INDISCIPLINES

Statistique et recherches interdisciplinaires

Implication d'une discipline sans objet

Francis Laloë

éditions Quæ

Statistique et recherches interdisciplinaires

Implication d'une discipline sans objet

 ISBN : 978-2-7592-2533-0 ISSN : 1772-4120

INDISCIPLINES

Statistique et recherches interdisciplinaires

Implication d'une discipline sans objet

Francis Laloë

éditions Quæ

INDISCIPLINES

La collection « Indisciplines » fondée par Jean-Marie Legay dans le cadre de l'association « Natures Sciences Sociétés-Dialogues » est aujourd'hui dirigée par Marianne Cohen. Dans la même orientation interdisciplinaire que la revue *NSS*, cette collection entend traiter des rapports que, consciemment ou non, les sociétés entretiennent avec leur environnement naturel et transformé à travers des relations directes, des représentations ou des usages. Elle mobilise les sciences de la terre, de la vie, de la société, des ingénieurs et toutes les démarches de recherche, éthique comprise. Elle s'intéresse tout particulièrement aux questions environnementales qui interpellent nos sociétés aujourd'hui, qu'elles soient abordées dans leur globalité ou analysées dans leurs dimensions les plus locales.

Le comité éditorial examinera avec attention toutes les propositions d'auteurs ou de collectifs qui ont adopté une démarche interdisciplinaire pour traiter de la complexité.

Sommaire

Introduction

Ce livre rend compte d'une expérience de chercheur statisticien à l'Institut de recherche pour le développement (IRD), un institut au sein duquel on ne pose pas *a priori* de questions de recherche en statistique.

Dans un tel institut, le statisticien peut se placer dans une situation inconfortable, voire intenable, s'il revendique que sa discipline, avec ses outils, ses concepts et ses démarches, soit jugée équivalente à celles qui posent ces questions de recherche *a priori*. Il convient, en première analyse au moins, de considérer la statistique comme une discipline sans objet propre, exerçant une activité de service justifiée par la recherche de réponses à des questions posées par d'autres[1].

C'est dans sa mise en œuvre que le champ de cette activité peut nécessairement s'élargir lorsque le service conduit à produire des résultats relatifs à un domaine plus vaste que celui défini par les questions initialement posées[2].

Cet élargissement est le fruit d'une dynamique dans laquelle une interdisciplinarité élargie (Jollivet et Legay, 1995) est nécessairement mise en œuvre pour permettre une confrontation à la réalité et considérer la représentation de systèmes complexes.

Cette interdisciplinarité élargie est double. Elle associe tout d'abord différentes disciplines thématiques avec leurs différents points de vue sur un objet commun. Par ailleurs, pour mettre en œuvre cette interdisciplinarité élargie, il s'agit aussi d'associer ces disciplines thématiques et des disciplines des sciences de l'information, dont la statistique.

C'est à la présentation du rôle de cette dernière dans ce contexte de double interdisciplinarité que ce texte est consacré.

---

1 C'est bien pour mener une activité dans un tel contexte que j'ai été recruté en 1979 comme « mathématicien informaticien en océanographie », au vu d'une formation hybride en statistique mathématique et en génétique quantitative, et que j'ai été affecté au centre de recherches océanographiques à Dakar pour collaborer aux recherches menées – par des biologistes – sur la pêche artisanale au Sénégal.

2 Dans la suite, ces questions seront qualifiées de thématiques, posées par des thématiciens relevant de disciplines elles-mêmes qualifiées de thématiques.

DOUBLE INTERDISCIPLINARITÉ ET COMPLEXITÉ

Dans une interdisciplinarité élargie, un objet commun a une existence propre, généralement attestée sous forme d'un enjeu ou d'un domaine général au sein même du titre des programmes interdisciplinaires. Cet objet n'est donc pas une construction *a priori* issue d'une ou plusieurs disciplines thématiques, comme par exemple dans la planification des expériences de l'époque fisherienne. Il s'agit d'un objet réel dont on ne peut pas faire à ce titre une représentation définitive et unique. On peut par contre en faire une multitude de représentations, chacune étant une construction. Dans le cadre d'une interdisciplinarité élargie, de telles représentations peuvent être produites, correspondant aux étapes d'une problématique progressive (Chassany et Jollivet, 1997) : chaque représentation est construite à partir des résultats obtenus lors de l'étape précédente et des questionnements qu'ils suscitent.

