Roland Capel
Denis Monod
Jean-Pierre Müller

De l'usage perverti des tests inférentiels en sciences humaines

In: Genèses, 26, 1997. pp. 123-142.

Résumé

■ Roland Capel, Denis Monod, Jean- Pierre Millier: De l'usage perverti des tests inférentiels en sciences humaines Les tests d'hypothèses sont très largement utilisés dans la recherche en sciences humaines, mais leur usage est cependant souvent jugé abusif, sinon . perverti, par de nombreux méthodologues dont les critiques restent pourtant sans effet depuis des décennies. Les . auteurs proposent diverses explications . de l'étonnante persistance de telles pratiques. L'étude du contexte historique et théorique dans lequel celles-ci ont pu se développer conduit à mettre en évidence les graves confusions liées à la notion de significativité. très prisée en sciences humaines.

Abstract

On the Abusive Use of Inferential Tests in the Human Sciences Testing hypotheses is very widely used in research in the human sciences, but its use has often been considered abusive, if not perverted, by a number of methodo- logists whose criticisms have nevertheless remained ineffectual for decades. The authors propose various explanations for the astonishing longevity of this practice. A study of the historical and theoretical context in which such tests developed brings to the fore the serious mix-ups pertaining to the highly-prized notion of meaninefulness in the human sciences.

De l'usage perverti des tests inférentiels en sciences humaines

Roland Capel, Denis Monod et Jean-Pierre Müller

Introduction

« L'usage abusif de la statistique en sciences humaines est bien connu [...] ». Cette remarque[1], formulée sans autre commentaire, soulève au moins deux questions. Faut-il y lire l'expression de ce sentiment souvent rencontré de « défiance envers les chiffres » propre, en général, aux personnes peu au courant en matière de statistiques et qui, pour cela, leur reprochent de « dire n'importe quoi » et de « tromper le monde » ? Ou devons-nous y voir plutôt l'expression d'une critique plus élaborée, invitant à rompre avec « [...] le positivisme ambiant et avec le fanatisme du nombre qui était, dans le sens fort du terme, le signe de la scientificité[2] » ?

La critique de Dacunha-Castelle, statisticien émérite, ne relève certainement pas de la simple défiance envers les chiffres puisque, loin de voir le démon caché derrière ceux-ci, son auteur propose de changer l'enseignement des statistiques de manière à les rendre un peu plus familières, de les appliquer largement, mais raisonnablement, dans des domaines bien précis, et parfois de les abandonner lorsque leur apport semble négligeable. Mais alors, à quels abus fait-il donc allusion ? Il nous a fallu franchir l'Atlantique pour trouver les premiers éléments consistants de réponse.

En novembre 1995, lors de sa dernière réunion bisannuelle, une sous-commission de l'APA[3], le Board of Scientific Affairs (BSA) a décidé de réexaminer la manière dont il est fait usage des techniques statistiques dans le domaine de la recherche en psychologie quantitative, celle-là même qui se pare souvent du titre de « scientifique ». Précisément, c'est la pratique des tests d'hypothèse qui y a été sérieusement mise en cause et, à travers elle, c'est l'approche inférentielle tout entière qui semble être visée. La déclaration du BSA[4] révèle qu'il existe un réel problème d'« over-reliance » – autrement dit d'excès de confiance

▶ ▶ ▶

1. Didier Dacunha-Castelle, *Chemins de l'aléatoire*, Paris, Flammarion, 1996, p. 84.

2. Michel Maffesoli, « Mythe, quotidien et épistémologie », in *Le mythe et le mythique*, Actes du colloque de Cerisy, Paris, Albin Michel, « Les cahiers de l'hermétisme », 1987, p. 92.

3. APA : *American Psychological Association*, organisation responsable, entre autres, des normes régissant l'écriture de toute publication scientifique en matière de psychologie. Son organe officiel est le *Publication Manual of the American Psychological Association*, 4th ed., 1994, Washington, American Psychological Association.

4. Lisible sur http://www. apa. org/psa/stats. html

Genèses 26, avril 1997, pp. 123-142

dans les tests d'hypothèse, qui semble toucher tous les domaines de la recherche en sciences humaines, et certains membres éminents de cet organisme n'hésitent pas à rejeter la responsabilité de ce phénomène sur les éditeurs de revues et leurs *rewiewers*.

Le réseau Internet a constitué notre seconde source d'information. On y trouve en effet des groupes de discussion (newsgroup: sci.stat.*) consacrés aux statistiques, et des échanges passionnés y ont lieu à longueur d'année entre des universitaires de tous niveaux qui dénoncent certaines pratiques qu'ils jugent pernicieuses, et ceux qui en prennent peu à peu conscience, non sans résistances. Peu à peu, une riche bibliographie à pu être constituée, preuve manifeste – il suffit d'en parcourir les titres – que « There is something wrong with statistics »[5].

Mais qu'est-ce qui ne tourne pas rond avec les statistiques ?

Le problème étant ainsi mieux cerné, voici les principaux reproches que l'on rencontre régulièrement dans les travaux des méthodologues les plus prestigieux Outre-Atlantique, tels Rozenboom, Campbell, Meehl, Carver, Cohen, et bien d'autres[6]:

1. Les tests inférentiels[7] sont souvent mal compris et utilisés par les chercheurs de sciences humaines peu au fait des fondements théoriques des techniques qu'ils utilisent. En conséquence, les résultats empiriques sont mal interprétés et conduisent à des conclusions inappropriées.

2. Des sommités respectées de la psychologie dite scientifique ont modifié le sens véritable du raisonnement inférentiel, qu'ils ont ensuite diffusé sous une forme que les méthodologues critiques considèrent comme « pervertie ». Depuis lors, ces conceptions erronées se perpétuent en conservant leurs incohérences miraculeusement intactes, d'enseignant

▶ ▶ ▶

5. Richard M. Royall, « The trouble with statistics », Troisième cycle romand de statistique et de probabilité appliquée, séminaire de printemps, Villars-sur-Ollon, 1995. (Communication non publiée, peut être obtenue auprès des auteurs).

6. Parmi les critiques les plus respectés dans le domaine William Rozenboom, « The fallacy of the null hypothesis significance test », *Psychological Bulletin*, vol. 57, 1960, pp. 416-428. – Paul E. Meehl, « Theoretical risks and tabular asterisks: sir Ronald, and the slow progress of soft psychology », *Journal of Consulting and Clinical Psychology*, vol. 46, 1978, pp. 806-834. – Jacob Cohen, « Things I have learned (so far) », *American Psychologist*, vol. 45, 1990, pp. 1304-1312. – Ronald P. Carver, « The case against statistical significance testing », *Harvard Educational Review*, vol. 48, 1978, pp. 378-399. – Donald T. Campbel, & Julian C. Stanley, *Experimental and quasi-experimental designs for research*, Chicago, Rand McNally, 1966.

