante du test que lorsqu'on peut exclure *a priori* l'un des deux sens de variation, lorsqu'il existe une raison physique impérieuse d'exclure l'un des deux sens de variation (par exemple : si l'expérience consiste à mesurer une collection de morceaux de bois, à en couper un bout, puis à les re-mesurer : il est

indiscutable que les longueurs mesurées dans la deuxième série doivent être inférieures à celles mesurées dans la première série ; chaque fois qu'un morceau de bois semblera s'être allongé, on saura avec certitude qu'il s'agit d'une erreur de mesure). Comme on peut s'en douter, ce genre de situation est très rare dans notre métier (il est rare de connaître à l'avance le sens de variation, avant même de faire l'expérience ; de toute ma carrière, je n'ai jamais eu à utiliser le *one-tailed t-test*). Outre qu'il est d'un usage très restreint, le *one-tailed t-test* est dangereux : il divise les *p-values* par deux, et peut donc faire apparaître comme significatives des différences qui ne le sont pas. Utiliser le *one-tailed t-test* dans un cas illégitime constitue donc une fraude scientifique : il faut donc bien s'assurer qu'il existe une raison physique indiscutable d'exclure *a priori* un sens de variation, avant d'utiliser le *one-tailed t-test* (et j'invite le lecteur à me décrire sa situation, s'il pense être dans un cas d'utilisation du *one-tailed t-test* : je serais curieux de connaître un exemple d'utilisation en biologie expérimentale !).

Quand utiliser le *t-test* ?

Nous avons vu toute l'utilité du *t-test* en biologie ; malheureusement, ce test n'est pas utilisable dans toutes les circonstances, et avant de faire le calcul, il faut vérifier que le *t-test* est applicable. La principale limite du *t-test* tient à la nature de la distribution des données : pour que le *t-test* soit utilisable, il faut que chacun des deux jeux de données suive une loi normale (également appelée «loi gaussienne», c'est la distribution en cloche que suivent les données idéales, représentées par les courbes bleues dans le panneau du milieu de la figure 1 de l'article précédent, à consulter sur http://www.sfbbm.fr/regards.php). La loi normale est décrite par une équation mathématique bien précise (toute courbe en cloche ne suit pas une loi normale !) ; bien entendu, aucun jeu de données issu du monde réel ne suit parfaitement la loi normale (voyez comme les histogrammes des jeux de données n°1 et 2 s'éloignent de la courbe bleue, sur le panneau du milieu de la figure 1 de l'article précédent). On n'exige donc pas que les jeux de données suivent rigoureusement une loi normale – juste, qu'ils ne s'en éloignent pas exagérément.

Il existe des tests statistiques qui permettent de contrôler qu'un jeu de données suit une loi normale (le test de Shapiro-Wilk, ainsi que le test de Kolmogorov-Smirnov, un peu moins précis). On fournit à ces tests un jeu de données, et il calcule une *p-value*, qui est la probabilité que ce jeu de données soit échantillonné à partir d'une population idéale, de taille infinie, qui suit une loi normale (il faut donc effectuer ce test sur chacun des deux jeux de données, et n'utiliser ensuite le *t-test* que si chacune de ces deux *p-values* était grande – en général, on estime qu'elle doivent être supérieures à 0,05 ; cependant, elles tendent à être facilement supérieures à 0,05 dès que le nombre de réplicats est petit : on aura d'autant plus confiance dans le résultat de ces tests, que le nombre de réplicats est grand).

Que faire si au moins l'un des deux jeux de données s'éloigne trop de la loi normale (*p-value* < 0,05 dans le test de Shapiro-Wilk ou le test de Kolmogorov-Smirnov) ? Deux solutions s'offrent à l'expérimentateur :

1. Utiliser un test plus robuste aux déviations à la normalité. Ce test s'appelle le «test de Wilcoxon» (ou : «test de Mann-Whitney»). Il s'utilise comme le *t-test* (il faut lui fournir les deux séries de réplicats, et il calcule la probabilité que les séries de données sont issues de populations d'effectif infini de même moyenne), mais il est applicable quand les jeux de données ne suivent pas une loi normale. Cette robustesse a un coût : le test de Wilcoxon est beaucoup moins puissant que le *t-test* (pour deux mêmes jeux de données, il détectera moins facilement les différences significatives) ; cette différence de puissance est particulièrement

gênante quand les jeux de données contiennent peu de réplicats (notamment, la *p-value* du test de Wilcoxon ne peut jamais descendre en-dessous de 0,1 quand chaque série contient trois réplicats).

