ENSEIGNER LE RECUEIL DES DONNÉES : EXPLORER LA VARIABILITÉ BIOLOGIQUE... AU CHAUD, DANS UNE SALLE DE COURS

Anne-Béatrice DUFOUR[1], Jean R. LOBRY[2] et Isabelle AMAT[3]

TITLE

Teaching data collection : exploring biological variability, comfortably installed in a university classroom

RÉSUMÉ

Enseigner la statistique en biologie, c'est rentrer dans le monde de la donnée tout en s'adressant à des étudiants de cultures différentes et en les confrontant à trois disciplines : biologie, statistique, informatique. La donnée relève d'une question biologique souvent complexe qu'il faut aller collecter : ce processus nécessite du temps, de l'argent, de la sueur. Celui-ci est associé à la construction des échantillons, aux plans d'expérience mis en place et à la définition des variables. Une fois les données recueillies, il faut les nettoyer, les explorer, les analyser.
L'objectif de cet article est : (1) de sensibiliser les étudiants au recueil des données ; (2) d'exposer quelques initiatives et expériences mises en place par les enseignants et les chercheurs du département de biologie de l'Université Lyon 1. Que les jeux de données soient réels ou simulés, petits ou grands, clefs en main ou recueillis par les étudiants, ils ont pour vocation de s'inscrire dans une double démarche de compréhension : les concepts statistiques et le pourquoi biologique.

Mots-clés : *données, biologie, statistique, informatique.*

ABSTRACT

Teaching statistics in biology must be linked to the data collection and students must understand three disciplines simultaneously : biology, statistics and computer science. The issue is a biological question that is often complex and requires to collect data : the process of data collection requires time, money, sweat. It is associated with sampling, experimental design and definition of variables. Once the data are collected, they must be cleaned, explored and analyzed.
The aim of this article is (i) to educate the students to the collection of the data, (ii) to expose some initiatives and experiments performed by teachers and researchers of the department of biology of the University Lyon 1. The data sets are real or simulated, small or large, turnkey or collected by the students. Their aim is to play a double approach to the statistical concepts and the biological framework.

Keywords: *data set, biology, statistics, computer science.*

1 Introduction

Historiquement, à l'Université Lyon 1, l'enseignement de la statistique en biologie est dispensé par des mathématiciens et des biologistes assurant leur recherche au Laboratoire de Biométrie

[1]Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR5558, anne-beatrice.dufour@univ-lyon1.fr
[2]Idem, jean.lobry@univ-lyon1.fr
[3]Idem, isabelle.amat@univ-lyon1.fr

et Biologie Evolutive (LBBE). Pour J.-M. Legay, son fondateur, l'acquisition des données est un enjeu important dans la démarche scientifique (Legay et Schmid, 2004). Tout mathématicien entrant dans son laboratoire se devait de suivre un enseignement intitulé « Mathématique Appliquée à la Biologie » qui comportait des travaux pratiques (TP) associés au recueil de données biologiques. Un TP consistait, par exemple, à mesurer plusieurs variables sur des fémurs humains afin (1) de caractériser leur sexe et (2) prédire le sexe des individus dont l'information était inconnue ; un autre TP visait à classer différentes variétés de pommes de terre, encore chaudes dans leur casserole de cuisson, selon leur qualité gustative.

Mais la sensibilisation à la donnée et donc le recueil des données doit-il être enseigné par le statisticien ou par le biologiste ? Cette question est complexe. Le statisticien, attaché aux concepts statistiques, souhaite les mettre en évidence à travers des exemples motivant les étudiants. Le biologiste, attaché à la problématique qu'il expose, souhaite développer ses hypothèses de travail et montrer aux étudiants comment les résoudre à l'aide des outils statistiques. Même si nous adoptons ce dernier point de vue, certaines difficultés subsistent car la biologie est multiple : écologie, microbiologie, botanique, génomique, etc.

Biologie et statistique sont liées. Les données réelles ne sont pas là seulement pour dire comment les analyser mais pourquoi on les analyse (Neumann *et al.*, 2013). Ceux qui enseignent la statistique aux étudiants de biologie à l'Université Lyon 1 sont à la fois biologiste et statisticien, l'un plus que l'autre. Enseigner la statistique en biologie sans partir des données est contre-productif. C'est l'inscrire dans un cadre d'abstraction entraînant une frustration qui détourne les étudiants de son apprentissage (Neumann *et al.*, 2013).