7. Pour être précis, le BSA vise ici toutes les techniques faisant appel à la notion de « signification », c'est-à-dire les tests de t, du « chi carré », de F et dérivés, de corrélations, ainsi que les tests multivariés (*cf.* régression, analyse discriminante, canonique), ainsi que les tests non-paramétriques.

8. Ce vocabulaire « psychanalytique » est emprunté à Gerd Gigerenzer, « The superego, the ego, and the id in statistical reasoning » *in* G. Keren & C. Lewis (eds), *A handbook for Data Analysis in Behavioral Science – Methodological Issues*, Hillsdale, Lawrence Erlbaum, 1993, pp. 311-339. Toutes les citations empruntées à cet auteur sont tirées de cet article, nous les avons traduites nous-mêmes, de même que celles des autres auteurs américains cités.

9. Le vocabulaire « religieux » est emprunté à David S. Salsburg, « The religion of statistics as practiced in medical journals », *American Statistician*, vol. 39, n°3, 1985, pp. 220-223.

10. J. W. Tuckey, « Analysing data : sanctification or detective work ? « *American Psychologist*, vol. 24, 1969, pp. 83-91.

11. D. Salsburg, *op. cit.*

12. Nigel G. Yoccozz, « Use, overuse, and misuse of significance tests in evolutionary biology and ecology », *Bulletin of the Ecological Society of America*, vol. 72, n°2, 1991, pp. 106-111.

13. Ces critiques ont été récemment synthétisées dans Frank Schmidt, « Statistical significance testing and cumulative knowledge in psychology : implications for training of researchers », *Psychological Methods*, vol. 1, n°2, 1996, pp. 124.

en enseignant et de manuel en manuel. On ne s'étonnera donc pas que la plupart des étudiants sortent de l'université sans avoir réellement compris les tenants et aboutissants de la méthode inférentielle.

3. La généralisation de ce mode de penser fallacieux débouche sur une pratique mécanique, *compulsive* et *obsessionnelle*[8], si ce n'est *religieuse*[9], des tests d'hypothèse, relevant davantage du rite propitiatoire que de l'esprit hypothético-déductif. Selon les critiques les plus acerbes, les résultats extraits de l'encensoir statistique n'auraient d'autre but que de satisfaire un rituel dont la finalité serait de sacraliser un résultat en lui permettant d'accéder au statut de vérité scientifique[10]. Les titres de certains articles sont évocateurs: «*La religion statistique telle qu'elle est pratiquée dans les publications de médecine*»[11] ou «*Usage, sur-usage ou mauvais usage des tests statistiques en biologie et en écologie*»[12] L'un des «réformateurs» les plus virulents, H. Rubin, professeur à Purdue University, n'hésite pas à déclarer: «De toutes les religions, la statistique est candidate à devenir celle que l'on pratique avec le plus de dévotion»...

Ces critiques ne surprendront pas ceux qui s'intéressent de près aux rapports entre la théorie et l'application des techniques statistiques, en psychologie aussi bien qu'en sociologie, biologie, médecine, géographie et bien d'autres disciplines. Les premières réflexions méthodologiques critiquant certains usages impropres des statistiques en sciences humaines, comme dans d'autres domaines, apparaissent en effet dès les années 1950. Or, la littérature consacrée à ce sujet suggère que cette mise en question radicale émerge dès la constitution, après la fin de la seconde guerre mondiale, d'un nouveau mode de penser que Gigerenzer nomme «logique hybride», consistant en une théorie syncrétique qui amalgame de manière surprenante des idées aux origines distinctes, voire conflictuelles. Cette théorie hybride, qui règne actuellement sur la quasi totalité du domaine de la recherche en sciences humaines, empêche tout progrès cumulatif du savoir en sciences humaines: la pratique compulsive des tests d'hypothèse qu'elle autorise en effet, dissimule le caractère illusoire des bénéfices que ceux-ci sont supposés rapporter[13].

Le présent article a pour but d'éclaircir ces questions. Nous présenterons tout d'abord le contexte théorique dans lequel la pratique des tests inférentiels – véritable paradigme actuel de la scientificité en sciences humaines – a pris racine. Nous analyserons ensuite les divers facteurs qui ont pu favoriser sa naissance, puis sa dérive vers une forme contestable; enfin, nous commenterons quelques hypothèses relatives à l'extraordinaire résistance de ce mode de penser qui, selon nous, n'est peut-être que l'avatar contemporain de l'antique fascination exercée par le nombre.

L'inférence statistique: un bref rappel

Pour comprendre et décrire l'origine des pratiques incriminées, il est nécessaire de fixer un certain vocabulaire, ce qui obligera le lecteur profane à faire quelques pas en direction du cœur de la théorie de l'inférence statistique.

Rappelons tout d'abord que les théories de l'inférence s'inscrivent dans le cadre général du problème de l'induction. Leur tâche est de répondre à la question: *que peut-on légitimement prédire d'un ensemble d'éléments lorsque les informations dont on dispose ne se rapportent qu'à quelques-uns de ses éléments*? En statistique, l'acte généralisateur, ou *inférence*, peut être défini comme la projection conjecturale et modalisée, sur une population *parente*, d'un savoir acquis sur l'un de ses échantillons[14]. Un tel échantillon est déclaré *représentatif* si chacun de ses éléments doit son appartenance à l'échantillon au *seul* fait

qu'il appartient à la population. Il est clair qu'une telle condition – dont le sens est de prévenir l'intervention d'un biais quelconque dans la constitution de l'échantillon – ne peut être mieux satisfaite que par l'application d'un procédé de sélection *aléatoire*, procédé relativement facile à concevoir en génétique végétale, mais dont l'application en sciences humaines est souvent condamnée à ne rester qu'un vœu pieux. La procédure d'échantillonnage a toujours pour fin la construction d'un microcosme dont les caractéristiques sont, à l'échelle près, aussi proches que possible de celles de l'univers qu'il est censé représenter.

L'étude de ce microcosme, ou *analyse descriptive*, fournit, pour autant que les mesures soient bonnes, des résultats chiffrés, précis et définitifs. On peut en ce cas calculer les valeurs *exactes* des grandeurs que l'on désire mesurer. Mais si l'on infère alors à la population parente les connaissances ainsi acquises – *et c'est ce que l'on fait chaque fois que l'on recourt à un test* – les valeurs obtenues perdent leur exactitude, et ne sont plus, en conséquence, que des valeurs estimées: ce ne sont plus des nombres qui les expriment alors, mais des grandeurs variables, sujettes aux aléas de l'échantillonnage, autrement dit *empreintes de probabilité*. Il se trouve donc, et on l'oublie parfois, que le gain fait par inférence en « généralisabilité » coûte son prix. Or il se trouve que les modalités de ce payement peuvent différer, ce qui explique l'existence de différentes approches, parfois conflictuelles.

