

Statistique et recherche interdisciplinaire - Implication d'une discipline sans objet



Francis LALOË

Statisticien – Ancien directeur de recherches à l'Institut de recherches pour le développement (IRD)¹

L'application de la statistique consiste en la recherche de réponses, sous forme de statistiques, à des questions posées en dehors de la discipline. La statistique appliquée est de ce point de vue sans objet propre, exerçant une activité de service. Mais cette « définition » ne suffit pas à caractériser la statistique appliquée. En effet, le champ d'application peut éventuellement s'élargir lorsque le service conduit à produire des résultats relatifs à un domaine plus vaste que celui défini par les questions initiales. La statistique participe alors à la dynamique du questionnement et son rôle dépasse le cadre de l'application pour entrer dans celui de l'implication.

Cet élargissement peut être légitime ou non, selon la nature et le contexte du service initialement demandé, service que le statisticien a accepté de rendre. Les choses peuvent ainsi être très différentes selon qu'il est rendu au sein d'un projet mono disciplinaire où la question relève du point de vue d'une discipline (ou d'une combinaison donnée a priori de disciplines), ou bien qu'il est rendu dans le cadre d'un projet interdisciplinaire dans lequel les points de vues de plusieurs disciplines sur un objet commun peuvent être confrontés.

Dans tous les cas, au départ, le service consiste en la production d'une synthèse, selon une statistique de dimension réduite par rapport à celle des données dont elle est une fonction. C'est ce que le grand statisticien Ronald Fisher (1890-1962) a résumé en des termes très généraux en précisant les qualités d'une telle synthèse :

« L'objet des méthodes statistiques est la réduction des données. Une certaine quantité de données, qui en général du simple fait de sa masse ne peut pas entrer dans l'esprit, doit être remplacée par un nombre relativement petit de quantités qui représenteront adéquatement l'ensemble, ou qui, en d'autres termes, contiendront la plus grande part possible de l'information contenue dans les données d'origine – idéalement, la totalité. »²

1. Ce texte présente, en reprenant son titre, mon livre publié en 2016 dans la collection « Indisciplines » des éditions Quae.

Voir : <http://www.quae.com/fr/r4969-statistique-et-recherches-interdisciplinaires.html>

2. « [...] 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 » Fisher R. A., (1922), On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, A, 222, p. 309-368.

L'objet de la statistique est donc de fournir une représentation d'information et non la représentation d'un objet ou d'un phénomène³. Cette représentation doit être la plus courte (synthétique) possible et contenir le plus possible de l'information, voire sa totalité⁴. Ces deux qualités sont en première analyse contradictoires dès lors qu'une synthèse ne permet pas la reconstitution exacte de l'observation en « gommant » les différences entre les jeux possibles de données dont les synthèses sont identiques⁵. D'une manière générale, ces différences sont d'autant plus nombreuses que la synthèse est courte.

Une situation optimale est celle dans laquelle une telle statistique est en lien direct avec la question posée. C'est ce qu'on recherche en construisant un protocole d'observation⁶ exclusivement dédié à cette question, auquel cas l'objet et la question sont en quelque sorte confondus. Par exemple, dans le cas d'une recherche relative à l'effet de diverses pratiques de culture sur le rendement de diverses variétés d'une plante, deux points de vue, sur les pratiques et sur les variétés, sont combinés a priori à l'intérieur d'un « plan d'expérience ». La théorie enseigne un résultat essentiel : si le plan est bien conçu, on peut répondre aux questions relatives aux sources de variation indépendamment les unes des autres et l'ensemble des estimations est bien relatif au questionnement qui est à l'origine de la construction de l'expérience. La qualité du protocole d'observation est donc essentielle.

