UNE APPROCHE DE TRANSPOSITION DIDACTIQUE POUR L'ENSEIGNEMENT UNIVERSITAIRE DU MODÈLE DE RÉGRESSION LINÉAIRE EN STATISTIQUE

Antoine Rolland

UNE APPROCHE DE TRANSPOSITION DIDACTIQUE POUR L'ENSEIGNEMENT UNIVERSITAIRE DU MODÈLE DE RÉGRESSION LINÉAIRE EN STATISTIQUE

Antoine Rolland*

A DIDACTICAL TRANSPOSITION APPROACH FOR UNDERGRADUATE TEACHING LINEAR REGRESSION MODELS IN STATISTICAL TRAININGS.

Abstract – Most of the didactic works about statistics in higher education focuses on the issue of introducing statistics to supposedly refractory students. On the contrary, we focus in this paper on statistics training as a matter of interest, by questioning the educational choices made by teachers in these specific field. Through the example of linear regression, we analyze these choices by studying course documents in six training courses at different levels. We mobilize the formal framework of didactical transposition. On the one hand, this study highlights the existence of a common body knowledge on the subject, and on the other hand it shows a "knowledge to be taught" in tension between generalist theory and practical application, which tension is specific to statistical science.

Key words: didactical transposition, statistics, higher education, linear regression

UN ENFOQUE DE TRANSPOSICIÓN DIDÁCTICA PARA LA ENSEÑANZA UNIVERSITARIA DEL MODELO DE REGRESIÓN LINEAL EN ESTADÍSTICA

Resumen – La mayor parte de los trabajos sobre la didáctica en estadística para la enseñanza superior se centra en el reto de presentar la disciplina a estudiantes a priori reticentes. En este documento el abordaje es otro, cuestionando las elecciones educativas realizadas por los profesores en estos campos específicos. A través del ejemplo de la regresión lineal, y utilizando el marco teórico de la transposición didáctica, analizamos estas elecciones mediante el estudio de los materiales didácticos de seis cursos de formación en diferentes niveles. Este estudio destaca por un lado la existencia de un "saber sabio" común, y por otro

* Laboratoire ERIC, EA 3083, et IUT, Université Lumière Lyon II, Université de Lyon, antoine.rolland@univ-lyon2.fr

lado un “saber para ser enseñado” resultante de una tensión entre teoría generalista y aplicación práctica, tensión propia de la ciencia estadística.

Palabras-claves: transposición didáctica, Estadística, enseñanza superior, regresión lineal

RÉSUMÉ

La plupart des travaux sur la didactique de la statistique dans l’enseignement supérieur se focalisent sur l’enjeu de présenter la statistique à des étudiants a priori réfractaires. Au contraire, nous interrogeons les choix pédagogiques effectués par les enseignants dans les filières de formation en statistique. A travers l’exemple de la régression linéaire, et en mobilisant le cadre de la transposition didactique, nous analysons ces choix par l’étude de documents de cours dans six formations de différents niveaux. Cette étude met en évidence d’une part l’existence d’un « savoir savant » partagé, et d’une autre part un « savoir à enseigner » témoin d’une tension entre théorie généraliste et application pratique, tension propre à la science statistique.

Mots-Clés : transposition didactique, statistique, enseignement supérieur, régression linéaire

INTRODUCTION

La didactique de la statistique, et plus largement les travaux de réflexion de praticiens sur l’enseignement de la statistique représentent un champ marginal (en quantité) mais cependant actif de la didactique mathématique, comme l’indique Rolland (2020) dans sa rétrospective de l’activité francophone de ce champ sur la période 2009-2019.

Il est intéressant de noter qu’il existe relativement peu de travaux d’analyse de l’enseignement de la statistique concernant directement l’enseignement supérieur, et très peu de revues s’intéressent à cette question : les revues *Statistique et Enseignement* en français[1] et *Journal of Statistic Education*[2] en anglais sont à peu près les seules à publier de manière substantielle de tels travaux. Ce sont des revues accueillant quelques travaux de

[1] Cette revue a cessé de paraitre en 2019 et constitue maintenant une rubrique de la revue *Statistique et Société*.

[2] Devenu *Journal of Statistics and Data Science Education* en 2021

recherches en didactique, mais surtout des articles de praticiens partageant études de cas, descriptions d'expériences et outils pratiques. La revue *Statistics Education Research Journal* s'intéresse également à l'enseignement de la statistique dans le secondaire et dans l'enseignement supérieur, essentiellement auprès d'étudiants non spécialistes de la statistique.

Parmi les travaux nous intéressant directement, nous pouvons citer Gattuso (2011) qui étudie le statut de la statistique comme discipline scolaire ou Régnier (2012) et Hahn (2015) qui proposent des états de l'art de la didactique de la statistique, ou encore Tishkovskaya et Lancaster (2012) qui présentent les défis auxquels est confronté l'enseignement de la statistique dans le supérieur.

La plupart des travaux s'intéressant à la didactique de la statistique étudient la question des blocages, des difficultés et des préjugés envers la statistique d'étudiants non statisticiens. Comme l'indiquent Tishkovskaya et Lancaster (2012, p. 4),

> *Over the years there has been strong anecdotal evidence that students at university develop antipathy towards statistics and, typically, students at all levels lack interest in learning when taking introductory statistics courses. This is closely related to the problematic area of service teaching statistics to non-specialists which is addressed by many statistics education researchers.*

Nous trouvons des exemples de telles études dans de nombreuses filières spécifiques : par exemple Bihan-Poudec (2012) s'intéressent à la représentation sociale initiale de la statistique pour 600 étudiants en sciences de l'éducation ; Carillo et al. (2016) font de même pour des étudiants en école de management ; mais aucune étude ne concerne des étudiants en statistique.

Dans la plupart des articles en français consultés, les auteurs privilégient le partage d'expérience, ou la description d'outils pédagogiques, mais ne mènent pas de réelle étude didactique. En sciences et techniques pour les activités physiques et sportives (STAPS), Genolini et Driss (2010) montrent comment ils essaient de motiver les étudiants à l'utilisation de la statistique via des exemples grands publics et non disciplinaires. En sciences de la vie et de la terre, et en épidémiologie, les partages d'expériences présentent des outils et techniques à même d'aider les étudiants.

Palm et Allagbe (2011) présentent un didacticiel pour mieux comprendre les notions de base en statistique. Senterre et al. (2011) décrivent un dispositif d'enseignement à distance avec tuteur dans le cadre des cours de statistique pour épidémiologiste. Dufour et al. (2017) et Jutand et al. (2017) s'intéressent au recueil des données, par la description d'une expérience mettant en avant la variabilité statistique pour les premiers, par la présentation des différentes stratégies pédagogiques (cours théoriques, exercices d'application, apprentissage par projet, lectures critiques) pour les secondes. Zendrera et al. (2017) s'intéressent également au recueil de données dans le domaine des sciences humaines et sociales, en pointant l'absence de sensibilisation à cette question comme une lacune faisant passer (à tort) la statistique pour un domaine purement calculatoire. En psychologie, Cañadas et al. (2012) étudient les compétences nécessaires pour effectuer le test du χ^2, test d'hypothèse très utilisé en psychologie, et en déduisent des implications pour l'enseignement de cette notion. Gélinas et al. (2018) présentent pour leur part une expérience permettant à des étudiants en psychologie de programmer des tutoriels interactifs pour les aider en compréhension des concepts statistiques.

Dans le même état d'esprit, l'état de l'art de Garfield et Ben-Zvi (2007) est dominé par l'apprentissage des bases de la statistique dans l'enseignement secondaire, l'analyse générale de Schwab-McCoy (2019) s'intéresse à l'enseignement de la statistique au service d'autres disciplines « clientes », ou les travaux de Nikiforidou, Lekkab et Pangec (2010) sur la définition d'un curriculum de statistique sont entièrement axés sur la « littéracie statistique », autrement dit la statistique pour les non-statisticiens.