Les premiers pas du test d'hypothèse

C'est à K. Pearson (1857-1936) que l'on doit une première solution formalisée au problème de l'inférence[15]. Sa théorie du « test d'ajustement » (*goodness of fit test*), qui s'apprête à fêter son premier centenaire, consiste en une comparaison entre distributions de fréquences théoriques et observées,

▶ ▶ ▶

14. À ce propos, William S. Peters (*Counting for something*, New York, Springer, 1987) donne (p. 2) une interprétation surprenante – quasi pulsionnelle – de l'inférence: « *lorsqu'une information est basée sur un échantillon de cas possibles, il existe un besoin de généraliser depuis les cas étudiés, sur l'univers de tous les cas possibles* ». L'induction peut donc être perçue comme une nécessité vitale, une propriété constitutive de l'esprit humain, comme substratum de la perception.

15. On trouve dans l'ouvrage de Alain Desrosières (*La politique des grands nombres, histoire de la raison statistique*, Paris, La Découverte, 1993), un passionnant historique de la construction du savoir statistique, de Pascal à nos jours.

comparaison qui se fait par l'intermédiaire d'une mesure de distance, rapportée à son tour à une loi de probabilité connue sous le nom de «chi carré». L'histoire des statistiques montre que cette procédure va inspirer la conception de tous les tests imaginés depuis, tels que les tests de t, de F, d'indices non paramétriques, etc. Ces procédures supposent toutes en effet, un état initial d'ignorance à propos d'une ou plusieurs caractéristiques d'une population bien définie, cible de la recherche. En présence d'un phénomène jugé digne d'intérêt, par exemple la relation entre deux mesures morphologiques X et Y sur une population de végétaux, le chercheur doit élaborer une hypothèse «nulle» (H_0) exprimant, par exemple, l'idée que les deux grandeurs X et Y sont *indépendantes*, autrement dit que la variation de X n'entraîne aucune variation prévisible de Y. S'il entend juger de la valeur de cette hypothèse, le chercheur devra la soumettre à l'épreuve de l'expérimentation sur un échantillon représentatif de la population parente. Le sens de cette expérimentation est précisément donné par une famille de procédures consistant à tester la crédibilité (ou la «recevabilité») d'une hypothèse en regard des données de l'expérience.

Il nous semble illusoire de vouloir exposer les principes du «test d'hypothèse» en quelques phrases rapides et superficielles, car l'enseignement de ce point constitue un défi pédagogique majeur, de l'aveu même de la plupart des enseignants chargés de le dispenser. Nous nous y risquerons néanmoins, de manière à introduire les concepts essentiels.
Partant d'une question que l'on peut se poser relativement à un certain aspect de la réalité, on définit une hypothèse «nulle» correspondant à la manifestation «attendue» d'une loi théorique (modèle) associée au phénomène qui nous intéresse, par exemple l'indépendance de deux caractéristiques dans une population, ou une loi génétique déterminée. L'expérimentation sur un échantillon nous conduira nécessairement – en raison de la variabilité interindividuelle – à observer un *écart* entre les données recueillies, et celles – *théoriques* – que le modèle fournit (H_0). C'est ici qu'intervient l'idée de test, car cet écart peut être attribué, *jusqu'à un certain point*, au hasard de l'échantillonnage, mais si cet écart devait dépasser un certain *seuil*, on serait amené à douter, voire à rejeter l'hypothèse nulle, car l'influence des *seuls* aléas d'échantillonnage ne peut plus, dans un tel cas, être déclarée suffisante pour expliquer son ampleur. Ce seuil devrait être en principe fixé à l'avance par l'expérimentateur, et on l'appelle généralement *seuil de signification* du test. On appelle *niveau de signification* du test (ou «**p-value**») la probabilité *conditionnelle* de trouver un écart égal ou supérieur à celui mesuré dans notre expérience particulière, entre ce que l'on a observé, et ce que l'on attendait. Le seuil de signification (choisi par l'expérimentateur) détermine la valeur du *risque de première espèce* (ou de type I), soit celui de rejeter à tort l'hypothèse nulle. Un enseignement plus strict insistera sur la nécessité de définir une autre hypothèse H_1, mise en compétition avec H_0. Cette hypothèse alternative peut, elle aussi, être rejetée à tort, ce qui constitue le risque dit de deuxième espèce. La connaissance de la probabilité de ce risque permet, le cas échéant, de calculer la puissance (sensibilité) du test, laquelle, combinée au seuil de signification choisi, permet de savoir combien il faut tirer d'individus de la population pour garantir une efficacité maximale du test.

Héritier spirituel de Karl Pearson, Sir Ronald A. Fisher, qualifié par d'aucuns de «père de la statistique moderne» n'enseigna jamais la théorie statistique pour elle-même. Professeur de génétique végétale, il inventa trois techniques inférentielles, dont une seule a fait école, quoique fort malmenée, comme nous le verrons par la suite.

Afin de bien comprendre la position de Fisher, il convient de redéfinir le plus précisément possible les termes introduits ci-dessus. Nous allons donc les reprendre un à un, et ainsi pourrons-nous rendre à Fisher ce qui devrait lui revenir en propre. Remarquons, tout d'abord, que Fisher n'a jamais parlé du «niveau de signification» d'un test en référence à la «p-value», et s'il a parfois utilisé l'expression «test d'hypothèse nulle», il dénomma toutefois clairement sa démarche «test de signification». De plus, si la «p-value» est effectivement pour lui une «probabilité de signification», il n'a cependant jamais préconisé de choisir avant le déroulement de l'expérience une valeur seuil

comparable au «seuil de signification» évoqué précédemment. À cet égard, Fisher a toujours milité en faveur d'une attitude flexible – devenue fort rare de nos jours – qui tienne compte d'autres données en jeu, comme l'expertise acquise par la longue fréquentation du domaine étudié, par exemple. À ce propos, Goodman commente: «[Pour Fisher] la p-value n'est pas interprétée comme une fréquence d'hypothétiques erreurs en cas d'expérimentations répétées. Elle représente une mesure d'«évidence», basée sur une seule expérience, reflétant la crédibilité *post hoc* d'une hypothèse, expérience faite»[16]. Dès lors que Fisher ne définit pas d'hypothèse alternative H_1 à l'hypothèse nulle H_0, les notions d'«erreur» de première et, a fortiori, de seconde espèce, n'existent donc pas chez lui, de même que celle de *puissance* d'un test. Fisher ne raisonne donc pas en termes de *risque de se tromper, mais bien plutôt de degré de conviction*, ou de «corroboration» (le terme anglais correspondant, difficile à traduire en français est: *évidence*, littéralement: *pièce à conviction*) en faveur de l'hypothèse nulle, la mesure de cette confiance étant donnée par la probabilité de signification (p-value)[17].

Fisher n'a jamais cherché à cacher les bénéfices limités de sa méthode. Dans son optique, la généralisation de résultats empiriques restera toujours du domaine de la *conjecture*: les techniques statistiques peuvent plus ou moins assurer le sérieux de ces conjectures en contrôlant au mieux les différentes sources de variation, tout en accumulant un faisceau suffisant d'éléments de corroboration (*évidence*) en faveur ou contre une hypothèse donnée, mais elles ne peuvent pas, et ne prétendent pas, muer une conjecture en connaissance[18].

