JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

MICHEL ARMATTE

Discussion de l'article de D. Denis « Sur les tests d'hypothèse : la véritable nature d'une méthodologie hybride »

Journal de la société française de statistique, tome 145, n° 4 (2004), p. 27-36

<http://www.numdam.org/item?id=JSFS_2004__145_4_27_0>

DISCUSSION DE L'ARTICLE DE D. DENIS

Sur les tests d'hypothèse : la véritable nature d'une méthodologie hybride

Michel ARMATTE *

RÉSUMÉ

Confirmant la nature hybride des pratiques de test statistique et leur écart à toute théorie qu'elle soit celle de Fisher, celle de Neyman et Pearson ou celle des bayesiens, l'auteur suggère qu'il n'y a pas à s'en offusquer, d'une part parce que les théories successives ont toutes été des hybrides et ont toutes fait preuve de dérives dans l'interprétation d'un test, d'autre part parce que les pratiques scientifiques dans lesquelles ces méthodologies ont pris place le sont encore plus, dans la mesure où elles ne sont pas indépendantes des spécificités disciplinaires (selon par exemple que celles-ci mobilisent des données d'observation ou d'expérimentation, qu'un modèle préalable des relations à tester existe ou non), pas indépendantes non plus des usages sociaux de ces procédures (des domaines d'application et des conséquences de la décision).

ABSTRACT

In this comments, we agree with the fact that "*Today's genre of null hypothesis significance testing (NHST) bears little resemblance to the model originally proposed by Fisher over seventy-five years ago*", but we argue first that its bears little resemblance to the other models originally proposed by Neyman and Pearson or by neo-bayesians, secondly that all successive theories are always hybrids of the preceding ones and of some practices, and so was for Fisher himself. Finally we discuss the author's thesis that the to-day hybrid theory of NHST is the same in different academic disciplines and in different contexts of utilisation, especially with reference to econometrics.

L'article de Daniel J. Denis s'attache à dénoncer le caractère hybride des procédures contemporaines de test d'hypothèse. Une fois de plus, puisque la critique des usages des tests dans les sciences expérimentales, à défaut d'être une tradition, est récurrente depuis la fin des années 1950. En France, c'est une idée qui a été discutée par Henri Rouanet dans plusieurs de ses écrits et dont Benjamin Matalon par exemple avait retracé la genèse dans ses séminaires des années 1980 à l'EHESS. Sans que ces controverses aient

* Université Paris Dauphine et Centre A. Koyré
michel.armatte@dauphine.fr

ralenti d'une quelconque manière la formidable diffusion de ces usages, parfois inconsidérés. Les écrits de Gigerenzer (1989, 1993) ont certainement eu un rôle déterminant dans le débat en allant au-delà de la critique de ces usages et en mettant en avant l'idée que les formes actuelles de la théorie statistique des tests qui sont enseignées et mobilisées par les chercheurs sont des hybrides entre les théories de Fisher, de Neyman & Pearson et des néo-bayesiens. La veine critique s'est alors enrichie de nombreux articles [1] qui ont « enfoncé le clou » chacun à leur manière. Et pourtant le texte de Denis est salutaire car il n'y a pas de changement notable dans la manière dont la théorie des tests est d'une part enseignée, et d'autre part mobilisée dans la recherche en sciences d'observation.

Denis cherche à rétablir une démarcation entre les deux traditions principales fisherienne ou neymanienne qui ont été aujourd'hui amalgamées. Il a le mérite de mettre parfaitement en évidence les dérives principales des usages contemporains en psychologie expérimentale par rapport au dogme fisherien : le rôle de la randomisation dans la construction de l'échantillon, la référence à une population infinie hypothétique, la considération de la seule hypothèse nulle, l'interprétation des résultats du test en terme de corroboration/invalidation de cette hypothèse, et d'une probabilité interprétable en niveau de confiance plutôt qu'en limite de fréquence, le choix d'un seuil de signification exact, ce sont autant de caractéristiques de l'approche de Fisher qui ont été dénaturées dans l'approche moderne. On pourrait chipoter sur quelques unes des affirmations de l'auteur en la matière : par exemple celle d'une mesure de *sensitivity* du test chez Fisher qui serait une sorte d'anticipation de la notion de puissance ne me semble pas complètement assurée ; il reste que le constat me convient tout à fait dans l'ensemble : « *Today's genre of null hypothesis signifiance testing (NHST) bears little resemblance to the model originally proposed by Fisher over seventy-five years ago* » (résumé).