Si le plan d'expérience est « déséquilibré », on ne peut pas estimer les effets des sources indépendamment les uns des autres. Cela peut parfois conduire à des malentendus lorsque la personne en charge du traitement des données doit conclure que la qualité du protocole d'observation ne permet pas d'estimer les effets d'une source sans faire d'hypothèse sur les effets des autres. C'est ce que Fisher a parfaitement résumé ainsi :

« Consulter le statisticien après la fin d'une expérience, cela revient souvent à lui demander seulement une autopsie. Il peut peut-être dire de quoi l'expérience est morte »⁷

Ici, la personne en charge du traitement peut être un statisticien qui a accepté de rendre un service ou bien la personne qui a posé la question et qui maîtrise, et parfois même conçoit et développe, l'outil statistique. Cette double compétence est ainsi affichée par certaines disciplines « hybrides » telles que la biostatistique ou encore l'économétrie... au prix d'une difficulté d'identification comme l'indique à propos de Fisher son collègue L.J.Savage :

« Je rencontre de temps en temps des généticiens qui me demandent s'il est vrai que le grand généticien R.A.Fisher était aussi un statisticien important »⁸.

-
3. Ce point de vue est développé dans : Varenne F., (2010), *Formaliser le vivant : Lois, Théories, Modèles ?* Visions des sciences, Hermann
 4. R. Fisher a identifié des situations dans lesquelles une synthèse de dimension réduite peut contenir toute l'information. Il définit ainsi la qualité d'exhaustivité (traduction de « sufficiency ») pour une statistique : *« [a sufficient statistic] is equivalent for all subsequent purpose of estimation to the original data from which it was derived »* Fisher R. A. (1925), *Theory of statistical estimation*, Proc. Camb. Philos. Soc., 22, p. 700-725. Cette qualité ne fait pas référence à la question initiale à laquelle on cherche à répondre mais elle établit que toute l'information contenue dans les données et relative à cette question pourra être exprimée comme une fonction d'une statistique exhaustive
 5. Cette qualité peut s'exprimer naturellement en référence à la définition d'une information selon Bateson : « une différence qui fait une différence ». Une synthèse gomme des différences qui ne font pas de différences ou, en d'autres termes, qui laissent indifférents. Voir : Bateson G. (1972), *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*, University of Chicago Press
 6. L'histoire et l'étymologie du mot observation (Dictionnaire historique de la langue française, Dictionnaires Le Robert, Paris, 2000) rendent bien compte de l'idée de service ; le mot 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...)
 7. *« 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 »* Citation extraite de : Fisher R.A. (1938) Presidential address to the first indian statistical congress *Sankhya* 4
 8. *« I occasionally meet geneticists who ask me whether it is true that the great geneticist R.A. Fisher was also an important statistician »* Savage L. J. (1976), *On rereading R. A. Fisher*, *The Annals of Statistics*, 4(3), 441-500

La collaboration au sein d'un projet interdisciplinaire

Dans ce qui précède, il s'agissait de répondre à une question explicite et de construire un objet (une expérience) dédié à cette question. La situation est très différente dans le cas d'une participation à un projet interdisciplinaire dans lequel les points de vue de plusieurs disciplines sur un objet commun peuvent être confrontés. Cet objet a une existence propre, souvent attestée sous forme d'un enjeu ou d'un domaine général au sein même du titre des programmes interdisciplinaires.

Dans ce contexte, la référence explicite à un tel objet commun est nécessaire et l'exemple d'une exploitation halieutique sera utilisé dans les propos qui vont suivre : Il s'agit d'un objet réel dont on ne peut pas faire à ce titre une construction définitive et unique. On peut, par contre, en faire une multitude de représentations, chacune étant une construction. Cette multiplicité est inéluctable dès lors que l'objet commun peut être observé depuis plusieurs points de vue, chacun privilégiant un ou plusieurs éléments particuliers. Les constructions sont autant de combinaisons de ces éléments ; elles sont nécessairement relatives à un système complexe selon la définition proposée par Jean-Marie Legay⁹ :

« Est complexe un système que la perte d'un de ses éléments fait changer de nature »¹⁰

Une procédure d'observation peut être associée à chaque construction et une observation peut par ailleurs être restituée selon plusieurs d'entre-elles¹¹ ; l'observation est une confrontation à la réalité. Il s'agit d'une expérience, selon la définition également proposée par Legay :

« [On appelle expérience] toute procédure organisée d'acquisition d'information qui comporte, dans la perspective d'un objectif exprimé, une confrontation avec la réalité »¹²

