QUEL USAGE DE LA STATISTIQUE DANS LES THÈSES?

Marthe-Aline Jutand[1]
[1] *Université de Bordeaux, ISPED, 146 rue Léo Saignat, 33000 Bordeaux*
marthe-aline.jutand@u-bordeaux.fr

Résumé.

La spécificité de la statistique, en tant que science d'analyse et de décision, suscite l'intérêt de l'ensemble des domaines scientifiques et sociétaux. Elle en devient ainsi une discipline protéiforme s'appuyant sur des dimensions partagées, mais aussi sur des formulations spécifiques. Dans le cadre du doctorat, le jeune chercheur doit développer des capacités à expliciter les étapes d'une démarche statistique et à communiquer de manière compréhensible à des fins de diffusion et d'appropriation. L'analyse des thèses met en lumière des régularités d'usage de concepts et de mots, mais aussi, de nombreuses spécificités disciplinaires. Cette étude de l'usage de la statistique dans les thèses peut alimenter la discussion concernant l'enseignement de la statistique à l'université et sur les stratégies pédagogiques à développer. Notre analyse met en évidence l'importance de former les étudiants à la diffusion et la traduction de résultats d'études statistiques. Tout ceci conduit à la question de la transformation des enseignements de la statistique à l'université afin de développer les temps et des méthodes d'apprentissage de diffusion, à visée d'explicitation, des savoirs statistiques et des résultats auprès de publics variés.

Mots-clés. Statistique, doctorat, transposition didactique.

Abstract.

Statistics, with its particularity being a science based on analysis and decision, generates interests from all scientific and social domains. Thus, statistics is a many-sided discipline based on shared dimensions, but also on specific formulations. Junior researchers must develop specific skills to explain each step of a statistical approach in situation studied: doctoral theses. Additionally, results need to be communicated clearly and precisely by a researcher in order to assure their comprehension and the appropriation of knowledge. The analyses of doctoral theses highlights a regularity in the use of statistical concepts and statistical language, but also numerous specificities for academic disciplines. This analyses aims to take part in the reflection about i) the teaching of statistics at the university and ii) the pedagogical strategies that needs to be developed by highlighting the teaching strategies proven to be useful for statistical education. Our analysis clearly shows the importance of qualifying students in the dissemination and communication of results from statistical analyses. Taken together, there is a clear need to develop new teaching strategies for statistics at the university aiming at developing teaching methods facilitating explicit communication of statistical expertise and results addressing a wide range of audiences.

Keywords. statistics, thesis, didactic transposition .

1. Introduction

La statistique est un domaine scientifique ayant la particularité de se développer dans différents espaces et pour différentes raisons. C'est cette double diversité, spatiale et fonctionnelle, qui en fait tout son intérêt et sa richesse, sa force et parfois sa faiblesse. Cette discipline a été longtemps rattachée principalement aux mathématiques, mais a pris, semble-t-il, depuis quelques décennies son autonomie et a acquis une réelle reconnaissance au sein de nombreux domaines.

Comme le souligne Desrosières dans la « *Politique des grands nombres* », *«la statistique et les calculs des probabilités occupent une place essentielle parmi les outils d'invention, de construction et de preuves des faits scientifiques dans les sciences de la nature comme les sciences sociales »* (Desrosières, 2000). La spécificité de la statistique, en tant que science d'analyse et de décision induite par l'existence de données et de démarche de compréhension de phénomènes observés, suscite l'intérêt de l'ensemble des domaines scientifiques et sociétaux. Elle en devient ainsi une discipline protéiforme s'appuyant sur des dimensions partagées, mais aussi parfois des formulations spécifiques. L'usage et le développement de la science statistique au sein de différents domaines scientifiques construisent ainsi des objets répondant à des besoins propres à un domaine disciplinaire. Ils prennent une dimension transverse en étant décontextualisés, et peuvent revenir avec force dans certains girons disciplinaires.