Si la démonstration semble à première vue assez convaincante, et corroborée, comme on l'a dit, par de nombreux observateurs, j'ai cependant envie de lui adresser deux critiques : a) on fait comme si la théorie de Fisher avait une consistance et une complétude qu'elle n'a certainement pas, b) on fait comme si les pratiques de test aujourd'hui étaient toutes fondées sur une et une seule théorie hybride parfaitement unifiée dans le temps et dans l'espace géographique et disciplinaire, ce qui n'est pas le cas non plus. Ce faisant, l'opposition des caractéristiques affichée pour chacune de ces « théories » n'est évidente que parce que les deux approches ont été typées et idéalisées comme deux théories. Or *théorie* fisherienne et *pratiques* contemporaines n'ont point le même statut épistémologique. Les comparer suppose que l'on associe à la théorie de Fisher des pratiques de recherche qui s'insèrent dans un certain milieu à une certaine époque – ce qui réclame un travail d'historien qui n'a pas été fait – et que l'on associe à des pratiques contemporaines bien identifiées (dans quelles disciplines ? dans quels lieux ?) un objet abstrait reconstruit bien identifié comme « théorie hybride ». Mais où cette dernière s'exprimerait-elle ? Dans des thèses, des articles, des cours ? Il faudrait être plus précis, et surtout

1. NDLR : voir la bibliographie de l'article de Jeff Gill dans ce dossier.

comparer ce qui est comparable, à savoir des paradigmes, c'est-à-dire au sens de Kuhn des configurations relativement stables et mixtes, à la fois cognitives et sociales, de notions, concepts et relations d'une part, groupes sociaux et usages historiquement situés d'autre part. Dans toutes les sciences empiriques, la méthodologie de traitement de l'information numérique et de construction de la preuve ne correspond jamais à aucun canon établi, *a fortiori* 150 ans plus tôt. Certes des écoles épistémologiques ont une existence et une influence, mais les *sciences studies* ont montré que la science « telle qu'elle se fait » obéit à un grand nombre d'autres injonctions qu'épistémologiques : demande sociale, organisation sociale de la recherche, régulations étatiques ou par le marché, agenda politique, environnement international (guerre et paix), échanges entre disciplines, etc.

Faute d'un raisonnement en terme de paradigme et d'une enquête plus approfondie, on peut craindre que l'opération de démarcation entre « today's NHST » et théorie fisherienne ne se ramène à une porte ouverte enfoncée. Oui les pratiques d'aujourd'hui diffèrent de la théorie d'avant-hier. D'abord parce que d'autres théories sont venues les concurrencer, ou les féconder, voire les hybrider. Ensuite parce que les pratiques sont influencées par bien d'autres choses que les théories, et pour revenir aux tests, par les conditions techniques et sociales de la recherche contemporaine (comment produit-on des « données », comment les traite-t-on avec un ordinateur et un logiciel, comment et pourquoi publie-t-on, quels liens y a-t-il entre connaissance publiée et action ou nouvelle recherche ?...). Oui les pratiques d'aujourd'hui ont des fondements hybrides. Mais quelles en sont les conséquences ? « Laisser dormir Fisher où il est » et ne plus se référer à lui dans la pratique des tests est la principale conclusion de l'auteur. Mais il n'a pas traité de façon symétrique les apports des courants de la décision et du bayesianisme. S'il l'avait fait il aurait certainement conclu aussi que leur théorie était dénaturée dans l'approche hybride, et qu'il fallait aussi laisser dormir Jerzy Neyman, Egon Pearson, ainsi que Bayes, Laplace, Jeffreys, etc... Cela fait beaucoup de morts à ne plus invoquer ! Et une discipline qu'on ne peut plus référer à aucune construction historique. Il me semble qu'il faut prolonger autrement le travail de démarcation engagé par Denis, en revenant sur la nature même des deux paradigmes du test fisherien et de l'hybride des psychologues aujourd'hui.