Dans ce cas, la confusion entre sources de variation ne résulte plus nécessairement d'une erreur de protocole. Dans le domaine halieutique on peut ainsi supposer, pour une espèce donnée, que les captures obtenues lors d'actions de pêche sont des réalisations de variables aléatoires dont l'espérance est positivement corrélée avec l'abondance. On peut même faire l'hypothèse que les rendements sont proportionnels à cette abondance en supposant que chaque poisson a la même probabilité d'être capturé lors d'une action de pêche (capturabilité constante). Mais il est possible que ces variations de rendements soient (aussi) liées au comportement des poissons qui peuvent être plus ou moins accessibles... et bien sûr les variations d'accessibilité peuvent conduire les pêcheurs à adapter leur pratique, soit pour maintenir le contact avec l'espèce qu'ils recherchent – auquel cas ils réduisent les variations d'accessibilité – soit pour rechercher une autre espèce plus accessible – auquel cas ils exacerbent ces variations.

Une série chronologique de rendements de pêche peut donc faire l'objet de synthèses selon différents points de vue, chacun offrant une interprétation cohérente¹³. On peut y voir des variations d'abondance ou des variations d'accessibilité de la ressource ou encore le reflet des décisions des pêcheurs dans le déroulement de leurs actions. Le problème est alors qu'en privilégiant un de ces points de vue on fait une hypothèse, généralement implicite, sur ce qui pourrait être vu à partir des autres : assimiler des variations de rendements à des variations d'abondance implique qu'on suppose que la capturabilité reste constante et que les pêcheurs font toujours la même chose... La confusion d'effets peut être masquée par le discours et être

9. Jean-Marie Legay biométricien français décédé en 2012

10. Legay J.-M. (1997), *L'expérience et le modèle. Un discours sur la méthode*, coll. « Sciences en questions », Inra.

11. La possibilité d'utiliser une observation selon plusieurs points de vue est sans doute une qualité propre à un observatoire.

12. Legay J.-M. (1993), Une expérience est-elle possible ? In J.-D. Lebreton et B. Asselain (eds), *Biométrie et environnement*, Masson, Paris, pp. 1-14

13. Legay (1997) considérait ainsi que la « complexité est d'une décision » dont une des conséquences est la perte des critères d'évidence » au profit de « réseaux de cohérence ».

source de malentendus, de manipulations, de négation de l'intérêt d'autres points de vue et d'incapacité de communication entre disciplines¹⁴ lorsqu'un point de vue disciplinaire conduit à masquer les variations perceptibles selon les points de vue d'autres disciplines.

Si la question n'était qu'une de celles évoquées ci-dessus, être confronté à cette difficulté indique que l'expérience est morte et qu'il faut reconsidérer le protocole d'observation. Ainsi pour estimer un indice d'abondance selon une moyenne de captures réalisées lors d'actions entreprises par des unités de pêche¹⁵, il faut utiliser les captures dont l'espérance est égale, à un facteur de proportionnalité près, à cette abondance (capturabilité constante). On doit donc privilégier l'observation des unités qui recherchent toujours la même espèce, avec la même efficacité¹⁶. Il se peut que cette qualité soit associée aux caractéristiques des unités de pêche selon, par exemple, qu'elles peuvent ou non poursuivre une espèce dans ses déplacements. Il sera alors légitime de privilégier l'observation des unités industrielles qui ont un rayon d'action supérieur à celui des unités artisanales¹⁷. Il en découle un problème de déontologie¹⁸ si l'objet de référence est l'exploitation halieutique et qu'on ne peut exclure a priori que toutes les unités de pêche puissent faire l'objet de questions légitimes. En effet, si un protocole construit en référence à une question particulière conduit à négliger l'observation de l'activité et des résultats de certaines unités de pêche, il en découle une moindre capacité de réponse aux questions relatives à ces unités. Ce problème traduit un paradoxe quant au profil professionnel de la personne qui construit le protocole et « produit » la synthèse. S'il s'agit de celle qui a posé la question et qui maîtrise l'outil statistique, elle aura mis son expertise en œuvre de façon à ce que le protocole d'observation qu'elle aura mis en place – ou au moins validé – produise un résultat dont la synthèse soit sans ambiguïté en relation avec la question qu'elle a posée ; il ne faut pas que d'autres questions soient soulevées par l'analyse des données... La situation est évidemment totalement différente si cette personne est un statisticien qui accepte de répondre à des questions, posées par d'autres, relatives à un objet ayant une existence propre.