La multiplicité des représentations possibles est inéluctable dès lors que l'objet commun peut être considéré selon plusieurs points de vue, chacun privilégiant un ou plusieurs éléments particuliers. Les représentations sont autant de combinaisons de ces points de vue et de ces éléments. Elles sont nécessairement relatives à des systèmes complexes selon la définition proposée par Legay (1997) :

> Est complexe un système que la perte d'un de ses éléments fait changer de nature.

Une procédure d'observation peut être associée à chacune de ces représentations ; il s'agit d'une confrontation à la réalité, qui participe à la construction des représentations à venir. Cette observation est une expérience, selon la définition également proposée par Legay (1993) :

> [Expérience] toute procédure organisée d'acquisition d'information qui comporte, dans la perspective d'un objectif exprimé, une confrontation avec la réalité.

Ces deux définitions, relatives aux systèmes complexes et à l'expérience, ont joué un rôle déterminant dans ma démarche personnelle en venant élargir le champ d'application de la statistique « classique ». En reprenant les termes utilisés par Legay, l'acquisition d'information (l'observation), la procédure qui lui est associée et l'analyse de ses produits en référence à l'objectif exprimé ont un rôle moteur dans la dynamique interdisciplinaire des représentations.

Ma thèse est que la participation de disciplines non thématiques, dont l'objet est l'information, est nécessaire pour tenir ce rôle, au-delà de l'expertise technique qu'elles apportent. Ces disciplines, et encore plus leurs méthodes, sont bien sûr nombreuses[3].

3. Un exemple en est la démarche construite et proposée par le groupe ComMod (Collectif ComMod, 2005) dans laquelle diverses méthodes en nombre non limité (modèles informatiques multi-agents, jeux de rôles…) sont au cœur de la démarche d'un groupe interdisciplinaire affichant une volonté de confrontation à la réalité dans le cadre de programmes comportant des objectifs finalisés d'aide à la décision. Dans cet exemple, une interdisciplinarité élargie est nécessaire pour la construction d'un objet commun avec des acteurs du développement.

La statistique est l'une d'entre elles, en s'intéressant à l'observation, sa conception, sa réalisation, sa restitution.

LA PLACE DE LA STATISTIQUE, DISCIPLINE SANS OBJET

La statistique est non thématique en n'ayant pas de point de vue sur l'objet réel étudié. Si par exemple cet objet réel est l'exploitation d'une ressource vivante, les biologistes vont poser des questions sur la biologie de cette ressource dans les conditions de son exploitation, et les questions des sciences humaines et sociales vont porter sur l'analyse de cette exploitation et de son contexte, conditionnellement à l'état et aux propriétés de la ressource exploitée. Dans les deux cas, l'observation de l'exploitation et sa restitution se feront au travers d'une représentation construite à partir d'un questionnement particulier. Si ces questionnements sont définis indépendamment les uns des autres, ils peuvent conduire à des procédures d'observation séparées. Mais il est également possible de restituer une même observation selon des représentations différentes ; cela se produit normalement dans le cadre de programmes interdisciplinaires lorsque les mêmes données peuvent être utilisées pour l'estimation de paramètres dont les définitions relèvent de points de vue éloignés ; cela peut être une source de tensions entre disciplines.

L'absence de point de vue statistique sur l'objet réel permet de ne pas le confondre avec une quelconque de ses représentations. Cette absence a un prix : contrairement à un chercheur thématicien, un statisticien ne peut pas mener seul, en tant que tel, une étude sur un objet réel.

Cette absence de confusion peut être mise à profit si la statistique ne reste pas obligatoirement confinée au seul rôle de service d'une discipline thématique. Dans ce dernier cas en effet, l'observation de l'objet a pour unique objectif de permettre de répondre à une question thématique et la critique du résultat ne peut être faite que de ce seul point de vue. L'objet peut donc être confondu avec la représentation définie par ce point de vue que sert la statistique. À l'inverse, cette confusion ne peut être maintenue si on doit pouvoir répondre à des questions relevant de points de vue différents, associées à des représentations différentes qui peuvent être légitimes.