Armatte (2010), à l'instar de ce qui existe en didactique des mathématiques plaide pour une approche historique dans l'enseignement de la statistique, combinée à l'approche axiomatique, l'approche expérimentale et l'approche problème. Cette tension entre approche formelle (axiomatique) et approche expérimentale (centrée sur les données et les pratiques) est également développée dans le contexte des écoles de management par Hahn et Stoltz (2013). Lahanier-Reuter (2012, p 29) met en évidence le décalage existant dans l'enseignement de la statistique et des autres branches des mathématiques dans le secondaire :

> *En statistique les élèves disent avoir appris moyenne, médiane (parfois), pourcentages et fréquences. Mais ce qui est*

> *plus intéressant peut- être est le fait que ces élèves ne citent jamais de théorèmes pas plus de lois ou tout autre contenu qui pourrait être assimilé à une règle établie, ce qui n'est pas le cas pour les réponses en mathématiques ou en analyse.*

Ce débat entre la prépondérance de l'aspect appliqué ou de l'aspect théorique de la statistique se retrouve dans l'enseignement de la statistique à des non-statisticiens, mais est également au cœur de la science statistique en elle-même. Le classement de la statistique dans la section 26 du Conseil National des Université « mathématiques appliquées et applications des mathématiques », témoigne de l'importance donnée à l'aspect pratique de la statistique, quand l'existence d'un groupe de travail nommé « statistique mathématique » au sein de la Société Française de Statistique montre au contraire son attachement à des fondements théoriques solides.

Nous n'avons trouvé que deux auteurs proposant des travaux dans le domaine des filières d'enseignement supérieur formant des statisticiens, ou à tout le moins des personnes à l'aise en statistique. Besse et Laurent (2016) proposent une analyse de l'évolution des enseignements au sein d'une filière statistique. El M'hamedi, à travers trois articles, présente quelques aspects de l'incompréhension des tests statistiques : difficultés posées par le vocabulaire et par le langage probabiliste des tests chez des licenciés scientifiques (El M'hamedi, 2019a), incompréhensions des tests statistiques par les enseignants stagiaires de mathématiques du cycle secondaire (El M'hamedi, 2019b), et plus récemment difficultés de compréhension du concept de niveau de signification par le même public (El M'hamedi, 2021). Nous constatons donc qu'aujourd'hui la question de l'enseignement de la statistique de manière spécifique pour des futurs statisticiens n'est que très peu abordée dans la littérature. Est-ce parce que les étudiants étant supposés être attirés par la matière, il n'y aurait pas à s'interroger sur la manière d'enseigner la statistique ? Nous pensons au contraire qu'une approche didactique spécifique mériterait d'être analysée et développée.

Nous nous proposons dans cet article d'explorer ce champ spécifique par un premier travail d'exploration, consistant en une analyse comparée de situations d'enseignements d'un même concept statistique dans des formations de différents niveaux. Cette première analyse nous permettra de dégager les invariants et les

différences dans la manière d'aborder le sujet considéré. Nous nous appuierons sur le concept de transposition didactique pour expliciter les choix conscients ou inconscients fait par chaque enseignant, et ainsi essayer de dégager des pistes de réflexions sur la manière dont inconsciemment ou non chaque enseignant adapte son enseignement au public spécifique auquel il s'adresse.

Nous avons choisi pour cette première approche de l'application du concept de transposition didactique à l'enseignement de la statistique de nous restreindre à un champ limité de cet enseignement. Plusieurs raisons nous ont guidés sur le choix de la régression linéaire :

- la diffusion large de la notion ; dans une perspective d'études comparées, il nous faut un panel conséquent de cours portant sur la même notion ;
- un enseignement reconnu comme classique ; la notion étudiée se doit d'être solidement établie, afin qu'un savoir savant standard existe (même de manière implicite) ;
- son enseignement à plusieurs niveaux de la formation, afin de mettre en avant les mécanismes différents de transposition didactique en fonction du public cible ;
- sa relative spécificité à un cursus statistique, afin de pouvoir mettre en avant, s'ils existent, les particularismes de l'enseignement de la statistique à des futurs statisticiens ;
- sa non trivialité, ou en tout cas son degré de complexité minimal, afin que l'exercice de transposition didactique soit nécessaire et réfléchi de la part de l'enseignant (même s'il est non formalisé).

Au vu de ces critères nous avons retenu la notion de régression, ou modélisation linéaire. Il nous semble que cette notion peut être un bon support à une première étude car le champ de la modélisation linéaire couvre une étendue de savoirs et savoir-faire allant de notions relativement simples pour le modèle de base à de nombreux développements complexes dans le cadre d'un modèle linéaire généralisé. En outre la modélisation linéaire est au fondement de la démarche de la modélisation statistique, et se trouve donc forcément enseignée dans une formation de statistique de niveau Bac+2 à niveau Bac+5.

A notre connaissance, il n'y a pas de travaux antérieurs portant spécifiquement sur l'enseignement de cette notion. Seuls Kasprowicz et Musumeci (2015) en font mention dans leur étude sur l'utilisation des points aberrants en régression linéaire.

Le travail présenté dans cet article est à considérer comme une première étape exploratoire d'un programme de recherche plus vaste visant à éclairer les enjeux de l'enseignement de la statistique aux futurs statisticiens. Nous ne nous intéressons pas ici à la réception des enseignements par les étudiants, mais nous étudions la construction de l'enseignement par l'enseignant. L'enseignement universitaire est représenté comme un lieu de grande liberté pédagogique, mais aussi, par contrecoup, comme un lieu de solitude[3]. Cette solitude dans la conception pédagogique des cours est à tempérer compte tenu de l'existence d'une communauté de pairs, formelle ou informelle, qui oriente le contenu des cours. Bosch et al. (2021) insistent par exemple sur la faible liberté réelle laissée à l'enseignant du supérieur dans la définition du savoir à enseigner. C'est cette tension entre le choix personnels et cadre commun extérieur dans la définition du contenu du cours que nous nous proposons d'étudier ici en mobilisant le cadre de la transposition didactique. Nous détaillons ce cadre dans la partie suivante, et nous explicitons également les concepts statistiques sur lesquels nous avons choisi de nous concentrer.

CADRE THÉORIQUE : TRANSPOSITION DIDACTIQUE ET RÉGRESSION LINÉAIRE

La transposition didactique

Avec Arsac (1992), nous adoptons la définition de la transposition didactique comme « le travail de fabrication qui permet de développer dans son originalité un savoir enseigné par rapport au savoir savant » (Arsac 1992, p. 11), compte tenu des diverses contraintes pesant sur l'enseignement.

Le concept de transposition didactique, développé par Chevallard (1985), permet de mettre en lumière l'existence simultanée de plusieurs niveaux de savoir correspondant à un domaine donné : d'une part le savoir savant, et d'autre part le savoir à enseigner et, ce qui lui est proche, le savoir enseigné.

Dans le cadre théorique défini par Chevallard, ce processus de transposition est le fait (en grande partie) de la noosphère, nom désignant un regroupement informel d'acteurs, pour la plupart institutionnels, qui vont délimiter le savoir savant, et figer et

[3] Poteaux (2013) relève que « *le travail d'équipe, quasiment incontournable en recherche, ne laisse-t-il pas place à la solitude dans les amphis ?* » (p. 11)

légitimer le passage du savoir savant au savoir enseigné. Dans le cadre universitaire, une noosphère formelle n'existe pas, ou simplement à l'état de trace. Les programmes, quand ils existent au-delà de l'intitulé du cours, sont écrits par ceux-là même qui vont les mettre en œuvre et se limitent généralement à une liste de quelques notions présentes sur la plaquette de la formation à destination des étudiants. Aucune préconisation pédagogique précise n'est liée au contenu des cours. Il n'y a pas, sauf rares cas comme le Diplôme Universitaire de Technologie de la spécialité « Statistique et informatique décisionnelle » (DUT STID[4]) d'instance nationale pilotant et tranchant le contenu de ces cours. Il n'existe pas de manuels scolaires mais des ouvrages de référence (du savoir savant), et des ouvrages destinés aux étudiants (cours et exercice) mais non rattachés à une formation ou un format de cours particulier[5].