Le test fisherien comme paradigme ne peut pas totalement être identifié à la seule œuvre écrite de Fisher parce qu'il est déjà lui-même un hybride. Pour une bonne partie de ses caractéristiques on pourrait par exemple remonter à Arbuthnot (1710) qui montre que le ratio de masculinité à la naissance ne s'écarte pas par hasard de 1/2 (il est en fait de l'ordre de 106/206), et que cela prouve l'existence d'une providence divine qui veille sur nous autres hommes, plus faibles que le sexe dit faible. On pourrait aussi le nommer test laplacien parce que Laplace est un des tout premiers non pas à introduire l'idée de test de signification, mais à la traduire dans les termes d'un calcul de probabilités. Comme Sheynin (1971) l'a suggéré, Laplace anticipe la méthode de l'hypothèse nulle. Or Laplace lui-même peut être pris en défaut de dérapage incontrôlé sur l'interprétation du résultat d'un test : se demandant par exemple si la pression atmosphérique est modifiée par le cycle jour nuit,

il écrit « *Pour déterminer avec quelle probabilité cette cause est indiquée, concevons que cette cause n'existe point et que la différence résulte des causes perturbatrices accidentelles et des erreurs d'observation* », et ayant montré que la somme des écarts est inférieure à 400 sous cette hypothèse (disons H_0) avec une très grande probabilité, il conclut : « *Nous pouvons donc regarder comme pratiquement certain que la somme des excès est inférieure à 400 s'il n'y a pas de cause constante : c'est-à-dire qu'il y a une forte probabilité en faveur de l'existence d'une cause constante* » (Laplace, 1886, L2, Ch.V, paragr. 25, p.356). Si X est l'échantillon, n'est-il pas un des premiers à inférer abusivement de la probabilité conditionnelle $p(X/H_0)$ à son inverse $p(H_0/X)$? N'est-il pas influencé par le raisonnement bayesien (qu'il a inauguré indépendamment de Bayes) selon lequel les probabilités des causes sont proportionnelles aux probabilités des résultats conditionnellement à ces causes (sous l'hypothèse – scabreuse aux yeux de Fisher – d'une équiprobabilité *a priori* de ces causes) ? N'est-il pas en train déjà de construire une théorie hybride ? Joseph Bertrand, un siècle plus tard, dans son grand traité anti-laplacien n'a pas manqué de railler les résultats de Laplace en ce qui concerne par exemple la masse de Jupiter, mais il reproduit la même interprétation bayesienne des tests de signification : « *Le rapport de probabilité des deux causes également probables a priori est celui des probabilités observées* » écrit-il à propos du test d'une pièce de monnaie lancée 500391 fois (Bertrand, 1889, page 277). Ronald Fisher, à peine 30 ans plus tard, construit l'ensemble de sa théorie inférentielle sur deux postulats : a) on ne peut pas identifier méconnaissance et équiprobabilité et toute application du théorème de Bayes qui utilise ce lemme est viciée, b) c'est la mesure de vraisemblance d'un échantillon qui fournit la bonne règle de décision. Si ce double principe lui a fourni en 1922 (en fait dès 1912) les bases d'une théorie de l'estimation qui a résisté au temps, on sait qu'il a été plus difficile pour lui d'y accrocher une théorie des tests : la notion de test dans l'œuvre même de Fisher évolue au gré de la construction des autres outils inférentiels. Denis note bien l'évolution chez Fisher du niveau de significativité choisi d'abord arbitrairement (5%) *a priori*, puis transformé en niveau exact de signification pour un échantillon donné (*p-value*). On peut aller plus loin et rapprocher cette évolution d'un changement d'intérêt dans les usages : impliqué par les recherches agronomiques dans la station de Rothamsted, où le thé qu'il prenait tous les jours « in the sample house » a fourni le premier exemple de test dans son ouvrage de 1925, Fisher s'est d'abord intéressé aux situations expérimentales et aux plans d'expérience. « *Each case came to him as a unique problem* » dit Joan Fisher-Box dans la biographie de son père. Mais ensuite, Fisher quitte Rothamsted dès 1933 pour University College où il reprend une partie de l'héritage de Karl Pearson. Il se trouve alors en face de données génétiques qui ne sont plus expérimentales, et se met en tête de construire une théorie générale de l'inférence dont un élément clé est la probabilité fiduciaire. L'idée même de probabiliser l'espace des paramètres est une concession faite aux approches bayesiennes bien qu'il insiste sur ce qui la distingue de la méthode de la probabilité inverse. Engagé dans une fin de carrière académique à Cambridge puis à Adélaide il s'éloigne de ce terreau expérimental d'où est sorti *Statistical Methods for Research*