Pour autant, l'éventuel besoin de dépasser le rôle de service ne signifie en aucun cas le rejet de l'idée de service. Il s'agit seulement de savoir au service de qui ou de quoi on se met. Ici, l'histoire du mot « observation » (Anonyme, 2000) est éclairante ; il peut tout autant être relatif à ce qui est observé (dans le sens d'une règle qu'on respecte) qu'à l'action même de l'observer (en le respectant ou le décrivant). Au-delà, la construction du mot renvoie à ce qui est « au devant de », « au vu de » (*ob*) et ce à quoi on est attentif (*servare* : préserver, sauver…) (Anonyme, 2000).

Dans l'interdisciplinarité élargie, la statistique, tout comme chacune des disciplines impliquées, est donc au service de l'objet réel, au travers – de façon concrète et transitoire – de diverses représentations. L'activité commence donc toujours par un service comme dans le cas d'une simple application mise en œuvre

pour répondre à une question dont la formulation reste inchangée. La nature de la contribution de la statistique sera ensuite différente selon que ce service peut être ou non considéré comme rendu sans avoir à reformuler le questionnement initial.

En résumé, l'activité statistique peut être menée dans au moins trois contextes différents :

– le premier est celui de la statistique pure. Il s'agit de produire des concepts et des méthodes. La statistique est une discipline qui relève des mathématiques, avec des concepts pouvant parfois être reliés à des considérations tout à fait générales dans le langage courant. Nous y reviendrons par exemple avec le terme d'indifférence qui peut être rapproché du concept d'exhaustivité (Fisher, 1925) qui est à mon avis fondateur de la statistique moderne. Même si ce contexte de la statistique pure n'est pas directement abordé dans le présent exercice, la question de la relation, au sein de programmes interdisciplinaires, entre statistique mathématique et données d'observation présente un certain intérêt. Elle est par exemple très différente de la relation, souvent présentée comme spécifique, entre mathématiques et physique (voir par exemple Lévy-Leblond, 1982) ;

– le second est celui de la statistique appliquée. Il s'agit de mettre en œuvre un outil, avec les concepts qui lui sont attachés, pour répondre à des questions thématiques non statistiques. L'objectif est de produire des statistiques, c'est-à-dire des fonctions de données – en général des quantités – qui vont constituer la contribution de ces données à la réponse aux questions posées. Dans ce cas, même si la statistique a développé des concepts permettant de caractériser la qualité de ces statistiques – et donc de ces réponses –, cette qualité est avant tout appréciée du point de vue thématique à partir duquel les questions ont été posées. Dès lors la démarche statistique est très logiquement perçue comme quantitative par les clients au service desquels elle est mise en œuvre ;

– le troisième est celui de la statistique impliquée, lorsque la mise en œuvre des outils et des concepts dans un contexte d'application conduit à montrer le besoin d'une reformulation des questions et à contribuer à cette reformulation. Dans ce cas, la qualité des réponses ne peut plus être appréciée du ou des seuls points de vue thématiques. Il faut pouvoir dire dans quelle mesure les données ont été bien traitées, indépendamment des questions posées et admettre la possibilité d'évaluations contradictoires. L'aspect qualitatif de la statistique doit être explicite pour que sa contribution soit effective.

Les frontières entre ces trois situations sont bien sûr très poreuses, nombre des résultats théoriques dans le domaine de la statistique pure résultent de questionnements identifiés dans le cadre des deux autres contextes.

L'EXHAUSTIVITÉ, UNE QUALITÉ STATISTIQUE FONDATRICE

Le choix d'un premier chapitre n'a pas été immédiat. Deux aspects en effet pouvaient être envisagés, l'un introduisant la question générale de la production de statistiques

dans un contexte d'implication, l'autre présentant les concepts proprement statistiques concrétisant l'autonomie de la démarche dans ce même contexte d'implication.

Il s'agit en effet de traiter de la question de produire des statistiques qui, une fois introduite, peut se décliner selon plusieurs aspects relatifs aux observatoires, aux indicateurs, au cas particulier de l'environnement, aux mésusages et aux pièges.

Mais, et peut-être surtout, dans un contexte de service, la qualité d'une statistique (et donc celles de l'analyse et du traitement de données dont elle résulte) doit être évaluée d'au moins deux points de vue, celui de la démarche statistique et celui de la question à laquelle elle apporte un élément de réponse. Cela conduit à afficher la spécificité de la démarche, impliquant qu'elle ne peut pas *a priori* être réduite au cadre de la ou des disciplines thématiques avec lesquelles la collaboration est entreprise. Cette réflexion s'appuie donc inéluctablement sur des concepts et des outils propres à la discipline qu'est la statistique, concepts et outils qu'il convient d'exposer dans une présentation qui ne soit pas saupoudrée au fur et à mesure de leur évocation.