De fait, le cadre pourtant ancien de la transposition didactique a été peu mobilisé dans l'analyse des pratiques enseignantes à l'université. Berthaume (2007) et Poteaux (2013) s'intéressent de manière large à la description empirique du savoir pédagogique disciplinaire des enseignants d'université. Plus récemment, de Husson et al. (2018) en physique, Bosch et Winsløw (2020) et Bosch et al. (2021) en mathématique ont mobilisé avec succès la notion de transposition didactique à l'université. Le constat général repose sur l'absence de noosphère formelle, et sur la grande liberté apparente mais faible liberté réelle laissée à l'enseignant dans le choix du savoir à enseigner. C'est aussi notre hypothèse pour la régression linéaire.

La délimitation du savoir savant de référence peut également être questionnée. Les travaux de Martinand (2001), cité par Léziart (2003, p. 84), ont montré que

> *la seule référence au savoir savant, dans la transposition didactique empêche de*

[4] Le DUT est devenu BUT (Bachelor Universitaire de Technologie) à la rentrée 2021, mais il existe toujours une commission pédagogique nationale chargée de la rédaction et de l'évaluation du programme pédagogique national.

[5] Par exemple dans l'avant-propos du livre de Husson et Pagès (2013), on lit *« Ce livre a été d'abord écrit pour les étudiants d'école d'ingénieurs, d'IUT ou BTS ou de l'université dans les filières des sciences de la vie, sciences sociales, sciences économiques, etc. »*, ce qui montre bien le côté bien peu spécifique de l'approche choisie.

> *penser certaines disciplines scolaires dans une perspective de formation générale.*

Dans cet esprit, il est donc indispensable d'élargir le savoir de référence aux pratiques de références sous toutes leurs dimensions : techniques utilisées, outils manipulés – par exemple pour nous les logiciels spécialisés –, mais aussi situations et problèmes-types, attitudes et habitus professionnel, etc. C'est ce que nous essayons de faire en prenant en compte comme savoir de référence non seulement un savoir savant développé dans les ouvrages de référence, mais également des pratiques standards issues du monde professionnel. Ceci est d'autant plus important que la science statistique se définit elle-même comme un champ des mathématiques appliquées. La transposition didactique est alors le passage de ce corpus de savoir savant dans sa double acception théorique et pratique au savoir à enseigner lui aussi dans sa double acception scientifique et technologique, comme l'indique Martinand (1989) cité par Arsac (1992, p.24) :

> *le rapport entre pratique de référence et activité scolaire correspondante constitue la transposition didactique, en un sens élargi puisqu'on ne se limite pas au seul savoir.*

Aux États-Unis, le rapport GAISE joue le rôle de prescripteur du savoir à enseigner pour l'enseignement secondaire. (voir GAISE (2016) pour la version la plus récente, et Fine (2013) pour une analyse en français de ce rapport). Parmi plusieurs exemples, Woodard et McGowen (2012) détaillent comment ils se sont appuyés sur le rapport GAISE pour mener à bien la refonte d'un cours d'introduction à la statistique en premier cycle universitaire, à destination d'étudiants non statisticiens. Plus largement, les manuels scolaires sont aussi constitutifs de cette noosphère informelle. C'est visible en filigrane chez Dunn et al. (2015) qui étudient les différences de définitions de notions statistique entre les manuels scolaires et les enseignants de lycée. Mais ces exemples ne concernent que le niveau secondaire, et confirment en creux qu'il n'existe pas de transposition didactique externe dans notre domaine d'étude, qui est l'enseignement universitaire de la statistique dans des filières de statistique.

Régression linéaire

En statistique, la régression (ou modélisation) linéaire consiste à considérer une variable d'intérêt Y comme le résultat d'une

fonction linéaire de plusieurs variables explicatives X_i. Formellement la modélisation linéaire d'une variable Y en fonction de p variables X_i s'écrit

$$Y = X\beta + \varepsilon$$

où X est le vecteur $(1, X_1, \ldots, X_p)$ des variables explicatives, β le vecteur des $p+1$ coefficients, et ε un terme aléatoire appelé écart ou résidu, modélisant la différence entre la valeur prédite et la valeur observée de la variable Y.

Dans le cas de la régression à une variable, le modèle s'écrit

$$y = \beta_0 + \beta_1 x + \varepsilon$$

et la droite $\hat{y} = \beta_0 + \beta_1 x$ est appelée « droite de régression ».

Dans le cadre du modèle linéaire à une variable sur un jeu de données connues, la régression est introduite comme un problème d'optimisation consistant à trouver la droite qui minimise la somme des carrés des écarts entre les valeurs observées y et les valeurs $\hat{y}$ estimées à partir du modèle. En utilisant une représentation graphique d'une droite dans un nuage de points (figure 1), et en indiquant visuellement où se situe l'écart (figure 2), on « voit » comment une droite peut être optimale au sens des moindres carrés.

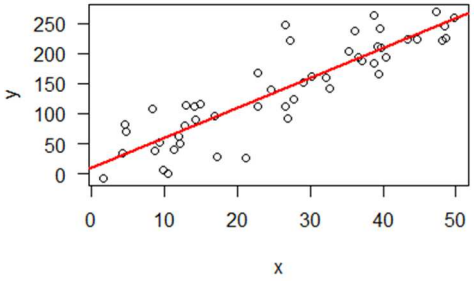

Figure 1-. Nuage de points et droite de régression de la variable y par rapport à la variable x. En rouge la droite de régression $\hat{y} = \beta_0 + \beta_1 x$

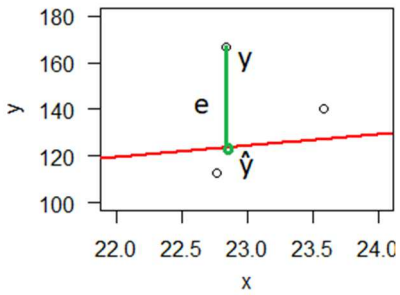

Figure 2.- Exemple visuel pour un point donné de la mesure de l'écart e entre y et ŷ (zoom de la figure 1). La droite de régression des moindres carrés est la droite qui minimise la somme des carrés des écarts pour le nuage de points.

Le cadre de la régression linéaire suppose que les résidus ε sont indépendants, et suivent une loi de distribution de variance stable tant vis-à-vis de la variable à expliquer que vis-à-vis des variables explicatives. Le cadre gaussien suppose en plus que la distribution des résidus suit une loi normale. Un certain nombre de résultats ne dépendent pas de la normalité des résidus ; c'est le cas par exemple de l'estimation des coefficients β du modèle obtenu par la méthode des moindres carrés, ou de l'estimation de la variance des résidus. D'autres résultats découlent du cadre gaussien : estimation des coefficients par la méthode du maximum de vraisemblance, loi de distribution des coefficients du modèle, etc. En particulier, le modèle gaussien permet de procéder à des tests de significativité des paramètres du modèle, c'est-à-dire de déterminer si les paramètres estimés sont statistiquement significativement différents de zéro ou non ; en d'autres termes, si les variables explicatives ont une influence statistiquement significative sur la variable que l'on cherche à expliquer.

Le coefficient de détermination R^2 d'un modèle de régression est un indicateur du pouvoir explicatif du modèle, c'est-à-dire l'explication de la variable Y par les variables X_i. Il est défini comme le ratio entre la somme des carrés des écarts expliqués par le modèle et la somme totale des carrés des écarts.