Reprochant à Fisher son interprétation des probabilités, les mathématiciens Jerzy Neyman et Egon Pearson proposent, dès 1928, d'abandonner les notions d'«évidence» ou de

▶▶▶

16. Steven N. Goodmann «P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate», *American Journal of Epidemiology*, vol. 137, n°5, 1993, p. 486. Nous reproduisons les citations que Goodman a tirées de Ronald A. Fisher, *The design of experiments* (8th ed. 1966). Edinburgh, Oliver & Boyd, 1935.

17. Les utilisateurs de *packages* statistiques savent bien que cet indice est très généreusement distribué par tous les logiciels statistiques modernes. Sur ce point, la conception de ces logiciels incite donc plutôt à un raisonnement fishérien.

18. Fisher (cité par Gigerenzer, *op. cit.* p. 319): «We may say that a nonsignificant result "confirms" but does not "establish" the null hypothesis». Fidèles à cette ligne de pensée, D. Campbell & J. Stanley écrivent «*Des résultats expérimentaux* [provenant de tests inférentiels] *ne prouvent jamais une théorie*», de même qu'aucune hypothèse nulle testée ne peut jamais être acceptée, ni confirmée. En accord avec Hume, ces auteurs pensent qu'en cas de non-rejet de celle-ci «[...] *l'hypothèse nulle a été exposée à l'infirmation, et a échappé à celle-ci.*» (*op. cit.* p. 35).

19. Fisher ridiculisa cette conception des probabilités qu'il considère comme une «*fiction de mathématiciens détachés de la réalité*», car de son point de vue, l'échantillonnage multiple dans une même population n'est pas concevable. Il lui semblait également insensé de choisir un seuil fixe de signification, ce qui revenait, selon lui, à confondre «*technology*» et «*knowledge*», et constituait à ses yeux une «*absurdité académique*». (d'après Gigerenzer, *op. cit.* pp. 316-317).

« conviction », difficiles à définir et par trop subjectives à leurs yeux.

Jerzy Neyman (1894-1981) naît en Russie, étudie en Ukraine puis en Pologne et se voit envoyé à l'*University College* de Londres par ses professeurs pour « publier avec le maître K. Pearson ou ne jamais revenir à Varsovie ». Ce qu'il fît en suivant les deux injonctions. Il devint assistant du grand Pearson en compagnie de Egon Pearson, son fils, avec lequel ils forgèrent leur théorie des tests d'hypothèse. Le mathématicien et chimiste W. S. Gosset (alias Student, 1876-1937) introduisit Neyman dans l'entourage de Fisher, de quatre ans son aîné, lequel succédera à K. Pearson en 1933 à la tête du *Galton Laboratory of Genetics*. À la même époque, E. Pearson succède à son père qui dirigeait également le département de *Applied Statistics*. En 1935 éclate le célèbre conflit opposant les deux héritiers de K. Pearson, l'un naturel: son fils associé à J. Neyman, et l'autre spirituel (Fisher). L'origine de la dispute n'est pas claire, William Peters (*Counting for something, op. cit.* pp. 183-189) suggère que Fisher, après la publication par Neyman et Pearson d'un compte rendu de recherche sur des données agronomiques, n'apprécia pas l'intrusion de mathématiciens dans un champ qui leur est en principe étranger. Loin de se régler par le dialogue, le conflit s'envenima au point que les deux collègues ennemis en vinrent à éviter soigneusement de se croiser dans la cafétéria de leur institut. D'après la petite histoire, cette situation dura des années, si bien qu'en 1937, Neyman quitte l'Angleterre pour ne revenir en Europe que lors de brefs séjours.

Selon leur théorie qualifiée depuis de « fréquentiste », le rôle d'un test n'est pas de justifier une confiance plus ou moins grande en une hypothèse, mais de fournir une règle de conduite permettant de savoir *comment se comporter* en face des résultats fournis par l'expérimentation. Ce que les gens désirent, c'est savoir s'il convient d'avaler ou non un certain médicament, d'utiliser tel engrais ou tel autre, d'appliquer telle mesure d'ordre social ou politique, etc.

On comprend sans peine l'importance de tels enjeux, en pleine crise économique de la fin des années vingt. Si l'on considère l'actualité de l'année 1996, on peut imaginer que Neyman et Pearson seraient moins soucieux de fixer le degré de confiance qu'il convient d'accorder à la thèse (H_0) de la non-transmissibilité de l'encéphalite spongiforme bovine (ESB) à l'homme, que de répondre à la question : « Faut-il tuer les vaches anglaises, suisses, etc., étant donné que cette maladie *risque* de se transmettre à l'homme » ?

Du point de vue formel, le raisonnement de Neyman et Pearson débute par la définition d'un plan expérimental dont les règles sont parfaitement définies.

Deux hypothèses rivales (par exemple : H_0 = « L'ESB ne peut pas se transmettre à l'homme » et H_1 = « La transmission est possible ») sont énoncées et mises en concurrence dans le cadre d'une procédure, appelée *test d'hypothèse*. Une statistique (variable) de décision S est définie sur la base de l'écart observé entre les données d'une expérience, et celles théoriquement fournies par le modèle correspondant à H_0. Cette valeur, comparée à la distribution théorique attendue sous H_0, permettra une décision simple et sans ambiguïté, selon que la valeur de S se trouve comprise, soit dans un intervalle préalablement défini comme *domaine d'acceptation* de H_0, soit dans un intervalle complémentaire, limité par un seuil (fixé) de signification, appelé *domaine de rejet* (ou région critique) de H_0. Dans ce dernier cas, c'est l'hypothèse alternative H_1 qui sera préférée à H_0. On est donc en présence d'une *règle de décision* dans laquelle il n'est plus question de *signification* ni d'*éléments de preuve* en faveur ou non de la crédibilité d'une hypothèse nulle. À ces notions fishériennes se substitue, chez Neyman-Pearson, celle de *risque d'erreur* (de décision). Ce risque est délibérément choisi lors de la définition du plan expérimental. En effet, l'argumentation fréquentiste postule que le chercheur, *avant d'entreprendre toute mesure et dès lors qu'il a lui-même choisi un seuil (par exemple a = 5 %)*, sait qu'en cas d'expérimentations répétées, menées sur des échantillons ayant tous la même taille mais dont la composition peut différer, 95 % des valeurs de la statistique S « tomberont » dans le domaine favorable à H_0, et 5 % en dehors. Les risques de se tromper sont donc clairement définis: le choix de H_1 au détriment de H_0 risque d'être malheureux 5 fois sur 100, et un raisonnement analogue permet, si H_1 est bien définie, de connaître le risque de se tromper en préférant H_0 à H_1. On retrouve ici la théorie des erreurs de première et de seconde espèce qui, comme nous venons de le voir, repose chez Neyman et Pearson, sur une conception de la notion de probabilité en termes de fréquence relative lors d'expérimentations répétées[19].