Cette présentation sera faite dans le premier chapitre. Le premier concept, considéré comme fondateur et que j'utiliserai à de nombreuses reprises comme référence, est celui de statistique exhaustive, proposé en 1925 par R. Fisher sous le terme de *sufficient statistic*. Si les utilisateurs de la statistique connaissent en général l'existence des tests de Fisher, peu d'entre eux ont entendu parler de statistiques exhaustives.

L'idée est de pouvoir remplacer un ensemble de données par une synthèse, c'est-à-dire par un nombre limité de fonctions de ces données, sans perte d'information. Si tel est le cas, on peut qualifier cette synthèse d'exhaustive selon la définition donnée par R. Fisher :

> Une statistique exhaustive est équivalente, pour toute opération d'estimation à venir, aux données dont elle est une synthèse.
>
> Traduction de l'anglais : *« [a sufficient statistic] is equivalent for all subsequent purpose of estimation to the original data from which it was derived. »*

Les conditions dans lesquelles cette qualité d'exhaustivité peut être présente doivent être précisées, impliquant le recours à des définitions et à un formalisme mathématiques. Ce formalisme est très utile en permettant de montrer qu'une synthèse est toujours associée, explicitement ou implicitement, à un modèle statistique associé à une représentation particulière. Il est également utile en montrant l'importance cruciale des protocoles d'observation, permettant ou non d'assurer de « bonnes propriétés », par exemple associées à l'indépendance des observations.

Mais il sera également intéressant de montrer à partir de quelques illustrations qu'il s'agit d'une qualité en relation directe avec le sens commun. Il en est de la

statistique comme de la prose de Monsieur Jourdain : on produit en permanence dans la vie courante des synthèses des observations qu'on a pu faire.

La présentation de l'exhaustivité conduit par ailleurs de façon naturelle à celle de l'estimation, qui est bien sûr un outil (avec un ensemble de méthodes), mais qui repose également sur des concepts qui peuvent mériter l'attention, en observant que l'estimation d'une ou de plusieurs quantités est une fonction de données. Une estimation est à ce titre la réalisation d'une variable aléatoire, un estimateur, qu'on peut entre autres choses caractériser par les espérances et les variances de ses éléments, ainsi que par les covariances et les corrélations entre ses éléments.

Cette présentation conduit également à celle d'un autre concept essentiel de la statistique – la vraisemblance – illustrant le fait qu'une synthèse n'est en définitive justifiable que si elle est associée, comme annoncé plus haut, à une représentation, un modèle statistique fondé sur la loi de la variable aléatoire dont les données disponibles sont une réalisation. Sans une telle représentation en effet, la seule synthèse exhaustive possible est l'absence totale de synthèse, c'est-à-dire la restitution fastidieuse et insipide de toutes les données disponibles.

Les termes modèle et représentation sont ici relatifs à la distribution d'une variable aléatoire. Ce n'en est pas le sens le plus courant. Nous y reviendrons avec le modèle linéaire qui associe de fait deux modèles de natures différentes, l'un rendant compte de l'espérance (non aléatoire) d'une variable aléatoire et l'autre de l'écart aléatoire entre cette espérance et les observations. Lorsqu'on parle de modèle de régression linéaire $Y = ax + b + \varepsilon$ on devrait ainsi se représenter à la fois la droite de régression et la distribution des écarts (les résidus ε). Ces deux modèles sont inextricables dès lors que, d'une part, les propriétés de la distribution des résidus définissent celles des distributions des estimateurs des paramètres de la droite de régression et que, d'autre part, une des qualités essentielles de la distribution des résidus (nullité de leur espérance) est garantie par la qualité de la droite ($E(Y) = ax + b$) en tant que représentation de l'espérance de la variable étudiée. La qualité d'un modèle de régression linéaire, et donc son utilité, ne peuvent donc pas être appréciées à partir de l'examen d'une seule de ces deux composantes. Cette question de modèle et de représentation sera également abordée dans la suite de cette introduction à propos de la production de statistiques dans la mesure où elle s'impose dans la discussion relative à la légitimité et à la spécificité de la démarche statistique.