Workers et se trouve embarqué à la fois dans une lutte violente avec les tenants d'une théorie de la décision que sont Neyman et Pearson – voir la citation de Fisher par Denis « *The situation is entirely different in the field of Acceptance Procedures,...* » qui se trouve déjà dans l'édition 1935 de The *Design of Experiment* – et dans des controverses autour de données d'enquêtes (par exemple dans ses tentatives de réconciliation des Mendeliens et des Darwiniens et dans la controverse sur tabac et cancer).

De ces quelques idées qu'il faudrait approfondir et des remarques nombreuses faites sur les propres contradictions de Fisher, on peut déduire que le paradigme du test fisherien est une nébuleuse à la fois plus large que la seule théorie présentée dans *Statistical Methods* et évolutive, y compris au cours de la carrière même de Fisher. Changeante elle l'est plus encore si l'on considère son hybridation rapide avec la théorie de Neyman et Pearson, du vivant même de Fisher, dès lors qu'elle s'incarne dans des pratiques situées. Et la cartographie de ces pratiques ne peut se réduire à la seule opposition entre connaissance savante et décision. En effet, cette distinction est fortement brouillée par le développement exponentiel, pendant la guerre et dans les deux décennies qui suivent, de procédures de gestion des grands systèmes socio-techniques qui articulent intensément connaissance et décision. Que l'on pense aux usages qui se développent à large échelle en agronomie, dans l'industrie pour le contrôle de fabrication, dans la prévision économétrique, ou encore dans la gestion des enquêtes épidémiologiques ou dans la recherche biomédicale et pharmacologique. Dans ces différents champs, la méthodologie des tests n'a pas été simplement un ingrédient de la recherche en laboratoire. Comme l'écrit Jean-Paul Gaudillère (2004) pour ce dernier secteur, « le recours aux statistiques a délimité un espace de négociation permettant aux cliniciens universitaires, aux industriels et aux autorités de santé publique de se mettre d'accord sur les propriétés du médicament et donc indirectement sur la régulation des usages médicaux ».

C'est ici qu'apparaît, me semble-t-il, une seconde difficulté dans le texte de Denis. L'hybride dont il est question n'est pas précisément décrit si ce n'est à travers un exemple tout à fait utile pour illustrer un type de pratique, mais pas suffisant pour en mesurer le degré de validité, de stabilité et de permanence pour toutes les activités scientifiques. L'affirmation que cet hybride est le même dans toutes les disciplines et en tout pays est présentée comme une évidence qu'il faudrait davantage étayer. Si la méthodologie hybride n'est pas une théorie que l'on peut référer à une science pure et unifiée de l'inférence et plus précisément à un auteur ou un groupe d'auteurs, mais plutôt une pratique sociale, il faut la décrire finement dans ses diverses configurations, en s'appuyant sur des études systématiques de corpus de manuels, comme le fait Huberty (1993) ou sur des enquêtes empiriques auprès des chercheurs comme le fait Poitevineau (1998) dans chaque discipline, ou encore par des enquêtes historiques dans chaque sphère d'usage comme l'ont fait par exemple Harry Marks (1997) pour la thérapeutique, ou Denis Bayart pour le contrôle de fabrication. Il faut alors faire quelques hypothèses sur les mécanismes au travers desquels elle se construit et se diffuse. Pour l'essentiel ces mécanismes ne relèvent pas seulement de la logique inductive mais aussi de logiques sociales

qui sont au carrefour de la recherche et de la gestion publique. Donnons quelques pistes d'une étude qui reste à faire.