14. Cette difficulté apparaît dans le titre même d'un article de Myers et Worm : « Rapid worldwide depletion of predatory fish communities. » paru dans la revue *Nature* (numéro 423 - 2003) où des données de rendements de pêche sont implicitement supposées parfaitement refléter des abondances. Un site web réunit un nombre important de réactions critiques http://www.soest.hawaii.edu/PFRP/large_pelagics/large_pelagic_predators.html

15. « Unité de pêche » consiste ici en une « entreprise » qui pratique la pêche dans le site considéré, avec un centre de décision, des moyens et des connaissances (pirogues, bateaux, outils, savoir faire...).

16. Plus précisément, ces moyennes peuvent être pondérées en admettant que les unités de pêche peuvent être caractérisées par des capturabilités différentes selon la puissance de leurs moteurs, la taille ou le type d'engin de pêche etc. On suppose alors que la capturabilité d'une action dépend de l'unité de pêche qui l'entreprend, mais que cette capturabilité reste constante pour une unité de pêche donnée.

17. Dans de nombreux cas aucune des unités de pêche n'est satisfaisante et le recours à des campagnes scientifique est privilégié, posant des difficultés liées aux coûts de ces campagnes...

18. Au sens d'un ensemble de règles qui régissent les rapports entre un professionnel et son ou ses clients

Un exemple

Ceci peut être illustré par un exemple « simple » issu d'une enquête sur les résultats obtenus par des unités de pêche lors de sorties quotidiennes réalisées au cours de six jours consécutifs en avril 1978 à partir du village de Kayar, un site majeur de pêche artisanale au Sénégal situé à une cinquantaine de kilomètres au nord de Dakar.

Tableau 1 : Effectifs des échantillons et captures moyennes par sortie (Kilogrammes) selon diverses espèces en fonction du jour

| Jour | Effectif des échantillons | Captures moyennes par sortie (kg) selon les espèces | | | | | | Total |
|------|---------------------------|-----------------------------------------------------|-----------|-----------|--------|-------|--------|-------|
| | | tiof | chinchard | tassergal | pageot | sarda | autres | |
| 1 | 96 | 1.25 | 2.66 | 25.07 | 3.28 | 3.18 | 6.71 | 42.15 |
| 2 | 141 | 2.37 | 3.61 | 21.05 | 2.07 | 0.09 | 2.56 | 40.75 |
| 3 | 143 | 4.40 | 2.66 | 5.11 | 3.12 | 10.39 | 3.35 | 29.03 |
| 4 | 123 | 3.22 | 4.42 | 0.24 | 9.05 | 14.53 | 4.31 | 35.77 |
| 5 | 123 | 2.76 | 5.02 | 1.47 | 8.80 | 10.62 | 6.35 | 34.94 |
| 6 | 105 | 2.17 | 8.34 | 1.32 | 12.64 | 3.11 | 8.27 | 35.84 |

Source : Laloë F., Bergerard P., Samba A. (1981), Contribution à l'étude de la pêcherie de Kayar : étude d'une partie des résultats du sur-échantillonnage de 1978 concernant les pirogues motorisées pêchant à la ligne, *Documents Scientifiques - CRODT*, 79, p. 45 p. multigr.

Ces observations étaient faites dans le cadre de la conception d'un système d'enquêtes en vue d'estimer les captures réalisées¹⁹ par la pêche artisanale. Dans le contexte d'un système à deux niveaux (sélection de jours d'enquêtes et sélection d'actions de pêche lors de ces jours), la présence d'effets jours importants (variance inter jours élevée) militait pour que les enquêtes aient lieu le plus grand nombre de jours possible. La question portait donc sur l'existence et l'importance de ces effets.

