Enseigner le recueil des données
Explorer la variabilité biologique ...
au chaud, dans une salle de cours.

Anne-Béatrice Dufour [1], Isabelle Amat [1] & Jean R Lobry [1]

[1] *Université de Lyon, Université Lyon 1, UMR CNRS 5558*
43 Bd du 11 novembre 1918, 69622 Villeurbanne Cedex, France
anne-beatrice.dufour@univ-lyon1.fr, isabelle.amat@univ-lyon1.fr,
jean.lobry@univ-lyon1.fr

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 : de la mathématique à la biologie en passant par l'informatique et vice versa. Enseigner le recueil des données relève de deux processus : le pour quoi et le comment. Le premier est biologique. Il donne toute sa puissance à la donnée, de par le temps, l'argent voire la sueur qu'on lui consacre. Le second est statistique. Il s'attache à bien définir la nature des variables, construire les échantillons ainsi que le plan de l'expérience. Une fois la collecte réalisée, une dernière tâche consiste à nettoyer les données avant analyse.
L'objectif de cette communication est de présenter, par quelques initiatives et expériences réalisées à l'université Lyon 1, comment sensibiliser les étudiants, de la licence au master, au recueil des données. Trois types d'expériences peuvent être distingués. Le premier est la réalisation d'exercices simples où la donnée est présente mais son intérêt biologique ténu. Le second est l'utilisation de jeux de données structurées, clés en main. Ces derniers sont obligatoirement associés au pour quoi et à la compréhension du contexte biologique. Le dernier type consiste à recueillir les données soi-même dans le cadre d'un cours de statistique ou d'une initiation à la recherche.
Pour conclure, tout cet apprentissage est rendu possible grâce à l'utilisation des logiciels et des nouveaux développements informatiques.

Mots-clés. enseignement de la statistique, recueil des données, échantillonnage, plan d'expérience, logiciel

Abstract. Teaching statistics in biology must be linked to the data collection for maths and biology students. The data collection is twofold: for what and how. The first one is biological. Data is valuable : it is time and money. The second is statistical: definitions of variables, samples and experimental designs.
The objective of this presentation is to show what is done at the University of Lyon 1 to raise awareness of students to the data collection. There are three distinct types of exercises. The first one is defined by classical exercises: the biological interest is small. The second one is the use of turnkey data sets. The students have to understand the

purpose of the biological question which led to that data. The last type of exercise is to collect its own data for a statistical course or an initial research assignment.
Finally, all this learning is made possible thanks to the use of the software and the new computing developments.

Keywords. teaching statistics, data collection, sampling, experimental design, software

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 et Biologie Evolutive. Pour J.M. Legay, son fondateur, l'acquisition des données est un enjeu important dans la démarche scientifique (Legay et Schmid, 2004 [1]). Tout mathématicien entrant dans son laboratoire, se devait de suivre un enseignement intitulé "Mathématique Appliquée à la Biologie" qui comportait des travaux pratiques associés au recueil de données biologiques, de la mesure de fémurs humains à la qualité gustative de différentes variétés de pomme de terre.

Puis le nombre d'étudiants en biologie augmentant régulièrement tandis que le nombre d'heures en statistique diminuait, la notion du recueil de données et les expériences associées ont petit à petit disparu. Emmener six cent cinquante étudiants sur le terrain n'est tout simplement pas possible. Cet enseignement a glissé de la licence vers le master 1 puis le master 2 voire pour certains, aujourd'hui, en thèse. Elle s'est diluée, s'est invitée dans d'autres disciplines.

Mais enseigner la statistique en biologie sans partir des données est contre productif. C'est l'inscrire dans un cadre d'abstraction qui entraîne une frustration chez les étudiants puis qui les détourne de son apprentissage. S'en suit alors une méconnaissance, une mauvaise utilisation dont nous sommes en partie responsable. C'est pour lutter contre cela que des enseignants-chercheurs, des chercheurs s'allient pour initier de nouvelles méthodologies de collecte de données afin d'explorer la variabilité biologique face à trente-cinq étudiants, bien au chaud, dans une salle de cours.

Notre objectif est d'exposer quelques expériences mises en place au près d'étudiants-test, en groupes de vingt ou trente et de montrer en quoi le dialogue entre enseignants et les développements informatiques récents peuvent aider à construire une nouvelle pédagogie. Nous distinguons trois types d'expériences, sans hiérarchie, tant sur le plan de la difficulté que sur celui de son enseignement à tel ou tel niveau de formation.

2 Comment parler des données ?

Des exercices élémentaires

En statistique, les notions d'échantillon et de population sont essentielles pour comprendre les principes de l'inférence. Mais comment extraire des individus d'une population, calculer les paramètres obtenus et discuter de la variabilité ?

Avec le logiciel R [2], cela peut être simple et ludique. La population est définie par des boules de couleur dans une urne-vecteur construit à cet effet. A l'aide de la fonction 'sample', chaque étudiant peut extraire un échantillon, constater que chacun a un résultat différent. Les enseignants peuvent alors poser les grands principes de l'inférence. Mais pour quelques étudiants, cela reste abstrait. Dématérialiser l'urne les rend non réceptifs. Afin de concrétiser l'expérience, des cercles de rayon peuvent être découpés dans une feuille de papier et symbolisés la circonférence des arbres d'une forêt. Les étudiants choisissent au hasard 10, 20 puis 30 cercles, calculent les moyennes et les variances obtenues. Ils peuvent aller écrire leurs résultats au tableau. L'enseignant donne les paramètres pour la population et initie l'échange.