La théorie des tests s'est d'abord étendue de l'agronomie expérimentale à différentes branches de la psychologie expérimentale. C'est là que sa diffusion a été la plus rapide et peut-être aussi la plus désastreuse si on en croit le procès fait par Gigerenzer (1993) aux auteurs qui en ont été les principaux vecteurs et interprètes comme Guilford (1942). C'est à ces années 1950 et à cette province de la Science qu'il faut attribuer les plus audacieux croisements qui permettent d'ajouter la notion de seuil critique à l'approche fisherienne et celle de *p-value* à l'approche de Neyman-Pearson. C'est également dans ce groupe de disciplines que le test d'hypothèse est devenu un point de passage obligé de la publication statistique et que s'est développé un usage mécaniste du niveau de significativité statistique, à la fois comme substitutif de l'intensité des effets que l'on veut mettre en évidence et comme principe de sélection des résultats d'expériences dignes de publications. Avec le fameux effet pervers bien connu du biais de publication qui en résulte : ne sont publiés que les résultats significatifs y compris les 5% correspondant à l'erreur de type I tandis que les résultas non significatifs ne le sont jamais (ce que Denis réfère sous le nom de *file drawer problem*). La course aux publications produit un flot d'énoncés empiriques validant des millions de micro-faits « prouvés » par la statistique dont on serait bien en mal de déduire une connaissance cumulative, un flot que les revues ne peuvent endiguer. Dans les années 1960, le *Journal of Experimental Psychology* exige que les tests soient au moins significatifs à un niveau de 1%, ce qui ne fait que renforcer l'assimilation abusive entre significativité et relation causale avérée, et qu'accentuer ce biais de publication. Le problème est si grave que récemment (1995) *l'American Psychological Association* a de nouveau mis le problème à l'étude via une de ses commissions (voir Capel *et al.*, 1997). Gigerenzer a interprété ce succès de la formule hybride en psychologie dans le théâtre même de la psychologie analytique par les figures du ça, du surmoi et du moi, comme la résultante d'un combat propre à chaque chercheur, entre le désir (bayesien) d'énoncé probabiliste sur les hypothèses, le surmoi (neymanien) qui exige de prendre toujours une décision, et le moi (fisherien) qui est l'auteur d'expériences et de publications. Il en résulte une figure du chercheur douloureusement tiraillé entre des injonctions différentes que l'on ne réduira pas de sitôt à l'harmonie, et qui s'en sort en tuant ses parents... qui ne sont autres que les pères fondateurs de la Statistique. Le meurtre est visible dans la pléthore de manuels qui reconstruisent avec de nombreuses contorsions un paradigme irénique et réunifié. La parabole est jolie et instructive, mais elle est aussi prisonnière de la discipline qui l'inspire. L'histoire sociale nous indique que ce théâtre d'ombres se double d'un autre théâtre plus visible à l'échelle de la société, qui est celui de la seconde guerre mondiale, puis celui de la guerre froide ; ces théâtres *d'opérations* fournissent le contexte dans lequel les sciences sociales sont profondément bouleversées par l'introduction des mathématiques appliquées à grande échelle, dans le cadre du paradigme dominant de la décision, et avec les moyens puissants des ordinateurs et des nouvelles méthodes de la simulation.