Cette procédure implique un net rejet de la théorie de Fisher, et Neyman et Pearson l'affirment avec force: « Aucun test basé sur la

théorie des probabilités ne peut, par lui-même, fournir une quelconque conviction pour ou contre une hypothèse[20]»; ils insistent également sur l'idée que seule une règle de décision peut gouverner notre comportement, à condition qu'elle nous assure de ne pas nous tromper trop souvent... Une autre différence fondamentale découle de ces divergences: bien interprétée, la démarche de Fisher n'aboutit jamais à des affirmations du type: «H_0 est vraie» ou «H_0 est fausse». Le test d'hypothèse de Neyman et Pearson ne présente pas ce type de délicatesse puisqu'il *force* le choix entre deux hypothèses rivales: le fait d'en rejeter une implique nécessairement que l'on accepte l'autre, c'est à dire que l'on agisse comme si elle était vraie. En ce sens on pourrait affirmer, en paraphrasant Goodman, que l'œuvre de Neyman et Pearson – dans la mesure où elle préconise une règle de pur *comportement* – représente *une tentative de complet rejet du raisonnement inductif.*

On pourrait encore se demander si la règle de décision mise au point par Neyman et Pearson constitue réellement un cadre conceptuel plus «objectif» que celui de Fisher, ainsi que le prétendent leurs auteurs. L'exemple des vaches montre à l'évidence qu'il n'en est rien, car si toute la procédure paraît limpide jusqu'à la définition de la valeur critique, c'est la problématique tout entière qui devient critique à ce point. Posée en termes humains, par exemple, la question du choix du seuil «critique» revient à se demander: combien de morts humaines dues à une transmission hypothétique de la maladie à l'homme peut-on mettre en balance avec la perte économique subie en cas de mise à mort de toutes les vaches (ou partie) du cheptel national? Qui peut en toute bonne foi prétendre juger de ce délicat problème? Qui, finalement, endossera la responsabilité, et payera éventuellement le prix de l'erreur liée à l'utilisation de la «machine à décider avec risques» de Neyman et Pearson? On voit immédiatement que le

▶▶▶

20. Jerzy Neyman & Egon Pearson, «On the problem of the most efficient tests of statistical hypothesis,» *Philosophical Transactions of the Royal Society*, vol. 231, 1933, p. 290. (Cité par S. Goodman, *op. cit.* p. 486).

21. Ces deux approches semblent être en conflit permanent, à tel point que lorsque deux théoriciens appartenant à ces deux écoles se retrouvent par un malencontreux hasard dans le même congrès ou séminaire, le ton monte rapidement, en même temps que le niveau passionnel du débat, et il n'est pas impossible d'entendre quelques invectives fort incongrues dans un tel cadre. Le lecteur peut imaginer notre surprise lors d'un mémorable séminaire de statistique, lorsque deux éminents statisticiens ont chacun jugé certaines idées de l'autre comme étant des «offenses à la science»...

22. On peut lire l'article original (Thomas Bayes, «An essay towards solving a problem in the doctrine of chance» *Philosophical Transactions of the Royal Society*, vol. 53, 1763, pp. 370-418) dans Egon. S. Pearson & M. C. Kendall (eds), *Studies in the history of statistics and probability*, London, Griffin, 1970.

cadre apparemment strict et rassurant du test fréquentiste ne résiste pas si l'on prend en compte le *coût* de l'erreur, cet aspect ayant toujours été traité de manière marginale par les statisticiens de cette école. La raison en est simple : ce point est crucial et marque les limites de leur compétence, car le choix du seuil «critique» (le bien nommé) n'est pas de leur ressort, mais bien de celui des commanditaires de l'enquête. En effet, ce sont ces derniers qui porteront *seuls* la responsabilité d'une mauvaise décision, étant donné que les conséquences réelles d'un choix malencontreux ont toujours un coût économique ou politique. Seul celui qui pose la question peut donner un sens à la réponse qu'il peut éventuellement trouver: les techniques de résolution de problèmes, pour leur part, ne s'en préoccupent pas et il serait aberrant de leur demander un jugement épistémique. Cette possible confusion des compétences relativement à l'interprétation d'un effet ou du choix d'un seuil conduit à un résultat extrêmement déplorable, mais parfois intéressé: celui de la *dissolution des responsabilités*. Si la décision finale déplaît au peuple, quoi de plus simple que d'attribuer sa responsabilité aux «experts» ou «aux statistiques»?

L'approche bayesienne de l'inférence

La comparaison entre les conceptions fishérienne et fréquentiste ne saurait être vraiment éclairante pour notre propos si on ne la situait dans la perspective plus générale de l'histoire du raisonnement inductif. Fisher, aussi bien que Neyman et Pearson, a connu l'œuvre de Bayes, mathématicien anglais de la fin du dix-huitième siècle. Fisher s'y réfère explicitement en rejetant l'idée qu'on puisse utiliser des probabilités *a priori* à propos de certaines hypothèses. Comme nous l'avons vu plus haut, Fisher comme K. Pearson, son maître, part d'une situation d'ignorance totale, quitte à diminuer un peu cette ignorance par la pratique du test de signification. Ce qui frappe dans la pensée de Neyman et Pearson, c'est leur violente opposition à toute idée d'induction, par leur attachement à une définition purement *fréquentiste* – pragmatique – de la notion de probabilité, ce qui les place aux antipodes de la conception bayesienne qui veut que cette notion de probabilité soit comme *essentiellement* attachée aux phénomènes réels. Cette opposition radicale[21], de nature philosophique, entre les approches fréquentistes et bayésiennes de la probabilité, a donné naissance à deux courants séparés, totalement cloisonnés, de la théorie statistique.

Bayes ne croit pas en la naïveté du chercheur, il admet et intègre dans sa logique inductive toutes les croyances préalables que celui-ci peut entretenir à propos de l'objet de sa recherche et des effets de celle-ci. Sa théorie[22] a été vigoureusement critiquée par le courant scientiste du XIX^{e} siècle, jusqu'à nos jours, principalement en raison de son postulat central de l'existence de probabilités *a priori* [*prior probabilities*]. Le choix de la valeur de cette probabilité *a priori* reste considéré par la plupart des tenants de la statistique «standard» comme une irruption intolérable de la subjectivité dans la démarche scientifique.

> Dans la logique bayésienne, les grandeurs avec lesquelles on travaille relèvent de deux types: celles qui sont connues de la personne réalisant l'inférence, et celles qui lui sont inconnues. Les quantités connues sont représentées par leurs valeurs, et celles qui sont inconnues par une distribution de probabilité conjointe. Les calculs sont de type probabiliste: si un certain modèle, donnant la probabilité des données selon un certain paramètre existe, et si, d'autre part, une distribution de probabilité *a priori* de ce paramètre est définie, alors il est possible de calculer à l'aide du théorème de Bayes la distribution *a posteriori* du paramètre, sur la base des données recueillies. Ainsi, toutes les incertitudes peuvent être évaluées à l'aide de la spécification du modèle *et* des probabilités *a priori*, et il est donc possible de calculer différents indicateurs statistiques. Les bayésiens justifient le choix de leur procédure en argumentant qu'elle est la seule à permettre la

résolution des problèmes d'inférence, à condition d'admettre que l'on puisse représenter *toute* incertitude par des probabilités; de plus, ils affirment que celle-ci est également la seule capable de satisfaire certaines conditions de cohérence définies à l'aide d'axiomes «naturels» sur l'incertitude[23]. (On ne s'étonnera pas que les fréquentistes divers prétendent que ceux-ci n'ont été définis que dans ce but...).