Les étudiants viennent en Sciences de la Vie par passion pour un modèle biologique (serpents, oiseaux, etc.), pour un métier (microbiologiste, bioinformaticien, écologue, etc.). Si le cours de mathématique et/ou de statistique ne s'articule pas avec la biologie, il devient disciplinaire, un passage obligé. Alors, pour certains, les concepts présentés sont mal assimilés, mal compris, oubliés. Les étudiants les retrouvent, en découvrent de nouveaux lors de stages en laboratoire, parfois tardivement dans leur cursus. Ils n'ont plus le même temps d'apprentissage. S'ensuit une mauvaise utilisation dont nous sommes en partie responsables. Equilibrer les concepts et les techniques n'est pas aisé. La donnée peut être le point d'ancrage.

Notre objectif est de témoigner de l'évolution de l'enseignement de la statistique en biologie grâce aux données réelles, qu'elles soient collectées par les étudiants eux-mêmes ou qu'elles proviennent de publications issues des activités de recherche. Cette évolution est liée à la démocratisation des moyens de calcul, à la croissance du nombre d'étudiants et de leur niveau. En effet, les données diffèrent de la licence au master. Elles restent assez générales et faciles à appréhender dans les premières années permettant ainsi de se focaliser sur la méthodologie statistique. Elles se complexifient dans les dernières années de formation car elles sont reliées à des thématiques de recherche : biologie et statistique s'entremêlent.

2 De la complexité structurelle

A l'Université Lyon 1, tous les étudiants en Sciences de la Vie reçoivent obligatoirement en première année de licence (L1) un cours de « Mathématiques pour les sciences de la vie » (cf. [9]) et en seconde année (L2), depuis la nouvelle habilitation 2016-2020, un cours de « Biostatistiques

et Bionformatique ». Il s'agit d'un enseignement de masse avec environ 1000 étudiants en L1 et 700 en L2. A la fin de la deuxième année de licence, les étudiants peuvent choisir une des orientations suivantes appelées parcours : (1) Biosciences, (2) Biochimie, (3) Génétique et biologie cellulaire, (4) Microbiologie, (5) Physiologie, (6) Sciences de la biodiversité et (7) Bio informatique, statistique et modélisation (noter ici que le terme Bio fait référence à la fois à l'informatique, à la statistique et à la modélisation mathématique). Nous intervenons principalement dans les parcours 6 et 7.

Quel que soit le niveau, quel que soit le choix au cours de son parcours universitaire, l'étudiant en biologie est parfois surpris par cet enseignement. Il lui faut attendre les premiers stages d'initiation à la recherche (troisième année de licence ou première année de master) pour comprendre l'importance de la statistique et de l'informatique dans la résolution de la question biologique qu'il étudie. C'est pourquoi nous défendons la démarche de travailler sur des données réelles dès le début de la formation afin de passer de la réalité biologique à la pratique statistique (figure 1).

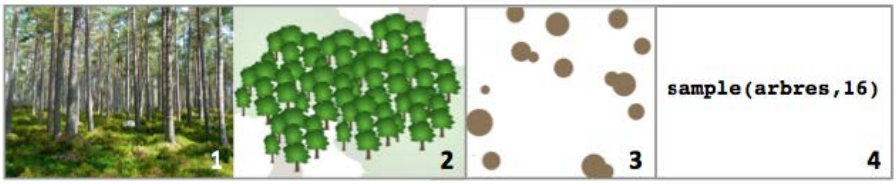

FIGURE 1 – *De la réalité biologique à la pratique statistique : de la forêt (1) à sa représentation (2) ; de la visualisation d'un échantillon de 16 circonférences d'arbres (3) à son expression informatique (4)*

Dans les deux premières années de licence, compte-tenu du nombre important d'étudiants, les travaux dirigés restent classiques : 36 étudiants dans une salle de classe normale, papier, calculette. Mais certaines notions, comme par exemple la distribution d'échantillonnage, sont travaillées sous la forme de travaux tutorés, en salle informatique. Exit les sommes des x, les sommes des x au carré, les séries statistiques à 10 ou 20 individus. Les étudiants échantillonnent des arbres dans des forêts par tirage aléatoire ou par stratification simple à l'aide d'images 2D (cf. image 2 de la figure 1), notent les circonférences des arbres retenus et calculent les moyennes et variances. Les résultats des différents groupes sont rassemblés et discutés. Pour ne pas complexifier l'apprentissage des concepts avec de la programmation informatique, une démarche interactive est rendue possible avec `Rmarkdown` (cf. [13]) par la création de documents dynamiques sous `R` (cf. [8]). Cette approche est appréciée mais elle nécessite un volume horaire important, l'intervention de nombreux encadrants et l'utilisation d'un grand nombre de salles informatiques.