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

Cependant, dans le cadre bivarié, le coefficient de détermination peut être aussi vu comme le carré du coefficient de corrélation, noté généralement ρ et défini comme le rapport entre la covariance de X et Y et le produit des variances de X et Y.

$$R^2 = \rho^2 = \frac{{\sigma^2}_{XY}}{{\sigma^2}_X {\sigma^2}_Y}$$

Dans le cas multivarié, le modèle de régression n'est pas unique et dépend des variables explicatives choisies : l'influence d'une variable X_i particulière sur la variable Y n'est pas indépendante de la présence ou non d'autres variables explicatives. Le choix d'un modèle parmi plusieurs modèles explicatifs possibles est alors une des tâches du statisticien. La démarche de choix d'un modèle de régression particulier est un exemple de savoir pratique, et non théorique, de la régression. Cette démarche fait appel à un certain nombre de points d'attention qui nécessitent une expertise métier de la pratique pour arriver à proposer le meilleur modèle possible, ou en tout cas un modèle acceptable. Au-delà de la simple vérification des hypothèses de modélisation, le choix de modèle se caractérise par des décisions à prendre en présence de critères subjectifs. Le savoir savant est ici un savoir-faire, nourri d'expériences et de bonnes pratiques autant que de fondements théoriques.

Face à un modèle de régression multivariée, le statisticien doit vérifier un certain nombre d'hypothèse pour que le modèle soit valide. Mais ces vérifications sont rarement évidentes : la normalité des résidus peut se vérifier via un diagramme quantile-quantile, ou via un test de normalité de Shapiro, qui ont chacun une part d'arbitraire (la valeur seuil retenue dans le cadre du test, par exemple) ; l'homogénéité de la variance est généralement vérifiée de manière empirique par une observation visuelle, ou par un test *ad hoc* aux mêmes limites que précédemment ; le choix de retirer ou non un individu aberrant de la population se fondent autant sur des justifications intrinsèques au modèle (distance de Cook, effet levier…) que sur des raisons liées au contexte du recueil des données. De même, il existe plusieurs indicateurs de qualité d'un modèle, mesurant le compromis entre le pouvoir explicatif du modèle et son caractère parcimonieux (R^2 ajusté, critère AIC, BIC ou Cp de Mallow[6]…). Ils ne convergent pas toujours, obligeant le praticien à justifier son choix de modèles par une analyse plus générale qu'une simple règle de comparaison de valeur d'un indicateur entre deux modèles. La pratique de la régression, à travers la démarche de modélisation ne peut pas être traitée de manière totalement indépendante de son contexte applicatif.

[6] Ces différents critères combinent la valeur du R^2 avec le nombre de variables (ou le logarithme du nombre de variables) nécessaire au modèle.

MÉTHODOLOGIE DE LA RECHERCHE

Les axes d'analyse retenus

En mobilisant le cadre de la transposition didactique, nous formulons deux axes d'analyse de l'enseignement de la statistique (et en particulier de la régression linéaire) dans des formations supérieures spécifiquement identifiée comme des cursus en statistique.

Le premier axe a trait à la cohérence : l'absence de noosphère formelle ne devrait pas nuire à l'homogénéité des contenus entre formation. Nous espérons, à travers l'étude des différents cours, mettre en lumière ce socle commun de notions consensuelles dans le passage du savoir savant au savoir enseigné.

Le deuxième axe d'analyse porte sur la tension entre apport théorique formel et application pratique concrète, tension qui parcourt la communauté des statisticiens. Notre intuition est que cette tension se retrouve au sein des formations et est influencée par le niveau de la formation, et son inscription dans un cursus professionnalisant ou non.

Afin de d'explorer ces axes d'analyse, nous travaillerons à partir de l'analyse de support de cours provenant de formations et d'enseignants variés, ainsi que par des interviews complémentaires de ces mêmes enseignants.

Les observables étudiés

L'analyse que nous menons nécessite de couvrir un large spectre de formations et d'enseignants, afin de pouvoir vérifier si les contenus sont cohérents entre formations. Les formations choisies doivent également être diverses dans leurs parcours, professionnalisant ou non, afin de permettre d'explorer notre hypothèse d'une importance différente donnée aux aspects théoriques et pratiques suivant la visée de la formation. Nous avons profité du fait que le spectre des formations en statistique est suffisamment étendu dans la région lyonnaise, ainsi que d'une connaissance personnelle des enseignants en charge du cours de régression linéaire, pour interroger un panel de cinq enseignants couvrant six formations (l'auteur de ces lignes compris). Nous avons pu ainsi considérer un large champ de formations, de la deuxième année de licence à la deuxième année de master, dans des cursus technologiques ou de formation générale :

1. Le Diplôme Universitaire de Technologie « Statistique et Informatique décisionnelle », (DUT STID) de l'IUT Lumière Lyon 2, formation technologique de niveau L2.
2. La licence Mathématiques appliquées aux Sciences Sociales (MASS) option Math-éco, de l'université de Grenoble Alpes, formation universitaire de niveau L2.
3. La licence Mathématiques et Informatique appliquées aux Sciences Humaines et Sociales (MIASHS) de l'université Lumière Lyon II, formation universitaire de niveau L3.
4. Le master 1 Mathématiques Appliquées, Statistique (MAS), de l'université Claude Bernard Lyon I, formation universitaire de niveau M1.
5. Le master de l'INSA Lyon, formation d'ingénieur de niveau M1.
6. Le master pro Mathématiques Appliquées, Statistique (MAS), de l'université Claude Bernard Lyon1, formation universitaire de niveau M2.

De manière schématique, on peut élaborer une double classification de ces cours, dans l'optique de notre problématique :

- Une opposition en fonction du niveau : niveau L (Licence) pour les formations DUT, L2 et L3, et niveau M (Master) pour les formations M1, M1 INSA et M2.
- Une opposition en fonction de la filière généraliste (formations L2, L3, M1, M2) ou technologique (DUT et M1 INSA).

À partir du choix de ces formations, nous avons mis en place deux modes d'étude du contenu des cours en lien avec nos axes d'analyse. Nous nous sommes concentrés dans un premier temps sur l'étude des supports de cours, que nous avons complété ensuite par des questions ciblées aux enseignants.

Les supports de cours analysés

Les supports de cours recueillis prennent la forme de diaporama et/ou de notes de cours rédigées. Nous les avons étudiés, annotés et nous avons reporté nos observations dans un tableau synoptique. Nous procédons dans la partie suivante à une analyse descriptive de ces supports de cours, avant d'essayer de mettre en lumière des choix de transposition didactiques pour chacun des niveaux étudiés.

Notons que nous n'avons aucune indication sur les prérequis de ces cours. En particulier, nous avons considéré que la position de l'enseignant est de considérer qu'il existe des étudiants (si ce n'est tous) dont c'est le premier contact avec la régression linéaire de manière formelle. On peut cependant supposer que tous les étudiants sont familiarisés avec la statistique descriptive à une et deux variables.

Les entretiens menés auprès des enseignants

Nous sommes dans un deuxième temps revenus vers les enseignants avec des questions complémentaires sur la manière dont ils ont appréhendé le contenu du cours, ou pour leur demander des précisions sur un point précis. Ces courts entretiens oraux n'ont pas fait l'objet d'enregistrement ou de transcription, mais simplement de compte-rendu.

ANALYSE DES RÉSULTATS DE LA RECHERCHE

L'étude des documents de cours permet de mettre en lumière le « socle commun » de notions abordées de manière incontournable dans le cadre d'un cours sur la régression linéaire. Cette étude permet également de remarquer des différences de traitement de ces notions, ainsi que la présentation ou non de notions complémentaires. Nous avons choisi de focaliser notre analyse sur six thèmes particuliers, deux d'entre eux se retrouvant dans toutes les formations de manière quasi identique, les quatre autres montrant des divergences. Ces points de divergence permettent de dessiner les choix effectués par chaque enseignant dans la manière de présenter les notions liées à la régression linéaire, et partant d'esquisser un processus de transposition didactique d'un savoir savant partagé à un savoir à enseigner dépendant du contexte d'enseignement.