Formuler avec clarté, précision et concision les démarches statistiques mises en œuvre et les résultats obtenus sont des éléments ayant un rôle social et stratégique pour le scientifique. Mais les spécificités de formulation dans les différents milieux concernant les objets eux-mêmes et leurs usages dans le cadre de la statistique sont à prendre en considération. Le scientifique a ainsi un rôle social dans l'accompagnement à la compréhension des résultats statistiques. Nous nous sommes donc intéressés tout particulièrement à l'usage de la statistique dans le cadre de thèses de disciplines très diverses.

2. Pourquoi s'intéresser aux thèses ?

La thèse est le premier objet de reconnaissance des qualités scientifiques du doctorant par ses pairs scientifiques ; la validation de son doctorat passera donc par l'évaluation par le jury de son écrit principalement. Il s'agit dès lors pour le doctorant de rédiger un écrit mettant en valeur ses acquis et ses compétences pour intégrer le milieu des chercheurs. La science statistique peut être présente dans le cadre de sa thèse, soit pour aider à développer des éléments d'argumentation de la question de recherche posée, soit dans le cadre de la mise en œuvre d'une démarche statistique lui permettant de répondre à sa question. Quels sont dès lors les usages de la statistique qu'il présentera dans le cadre de ce travail comme point d'étape et de passage entre le milieu du novice à celui du monde des chercheurs ? Peut-on ainsi reconnaitre des régularités d'usage inter ou intra discipline ? La statistique est-elle à ce point protéiforme pour en reconnaitre les besoins et les outils ?

Dans le cadre des thèses, la communauté dans laquelle se déroule la transposition est constituée de scientifiques du même environnement, nous avons dans ce cas une communauté constituée autour de savoirs partagés. Le but du doctorant consiste à faire reconnaitre son travail personnel, sa thèse, pour être lui-même reconnu dans cette communauté non plus en tant qu'étudiant mais en tant que chercheur. Il se doit dès lors d'utiliser le cadre de transmission de son savoir usité dans le domaine dans lequel il souhaite se faire reconnaitre.

Approcher la question des usages et de la formulation de la science statistique par l'étude des corpus

de thèse était un choix effectué pour approcher la pluridisciplinarité sans avis *a priori* des usages mais dans un contexte de recherche. Comme l'indique Latour (2008) (2, il est intéressant d'étudier les sciences en construction et non sanctionnées car ainsi nous pouvons palper l'hétérogénéité de la production avant le formatage induit par les cultures des communautés de recherche et les contraintes éditoriales et de diffusion pouvant étouffer et normaliser les créations rédactionnelles. Les thèses sont donc les écrits présentant la production scientifique de novices souhaitant intégrer la "cour des grands". Ils sont dans l'attente de la reconnaissance de leur travail par leurs pairs scientifiques respectifs, et donc de la reconnaissance de leurs productions scientifiques comme pouvant venir nourrir l'ensemble des savoirs de leurs communautés de rattachement souhaité (Lonka et al. 2014) (3). Etudier les usages dans les thèses, c'est aussi voir le résultat de tout un processus de formation mis en œuvre dans le cadre de la formation du supérieur, et cela peut soulever des pistes de recommandations pour la construction de l'intégration de l'enseignement de la statistique dans l'offre de formation.

3. Méthode

Les éléments de statistique présents dans les thèses, sont des indicateurs de la place que joue la statistique dans la discipline ou au sein de la communauté scientifique de rattachement du doctorant et de son directeur de thèse.

Il s'agissait de repérer, au sein de chaque thèse, les éléments se référant à la science statistique et d'étudier les régularités et les dissemblances d'usage de démarche statistique, mais aussi de repérer les spécificités disciplinaires. Cette étude est descriptive des démarches et techniques statistiques utilisées, ainsi que des outils utilisés et des niveaux de preuve appréciés dans le cadre de mise en œuvre. Nous avons aussi relevé les usages de terminologies et les spécificités linguistiques en fonction des disciplines, ainsi que la formalisation des présentations.

Le terrain d'étude correspond aux thèses des écoles doctorales des établissements sur le site de Bordeaux. Les thèses sélectionnées ont été choisies dans la population de thèses soutenues entre janvier 2012 et décembre 2013 et déposées sur "theses.fr" quel que soit le domaine disciplinaire.