L'examen du tableau montre que ces différences sont flagrantes et considérables. Pour l'espèce tassergal le rendement observé est divisé par 100 entre le premier et le quatrième jour... Une interprétation cohérente (confortée par d'autres éléments tels que l'évolution de la fréquentation des lieux de pêche) met en relation cette chute avec l'augmentation des rendements en pageot et chinchard : une diminution de l'accessibilité du tassergal, espèce de pleine eau pêchée en poursuivant les bancs de poissons, a conduit les pêcheurs à rechercher, en ancrant leurs pirogues, des espèces vivant sur le fond. Le choix de l'une ou de l'autre de ces méthodes est une décision et les captures peuvent donc être des réalisations de variables aléatoires de distributions différentes. Si on suppose que les pêcheurs ne disposent pas d'un tel choix, leurs résultats sont des réalisations de variables ayant toutes la même distribution, dont les espérances peuvent éventuellement dépendre du jour de pêche. Dans ce cas les moyennes par jour sont des synthèses qui contiennent beaucoup d'information : si les captures réalisées pour une espèce lors des sorties faites le jour i sont des réalisations de lois normales indépendantes²⁰ d'espérance m_i et de variance S_i^2 , les moyennes présentées dans le tableau 1 sont les meilleures estimations possibles des m_i et elles constituent une partie importante d'une statistique exhaustive minimale ; il y manque les éléments relatifs à l'estimation des variances... Il y a donc une adéquation entre les trois éléments que sont la question de l'existence d'effets

19. Ce système est stratifié, chaque strate réunissant les actions de pêche réalisées à l'aide d'un engin donné, pendant une période de temps donnée et à partir d'un site donné. Les données utilisées ici relèvent toutes d'une même strate (lignes à main à partir de Kayar au cours d'une même semaine) et on « aimerait » à ce titre pouvoir les considérer comme un échantillon aléatoire simple... La présence d'effets jours met à mal cette hypothèse.

20. L'indépendance peut être assurée par le respect des règles de l'échantillonnage aléatoire simple lors de la sélection des échantillons quotidiens d'actions de pêche... Ce qui est loin d'être évident au vu des contraintes de terrain.

jours dans le contexte de la mise en place d'un système d'enquêtes, la forme des synthèses par espèce présentées au tableau 1 et une hypothèse de distribution des observations dont rend compte l'équation :

$$Y_{ik} = m_i + e_{ik}$$

(e_{ik} est l'écart entre la $k^{\text{ième}}$ observation du jour i et l'espérance m_i). En ce sens l'expérience conduisait à conclure en faveur de l'affectation d'un enquêteur à plein temps pour estimer les captures, et à estimer les indices d'abondance à partir des résultats d'autres composantes de la pêche.

Mais la discussion sur les moyennes observées, avec l'hypothèse cohérente d'un changement rapide d'accessibilités, auquel les pêcheurs s'adaptent en changeant d'espèces cibles, soulève une question « délicate ».

S'il s'agit de bien connaître la dynamique de la ressource exploitée, entre autres pour permettre sa gestion « rationnelle », il convient, comme évoqué plus haut, de privilégier l'observation des unités de pêche dont le comportement conduit à stabiliser la capturabilité. Ceci est d'autant plus utile que l'activité de ces unités peut être directement liée à la mortalité qu'elles provoquent et qu'elle peut donc être une variable de contrôle efficace de cette mortalité.

Par contre, si l'objet est l'exploitation halieutique, la présence de différentes flottes de pêche ne peut plus a priori être discutée du seul point de vue de la dynamique de la ressource qui amène à porter des jugements de valeurs négatifs sur les unités dont les pratiques les conduisent à générer un impact variable. Comprendre cette variabilité conduit à poser la question de la variabilité de la capturabilité causée par les décisions des pêcheurs. Cette question est orpheline dans la mesure où elle conduit pour les sciences de la vie à des difficultés de description de leur objet et où, en portant sur la dynamique de la ressource biologique, elle n'est pas directement relative à l'objet des sciences humaines.