Désormais, les jeux de données peuvent être simulés ou réels, petits ou grands. Tous ont leur utilité. Démontrer un concept statistique avec des données réelles peut demander un temps de préparation très important, les simuler, quelques secondes. Comme le soulignent Singer et Willet (1990), construire une variable normalement distribuée avec une ou deux valeurs aberrantes n'est qu'une question simple de programmation ; trouver un jeu de données avec ces mêmes caractéristiques peut prendre plusieurs heures. Si les petits jeux de données permettent de mieux appréhender la valeur de la mesure, les données réelles motivent les étudiants et faci-

litent leurs apprentissages (Unwin, 2010). Mais d'où proviennent-elles ?

3 Des données réelles

Parmi les chercheurs et les étudiants que nous formons, certains recueillent des données sur le terrain (dans les Alpes, en Afrique...), d'autres en laboratoire. D'autres encore ne voient jamais la réalité de cette quête, assis derrière un ordinateur, face à un ensemble de données ou une base de données. Il nous paraît important que tous les étudiants soient confrontés, au moins une fois dans leur cursus universitaire, à la collecte des données : quelle mesure choisir pour répondre à la question posée ? Comment la prélever ? Quelle variabilité comporte-t-elle ? Tout ceci avant qu'elle soit analysée statistiquement. Après ce travail (entre 3 et 6 heures), les étudiants comprennent qu'une valeur ou une modalité dans un tableau est plus qu'une mesure ou une observation. Voici deux expériences que nous proposons aux étudiants de L3 et de M1 (figure 2).

Dans la librairie `MASS` du logiciel `R` (R Core Team, 2016) se trouve un jeu de données recueillies auprès de 237 étudiants de l'Université d'Adelaïde (Australie) portant sur la mesure de quelques variables morphologiques (âge, poids, taille, empan) et l'observation de quelques mouvements associés à la latéralité (main d'écriture, lancer de ballon, etc.). Nous proposons à des étudiants en troisième année de licence de recueillir ces données pour leur année de promotion (entre 20 et 30 individus), de les ajouter aux promotions précédentes et de les stocker dans un tableur. Se posent alors les questions du comment recueillir, comment stocker et quelles unités conserver (Dufour et Lobry, 2008). Puis nous proposons de comparer les étudiants des deux universités. Depuis deux ans, nous avons initié une autre procédure pour le recueil de ces données. Chaque étudiant dessine l'empreinte de sa main, posée bien à plat sur une feuille, les doigts écartés au maximum. Puis chacun mesure (1) l'empan de sa main, c'est-à-dire la distance entre le pouce et l'auriculaire, puis (2) l'empan de la main dominante de chacun de ses camarades. Cela permet alors d'aborder des questions sur les instruments de mesure, la variabilité intra- et inter-opérateurs.

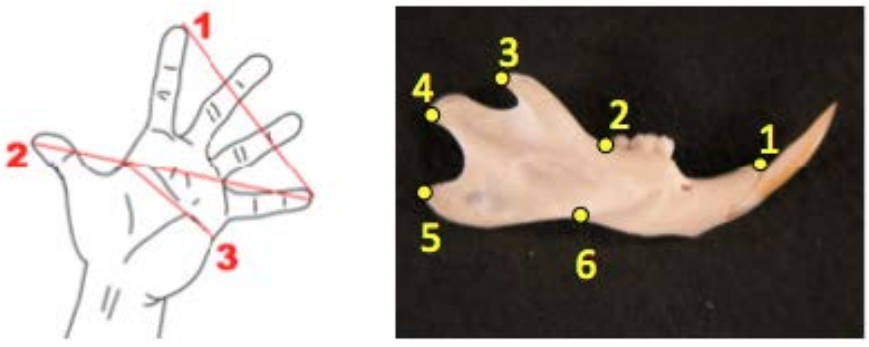

FIGURE 2 – *Recueil de données : mesure de l'empan chez l'homme (in Wikipedia) et caractérisation de la mandibule chez le mulot (Renaud et Millien, 2001)*