2. Utiliser une transformation mathématique. Lorsqu'un jeu de données ne suit pas une loi normale, il est fréquent que le logarithme des données suive une loi normale (le logarithme, qui est une fonction croissante dont la croissance est très lente, tend à écraser les différences, donc à rapprocher les valeurs de leur moyenne – les mesures aberrantes, souvent responsables des déviations à la normalité, ont alors moins d'influence sur la forme générale de la distribution). On applique alors le *t-test* sur le logarithme des données, plutôt que sur les données elles-mêmes (et si les logarithmes sont significativement différents, alors les valeurs non transformées le sont nécessairement ; simplement, grâce au logarithme, on aura pu estimer plus précisément la *p-value* de cette différence). D'autres transformations mathématiques peuvent avoir la même vertu (la racine carrée, l'arcsinus, …), et leur usage s'apparente parfois à une cuisine un peu anarchique, dans laquelle l'utilisateur cherche une transformation qui tordra ses jeux de données dans le bon sens … Il est donc préférable d'utiliser une transformation qui ait un sens physique (le logarithme d'une concentration en réactif, par exemple, a un sens physique : il sert à calculer les pH et les pKa ; il est moins logique de calculer le logarithme d'une grandeur qui ne serait pas multiplicative, mais additive).

La démarche à suivre est résumée en figure 1. Il faut savoir que la plupart des tests statistiques cités dans cette figure (tests de Shapiro-Wilk, de Kolmogorov-Smirnov, de Levene, de Wilcoxon) ne sont pas disponibles dans le programme Excel. Des logiciels libres («R», téléchargeable sur http://www.r-project.org/ ; «PAST», téléchargeable sur http://folk.uio.no/ohammer/past/) proposent tous ces tests.

Il faut aussi mentionner une autre limite du *t-test*, qui, si elle est évidente en premier abord, est souvent négligée : ce test permet de mesurer la confiance dans les différences mesurées entre deux jeux de données, rien de plus ; il ne permet pas, par exemple, de comparer deux distributions de données.
Imaginons que l'expérience consiste à comparer la répartition de cellules dans les différentes phases du cycle cellulaire (mitose, phase G1, phase S, phase G2). On dispose de deux séries de comptage, dans deux conditions expérimentales (par exemple, un sauvage et un mutant). Comme toujours, on ne trouvera pas rigoureusement la même distribution entre les différentes phases du cycle, dans les deux conditions – et on aimerait savoir si cette différence est significative. L'expérimentateur pourrait être tenté d'utiliser le *t-test* pour comparer, un par un, les effectifs dans chacune des phases du cycle cellulaire dans les deux conditions. Deux problèmes s'annoncent :

1. Comme on l'a vu, le *t-test* réclame plusieurs réplicats de chaque mesure. Ici, la mesure consiste en un comptage de cellules (probablement plusieurs centaines) : d'une certaine manière, l'expérience a déjà été réalisée un grand nombre de fois (dans chaque catégorie – ici : dans chaque phase du cycle cellulaire – l'effectif compté est le résultat de l'analyse de centaines de cellules), et il semble curieux de devoir faire plusieurs réplicats d'une expérience … qui contenait déjà plusieurs centaines de comptages.
2. En admettant que ce problème soit réglé (l'expérimentateur a effectivement fait plusieurs comptages indépendants, de plusieurs centaines de cellules chacun), le *t-test* pourra donner des résultats aberrants : il pourra, par exemple, montrer que le nombre de cellules en phase G1 est significativement différent entre les deux conditions, sans détecter de différence significative pour les trois autres phases du cycle (ce qui est possible, si la différence qui compense celle de la phase G1 se

répartit relativement uniformément entre les trois autres phases). Comment, alors, interpréter ce résultat : la distribution des cellules au cours du cycle cellulaires est-elle affectée par les conditions testées ?

Ce deuxième problème potentiel illustre la nature de l'erreur qui a été commise : ici, il ne s'agissait pas de comparer, catégorie après catégorie, les effectifs dans les deux conditions. Il s'agissait de comparer les deux distributions, dans leur ensemble.

Ce type de problème est inaccessible au *t-test* ; il faut utiliser un autre test : soit le test du χ^2 de Pearson (si les effectifs sont suffisants ; on considère en général qu'il faut que chaque effectif, dans chaque catégorie, soit supérieur à 5) ; soit le test exact de Fisher (qui n'a pas cette limitation, mais qui est un peu plus long en temps de calcul – ce problème n'en est plus un depuis que les ordinateurs sont devenus monstrueusement rapides). Il faut fournir à ces tests les séries de comptage, dans chaque condition expérimentale, et la *p-value* calculée est la probabilité que ces deux séries de comptage dérivent de populations idéales, de taille infinie, qui auraient la même distribution.
Il est particulièrement important de retenir que ces deux tests ne doivent être appliqués qu'à des nombres d'observations, des effectifs, qui n'auront pas été normalisés (on n'applique pas, par exemple, ces tests sur des pourcentages, dont la somme aura été arbitrairement amenée à 100 en multipliant tous les effectifs par un facteur de normalisation). En effet, ces deux tests tiennent compte des effectifs pour mesurer la confiance à apporter dans les différences mesurées (ils considèrent que 500 est très différent de 450, alors que 50 est peu différent de 45) : cette sensibilité aux effectifs, qui est une qualité utile, est perdue si on augmente ou diminue arbitrairement les données en les multipliant par un facteur de normalisation.