Tentons d'illustrer les différentes démarches exposées ci-dessus par un exemple simple. Supposons qu'une pièce de monnaie soit présentée à trois chercheurs de stricte obédience, représentant les trois doctrines: fishérienne, fréquentiste (Neyman et Pearson) et bayésienne. Chacun d'eux doit décider si cette pièce est équilibrée ou non.

L'adepte de Fisher va postuler que la pièce est équilibrée, puis mettra cette hypothèse (H_0) à l'épreuve du *test de signification* en la lançant en l'air un certain nombre de fois. Selon le résultat de cette expérience, il calculera une p-value qui influencera son degré d'adhésion à H_0. Une différence jugée par lui *trop grande* entre le nombre de jets «pile» et de jets «face», ébranlera sa confiance en H_0 et il dira que la pièce est probablement déséquilibrée, sinon il dira que cette unique expérience n'a pas entamé sa confiance dans l'hypothèse nulle. On remarque que le test fishérien est essentiellement prudent et de pouvoir heuristique faible: que l'on déclare la pièce truquée ou non, ne met pas ce jugement à l'abri d'une autre expérience, éventuellement contradictoire.

De son côté, le pur fréquentiste, disciple fidèle de Neyman et Pearson, considère le problème sous un autre jour: cette pièce peut-elle être utilisée ou non, pour acheter une marchandise sans risque de poursuites? Admettant qu'une pièce équilibrée est en principe vraie, il opposera deux hypothèses («la pièce est vraie = H_0» et «la pièce est fausse = H_1») dans le cadre d'un *test d'hypothèse*. Il va ensuite définir les modalités exactes de l'expérience, autrement dit: fixer la

▶▶▶

23. Voir D. V. Lindley, «Bayesian inférence», *in Encyclopedia of statistical sciences*, vol. 1 «A to circular error», New York, Wiley, 1982, pp. 197-204.

24. Comme nous l'avons vu plus haut, ce choix est loin d'être aussi innocent que le laisse supposer l'usage mécanique des tests, car il s'agit bien, pour en rester à cet exemple, de trouver un moyen terme entre l'intérêt d'utiliser la pièce et le risque de se voir puni pour l'avoir fait. Le problème peut paraître secondaire si l'on suppose la pièce composée de laiton et le châtiment léger, mais si elle était faite d'or, ne serait-il pas plus douloureux de rejeter H_0? Et si l'on risquait la peine de mort, ne devrait-on pas fixer une puissance maximum? Ces questions ne semblent jamais effleurer les multiples utilisateurs de tests d'hypothèse en sciences humaines qui fixent machinalement, «par convention», leur seuil de rejet à 5 ou 1 %.

25. Braxton Alfred, *Elements of statistics for the life and social sciences*, New York, Springer, 1987.

26. Lire à ce propos, James O. Berger & Donald A. Berry, «Statistical analysis and the illusion of objectivity», *American Scientist*, vol. 76, 1988, pp. 159-165.

27. Carl J. Huberty, «Historical origins of testing practices: the treatment of Fisher versus Neyman-Pearson views in textbooks», *Journal of Experimental Education*, vol. 61, n°4, 1993, pp. 317-333.

frontière (seuil) de la région critique[24], décider de la sensibilité (puissance) du test et calculer le nombre de jets nécessaires à la mise en œuvre de la procédure dont il a ainsi lui-même défini les critères de fiabilité. Il procède alors à l'expérimentation dont le résultat est sans appel: l'une des deux solutions est choisie et le problème du doute ou de la confiance dans le résultat n'existe pas: l'expérimentateur connaît exactement les conséquences de sa décision, elles se calculent en termes de *risque délibérément calculé et consenti*: c'est – en principe – le cœur léger qu'il jette la pièce dans le fleuve ou qu'il la présente sur le comptoir du marchand. Mais qu'en est-il alors de la connaissance de la nature propre de la pièce? Celle-ci reste inexistante, la règle de décision utilisée dans ce cas ne s'en préoccupe pas.

L'approche bayésienne quant à elle, s'intéresse davantage à la nature des choses: le statisticien bayésien qui reçoit la pièce énigmatique va tout d'abord la soupeser, la palper et l'inspecter, s'informer de son origine, analyser le visage sérieux ou goguenard de celui qui pose le problème, etc. Il va ainsi se forger une idée *a priori* et, selon ses convictions, il décidera par exemple que la pièce qu'il tient dans la main est probablement déséquilibrée en faveur des occurrences de «Face». Il quantifiera cette estimation en définissant une densité de probabilités (en langage bayésien: *priors*) «avant expérience» privilégiant l'événement «Face». Après expérience, le théorème de Bayes lui permettra d'obtenir une nouvelle distribution de probabilités appelées *posteriors*, et son nouveau savoir dérivera des caractéristiques de cette distribution.

Précisons encore que la démarche bayésienne demeure essentiellement suspecte aux yeux de certains, comme en témoignent par exemple les deux dernières phrases de l'ouvrage consacré par Alfred aux éléments de statistiques: «Excepté dans des cas spéciaux, [le théorème de Bayes] ne doit pas être considéré comme un outil scientifique, dès lors que les distributions de probabilité a priori de l'ensemble des hypothèses sont requises. Ceci est typiquement le but de la recherche scientifique[25]». Fortement imprégnées de scientisme, les idées d'Alfred montrent à quel point l'intervention explicite de la subjectivité heurte une conception de la science obnubilée par des critères d'objectivité importés des sciences dures[26]. Ce fait explique sans doute que les bayésiens restent encore peu nombreux, quand bien même leur approche semble plus naturelle que celle des statistiques qualifiées aujourd'hui de «standard».

Genèse des statistiques «standard»

Que faut-il entendre par «statistiques standard» et, plus précisément, quels éléments des trois doctrines originales décrites ci-dessus, retrouve-t-on dans les enseignements et la pratique actuelle des statistiques en sciences humaines? Selon certains auteurs, la pratique statistique qui sévit actuellement dans la recherche quantitative en sciences humaines est le produit d'une évolution complexe, au cours de laquelle les idées des fondateurs ont été en partie *occultées* et *mélangées*, ce qui a donné naissance à un mode de penser «hybride» qui, malgré la nature inconciliable des ses éléments, perdure depuis bientôt un demi-siècle.