Le nombre de thèses soutenues, entre le 1 janvier 2012 et le 30 décembre 2013 dans une des écoles doctorales des 4 universités de Bordeaux était de 612. Une sélection a été réalisée avec un taux de sondage global de 12% en stratifiant selon les écoles doctorales afin d'assurer l'hétérogénéité disciplinaire au sein de l'échantillon d'étude.

Dans un premier temps, nous avons appliqué une méthode d'interrogation de ce corpus de thèses. Et nous avons localisé des extraits illustrant notre thème « science statistique » à partir des formes : *statis, échantillon,* m*oyenne, varia, regress* et *correl, Gauss, parametre* et *estim, proba* et *aleatoire, représenta, significat.*

4. *Résultats*

La proportion de 55 thèses sur 73 se référant, selon les thèses, à des éléments statistiques ou à une totale démarche statistique est un marqueur de la place de cette discipline dans la construction de thèses développées par les chercheurs de nombreuses disciplines. Les thèses de sciences expérimentales, démographie, santé publique, et économie sont tout particulièrement impactées par ces usages mais il ne faut pas négliger la part occupée par la statistique dans les autres thèses.

Une grande hétérogénéité a ainsi été soulignée quant au niveau d'explications et de précision lors de la présentation de l'échantillonnage et du recueil de données d'intérêt pour l'analyse statistique. Il est à souligner de réelles spécificités disciplinaires. Pour les disciplines de sciences et techniques, et de sciences de vie et de la terre, les processus de recueil, associés souvent à des situations expérimentales, sont présentés avec beaucoup de précision ; cette étape étant considérée pour ces

disciplines comme centrale de la démarche statistique.

Quelles que soient les disciplines, les termes les plus classiquement connotés (moyenne, écart-type, corrélation) sont plutôt utilisés et formulés de la même façon mais rarement redéfinis, induisant dès lors que les lecteurs comprennent et partagent le même sens de manière implicite. Ces notions sont vues dans tous les enseignements de bases en statistique ou ont une existence en dehors des murs scolaires, ce qui peut ainsi en justifier cet usage.

La moyenne, dans la majorité des thèses, caractérise de manière implicite la moyenne arithmétique, sauf lorsqu'il est précisé qu'il s'agit d'une autre formule de moyenne. Le terme "*moyenne*" est utilisé de manière indifférenciée pour représenter soit une statistique au sens fonction mathématique de données issues d'un échantillon, soit le résultat de cette fonction pour un échantillon spécifique, c'est-à-dire au sens d'indicateur. Les résultats de moyenne sont le plus souvent présentés accompagnés d'un indicateur représentant la notion de variabilité. Selon les habitudes disciplinaires cette présentation peut être une barre d'erreur représentant la notion de ± un écart-type, représentant ainsi l'information du niveau d'hétérogénéité de la distribution de la variable étudiée. Ce type de présentation est particulièrement présent dans le cadre des thèses de l'école doctorale sciences physiques et de l'ingénieur, et principalement en physique et chimie. D'autres habitudes disciplinaires consistent à représenter l'information avec l'erreur standard de la moyenne. Dans cette situation, il ne s'agit pas de décrire la distribution de la variable initialement recueillie, mais de déterminer un niveau de précision de l'estimation de la moyenne ; il s'agit d'ailleurs d'une situation plus proche de la présentation que l'on trouverait en utilisant un intervalle de confiance. Le choix de telle ou telle forme de représentation n'est pas souvent justifié par les auteurs, mais est implicite pour les lecteurs de la même communauté. Cependant, il est à souligner que la rédaction des articles, pour les thèses sur articles, est beaucoup plus précise et suggère moins d'implicite. Cela peut suggérer que les codes disciplinaires ne passent peut être pas les frontières, ou que les revues

scientifiques souhaitent offrir leurs éditions à un public plus large qu'une simple communauté.