Mais si cette question est légitime, il faut l'adopter et traiter « en ce sens » les données qui ont conduit à la soulever. Elle est légitime ici parce que relative à des unités de pêche dont la présence est essentielle pour de nombreuses raisons sociales et économiques et parce que l'activité de ces unités peut être viable parce qu'elles peuvent à tout moment rechercher une espèce à leur portée en choisissant la méthode qu'elles estiment être la plus utile parmi celles dont elles disposent. Chacune de ces méthodes j peut être caractérisée par une distribution d'espérance m_{ij} le jour i . En étant choisie avec une probabilité p_{ij} le modèle de distribution peut s'écrire :

$$Y_{ik} = \sum_{j=1}^J p_{ij} m_{ij} + e_{ik}$$

Les p_{ij} résultent des décisions des pêcheurs, engendrant ainsi la variabilité de leur impact.

Selon ce modèle de distribution, les moyennes simples du tableau 1 sont de qualité nettement moindre en termes de contenu d'information et la construction d'un modèle²¹ est proposée, permettant d'exprimer les probabilités p_{ij} sous forme de fonctions de ses paramètres. Il articule la dynamique d'une ressource plurispécifique, à l'aide de modèles de production, avec la dynamique de l'exploitation menée par des unités de pêche réunies selon des flottes de pêches caractérisées par les ensembles de méthodes disponibles. La probabilité de choisir une méthode est estimée à l'aide d'un « modèle d'utilité aléatoire »²². Les paramètres du modèle

21. Le terme de modèle peut être appliqué à plusieurs choses différentes. Peut-être le terme de cadre serait-il préférable. Cette construction est décrite dans le livre que ce texte présente.

sont estimés en recherchant les valeurs qui conduisent à reconstituer des données d'activités et de rendements de pêche les plus proches possibles (selon un critère de moindres carrés) de celles résultant des observations collectées dans le cadre du système d'enquêtes.

En guise de conclusion

A l'issue de cette opération, il est possible de répondre, sous la forme de fonctions des estimations des paramètres, à des questions faisant intervenir les décisions des pêcheurs. Quelques exemples sont décrits dans le livre présenté ici.

En termes statistiques, on peut considérer que les estimations de l'ensemble des paramètres du « modèle » constituent une statistique contenant le plus possible de l'information présente dans les données « traitées ». Si certains des paramètres peuvent avoir une interprétation directe (des prix, des coûts, des capturabilités, des effectifs de flottes de pêche...), l'intérêt de cette statistique est de pouvoir être utilisée pour répondre, sous forme de fonctions de ses éléments, à des questions qui peuvent ne pas être toutes identifiées au départ. En ce sens, cette statistique peut être vue comme une première restitution de l'information collectée dans le cadre d'un observatoire. L'idéal serait de produire une statistique exhaustive minimale, en étant alors certain que toute meilleure réponse à une quelconque question serait une fonction de ses éléments... Cet idéal constitue une référence et un outil critique très utiles même si, dès lors qu'il n'y a pas de représentation unique et définitive, une telle statistique n'existe pas.

Une analogie avec le système judiciaire peut être utile : Le juge d'instruction instruit à charge et à décharge. Il est censé réunir, organiser et mettre à disposition des différentes parties toute l'information utile relative à l'affaire qui doit être jugée. Les avocats vont disposer de ces informations pour les restituer, chacun selon l'intérêt de la partie qu'il représente (il s'agit bien d'une représentation !). Le statisticien doit au moins assumer le rôle du juge d'instruction... mais il peut aussi restituer l'information selon un point de vue particulier. Dans le cadre d'un projet monodisciplinaire, ces rôles doivent pouvoir être confondus grâce à la mise en place d'un protocole d'observation adéquat. A l'inverse, dans celui d'un projet interdisciplinaire, ces deux rôles ne peuvent être confondus ; et si le statisticien est amené à les tenir, il est déontologiquement nécessaire qu'il précise en toute circonstance lequel de ces deux rôles il tient, et au service de qui.

22. Random Utility Model (RUM) dont une description a été proposée par Mac Fadden. On considère que chaque décideur réalise une estimation de l'utilité de chacune des options à sa disposition et qu'il choisit celle dont l'estimation est la plus élevée. Voir : McFadden D. (1973) Conditional logit analysis of qualitative choice behavior. In Zarembka P., éditeur *Frontiers in econometrics*, pages 105 - 142. Academic Press, New-York.