Au-delà de ces aspects, une situation particulièrement intéressante est celle où un ensemble d'estimateurs en nombre aussi réduit que possible constitue par ailleurs une statistique exhaustive. Cette situation est d'autant plus intéressante que ces estimateurs peuvent être ceux des paramètres d'un modèle. C'est le cas du modèle linéaire gaussien auquel une partie de ce chapitre sera consacrée. Le modèle linéaire gaussien permettra d'illustrer très concrètement quelques principes liés à la construction de protocoles d'observation, associés à des modèles, en recherchant

dans la mesure du possible à produire des estimateurs de leurs paramètres (ou de groupes de leurs paramètres) indépendants les uns des autres. Cette indépendance entre (groupes d')estimateurs est une propriété du protocole d'observation qui doit être à ce titre contrôlé ; elle offre un confort précieux d'interprétation (pouvoir faire des hypothèses sur l'estimation des paramètres d'un groupe indépendamment de celles faites sur d'autres). À l'inverse, les problèmes d'interprétation liés aux corrélations entre estimateurs pourront être exposés ; ils sont particulièrement importants ici parce que souvent liés aux difficultés de communication entre les disciplines, difficultés qui peuvent être irréductibles lorsque les plans associés à des données de terrain correspondent à des plans d'expérience non contrôlés et de fait non orthogonaux engendrant une dépendance entre les estimateurs de paramètres relevant de différents points de vue. Ces difficultés ont été par exemple abordées par Searle (1971) dans une partie consacrée au traitement des données d'enquêtes au sein d'un ouvrage sur le modèle linéaire.

En parlant de vraisemblance et de la loi de la variable aléatoire dont les données disponibles sont issues, on fait référence à l'un des fondements de la démarche statistique. Un ensemble de données n'a pas de sens en lui-même. Il n'est que la concrétisation d'une procédure d'observation : un protocole a été mis en place conduisant à sélectionner des individus dont l'observation conduit à disposer d'un ensemble (jeu) de données. Cet ensemble est aléatoire : si deux observateurs appliquent exactement les mêmes consignes, ils n'obtiendront pas les mêmes données. Un jeu de données n'est donc qu'une réalisation, parmi d'autres possibles, qui avait une certaine probabilité d'être réalisée. L'art de la statistique est de manipuler ces données, en produisant des statistiques, dont la qualité sera évaluée à partir des probabilités d'occurrence des jeux de données possibles… dont un seul est disponible. Le paradoxe lié à cette unicité d'observation peut être levé à partir des qualités des protocoles d'observations[4]

Une des difficultés de l'enseignement de la statistique est de bien faire ressentir cette idée de réalisation de variable aléatoire, signifiant que ce qu'on observe n'est qu'un événement particulier parmi ceux qui auraient pu se produire. Cela peut s'exprimer avec un formalisme mathématique (celui des espaces probabilisés) mais des textes littéraires peuvent être tout aussi efficaces, comme par exemple cet extrait du livre de Marguerite Yourcenar *L'Œuvre au noir* :

> L'avenir est gros de plus d'occurrences qu'il n'en peut mettre au monde. Et il n'est point impossible d'en entendre bouger quelques-unes au fond de la matrice du temps. Mais l'événement seul décide laquelle de ces larves est viable et arrive à terme. Je n'ai jamais vendu au marché de catastrophes ou de bonheurs accouchés d'avance.

4. Par exemple, si n individus d'une population donnée sont sélectionnés en appliquant les règles de l'échantillonnage aléatoire simple, on peut considérer qu'on dispose de n réalisations de variables indépendantes ayant toutes la même loi.

Si la statistique consiste pour une large part à produire des synthèses dont la qualité doit être analysée en référence à la distribution dont les données traitées sont issues, on peut inverser ce processus et rechercher, pour une synthèse particulière de données, quelle(s) serait (seraient) la (les) représentation(s) qui rendrait (rendraient) cette synthèse exhaustive. On dispose alors d'un outil pour participer à l'explicitation – et à la critique – des représentations implicitement définies par des synthèses particulières de données. Ici encore, un formalisme mathématique pourra être utilisé sur un exemple simple.