Pour vérifier cette affirmation, le méthodologue américain Huberty a cherché dans 28 manuels de statistique (édités entre 1910 et 1992) les traces des idées propres à Fisher et à Neyman et Pearson. Il ressort de ses travaux[27] que jusqu'aux années trente, la statistique inférentielle s'appliquait surtout à estimer des paramètres et calculer des distributions de probabilités. Dans la période des années 1930 à 1950, apparaissent de nombreux ouvrages exposant, soit la théorie de Fisher, soit celle de Neyman et Pearson, mais le plus souvent

sans mention du nom des auteurs. Les premières confusions y apparaissent: la logique fishérienne se retrouve soudain «enrichie» d'un seuil critique, alors que celle de Neyman et Pearson doit s'accommoder chez certains auteurs d'un *seuil de signification* qu'accompagnent des considérations sur la «significativité» variable d'une p-value. Du coup, l'alternative H_0 / H_1 disparaît, si bien que les erreurs de première et de seconde espèce ne peuvent plus être définies, de même que la notion de puissance qui disparaît quasiment des plans expérimentaux. L'occultation – par mélange – des conceptions originelles a donc débuté très tôt, largement du temps des années professionnellement actives des pères fondateurs.

Huberty a ensuite analysé l'évolution des conceptions de cinq «auteurs-enseignants au long cours» de la statistique, de 1940 à la période contemporaine (1992). Cette recherche donne lieu à d'intéressantes observations: il découvre de purs adeptes de Fisher, sans compromis; d'autres restituent les idées de Neyman et Pearson, mais en amputant leur théorie des points les plus malcommodes, comme la définition d'hypothèses alternatives, le choix de la puissance du test et de la région critique. D'autres enfin, plus redoutables, font preuve de «créativité» et réinterprètent à leur manière le message des anciens. Pour Huberty, l'année 1956 marque la naissance officielle de la logique hybride, sa conception étant attribuée au psychologue Guilford qui, dans la troisième édition de son manuel[28], définit et consacre les concepts qui essaimeront rapidement dans toutes les disciplines des sciences humaines pour fonder la pratique fallacieuse actuelle, dont l'APA cherche aujourd'hui à limiter les dégâts[29].

Selon Huberty, les conceptions infidèles de Guilford permettent d'utiliser les tests d'hypothèse en vue d'un nouvel objectif, à savoir celui de *décider si une différence observée est, en soi et sans référence à une hypothèse nulle,*

▶▶▶

28. J. P. Guilford, *Fundamental statistics in psychology and education*, (3rd ed. 1956), New York, Mc Graw Hill, 1942.

29. Même le lecteur le plus indulgent reconnaîtra que Guilford mélange tout. Pour ne prendre que quelques exemples, et toujours selon Huberty: ni l'analyse de variance (ANOVA), ni la discussion à propos d'une différence de deux moyennes observées ne sont pour lui des tests d'hypothèse, ces techniques trouvent place dans les chapitres consacrés à la mesure et aux questions de fidélité (reliability); ailleurs, le «seuil critique» de Neyman et Pearson devient un «seuil de confiance», notion qui réunit les deux statisticiens ennemis du *London College* pour une bien surprenante lune de miel...

30. Robert Abelson, *Statistics as principled argument*, Hillsdale (NJ), Lawrence Erlbaum, 1995, p. 40.

31. Cf. Karl Danziger, *Constructing the subject.* Cambridge, Cambridge University Press, 1990.

32. A. W. Melton, «Editorial», *Journal of Experimental Psychology*, vol. 64, 1962, pp. 553-557. (cité par Gigerenzer, *op. cit.* p. 313)

«grande» ou «petite». Remarquons que cette singulière erreur constitue une perversion au sens propre du terme, dans la mesure où la fonction originaire de l'outil est littéralement détournée de son objectif, orientant l'instrument vers des fins pour lesquelles il n'était pas destiné. Dans un ouvrage publié récemment, le méthodologue Abelson écrit à ce propos: «Une confusion fréquente consiste à utiliser le niveau de signification comme un indicateur du mérite du résultat[30]». Tout semble donc indiquer que, utilisé dans le cadre de la logique hybride, *le test statistique est devenu un outil magique permettant de savoir si ce que l'on observe est intéressant (scientifique?) ou non.*

Cherchant à expliquer le développement de ce qu'il a appelé la «logique hybride», Gigerenzer invoque deux facteurs étroitement imbriqués, l'un psychologique et l'autre lié à ce qu'il appelle la «révolution inférentielle» intervenue dès l'après-guerre aux États-Unis. Rappelons brièvement le contexte. Avant la seconde guerre mondiale, les hommes de science pratiquaient avant tout la méthode expérimentale ou l'observation systématique, mais cela ne signifie pas que les techniques inférentielles étaient alors inconnues. En effet, dans un ouvrage publié en 1897, Fechner aborde divers thèmes méthodologiques, dont quelques techniques inférentielles, mais celles-ci ne constituent pas la méthode scientifique, elles en font partie au même titre que bien d'autres. Comme nous l'avons vu, la technique de Fisher apparaît dès 1930 dans les manuels, celle de Neyman et Pearson un peu plus tardivement, mais le fait est qu'avant 1940, fort peu d'articles scientifiques font mention de tests d'hypothèse, alors qu'en 1955 on note 80% d'articles présentant des résultats de tests. De nos jours, toujours selon Gigerenzer, ce taux avoisinerait 100%.

Une véritable révolution méthodologique, particulièrement évidente aux États-Unis, a donc eu lieu, reflétant une profonde mutation de la pratique expérimentale. Dès la seconde moitié du XIX[e] siècle et jusqu'aux travaux du psychologue behaviouriste Skinner, on étudiait de préférence des cas uniques, dans des conditions d'expérience en principe parfaitement contrôlées. Dans la période de l'entre-deux guerres et sans doute sous la pression de contraintes utilitaristes, les psychologues américains ont cru devoir légitimer socialement leurs recherches en fournissant des résultats applicables à des groupes, et non plus à des individus isolés. Ils pouvaient ainsi justifier plus facilement leur discipline dans divers domaines d'utilité publique comme l'éducation, la sélection et l'orientation professionnelle, et non plus seulement dans l'armée comme ce fut le cas pendant la guerre, par exemple. Ainsi, l'expérimentation sur des groupes prend un essor fulgurant: entre 1915 et 1950, le pourcentage d'études de cas uniques chute de 70% à 17%, alors que les études collectives passent de 25% à 80%[31]. Parallèlement, les éditeurs adaptent leurs critères: par exemple Melton, éditeur du *Journal of experimental psychology*, exige que les tests d'hypothèse soient au moins «significatifs» à un niveau de .01 car, selon lui, «les résultats situés entre .05 et .01 prennent la place de résultats de meilleure qualité»[32].

On peut faire l'hypothèse que dans un climat de compétition du type **publish or perish** de plus en plus contraignant, certains psychologues ont cru bon de remédier aux faiblesses des théories de Fisher, trop modestes à leurs yeux, en les dépouillant de leurs «*relents agricoles et de leur complexité mathématique*»[33] afin de les unifier en une seule et unique méthode efficace de production de faits scientifiques. C'est donc sous cette forme que la nouvelle statistique inférentielle, importée (encore intacte) d'Angleterre par Snedecor, Hotelling et bien d'autres, se répandit sur le sol américain. Investie par les sciences humaines, elle changea complètement de fonction: de méthode «naïve» d'investigation

de la réalité (parmi d'autres) elle devint un outil privilégié d'auto-validation scientifique. Son succès fut, on s'en doute, foudroyant.