Lien entre modèle simple et modèle multivarié

Le modèle général de la régression linéaire est le modèle à plusieurs variables explicatives. Cependant, le modèle simple à une seule variable explicative est généralement connu des étudiants car présenté en cours de statistique descriptive bivariée. Il est donc possible de présenter soit le modèle multivarié comme une généralisation du modèle simple, soit, à l'inverse, le modèle simple comme un cas particulier du modèle multivarié. Nous constatons

que tous les cours commencent par présenter le modèle simple[7], de manière illustrée avec des graphes permettant de visualiser le nuage de points et la droite de régression. L'estimation des paramètres du modèle s'effectue toujours par un calcul explicite dans le cadre à une variable, ainsi que tout le processus d'étude des résidus et de test de la significativité. Le modèle est ensuite généralisé au cas à plusieurs variables explicatives, le calcul des paramètres et de la qualité du modèle faisant généralement explicitement référence au cas simple : « comme dans le cas à une variable… »

Nous sommes donc en présence ici d'un exemple de transposition didactique partagée entre tous les enseignants, de manière non concertée. Le savoir savant n'impose aucun traitement particulier au cas simple : tout le matériel théorique de la régression linéaire multiple, toutes les formules de calcul des coefficients, des statistiques de test, d'études des résidus sont les mêmes quelle que soit la valeur de p, le nombre de variables, y compris pour $p=1$. Mais tous les enseignants considèrent que le cas $p=1$ nécessite d'expliciter spécifiquement le modèle, créant ainsi un savoir à enseigner différent du savoir savant. D'où vient que ce savoir à enseigner soit si répandu, voire exclusif ? Il n'y a aucune injonction extérieure à commencer par un modèle simple. Cependant une progression pédagogique évidente consiste à aller du plus simple au plus complexe ; dans cette perspective l'exemple du cas à une variable explicative est bien plus simple que le cas multivarié, et permet d'introduire toutes les notions importantes à un moindre coût de calcul. De plus, le modèle simple est très utilisé pour lui-même et donc ses applications sont dignes d'intérêt pour la plupart des étudiants.

Estimateurs des moindres carrés

Le critère des moindres carrés est presque toujours introduit de manière empirique dans les cours, comme une distance classique, pratique à utiliser, pour « bien prendre en compte l'éloignement des points à la droite de manière similaire »[8]. le critère des moindres carrés est justifié par le théorème dit de Gauss-Markov, qui indique que les estimateurs de moindres carrés de la régression sont les meilleurs estimateurs sans biais possibles. Seul le cours dispensé en L3 le mentionne. Pour les autres cours, nous voyons

[7] Le cours de l'INSA ne traite pas de régression linéaire multiple.

[8] Cours de M1 INSA

très clairement le passage d'un savoir savant – le modèle est justifié par un théorème d'optimalité – à un savoir à enseigner qui se rapproche plus d'une intuition – il faut visualiser la distance que l'on cherche à minimiser –.

Les démonstrations

La place des démonstrations varie grandement dans les cours étudiés. Nous nous restreindrons à trois résultats clés du cours qui peuvent faire l'objet de démonstrations : la formule de décomposition de la variance ; le calcul des paramètres du modèle linéaire par la méthode des moindres carrés ; le fait que la loi du ratio des carrés moyens est une loi de Fisher sous l'hypothèse de normalité des résidus.

L'étude des documents de cours montre que pour chacun des résultats-clés ci-dessus, la démonstration se trouve dans le cours destiné aux niveaux M1 et M2 ; la démonstration est laissée en exercice pour les niveaux L2 et L3 ; le résultat est donné sans démonstration pour le DUT et le M1 INSA. Il existe donc une différence évidente en terme de démonstration et de présentation entre les cursus appliqués (DUT, INSA) et les cursus plus théoriques (L, M1, M2). La statistique, en tant que mathématiques appliquées, est toujours en tension entre un pôle mathématisant, qui insiste sur le caractère rigoureux de la science statistique, et un pôle applicatif qui insiste sur la pratique du calcul statistique. Le savoir savant réside ici dans ce caractère rigoureux des formules et des calculs statistiques. Le statisticien sait que la statistique est une branche des mathématiques assise sur des résultats précis démontrés à l'aide de théorèmes. Refaire la démonstration d'un résultat lors d'un enseignement, sans que cette démonstration ne doive nécessairement être apprise et retenue, permet d'ancrer dans l'esprit de l'étudiant qu'un résultat statistique est d'abord dû à un théorème dont on vérifiera les hypothèses avant de l'appliquer dans un cas pratique très précis. De plus, démontrer un théorème permet de mettre en lumière les mécanismes mathématiques à l'œuvre derrière un résultat. Au-delà de l'objectif de convaincre, par un raisonnement strict, l'étudiant de la véracité d'un résultat, la démonstration dans un cours de régression vise aussi à permettre à l'étudiant d'imaginer et de mettre en œuvre d'autres modèles, dans d'autres situations. Par exemple, on peut imaginer que démontrer que la statistique[9] du test de significativité du R^2 suit une loi de

[9] Appelée également statistique de l'ANOVA (pour « ANalysis Of VAriance »).

Fisher si les résidus suivent une loi normale permet à l'étudiant d'imaginer quelle loi suivrait cette statistique si les résidus ne suivent pas une loi normale mais une autre loi donnée.

Au contraire, dans le cas où les démonstrations ne sont pas enseignées, l'enseignant insiste plus sur le côté applicatif de la modélisation. Le savoir à enseigner est ici non pas la validité théorique d'un résultat mais son utilisation pratique. L'objectif n'est pas que l'étudiant comprenne mathématiquement la méthode, mais qu'il connaisse les hypothèses à vérifier en pratique pour que la méthode fonctionne. Par exemple la formule de décomposition de la variance utilise comme résultat clé le fait que les résidus sont centrés sur zéro. Tous les cours indiquent que les résidus sont centrés par construction, mais seuls ceux qui en font la démonstration montrent comment cela influe directement sur la décomposition de la variance.

Nous pouvons faire ici un aparté sur le cours de L2. Certains résultats (les plus faciles) sont démontrés dans ce cours. Les résultats les plus importants, tels que ceux listés ci-dessus, sont démontrés en exercices (et non « laissés à titre d'exercices »). L'approche adoptée en L2 est donc plus proche de celle des M1 et M2 que de celles des autres formations étudiées. De ce point de vue, la formation en L2 se situe très nettement dans une perspective de cursus théorique long, et non comme un cursus appliqué, malgré l'affichage de la filière « mathématiques appliquées au sciences sociales ».

Normalité des résidus

Comme indiqué précédemment, l'hypothèse de normalité des résidus, c'est-à-dire le cadre gaussien du modèle, est indispensable pour un certain nombre de résultats, mais elle n'est pas nécessaire pour toutes les propriétés du modèle de régression. Cette distinction entre hypothèse nécessaire et hypothèse suffisante n'est pas traitée de la même manière dans tous les cours. Par exemple, le cours de DUT affiche que le cas non gaussien sort du cadre du cours ; la situation en cas de résidus non gaussien est donc supposée non traitable par les étudiants. De manière générale, les cours de DUT, de L3 MIASHS et de M1 INSA supposent la normalité des résidus dès le début du cours. Les cours dispensés en L2, M1 et M2 font explicitement le distinguo entre les résultats qui ne dépendent pas de la normalité des résidus et ceux qui la supposent. On retrouve donc ici encore une opposition entre cursus applicatifs et cursus théoriques.