L'intuition de Denis, que les choses se passent à peu près de la même façon dans toutes les sciences me semble en partie exacte parce que la reformulation de la théorie des tests a une même origine historique. Le Statistical Research Group fondé au début des années 1940 dans le cadre de l'OSRD américain et de l'Applied Mathematical Panel a été le cadre de l'émergence d'une théorie de la Décision et d'une théorie des Jeux qui ont envahi les principaux *think tanks* des années 1950 comme la Cowles Commission, et la Rand Corporation, et qui ont fécondé toutes les disciplines, de la psychologie (très présente à Santa Monica) à l'économie et la recherche opérationnelle (l'école de Chicago, la Cowles Commission) en passant par la géographie (Armatte, 2004). Un signe important parmi d'autres : le privilège accordé aujourd'hui à l'utilisation de la *p-value* (associée à l'échantillon et l'approche fisherienne) plutôt qu'à la détermination d'une valeur critique (associée à une hypothèse et à l'erreur de premier type) doit beaucoup plus à l'ordinateur qu'à la bataille entre Fisher et Neyman ; si les tables ne permettaient qu'un calcul pour quelques valeurs de alpha, l'ordinateur fournit tout de suite une valeur de la *p-value* pour n'importe quelle valeur de la statistique calculée pour le test.

À y regarder de plus près, l'histoire propre du développement des différentes disciplines après guerre joue cependant un rôle déterminant dans la construction de la méthodologie hybride et ne permet pas de dire que la situation est la même dans les différentes sciences. On sait par exemple que la pratique des tests, timidement introduite en sociologie empirique dans la mouvance des travaux de Lazarsfeld, a été fortement rejetée par la plupart des sociologues sous le motif que la randomisation y est exclue et qu'une enquête n'est pas un plan d'expérience. Des articles comme ceux de Lipset et de Selvin dans les années 1950 ont semble-t-il joué un rôle de contre-feux et empêché que la littérature se remplisse de résultats de tests. La seule exception est tout de même l'usage souvent immodéré des tableaux croisés qui a conduit les sociologues de ce courant à trier trop mécaniquement les bons tableaux des mauvais par un test du χ^2.

En économétrie au contraire, le rôle des tests a été déterminant dans le dépassement des nombreuses apories qui ont émergé d'une utilisation abusive de la notion de corrélation dans la statistique économique de l'entre deux guerres. Aux baromètres économiques qui fondaient la prévision conjoncturelle sur la covariation de quelques indicateurs ont succédé, après la crise de 1930, des modèles macroéconomiques traduisant les relations structurelles d'une économie par des équations avec un terme d'erreur aléatoire. Tester une théorie est devenu après le texte-manifeste d'Haavelmo (1944) synonyme de tester une hypothèse probabiliste qui prend forme d'un modèle, et ce au vu d'observations statistiques chronologiques non répétables et non expérimentales. On est dans une toute autre configuration que Fisher à Rothamsted et que les psychologues qui montent un plan d'expérience. La randomisation n'a aucun sens sur des données historiques et agrégées. La référence à une décision existe, beaucoup plus qu'en psychologie expérimentale, mais elle n'est pas similaire à la décision d'embauche qui résulte d'un test psychotechnique, ou encore à la décision d'accepter ou refuser un lot en contrôle de fabrication. La décision économique et politique n'est pas de gestion mais

de choix de société. Elle n'est pas directement associée au test statistique comme l'achat d'un lot mais elle est médiatisée par un instrument d'aide à la décision plus complexe qui est un modèle macroéconomique. La médiation du modèle est plus ou moins directe, longue en macroéconomie où elle passe par des organismes de prévision, planification et décision politique, plus courte dans des modèles de recherche opérationnelle qui visent à optimiser tel ou tel dispositif, quasiment immédiate en finance de marché où le modèle agit directement sur le marché. La référence à un modèle théorique est ce qui différencie le plus la situation économétrique de la psychologie expérimentale. Dans les années 1950, le modèle place d'entrée de jeu le chercheur dans une situation de corroboration ou de falsification d'hypothèses théoriques, ce que renforce encore une adhésion à l'épistémologie poperienne. Mais le modèle dans sa totalité est rarement invalidé par la batterie de tests. Un test de Student ou de Fisher-Snedecor ne permet pas autre chose que d'invalider une hypothèse de nullité d'un ou plusieurs coefficients du modèle. Dans bien des cas, une hypothèse alternative peut facilement être formulée dans ce cadre – par exemple, dans une régression, le test de non autocorrélation des erreurs a comme hypothèse alternative une autocorrélation d'une forme connue, un processus AR(1) par exemple – mais la nature composite de l'hypothèse alternative empêche souvent le calcul simple d'un risque de seconde espèce. Les *p-values* ont depuis peu remplacé les valeurs critiques dans les manuels après l'avoir fait dans les logiciels. Il faut dire aussi que la valeur arbitraire de 5% qui convient bien à un tel ouvrage, c'est-à-dire un endroit ou la théorie est assez désincarnée et scolaire, n'a plus aucun sens dès que les risques sont concrets (en épidémiologie, en essais thérapeutiques, en calcul de structures mécaniques, en décision économique et sociale).