Suite à un cours sur l'évolution proposée par S. Renaud, chercheur au LBBE et un cours

d'analyse multivariée comprenant l'analyse en composantes principales (ACP) donné par A.-B. Dufour, des « travaux dirigés » (TD) ont été proposés à une trentaine d'étudiants. Les deux intervenantes étaient dans la même salle : l'une expliquant les objectifs de la recherche et la façon de recueillir les données, l'autre, la façon de les analyser. L'objectif général de l'étude est la caractérisation des patrons de différentiation morphologique chez deux espèces de mulots via l'étude des contours de leur mandibule (Renaud et Millien, 2001). Comme le montre la figure 2, les étudiants doivent sélectionner des points sur les mandibules, choisir leur position et leur nombre pour bien décrire la forme. Une fois les données recueillies, ils réalisent une ACP dont l'interprétation est discutée conjointement par les enseignants et les étudiants. Cette initiative, très constructive et très motivante, possède un coût (deux enseignants dans une même salle) qu'il est difficile de reconduire régulièrement à l'Université.

Comme le montre cet exemple, les données utilisées pour nos propres recherches ou celles de collègues sont le point d'ancrage de l'interaction entre la statistique et la biologie dont nous avons tant besoin pour animer nos TD ou TP. Une autre solution est de travailler ensemble à la construction de TD. Quelques fiches sont proposées sur le site d'enseignement de la statistique en biologie (cf. [11]) comme par exemple une familiarisation à la planification expérimentale autour des insectes vecteurs de la maladie de Chagas (Menu *et al.*, 2010). Les étudiants peuvent alors s'approprier les problématiques des études proposées sans recours aux articles originaux généralement écrits en anglais.

Les données réelles présentent également l'intérêt de mettre les étudiants en phase avec leur vie professionnelle future. Ils se confrontent ainsi à des données brutes comportant des valeurs aberrantes, des valeurs manquantes et nécessitant une réflexion sur leur qualité, c'est-à-dire la façon dont elles ont été obtenues et le pourquoi. Des jeux de données peuvent ainsi être trouvés sur Internet (par exemple, météorologie, pollution de l'air), dans les revues scientifiques qui exigent, pour la publication, de les fournir ou encore dans des librairies comme DASL (Data and Story Library), dans des ouvrages dédiés (Andrews et Herberg, 1985 ; Hand *et al.*, 1984). D'autres sont proposés, par exemple, dans les librairies du logiciel R ou d'autres logiciels dédiés à la statistique.

Enfin, sur le site pédagogique d'enseignement de la statistique en biologie (cf. [11] et Dufour, 2012), de nombreuses méthodes statistiques sont illustrées par des données accessibles à tous. Celles-ci sont fournies gracieusement par les chercheurs, en totalité ou en partie. Des fiches explicatives décrivent la problématique et les variables de l'étude ; elles sont nommées « problèmes pratiques de la statistique » (pps) et sont stockées dans le menu « Données » du dit site.

4 Conclusion

Des cours d'enseignement de recueil des données existent dans les formations universitaires ; leur mise en œuvre reste un défi à relever compte tenu de l'augmentation du nombre d'étudiants (les groupes de TD ne pouvant être démultipliés à l'infini) et de la diminution du volume horaire des cours de statistique en biologie. Mais c'est un défi que de nombreux enseignants sont prêts à relever pour augmenter l'intérêt et la motivation des étudiants. Des échanges interdisciplinaires se mettent en place, créant un enrichissement réciproque entre statistique, ou plus largement mathématique, et biologie (Chessel, 1992 ; Lange, 2000). Sous l'impulsion d'I. Amat, nous

sommes en train de construire un nouvel enseignement où les étudiants vont recueillir des données de séquençage haut débit pour décrire des communautés bactériennes avec un enseignant et vont les analyser avec un autre.