Introduction du coefficient de détermination

Le coefficient de détermination est introduit, suivant les supports, soit par sa définition formelle, soit à partir du coefficient de corrélation. Cette dernière approche est plus intuitive, le coefficient de corrélation étant généralement connu des étudiants pour être abordé dans le cadre des cours de statistique descriptive à deux variables. Nous avons vu *supra* que tous les cours étudiés commencent par présenter la régression dans sa version à deux variables avant de la généraliser à plusieurs variables. Cette unanimité ne se retrouve pas lors de l'introduction du coefficient de détermination, qui se fait de manière très variée suivant les niveaux. Ainsi, les cours de DUT et de l'INSA définissent d'abord le R^2 comme le carré de ρ, qui est lui-même défini à partir de la covariance. Il est ensuite indiqué (pour le cours de DUT) que le R^2 obtenu à partir de la décomposition de la variance est le même que le carré de ρ, avant sa généralisation au sein de la régression multiple. Le cours de M2 définit le coefficient de corrélation, ou plutôt en rappelle la définition supposée connue, dans le cas simple, et indique simplement que le carré de ρ s'appelle le coefficient de détermination. Ce même coefficient de détermination est ensuite défini dans le cas général par le ratio des carrés des écarts, mais pas dans le cas bivarié. Le cours de L3 définit de manière générale le R^2 par le ratio, avant définir d'autorité ρ comme la racine de R^2 ayant même signe que le coefficient directeur de la droite de régression. La mention « on montre que » avant cette définition indique là aussi que le coefficient de corrélation, et sa définition à partir de la covariance, est supposé connu des étudiants. Enfin le cours de L2 comme celui de M1 ne parlent pas du tout du coefficient de corrélation.

Le lien entre coefficient de corrélation et coefficient de détermination est donc un bon exemple de l'adaptation des notions à enseigner effectuée par l'enseignant, c'est-à-dire la transposition didactique du savoir savant au savoir à enseigner. La démarcation se fait ici en fonction du profil des étudiants, du plus applicatif (DUT) au plus théorique (L2/M1), en passant par une formation théorique prenant appui sur des connaissances appliquées antérieures (L3, M2).

Démarche de modélisation

L'examen des résidus est traité de manière importante en DUT, en L3 (sous forme d'un chapitre à part), à l'INSA et en M2. L'utilisation de graphes et de sorties de logiciels, indique que l'accent est mis sur la pratique, et les « bonnes pratiques » de la

modélisation. Le cours de L2, comme celui de M1, insistent moins sur ces aspects. Même si les hypothèses sur les résidus ont été introduites, la manière pratique de tester ces hypothèses n'est pas mise en avant. On peut supposer que contrairement aux étudiants de filières appliquées (DUT, INSA) ou de formations de fin de cycles (L3, M2), les étudiants de L2 ou M1 ne sont pas confrontés directement à la pratique de la régression. Il apparait donc moins important d'insister sur les aspects pratiques que sur les aspects théoriques de la modélisation dans le cours.

Par contre, à part dans le cours de l'INSA (qui n'aborde pas la régression multiple), les questions de choix de modèles sont abordées de manière approfondie avec des méthodologies (sélection forward ou backward par exemple) très appliquées et utilisables par les étudiants. Le savoir-faire enseigné ici rejoint systématiquement le savoir-faire de référence, utilisant des méthodes éprouvées pour résoudre des problèmes concrets. Il semble y avoir unanimité pour considérer que la méthodologie d'un choix de modèle de régression multiple est au cœur du savoir du statisticien modélisateur.

Analyse complémentaire des entretiens auprès des enseignants

Nous avons complété cette analyse des supports de cours par des entretiens avec les enseignants. En effet, la simple analyse des supports de cours ne suffit pas à identifier l'intention de l'enseignant. Par exemple, dans un article présentant un cours relatif à un autre sujet que la modélisation linéaire, Rakotomalala (2017) indique qu'il possède un unique support de cours qu'il utilise, de manière différenciée, pour cinq formations différentes en adaptant son discours à l'oral. Il ressort des entretiens menés que les choix présidant à l'élaboration des cours ont été des choix très individuels (à l'exception du DUT, par l'existence d'un programme pédagogique national). Les principes guidant les cours ont été choisis par les enseignants après leur analyse de la situation. En L3, le contenu est guidé par les poursuites d'études potentielles des étudiants : que doit savoir un étudiant pour poursuivre en master d'économétrie ? En M1 INSA le contenu est fortement contraint par le faible volume horaire – 6h – dédiés à la statistique dans une formation qui se veut pourtant orientée vers le traitement de données. Le choix a été délibérément fait d'un cours orienté vers la pratique plutôt que vers la justification théorique. En M1 généraliste, la volonté est de garder une approche mathématique mais de prendre le temps de rendre les développements mathématiques accessibles au plus grand

nombre d'étudiants. Dans tous les cas, les enseignants témoignent d'une construction du cours à partir de nombreuses sources (livres de référence, cours et exercices trouvés sur internet), mais surtout à partir de leur *compréhension personnelle* du sujet d'une part, et des enjeux du module au sein de la formation d'autre part. Les enseignants interrogés témoignent donc d'un réel effort d'adaptation à la situation, à travers une prise en compte non seulement du public enseigné, mais surtout de la place de l'enseignement dans la trajectoire de formation de ce public : au-delà du niveau (licence ou maîtrise) et du type de filière (générale ou technologique), c'est aussi la perception par l'enseignant des enjeux du cours donné au sein de la filière qui détermine l'équilibre entre maîtrise de la pratique et généralisation théorique. Les entretiens avec les enseignants viennent corroborer cette dichotomie entre approche centrée sur la pratique et approche centrée sur la généralisation. Dans certaines formations l'important est que l'étudiant soit capable d'appliquer les techniques de régression :

- dans des cas simples, à titre d'ouverture vers les techniques statistiques (M1 INSA) ;
- dans des cas typiques, afin de pouvoir être autonome en tant qu'exécutant de techniques éprouvées (DUT) ;
- dans des cas typiques, afin de posséder les bases nécessaires à une poursuite d'étude avancée (L3).

Dans les autres formation, l'important est que l'étudiant soit capable de comprendre la théorie mathématique sous-jacente à la régression linéaire, en l'accompagnant comme en L2 ou en M1 ou en visant le plus de précision comme en M2.

RÉSULTATS : DES CHOIX DE TRANSPOSITION DIDACTIQUE EN RÉGRESSION

Le savoir à enseigner : un socle partagé

Nous avons proposé ci-dessus l'analyse comparée de six supports de cours différents suivant six axes différents. Ces analyses permettent de dégager une vue d'ensemble des choix didactiques effectués par les enseignants à propos de la modélisation linéaire. Un savoir à enseigner a été identifié par déduction, tant dans sa partie théorique que pratique. Il contient les estimateurs des moindres carrés et les tests possibles sur ces estimateurs, la définition du coefficient de détermination, ainsi que l'analyse des

conditions d'application du modèle et le choix du meilleur modèle possible.

Ce savoir à enseigner, naturellement critiquable, nous semble cohérent avec les ouvrages spécialisés consultés, par exemple le livre de Saporta (2011). L'apparition de ce savoir à enseigner a été observé via les traces écrites des supports de cours, supposés être plus proches de ce que l'enseignant souhaite enseigner que de ce qu'il a réellement enseigné en classe.

L'absence de noosphère externe formelle n'empêche donc pas l'apparition d'un socle de savoir à enseigner partagé, confirmant ainsi notre intuition d'observer également dans le domaine de la statistique le constat évoqué par Bosch et al. (2021) d'une faible liberté réelle laissée aux enseignants dans la définition du savoir à enseigner.

Théorie ou pratique ?

Cependant, si les notions abordées sont relativement similaires sur le fond, l'approche de ces notions varie selon les formations étudiées. Le savoir à enseigner apparait alors en tension entre une approche centrée sur la justification des résultats, conférant à la modélisation statistique une rigueur mathématique, et une approche centrée sur la technique permettant à l'étudiant de mettre en pratique concrètement ces outils de modélisation.