En dehors de tout formalisme mathématique, l'exhaustivité peut être rapprochée de la définition d'une information – « une différence qui fait une différence » – telle que proposée par G. Bateson (1972). En effet, si on prétend qu'une synthèse peut être substituée sans perte d'information au jeu de données dont elle est issue, cela implique que les différences entre jeux de données possibles (issus de la même distribution) ayant des synthèses identiques ne contiennent pas d'information : ces différences ne font pas de différence. Cela montre à nouveau qu'un jeu de données n'est rien d'autre qu'une réalisation particulière d'une variable aléatoire parmi une multitude de réalisations possibles : les différences entre les jeux de données qui nous disent la même chose sur cette variable ne font pas de différence.

Et il est remarquable enfin d'observer que la définition de Bateson se retrouve en négatif dans le langage courant lorsqu'on peut dire de certaines choses (certaines différences) qu'elles nous laissent indifférents.

Cette définition d'une information est particulièrement intéressante en permettant entre autres de faire le lien entre information et communication. Cela méritera discussion par la suite.

Un exemple simple permettra d'illustrer en situation nombre des considérations présentées. Il est fondé sur l'examen d'une synthèse de données collectées au cours de six jours consécutifs en 1978 à partir du village de Kayar au Sénégal. Cette synthèse consiste en un tableau de captures moyennes obtenues pour certaines espèces de poissons lors de sorties de pêche faites avec un même engin (lignes à main). Ces moyennes sont perçues comme des estimations d'indices d'abondances dont l'évaluation constituait l'objectif de l'observation mise en place. Ces « simples » moyennes constituaient par ailleurs la synthèse la plus directe des données collectées et leur qualité aurait à ce titre été d'autant plus satisfaisante qu'elles auraient pu être directement liées à une statistique exhaustive. Ce souhait pouvait être satisfait en considérant un modèle statistique avec ses hypothèses sur la distribution des variables dont les captures étaient des réalisations. Ces hypothèses consistaient pour l'essentiel à admettre que les espérances de ces variables sont proportionnelles aux abondances des espèces considérées. L'examen des distributions des observations a conduit à remettre en cause ces hypothèses. Cette remise en cause permettra

d'illustrer la question de la production de statistiques et justifiera une reconstruction du modèle statistique sous-jacent.

Cet exemple constituera à ce titre la première étape d'un fil rouge qui sera suivi tout au long de ce texte.

PRODUIRE DES STATISTIQUES

Le second chapitre sera une discussion générale autour de la question de la production de statistiques.

La qualité d'une statistique produite pour répondre à une question peut être évaluée en considérant deux conditions, toutes deux nécessaires mais pouvant être abordées séparément. La première relève de la démarche statistique en vérifiant que cette statistique peut être une partie d'une synthèse contenant un maximum possible de l'information – dans l'idéal une statistique exhaustive. Cette qualité doit être appréciée en référence au modèle statistique dont sont issues les données traitées. La seconde est d'ordre thématique en vérifiant que les paramètres de ce modèle statistique peuvent être interprétés en relation avec la question posée.

La spécificité de la démarche statistique au sein d'un programme interdisciplinaire conduit alors à poser la question de savoir « qui » produit des statistiques. Cette question n'est pas anodine. Elle est généralement posée en termes de compétence technique et il n'est alors pas possible d'y répondre. Au-delà des qualités relevant de la compétence technique, elle doit aussi porter sur la définition de l'ensemble des questions thématiques auxquelles les statistiques pourront apporter des éléments de réponses. Plus précisément, il s'agit de savoir si cet ensemble évolue au cours du déroulement d'un programme de recherches et si le statisticien peut jouer en tant que tel un rôle actif dans cette évolution, c'est-à-dire participer en tant que statisticien à la progression d'une problématique thématique en se référant à nouveau à l'idée de problématique progressive (Chassany et Jollivet, 1997),

Lorsqu'on s'en tient à la seule application, un thématicien peut très bien acquérir l'expertise technique nécessaire et la mettre en œuvre pour répondre au mieux à une question qu'il se pose dans son domaine thématique (on n'est jamais mieux servi que par soi-même). Il n'y a donc sans doute pas de réponse définitive à la question de savoir si, pour appliquer la statistique à un domaine, il convient de faire appel à un statisticien qui doit « accepter » la question à laquelle on lui demande de répondre ou à un thématicien qui possède une bonne compétence technique et qui pourra mettre en œuvre un outil pour répondre à sa question. Dans un contexte d'application, les deux solutions sont possibles, avec même l'apparition de disciplines hybrides dont le nom associe explicitement une thématique et une méthode se mettant à son service, telles que la biométrie (ou la biostatistique) et l'économétrie. Ronald Fisher en est sans doute un des représentants les plus emblématiques, mais

au prix d'une difficulté d'identification, en étant clairement adopté, et revendiqué, par plusieurs camps. C'est ce dont rend compte Savage (1976) :