Conclusion

Les réflexions et les illustrations présentées ici n'ont pas la prétention de couvrir tous les usages possibles des statistiques pour les biologistes ; elles visent principalement à expliquer des concepts qui sont beaucoup plus généraux, et à dénoncer quelques erreurs fréquentes dans les publications de biologie expérimentale. Elles devraient aider le lecteur à comprendre les lectures qui lui seront nécessaires pour traiter des problèmes plus complexes que ceux qui ont été cités ici.

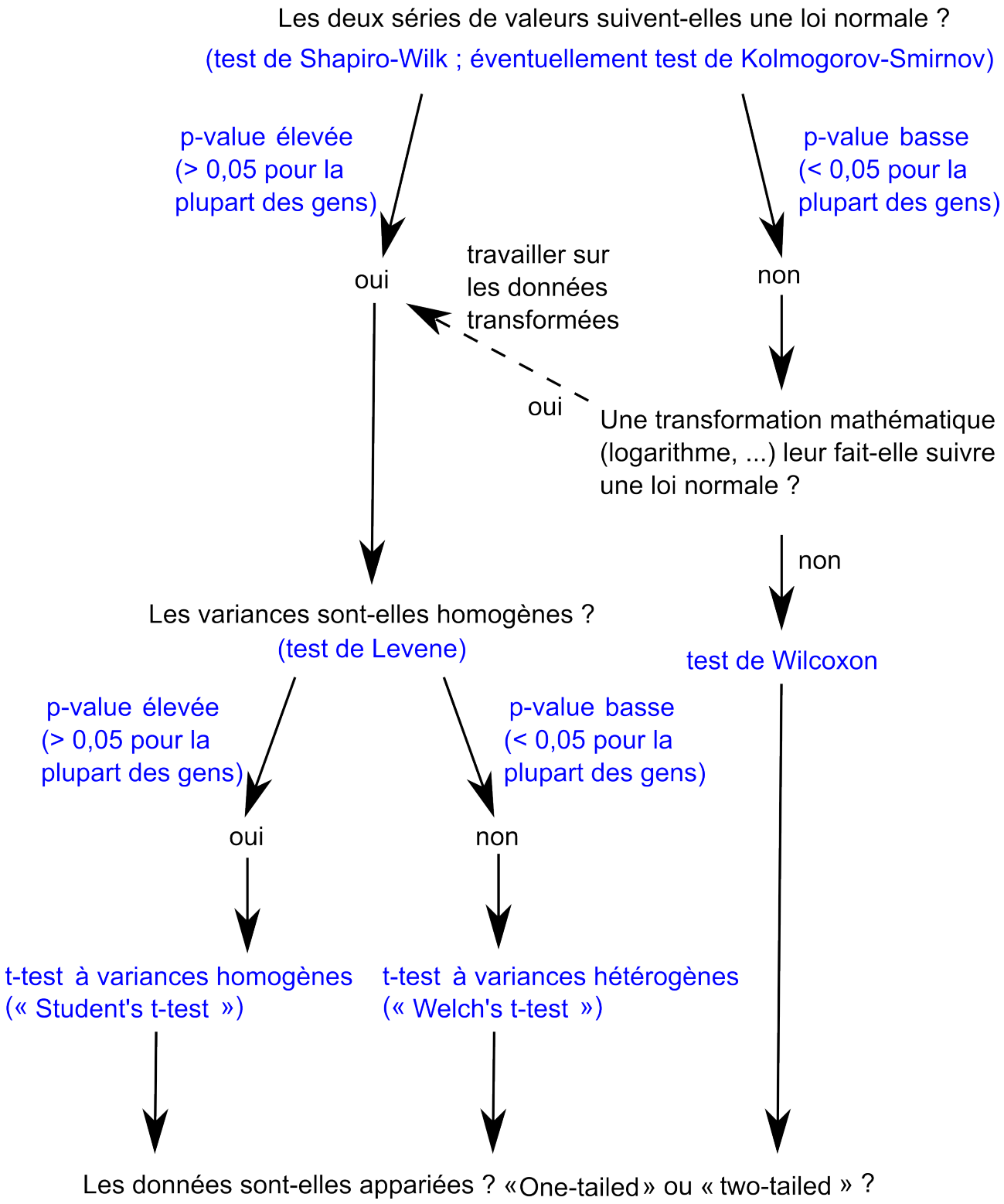

Figure 1 :
Cette figure résume la démarche à suivre pour comparer deux jeux de données. Noter que plusieurs tests sont disponibles pour la plupart des questions (la normalité peut se tester avec les tests de Shapiro-Wilk, et de Kolmogorov-Smirnov ; les deux donnent généralement le même résultat, mais en cas de désaccord, privilégier le test de Shapiro-Wilk, plus précis ; l'homogénéité des variances peut se tester avec d'autres tests que le test de Levene – ils donnent généralement la même réponse).