Reprenons les six axes d'analyses des supports de cours sous cet angle :

- Modèle simple et modèle multiple : l'approche orientée vers la théorie voit dans le modèle simple un cas particulier du modèle multiple (aucune formation ne présente les choses ainsi), quand l'approche orientée vers la pratique introduit le modèle multiple comme une généralisation du modèle simple (toutes les formations le présentent ainsi).
- Estimateur des moindres carrés : dans l'approche orientée vers la théorie l'estimateur est justifié par le théorème de Gauss-Markov (L3), quand dans l'approche orientée vers la pratique l'estimateur est justifié par le dessin et l'intuition (DUT, L2, M1, M1 INSA, M2)
- Démonstrations : on s'attache à présenter des démonstrations rigoureuses des principaux résultats dans l'approche orientée vers la théorie (L2, M1, M2), là où les résultats sont non démontrés dans l'approche orientée vers la pratique (DUT, L3, M1 INSA).
- Normalité des résidus : l'approche orientée vers la théorie présente le cas gaussien comme un cas particulier, nécessaire

pour certains résultats seulement (L2, M1, M2) ; l'approche orientée vers la pratique suppose la normalité des résidus toujours vérifiée pour pouvoir effectuer la modélisation (DUT, L3, M1 INSA).

- Coefficient de détermination : dans l'approche orientée vers la théorie le coefficient est défini comme le ratio des carrés moyens (L2, M1 – et en partie L3 et M2), alors que dans l'approche orientée vers la pratique le coefficient est introduit à partir du coefficient de corrélation puis généralisé (DUT – et en partie L3 et M2).
- Démarche de modélisation : l'approche orientée vers la théorie présenterait une justification théorique des indices de qualité (aucune formation ne procède ainsi), contrairement à une approche centrée sur les bonnes pratiques (toutes les formations).

Le tableau 1 ci-dessous résume les différences d'approche entre les formations traitées.

Axe d'étude	Approche orientée théorie	Approche orientée pratique
Modèle simple et modèle multiple	Aucune formation	Toutes les formations
Estimateur des moindres carrés	L3	DUT, L2, M1, M1 INSA, M2
Démonstrations	L2, M1, M2	DUT, L3, M1 INSA
Normalité des résidus	L2, M1, M2	DUT, L3, M1 INSA
Coefficient de détermination	L2, M1 – et en partie L3 et M2	DUT – et en partie L3 et M2
Démarche de modélisation	Aucune formation	Toutes les formations

Table 1.- répartition des formations par approche par axe d'étude.

Nous envisagions que la tension entre théorie et pratique au sein de la discipline statistique pourrait se retrouver au sein du panel de formations, et être influencée par le niveau de la formation et son inscription dans un cursus professionnalisant ou non. L'analyse des corpus de cours relatifs à la régression linéaire (table 2) permet de constater que cette tension entre présentations

théorique et pratique existe bien, mais que les oppositions ne sont pas univoques :

- Toutes les formations considèrent, au moins pour trois axes, une approche orientée pratique : c'est donc le choix d'un équilibre spécifique entre approche théorique et approche pratique qui distingue les formations entre elles.
- L'opposition n'est pas spécifiquement une opposition de niveau, la L2 se comportant comme le M1 et le M2, et le DUT comme le M1 INSA.
- L'opposition n'est pas non plus uniquement entre filière généraliste et technologique, la L3 se retrouvant souvent avec le DUT et le M1 INSA.

Notre analyse des documents ne montre donc pas une opposition franche entre théorie mathématique et pratique empirique. L'analyse des entretiens permet alors de proposer une grille de lecture complémentaire résultant de la perception qu'a l'enseignant des enjeux du cours donné au sein de la filière. L'opposition (relative) se situe alors entre une approche centrée sur la maîtrise de la pratique dans un environnement donné (DUT, L3, M1 INSA) et une approche insistant sur la généralisation possible des techniques utilisées (L2, M1, M2). De ce fait, on attend d'abord de l'étudiant une capacité d'application dans des cas-types dans le cas des DUT, L3, M1 INSA. Au contraire, on attend une compréhension des mécanismes mathématiques permettant une généralisation, dans le cas des filières L2, M1, M2. La tension entre application et théorie générale dans la transposition didactique des enseignements étudiés intègre pleinement la place spécifique de la filière dans le parcours de formation du futur statisticien.

Autres facteurs explicatifs

Notons que d'autres facteurs non recensés ici peuvent avoir un effet sur le contenu de l'enseignement. La formation initiale de l'enseignant ainsi que son activité de recherche influencent son approche vraisemblablement autant que le type de formation dans laquelle il enseigne. Le volume horaire affecté au cours peut également amener l'enseignant à faire des choix qui ne résultent pas de considérations didactiques. Enfin le temps disponible de l'enseignant pour la préparation de ses cours joue peut-être aussi. Par exemple, le cours de L2 et de M1, donnés par le même enseignant, ne diffèrent que très peu l'un de l'autre. Est-ce dû à une analyse des besoins des étudiants par l'enseignant, ou au fait

qu'à partir d'un même support – gain de temps et d'énergie – l'enseignant est capable de mener deux cours différents ?

CONCLUSION

L'absence d'une noosphère formelle dans l'enseignement supérieur ne signifie pas que l'enseignant a toute liberté pour décider du contenu de son cours (Bosch et al. (2021)). Il existe cependant un espace de possibilités de transposition du savoir savant au savoir à enseigner. Ces possibilités sont plus ou moins investies par les enseignants, certains faisant des choix pédagogiques explicites et personnels, d'autres se reposant sur des « standards de fait » à travers la reprise de propositions de cours existants par ailleurs. Le caractère collectif de la transposition didactique est peu marqué : au mieux, il reste implicite par les échanges de documents de cours, mais le plus souvent l'enseignant ne peut compter que sur sa propre compréhension des enjeux du cours au sein de la formation pour décider de ses choix pédagogiques. Il en ressort, dans le cas de l'enseignement de la régression linéaire, une tension entre présentation de méthodes pratiques spécifiques et effort de généralisation théorique.

Il est à noter que contrairement à de nombreux domaines des mathématiques pouvant être étudiés dans une perspective purement abstraite, cette tension entre théorisation et application est intrinsèque à la statistique en tant que discipline savante. Il est donc normal que cette tension apparaisse également dans les cours de statistique destinés à des statisticiens.

Cette première approche analytique de la transposition didactique dans l'enseignement de la statistique dans des filières spécialisées nécessite d'être poursuivie pour mettre en évidence le rôle personnel joué par l'enseignant dans cette transposition. En effet, l'analyse effectuée dans cet article est limitée par le fait que nous avons étudié des couples formation/enseignant, sans pouvoir isoler l'effet de la formation de celui de l'enseignant au sein du dispositif. Une étude conduite sur des cours identiques dans des formations comparables devrait permettre d'affiner notre compréhension des enjeux de la transposition didactique dans les cours de statistique pour statisticiens.

REMERCIEMENTS

L'auteur remercie sincèrement Irène Gannaz, Cécile Mercadier, et Ricco Rakotomalala pour la communication de leurs documents de cours et leur aide dans l'élaboration de cet article. Cet article est particulièrement dédié à François Wahl, décédé avant la parution de cet article. L'auteur remercie également les relecteurs des versions antérieures de cet article pour leurs nombreux retours et commentaires très riches et très profitables.

BIBLIOGRAPHIE

Armatte, M. (2010). Le rôle de l'histoire dans l'enseignement de la statistique. *Statistique et Enseignement*, *vol 1*, n°2, pp. 23-47.

Arsac, G. (1992). L'évolution d'une théorie en didactique : l'exemple de la transposition didactique. *Recherches en Didactique des Mathématiques, Vol. 12,* n°l, pp. 7-32.

Berthiaume, D. (2007). « Une description empirique du savoir pédagogique disciplinaire des professeurs d'université », dans les *Actes du colloque de l'AIPU : regards sur l'innovation la collaboration et la valorisation*, Montréal, p.179-181.

Besse, P., & Laurent, B. (2016). De statisticien à data scientist, développements pédagogiques à l'INSA de Toulouse. *Statistique et enseignement*, *vol. 7,* nº 1, pp. 75–93.