> Il m'arrive de rencontrer des généticiens qui me demandent s'il est vrai que le grand généticien R. A. Fisher était aussi un statisticien important.
>
> Traduction de l'anglais : *« I occasionally meet geneticists who ask me whether it is true that the great geneticist R. A. Fisher was also an important statistician. »*

Mais, bien sûr, n'est pas Fisher qui veut. On peut simplement dire que Fisher était un statisticien qui a formulé, en développant et en mettant en œuvre sa démarche de statisticien, des questions et des réponses originales en génétique (et en agronomie) et en statistique. Ainsi en est-il de ses recherches sur les plans d'expériences, initiées à la suite de son recrutement, au départ pour une durée limitée, au centre de recherches de Rothamsted pour réaliser le traitement de jeux de données de grande taille (Droesbeke *et al.*, 1997).

Pour que le service soit rendu, l'outil doit être utilisé avec la rigueur nécessaire pour que soit satisfaite la première condition énoncée plus haut, relative à la mise en œuvre de la démarche statistique. Cette rigueur doit s'exercer tout au long du processus de l'application, depuis l'établissement des protocoles d'observation jusqu'à la synthèse des données collectées. Cela semble évident mais il n'est pas exceptionnel pour un statisticien d'être sollicité pour traiter des données après qu'elles ont été collectées, donc sans avoir participé à la construction du protocole d'observation dont elles sont issues. Les statisticiens redoutent cette situation qui peut souvent conduire à ne pas pouvoir répondre à la question posée, et donc engendrer des situations conflictuelles. C'est ce qu'exprimait de manière on ne peut plus explicite Ronald Fisher (1938) :

> Consulter le statisticien après la réalisation de l'expérience consiste souvent à lui demander d'en faire l'autopsie. Il peut peut-être dire de quoi l'expérience est morte.
>
> Traduction de l'anglais : *« To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of. »*

Mais, une fois encore, n'est pas Fisher qui veut[5].

Si l'outil n'a pas été appliqué avec la rigueur nécessaire, il devient difficile d'établir les qualités des statistiques produites en relation avec un modèle statistique et le risque est de ne pas pouvoir rendre le service attendu ou pire de produire des faux-

5. Ainsi, un statisticien, débutant théoricien désirant appliquer la méthode qui faisait l'objet de sa recherche, a proposé, sur mon conseil, ses services à des collègues écologues qui lui ont fourni un jeu de données avec la question à laquelle ils souhaitaient répondre. Je lui ai bien précisé qu'il fallait accepter ces données et la question, ce dont il est convenu. Le problème est ensuite que la réponse n'a pas entièrement satisfait ses interlocuteurs et mon jeune collègue m'a expliqué que, si la question lui avait bien été posée – comme je le lui avais indiqué –, j'avais omis d'ajouter que la réponse pourrait lui être aussi pour une large part « proposée ».

sens ou des contresens en voulant répondre à tout prix à la question posée et faire ainsi dire, selon le point de vue, plus ou moins qu'elles ne peuvent aux données.

Si le fait de ne pas rendre le service attendu est un échec dans le contexte d'une application, ce n'est plus nécessairement le cas dans celui d'une implication. Cela peut montrer le besoin de reformuler le questionnement qui peut alors ne plus relever du seul domaine thématique initial. La situation d'un thématicien compétent en statistique peut alors être inconfortable parce que la technique qu'il maîtrise n'apporte pas de réponse à son problème, mais conduit, bien au contraire, à le reformuler au-delà de son cadre thématique initial.