De par l'histoire du Laboratoire de Biométrie et Biologie Evolutive, la question des données est au cœur des enseignements. L'objectif de cet article était de proposer des exemples de mises en situation expérimentées au sein de la filière biologie de l'Université Lyon 1. La richesse des données accumulées sur le site d'enseignement de la statistique en biologie peut servir à d'autres enseignants. Les premiers développements interactifs en `Rmarkdown` peuvent inspirer d'autres expériences et semblent une bonne alternative notamment pour les premières années d'université.

Avec l'explosion d'internet, les données sont plus accessibles. Avec les développements technologiques, les jeux de données deviennent de plus en plus importants. Nous sommes confrontrés aux « big data », données ne pouvant être stockées sur une même machine et dont les traitements eux-mêmes nécessitent plusieurs machines (Saporta, 2017). A la biologie et la statistique, s'ajoute la mise en œuvre informatique de plus en plus complexe. Il nous faut préparer les étudiants à ces nouveaux défis (Saporta, 2012) par une réflexion sans cesse argumentée sur la nature et l'usage des données.

En conclusion, enseignants-chercheurs et chercheurs se mobilisent pour initier de nouvelles méthodologies pour le recueil des données afin d'explorer la variabilité biologique, bien au chaud, dans une salle de cours. Par ce clin d'œil, nous remercions tous les collègues qui ont accepté et acceptent de rendre publique une partie de leurs données pour l'enseignement.

Références

[1] Andrews, D.F. and A.M. Herzberg (1985), *Data. A Collection of Problems from Many Fields for the Student and Research Worker*, Springer Series in Statistics, New-York.

[2] Chessel, D. (1992), *Echanges interdisciplinaires en analyse des données écologiques*, Mémoire d'Habilitation à diriger des recherches, Université Claude Bernard - Lyon 1.

[3] Dufour, A.B. et J.R. Lobry (2008), tdr316 : Variables estudiantines (examen de contrôle continu depuis 2003), *in* Enseignements de statistique en biologie, `http://pbil.univ-lyon1.fr/R/enseignement.html`

[4] Dufour, A.B. (2012), La part du logiciel R dans l'enseignement de la statistique en biologie. Le site web de Lyon, *Statistique et Enseignement*, **2**(2), 41–47.

[5] Hand, D.J., F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski (1984), *A Handbook of Small Data Sets*, Chapman & Hall, London.

[6] Lange, J.M. (2000), Les relations biologie / mathématique interrogent l'enseignement des sciences de la vie, *ASTER*, 30, 123–142.

[7] Legay, J.M. et A.F. Schmid (2004), *Philosophie de l'interdisciplinarité*, Editions PETRA, Paris.

[8] R Core Team (2016), *R : A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

[9] Mathématiques pour les sciences de la vie, `http://pbil.univ-lyon1.fr/mathsv/`

[10] Neumann, D.L., M. Hood, and M.N. Neumann (2013), Using Real-Life Data when Teaching Statistics : Student Perceptions of this Strategy in an Introductory Statistics Course, *Statistics Education Research Journal*, **12**(2), 59–70.

[11] Enseignements de statistique en biologie, `http://pbil.univ-lyon1.fr/R/enseignement.html`

[12] Renaud, S. and V. Millien (2001), Intra- and Interspecific Morphological Variation in the Field Mouse Species Apodemus Argenteus and A. Speciosus in the Japanese Archipelago : the Role of Insular Isolation and Biogeographic Gradients, *Biological Journal of the Linnean Society*, 74, 557–569.

[13] R Markdown. Dynamics documents for R, `http://rmarkdown.rstudio.com/`

[14] Menu, F., A.B. Dufour, E. Desouhant et I. Amat (2010), tdr335 : Densité de population et ingestion de nourriture chez un insecte vecteur de la maladie de Chagas, `http://pbil.univ-lyon1.fr/R/enseignement.html`

[15] Saporta, G. (2012), Il faut pouvoir répondre à l'invasion des données, *Sciences et Avenir*, 42–45.

[16] Saporta, G. (2017), Quelle statistique pour le Big Data ?, *Statistique et Société*, **5**(1), 31–36.

[17] Singer, J.D. and J.B. Willet (1990), Improving the Teaching of Applied Statistics : Putting the Data Back Into Data Analysis, *The American Statistician*, **44**(3), 223–230.

[18] Unwin, A. (2010), Workshop report. Datasets on the Web : a Resource for Teaching Statistics ?, *MSOR Connections*, **10**(3), 38–41.