Chez les économètres aussi le biais de publication existe mais il est moins flagrant parce que ce qui compte n'est pas un effet mais un ensemble d'effets traduits par un modèle. Il peut cependant prendre une forme très perverse de ce fait même. Des économètres (Leamer, 1983; Lovel, 1983; Charemza and Deadman, 1997) proches de l'école de Hendry, à la London School of Economics, et dont il faut rappeler la devise principale : « *econometrics is testing, testing and testing* », ont dénoncé les pratiques de *data mining* consistant à renouveler le modèle par ajout de variables puisées dans une base de donnée, ou par modification de la forme analytique de la relation, jusqu'à trouver la spécification du modèle qui satisfasse à tous les tests. Cette stratégie de recherche tout autant que de publication, conduit à renforcer très sérieusement la valeur effective du risque de première espèce publié. Supposons par exemple qu'un chercheur cherche à expliquer une certaine variable endogène y par 2 des 10 variables exogènes d'une base de données, supposées orthogonales pour simplifier, et dont aucune n'est corrélée avec y. S'il lance 10 régressions de y sur chaque variable exogène, il a une probabilité de rejeter toutes les variables de $0.95^{10} = 0.5987$ pour un niveau de confiance déclaré de 5%. Mais en sélectionnant le modèle aux 2 variables qui ont le plus grand Student, le *vrai* niveau de confiance de la procédure (la probabilité de rejeter l'hypothèse nulle qu'aucun coefficient n'est significatif) n'est pas 5% mais une certaine valeur α^* telle que $(1 - \alpha^*)^2 = 0,5987$, soit

$\alpha^* = 0,226 = 22,6\%$. La probabilité de considérer comme significative une variable qui ne l'est pas devient ainsi très importante et la confiance que l'on peut attribuer aux modèles sélectionnés par cette pratique du *data mining* ne dépend plus seulement du modèle et des données mais aussi de la stratégie de recherche elle-même, le plus souvent non visible dans la publication. L'École de David Hendry a donc proposé comme solution toute une méthodologie de sélection descendante, du modèle le plus complet au modèle le plus spécifique, pour contrer ce biais.

La méthodologie des tests d'hypothèse en économie est tout autant hybride qu'en psychologie mais l'hybride construit ne me semble pas être le même. On trouverait d'ailleurs un résultat analogue si on s'intéressait à d'autres objets statistiques. Leur unicité formelle, leur neutralité apparente, cachent des différences profondes de signification, d'usage et d'efficacité quand on les transporte d'une discipline à une autre, d'un champ à un autre, comme nous l'avons montré pour le coefficient de corrélation (Armatte, 2001). La tâche des statisticiens et des historiens est difficile. Comme l'écrivent Capel, Monod et Müller (1997), « à la fois signe de légitimité et cause de rejet, l'argumentation statistique ne peut pas, en l'état, satisfaire simultanément les exigences de la recherche en sciences humaines et les attentes du praticien »... ajoutons « et celles du mathématicien qui recherche une forme canonique de tests de signification ». Dès que l'on sort de la théorie reconstruite des manuels il n'y a plus que des méthodologies évidemment hybrides, adaptées à l'environnement de chaque province de la science. L'hybridisation n'est pas une tare épistémologique, c'est une nécessité sociale des pratiques de la recherche appliquée.

Références

ARBUTHNOT J. (1710). An argument for Divine Providence, taken from the constant regularity observed in the births of both sexes, *Phil. Trans. Of the Royal Society*, 27, p.186–190.