Bihan-Poudec, A. (2012). Un contrôle surprise pour l'enseignant ! L'évaluation comme révélateur des préconceptions de la statistique chez les étudiants. *Statistique et enseignement*, *vol. 3,* n° 1, pp. 63–72.

Bosch, M., T. Hausberger, R. Hochmuth, M. Kondratieva & C. Winsløw (2021). External Didactic Transposition in Undergraduate Mathematics. *International Journal of Research in Undergraduate Mathematics Education*, *7* (1), pp.140-162.

Bosch, M., & Winsløw, C. (2020). The external didactic transposition of mathematics at university level: dilemmas and challenges for research. *Educação Matemática Pesquisa*, *v. 22* n. 4, pp. 373-386.

Cañadas, G., Batanero, C., Diaz, C., & Roa, R. (2012). Psychology students' understanding of the chi-squared tests. *Statistique et enseignement*, *vol. 3,* nº 1, pp. 3–18.

Carillo, K., Galy, N., Guthrie, .& Vanhems, A. (2016). "J'aime pas les stats !" : mesure et analyse de l'attitude à l'égard du cours de statistique dans une école de management. *Statistique et enseignement*, *vol. 7*, nº 1, pp. 3–31.

Chevallard Y. (1985). *La transposition didactique – Du savoir savant au savoir enseigné*, La Pensée sauvage, deuxième édition augmentée, 1991.

Dufour, A.-B., Lobry, J., & Amat, I. (2017). Enseigner le recueil des données : explorer la variabilité biologique, au chaud, dans une salle de cours. *Statistique et enseignement*, *vol. 8*, nº 2, pp. 79–85.

Dunn, P. K., Marshman, M., McDougall, R,. & Wiegand, A. (2015), Teachers and Textbooks: On Statistical Definitions in Senior Secondary Mathematics. *Journal of Statistics Education*, 23:3.

El M'hamedi, Z. (2019a). Effets du vocabulaire et de l'ambiguïté linguistique sur la compréhension des tests statistiques. *Annales de Didactiques et de Sciences Cognitives*, *vol. 24*, pp 133-181.

El M'hamedi, Z. (2019b). Quelques aspects de l'incompréhension des tests statistiques. *Recherches en didactique des mathématiques*, *vol. 39/2*, pp 167-211.

El M'hamedi, Z. (2021). Difficultés de compréhension du concept de niveau de signification., *Éducation et didactique*, 15-3

Fine, J. (2013). Le rapport GAISE (US) cadre d'un curriculum statistique de la maternelle à la terminale. *Statistique et enseignement*, *vol. 4*, no 1, pp. 25–54.

GAISE College Report ASA Revision Committee, *Guidelines for Assessment and Instruction in Statistics Education College Report 2016*, http://www.amstat.org/education/gaise.

Garfield, J. & Ben-Zvi, D. (2007). How Students Learn Statistics Revisited: A Current Review of Research on Teaching and Learning Statistics. *International Statistical Review / Revue Internationale De Statistique*, *vol. 75*, no. 3, pp. 372–396.

Gattuso, L. (2011). L'enseignement de la statistique : où, quand, comment, pourquoi pas ? *Statistique et enseignement, vol 2*, n° 1, pp. 5–30.

Gélinas, S., Berger, E. R., Balbinotti, M., Lalande, D.& Cantinotti, M. (2018). L'apprentissage par la pratique : vécus d'étudiants en psychologie impliqués dans la création de tutoriels informatisés en méthodes quantitatives. *Statistique et enseignement, vol. 9*, nº 1, pp. 23–41.

Genolini, C. & Driss, T. (2010). Eveiller l'intérêt pour la statistique par l'exemple. *Statistique et enseignement, vol. 1,* nº2, pp. 49–57.

Hahn, C. (2015). La recherche internationale en éducation statistique : état des lieux et questions vives. *Statistique et enseignement, vol. 6,* nº 2, pp. 25–39.

Hahn, C. & Stoltz, G. (2013). Savoir académique, savoirs pratiques : tensions et recherche d'équilibre. *Statistique et enseignement, vol. 4,* nº 2, pp. 19–52.

de Hosson, C., Manrique, A., Regad, L. & Robert, A. (2018). Du savoir savant au savoir enseigné, analyse de l'exposition des connaissances en cours magistral de physique : une étude de cas. *Revue internationale de pédagogie de l'enseignement supérieur*, 34 (1)

Husson F. & Pagès, J. (2013) *Statistiques générales pour utilisateurs. 2 - Exercices et corrigés*. Presses Universitaires de Rennes

Jutand, M.-A., Leffondré, K., Savès, M. & Kiewsky, V. (2017). Enseigner le recueil des données : étude de cas en épidémiologie. *Statistique et enseignement, vol. 8,* nº 2, pp. 87–101.

Kasprowicz, T. & Musumeci, J. (2015). Teaching Students Not to Dismiss the Outermost Observations in Regressions. *Journal of Statistics Education*, 23:3.

Lahanier-Reuter, D (2012). La statistique est-elle une discipline scolaire ? *Statistique et enseignement, vol. 3,* n°2, pp 23-32.

Léziart, Y. (2003). Transposition didactique et savoirs de référence : illustration dans l'enseignement d'une pratique particuliere de saut, le Fosbury-flop. *Movement & Sport Sciences*, nº 50, pp 81-101

Martinand J. L. (1989). Pratiques de référence, transposition didactique et savoirs professionnels en sciences et techniques. *Les sciences de L'éducation,* 2/1989, pp.23-29.

Martinand, J.L. (2001). Pratiques de référence et problématique de la référence curriculaire. Dans A. Terrisse. *Didactique des disciplines*. De Boeck Université.

Nikiforidou, Z., Lekkab, A. & Pangec, J. (2010). Statistical literacy at university level: the current trends., *Procedia Social and Behavioral Sciences*, 9, pp 795–799.

Palm, R. & Allagbe, G. (2011). Simuler pour comprendre : un didacticiel pour l'apprentissage de notions de base en statistique inférentielle. *Statistique et enseignement, vol. 2,* nº 1, pp. 77–84.

Poteaux, N. (2013). Pédagogie de l'enseignement supérieur en France : état de la question. *Distances et médiations des savoirs*, 4.

Rakotomalala, R. (2017). Introduction à l'apprentissage supervisé. *Statistique et enseignement, vol. 8,* nº 2, pp. 43–58.

Régnier, J.-C. (2012). Enseignement et apprentissage de la statistique : entre un art pédagogique et une didactique scientifique. *Statistique et enseignement, vol. 3,* n° 1, pp. 19–36.

Rolland, A. (2020). 2009-2019 : dix ans de publications sur l'enseignement de la statistique en France. *Statistique et Société, vol. 8,* n°1, pp 55–71.

Saporta, G. (2011). *Probabilités, analyse de données et statistique*, Ed Technip.

Schwab-McCoy, A. (2019). The State of Statistics Education Research in Client Disciplines: Themes and Trends Across the University. *Journal of Statistics Education*, 27:3, pp 253-264.

Senterre, C., Coppieters, Y., Levêque, A., & Dramaix, M. (2011). Présentation et analyse d'un dispositif d'apprentissage en analyse multivariable appliquée à l'épidémiologie. *Statistique et enseignement*, *vol. 2*, nº 1, pp. 61–75.

Tishkovskaya, S. & Lancaster, G. A. (2012). Statistical Education in the 21st Century: A Review of Challenges, Teaching Innovations and Strategies for Reform. *Journal of Statistics Education*, 20:2.

Woodard, R. & McGowan, H. (2012). Redesigning a Large Introductory Course to Incorporate the GAISE Guidelines. *Journal of Statistics Education*, 20:3,

Zendrera, N., Dubreil-Frémont, V., Marion, J.-M. & Bihan-Poudec, A. (2017). Les données et leur production : réflexion sur une lacune paradoxale en éducation statistique. *Statistique et enseignement, vol. 8*, nº 2, pp. 59–78.