Il est à souligner une plus forte hétérogénéité dans le cadre des représentations graphiques, tant pas le contenu que par la forme. Le but premier de l'utilisation de représentations graphiques est d'assurer une expression claire et compréhensible d'informations qui pourraient nécessiter un texte long et difficile. Cependant, la lecture des différentes thèses permet de mettre en évidence que l'usage de ce type d'outils de diffusion de résultats est difficile à appréhender, les graphiques proposés par les doctorants n'étant globalement pas jugés comme aidant pour la compréhension

Les logiciels statistiques utilisés sont multiples et souvent différents d'une discipline à l'autre. Une culture disciplinaire d'usage des outils informatiques est réellement présente, et il est à noter que certains logiciels ont été directement développés par des chercheurs de la discipline, ceci marquant clairement l'existence d'outils statistiques propres à certaines disciplines nécessitant des développements informatiques spécifiques. Il est cependant important de souligner que le logiciel libre R, le plus cité dans des thèses de disciplines différentes permet de couvrir un champ très large d'usages de par le développement de paquets spécialisés par toute une communauté de chercheurs mettant en accès les programmes qu'ils développent dans le cadre de leur recherche.

La présentation des méthodes statistiques est plus ou moins explicitée selon les thèses et il en est de même de l'explication des résultats obtenus. L'analyse des contenus des différentes thèses suggèrent que le développement de la partie méthode statistique est proportionnel à l'habileté du doctorant dans ces usages et dans sa compréhension des objets. Ainsi, lorsque certains doctorants intègrent de façon brute les sorties de logiciels, on peut s'interroger sur le manque d'effort didactique. Est-ce une pratique due à l'inquiétude de ne pas tout transmettre au lecteur pour une analyse complète de la situation ? Cependant, ce type de comportement peut suggérer un manque de compréhension de la

totalité de la sortie du logiciel et possiblement un défaut de compétence.

La place importante de l'activité de modélisation est à souligner dans les thèses, avec une place assez importante des modèles de régression classiques dans les usages toutes disciplines confondues. Plusieurs doctorants signalent s'être intéressés à des modèles utilisés dans d'autres communautés que leur communauté, et cela traduit réellement un effet de changement d'habitat des modèles qui, dès lors, évoluent souvent différemment dans chaque environnement mais cependant avec des bases et cadres communs. Il faut souligner cependant que les thèses utilisant des modèles plus sophistiqués ou moins classiques sont des thèses ayant dans leur ensemble une approche statistique assez riche, avec des développements d'écriture mathématique.

5. Conclusion

Au moins trois facteurs peuvent influencer l'importance de la place de la statistique dans la construction de l'argumentaire d'une thèse : la formation antérieure du doctorant et son propre jugement de ses compétences dans ce champ disciplinaire, la représentation par son directeur de thèse de la statistique comme discipline indispensable dans la mise en œuvre d'une recherche scientifique et, enfin, la reconnaissance par la communauté scientifique de la légitimité de la statistique.

« *Dans de nombreuses situation les chercheurs ne cherchent pas à travailler sur une dialectique commune* » disait G. Brousseau lors d'un séminaire organisé par l'équipe de recherche ADES du 14 octobre 2013. Ce constat peut expliquer la difficulté de la mise en place de projets de recherche interdisciplinaires.

L'ensemble de ces éléments sont ainsi à prendre en considération dans la réflexion de la formation universitaire et en tenant compte des changements de paradigmes que semble vivre l'université. Les réflexions disciplinaires nourrissent ainsi la réflexion de la pédagogie universitaire, mettant ainsi en connexion les changements environnementaux de la formation universitaire, mais aussi des publics accueillis, les évolutions sociétales et professionnelles.

Bibliographie

[1] Desrosières, A. (2000). *La politique des grands nombres : histoire de la raison statistique*. Paris: La Découverte. 456p.
[2] Latour, B. (2008). *La vie de laboratoire : la production de faits scientifiques*. Paris: La Découverte. 300p.
[3] Lonka, K., Chow, A., Keskinen, J., Hakkarainen, K., Sandström, N., & Pyhältö, K. (2014). How to measure PhD. Students' conceptions of academic writing and are they related to well-being? *Journal of Writing Research*. pp 245—269.