Mais d'aucun pourrait voir dans ces exemples un manque de modernité. C'est certain. Dans l'enseignement de Mathématique pour les sciences du vivant dit MathSV [3], des travaux tutorés proposent aux étudiants de première année des animations flash pour comprendre le phénomène biologique et construire la démarche statistique, démarche qui peut être rendue interactive avec Rmarkdown par la création de documents dynamiques sous R [4].

Aux jeux de données clés-en-main

Lors de la construction du site pédagogique d'enseignement de la statistique en biologie [5], D. Chessel insistait sur le fait que chaque méthode devait être illustrée par des données accessibles à tous. Cette démarche volontaire pose le problème de la propriété des données et le respect que chaque utilisateur doit à celui qui les a recueillies.

Impliquer les étudiants dans l'étude d'un jeu de données est une expérience riche, rendue possible par le logiciel R (Dufour, 2012 [6]) L'enseignant décrit le contexte ayant conduit au recueil des données, suscite la curiosité des étudiants et peut répondre à celle-ci par des représentations graphiques, quelques calculs, élementaires ou non, selon le niveau d'étude.

Dans le cadre de travaux pratiques, il est parfois demandé aux étudiants de s'approprier un problème biologique pour lequel ils ont un intérêt, de rechercher sur le net des données et de les analyser. L'exercice est périlleux. Nous ne parlons pas de recueil de données

à proprement parlé mais de recueil de l'information ayant conduit à générer le tableau qu'ils ont téléchargé.

2.1 En passant par le recueil des données

Le respect de la donnée passe par le fait de la recueillir soi-même et de comprendre toute la difficulté du choix de la mesure, de la mesure elle-même et de la variabilité qu'elle comporte avant qu'elle soit analysée statistiquement.

Dans la librairie MASS du logiciel R, se trouve un jeu de données recueillies au près 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, 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 conservées. Puis nous proposons de comparer les étudiants des deux universités.

Depuis deux ans, nous avons initié un autre usage de recueil des 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 ces camarades. Cela permet alors d'aborder des questions sur les instruments de mesure, la variabilité intra et inter opérateurs.

L'association au sein d'un même cours de deux enseignants, l'un biologiste, l'autre statisticien, reste certainement la conjoncture la plus enrichissante. Deux expériences ont été réalisées. La première résulte de l'interaction entre un cours de biologie évolutive et un cours d'analyse multivariée ; l'autre résulte d'un échange autour de la construction d'un plan d'expérience pour une analyse de données microbiologiques.

3 Conclusion

Enseigner le recueil des données est une gageure compte tenu de l'augmentation du nombre d'étudiants (les groupes de travaux dirigés ne pouvant être démultipliés à l'infini) et de la diminution horaire des cours de statistique en biologie. Mais c'est un défi que de nombreux enseignants sont prêts à relever dans le cadre des nouvelles habilitations. Un enseignement des nouvelles technologies d'acquisition des données sera proposé dans le master Biodiversité, Ecologie, Evolution de Lyon. Des échanges interdisciplinaires se mettent en place. Comme le soulignait D. Chessel en 1992 [7], chacun arrive avec ses

connaissances pour générer un dialogue avec un autre dont le langage est inconnu. Cela reste vrai et il est illusoire de penser qu'un enseignant, seul, peut tenir aujourd'hui les deux disciplines ensemble.

Notre objectif était de proposer des exercices, des mises en situation expérimenter au sein de la filière biologie de l'université Lyon 1 et du laboratoire de Biométrie. 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. L'exposé a pour vocation d'ouvrir le débat et d'initier une discussion sur la donnée d'aujourd'hui.

En effet, avec l'explosion d'internet, la question du recueil des données devient centrale. L'enseigner, c'est faire comprendre les enjeux économiques et politiques qu'elles contiennent. Comme le soulignait J.M. Legay en 2004, on se bat pour ces informations, on les achète, on les échanges, on les vole s'il le faut. Enseigner le recueil des données, c'est aussi rappeler qu'elles sont inscrites dans un contexte biologique. Que valent-elles hors de ce contexte ? Les innovations technologiques conduisent à générer des masses de données. Ne faut-il pas préparer nos étudiants à répondre à l'invasion de ces dernières (Saporta, 2012, [8]) par une réfléxion argumentée sur leur nature et leur usage ?

Bibliographie

[1] Legay, J.M. et Schmid, A.F. (2004) *Philosophie de l'interdisciplinarité*, Editions PETRA, Paris.
[2] R Development Core Team (2010), *R: A language and environment for statistical computing.*, R Foundation for Statistical Computing, Vienna, Austria.
[3] Mathématiques pour les sciences de la vie. *http://pbil.univ-lyon1.fr/mathsv/*
[4] R Markdown. Dynamics documents for R. *http://rmarkdown.rstudio.com/*
[5] Enseignements de statistique en biologie. *http://pbil.univ-lyon1.fr/R/enseignement.html*
[6] 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
[7] Chessel, D. (1992), *Echanges interdisciplinaires en analyse des données écologiques*, Mémoire d'Habilitation à diriger des recherches, Université Claude Bernard - Lyon 1.
[8] Saporta, G. (2012), Il faut pouvoir répondre à l'invasion des données, *Sciences et Avenir*, 42–45.