Pour pouvoir identifier et justifier ce besoin, deux conditions sont nécessaires. Il faut d'abord disposer d'une démarche indépendante des points de vue thématiques. Il faut, pour mettre en œuvre cette démarche, disposer d'outils permettant de faire le lien entre le protocole d'observation de l'objet réel, les données résultant de l'observation et la restitution de ces données selon une synthèse – fonction des données, c'est-à-dire par définition une statistique – dont la qualité doit pouvoir être définie indépendamment des points de vue à partir desquels la question a été posée. Comme annoncé plus haut, il s'agit donc bien alors avant tout d'une démarche qualitative, dont la qualité, contrairement au contexte de l'application, ne peut pas être appréciée, ni donc définie, à partir des seuls points de vue thématiques puisqu'il s'agit de pouvoir les dépasser, s'il y a lieu.

En effet, identifier un problème de formulation ne suffit pas à justifier le besoin d'une reformulation. Il peut simplement ne s'agir que de l'identification de la cause du décès de l'expérience en reprenant la citation de Fisher donnée plus haut. Il faut donc – et c'est la seconde condition – que cette reformulation intègre une nouvelle question légitime, constituant ainsi une nouvelle étape de la problématique progressive définie par Chassany et Jollivet (1997).

La réunion de ces deux conditions, démarche indépendante et question légitime, conduit à revenir de façon plus globale sur la spécificité et la légitimité de la démarche statistique en tant que discipline relevant des mathématiques. Cela conduit entre autres à préciser la relation avec la modélisation.

Une démarche indépendante

La qualité d'une statistique doit donc pouvoir être définie indépendamment de toute question thématique, en référence aux seules données dont elle est une fonction. Bien entendu, la qualité de cette statistique devra pouvoir par ailleurs être jugée sur sa pertinence pour la question posée. L'objectif peut être alors de fournir une synthèse – courte – qui soit équivalente à l'ensemble de ces données dont le nombre peut être considérable. C'est très exactement ce qu'exprime R. A. Fisher en 1922, cité par Varenne (2010) :

> […] dans sa forme la plus concrète, l'objet des méthodes statistiques est la réduction de données. Une quantité de données qui, en général, de par sa simple masse est incapable d'entrer en l'esprit, doit être remplacée par des quantités en nombre relativement faible et qui doivent représenter adéquatement le tout ou qui, en d'autres termes, doivent contenir le plus possible – idéalement la totalité – de l'information pertinente contenue dans les données originelles (traduction de F. Varenne).
>
> Traduction de l'anglais : « […] *in its most concrete form, the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.* »

Varenne (2010) tire une conclusion essentielle de cette assertion :

> Ce qui doit être représenté, c'est le tout non pas du phénomène, mais de l'information. L'objet de la statistique est donc de fournir une représentation d'information et non la représentation d'un objet ou d'un phénomène naturel.

Dans la citation qui vient d'être donnée de Fisher, il est annoncé que le nombre relativement faible de quantités (c'est-à-dire une statistique de dimension relativement faible) remplaçant les données originelles devrait idéalement contenir toute l'information pertinente contenue dans ces données. Fisher a proposé avec l'exhaustivité décrite plus haut un cadre permettant d'atteindre cet objectif dans certains cas, relevant pour l'essentiel des plans d'expérience contrôlés. En permettant de caractériser certaines conditions dans lesquelles on peut, ou non, rechercher une statistique pouvant être substituée sans perte d'information aux données dont elle est fonction, l'exhaustivité est bien une qualité de référence absolument non triviale et indépendante de tout point de vue thématique. Cette indépendance est relative à la qualité d'exhaustivité dont la définition ne fait pas intervenir celle des questions auxquelles elle permettra d'apporter des éléments de réponses. Cela ne signifie bien sûr pas qu'une statistique exhaustive soit indépendante de tout point de vue thématique puisque des éléments de réponse à des questions thématiques seront produits à partir d'une statistique qui peut être exhaustive ou non. C'est dans une très large mesure la qualité du protocole d'observation qui peut assurer que les deux qualités, statistique et thématique, soient toutes deux satisfaites.

Mais, bien sûr, cette qualité d'exhaustivité n'est en général pas accessible et, en reprenant la citation de Fisher, l'objectif est de produire une synthèse de dimension relativement faible contenant un maximum possible de l'information contenue dans les données. Il s'agit alors de rechercher un compromis. Une telle recherche peut être faite par exemple à l'aide d'analyses des données multivariées telles que les analyses en composantes principales ou les analyses factorielles des correspondances. Il s'agit de méthodes qui font largement référence à la géométrie euclidienne. Les mathématiques jouent donc encore un rôle ici et la question de la relation entre mathématiques, statistique, données d'observation et disciplines thématiques se