ARMATTE M. (1988). La construction des notions d'estimation et de vraisemblance chez Ronald A.Fisher, *Journal de la Société de Statistique de Paris*, tome 129, N°1–2, 1988.

ARMATTE M. (2001). Le statut changeant de la corrélation en économétrie (1910-1944), *Revue Économique*, 52-3, p. 617–631.

ARMATTE M. (2004). Les sciences économiques reconfigurées par la pax americana, in Pestre & Dahan (eds), *Les sciences dans et pour la guerre*, 1940-1960, Paris, Presses de l'EHESS, p. 129–174.

BERTRAND J. (1971, 1889). *Calcul des probabilités*, 3ème édition, New York, Chelsea, XLIX + 317 p.; 1ère éd. 1889; 2ère éd. 1907.

CAPEL R., MONOD D., MÜLLER J.-P. (1997). De l'usage perverti des tests inférentiels en sciences humaines, *Genèse*, 26, p.123–142.

CHAREMZA W., DEADMAN D.F. (1997). *New Directions in Econometric Practice*, Edward Elgar.

FISHER R. A. (1935). « The logic of inductive inference », *Journal of the Royal Statistical Society*, 98, p. 39–82, avec discussion, [*Collected Papers* 124].

FISHER R. A. (1935). *The Design of Experiment*, Edinburgh, Oliver and Boyd.

FISHER-BOX J. (1978). *R.A. Fisher, the life of a scientist*, N.Y., John Wiley & sons.

GAUDILLERE J.-P. (2003). *Inventer la biomédecine : la France, l'Amérique et la production des savoirs du vivant* (1945-1965), Paris, La Découverte.

GAUDILLERE J.-P. (2004). Mobiliser les sciences pour vaincre le cancer, in Pestre & Dahan (eds), *Les sciences dans et pour la guerre*, 1940-1960, Paris, Presses de l'EHESS, p. 343-368.

GIGERENZER G., MURRAY D. J. (1987). *Cognition as Intuitive Statistics*, Hillsdale/ London, Lawrence Erlbaum associates.

GIGERENZER G. (1993). The Superego, the Ego, and the Id in Statistical Reasoning, in Keren G. et Lewis C., *A handbook for data analysis in the behavorial sciences : Methodological Issues*, Londres, Lawrence Erlbaum Ass. Ed.

GUILFORD J.-P. (1942). *Fundamental Statistics in Psychology and Education*, New York, McGraw-Hill.

HAAVELMO T. (1944). *The probability Approach in Econometrics*, Supplément à *Econometrica*, 12, p. 115.

HUBERTY C. J. (1993). Historical origins of testing practices : the treatment of Fisher versus Neyman-Pearson views in textbooks, *Journal of Experimental Education*, 61 (4), p. 317–333.

LAPLACE P. S. (1886). *Théorie analytique des probabilités*, Œuvres Compl., VII; 1ère éd. 1812, 2ème éd. 1814 avec préface; 3ème éd.1820 avec trois suppléments.

LEAMER (1983). Let's take the con out of econometrics, *American Economic Review*, 23, p. 31–43.

LOVEL (1983). Data mining, *Review of Economics and Statistics*, 65, p. 1-12.

MARKS H. (1997). *The Progress of Experiment, Science and Therapeutic Reform in the United States*, 1900-1990, Cambridge, Cambridge University Press.

POITEVINEAU J. (1998). *Méthodologie d'analyse des données expérimentales... Étude de la pratique des tests statistiques chez les chercheurs en psychologie, approches normative, prescriptive et descriptive*, Thèse, Université de Rouen.

ROUANET H. (1991). Les tests statistiques revisités. *In* H. Rouanet, M.-P. Lecoutre, M.-C. Bert, B. Lecoutre, J.-M. Bernard & B. Leroux (Eds.), *L'inférence statistique dans la démarche du chercheur*. Berne : Peter Lang.

SHEYNIN O. B. (1971). Newton and the classical theory of probability, *Archiv Hist. Exact Sciences*, vol.7, N°3, p.217–243.