Structure d'un dogme

La conception de la nouvelle théorie syncrétique peut être expliquée, selon nous, par le désir de satisfaire deux exigences liées à une représentation erronée et proprement fantasmatique de la « scientificité » que les chercheurs de sciences humaines crurent bon d'imposer dans leur discipline.

Premier fantasme : une méthode « scientifique » doit permettre de valider une hypothèse en fournissant une estimation (si possible chiffrée) de la probabilité de sa vérité.

Second fantasme : une méthode « scientifique » doit permettre de connaître le degré (si possible chiffré) de réplicabilité d'un résultat.

Sur un squelette fishérien (H_0, p-value), se greffe une règle de décision de type Neyman et Pearson (seuil, domaine de rejet, risque) et le tout s'exprime en un langage quasi bayésien (posterior probability). Gigerenzer dissèque sans pitié la rhétorique de Guilford ; pour notre part, contentons-nous de décrire les principaux avantages et inconvénients de ce nouveau mode de « raisonnement », aux promesses si séduisantes, et aux vices si bien cachés.

Chez Guilford, la *p-value* (niveau de signification) de Fisher se mue littéralement en pierre philosophale de la recherche scientifique : dans l'optique de cet auteur, elle est tout simplement la probabilité que l'hypothèse nulle soit vraie[34], elle serait donc une mesure de sa véracité (cf. satisfaction du premier fantasme). Plus grave : chez d'autres auteurs[35], le niveau de signification de Fisher hérite encore d'une autre propriété, celle de constituer une *valeur chiffrée du degré de réplicabilité d'un résultat* (cf. satisfaction du second fantasme). On peut montrer que ces

▶▶▶

33. G. Gigerenzer, *op. cit.* p. 323.

34. Selon Gigerenzer (op. cit. p. 323) : « Entre les mains de Guilford, la p-value, qui spécifie : p (D / H_0), i.e. la probabilité (conditionnelle) d'observer nos données, étant donné H_0, devient miraculeusement p(H_0 / D), probabilité bayésienne a posteriori de l'hypothèse, étant donné les observations ».

35. Gigerenzer cite les exemples de Anne Anastasi (*Differential psychology*, (3rd ed.), New York, Macmillan, 1958, p. 9) : « Le problème de la signification statistique réfère à celui de connaître la probabilité d'observer un résultat similaire en cas de répétition de l'expérience » ; ainsi que celui de J. C. Nunnally (*Introduction to statistics for psychology and education*, (6th ed. 1978), New York, McGraw-Hill, 1975, p. 195) qui écrit « Si la significativité statistique est au niveau .05, [...] alors l'expérimentateur peut espérer retrouver avec 95 chances sur 100 des différences semblables dans des expériences ultérieures ».

36. Selon D. Bakan (« The test of significance in psychological research », *Psychological Bulletin*, vol. 66, 1966, pp. 423-437), Michael Oakes (*Statistical inference: a commentary for the social and behavioral sciences*, New York, Wiley, 1986), et plusieurs auteurs auxquels Gigerenzer fait référence, cette croyance (*belief*) est partagée par 96 % des psychologues américains de niveau académique. L'erreur consiste à « oublier le cadre probabiliste de l'expérience sitôt qu'elle fournit ce qui était espéré » (R. Abelson, *op. cit.*).

37. G. Gigerenzer, *op. cit.* p. 326.

38. *Ibid.*

39. Gigerenzer cite une des nombreuses rééditions de W. L. Hays, *Statistics (for psychologists)*, New York, Holt, Rinehart & Winston, 1963, p. 287.

40. D. S. Salsburg, *op. cit.* p. 220.

41. R. Abelson, *op. cit.* p. 77.

42. Salsburg, *ibid.*

conceptions sont fausses[36], mais peut-on cependant les attribuer à une simple incompréhension? Gigerenzer, qui n'accuse pas les psychologues d'incompétence, attribue cette déviance à une *distorsion volontaire*, que nous pensons motivée par les fantasmes de scientificité évoqués plus haut.

Les conséquences psychologiques de cette « profanation » ne manquent pas d'intérêt. Jugeant que l'héritage des pères a été transmis de manière pour le moins tendancieuse – *le degré de conviction* de Fisher se muant, dans la logique hybride, en un *degré de véracité* – Gigerenzer fait l'hypothèse que les artisans de cette hérésie n'ont pas enfreint la Loi sans quelques remords... Les modalités de cette trahison sont d'ailleurs plus complexes car, outre la théorie de Fisher, la « théorie hybride » a également intégré, on l'a dit, la structure rigide de la mécanique décisionnelle de Neyman et Pearson, auteurs dont les idées sont, comme nous l'avons vu, notoirement opposées à celles de Fisher. Né d'une union aussi étrange, véritable transgression d'un *tabou épistémologique*, le nouveau mode hybride de « raisonner » possède dès lors toutes les caractéristiques d'une chimère: il se révèle en effet infiniment séduisant et, force est de le constater, indestructible. On peut imaginer, avec Gigerenzer, que la faute d'une conception aussi monstrueuse va nécessairement retomber sur ses auteurs et leurs héritiers, désormais habités par des « Sentiments de malhonnêteté et de culpabilité pour avoir violé les règles »[37]. Parallèlement, des mécanismes de défense n'ont pas tardé à s'installer, si bien que, toujours selon cet auteur: « La logique hybride a tenté de résoudre le conflit de ses parents en leur déniant le statut de parents »[38].

À l'appui de cette thèse, Gigerenzer examine trente manuels de statistique modernes et constate que vingt-cinq d'entre eux ne mentionnent *jamais* Neyman et Pearson, alors que le nom de Fisher apparaît plus fréquemment, mais uniquement en tête des tables statistiques dont il est l'auteur. Il constate encore que la logique « hybride » ne s'arrête pas là: au déni de l'existence des parents, s'ajoute encore le *déni de leur conflit*: le seul auteur[39] qui, dans ses manuels, les mentionne nommément, présente la théorie de Neyman et Pearson comme un simple progrès par rapport à celle de Fisher. Gigerenzer note que dans *aucun* de ces trente ouvrages, il n'est fait mention que la théorie statistique a toujours été un lieu propice aux controverses. Bien au contraire, la culpabilité liée au « meurtre des pères ennemis » a contribué à créer un climat de dogmatisme extrêmement contraignant: le respect absolu du rituel statistique est devenu de nos jours la voie obligée menant au ciel de la publication.

Le *paradigme statistique* s'est ainsi imposé. Essentiellement syncrétique et incohérente, la « logique hybride » demeure, de par sa nature même, nécessairement inaccessible aux efforts de compréhension de son utilisateur. Si celui-ci disposait du temps et de l'énergie nécessaire pour suivre la voie des méthodologues critiques que nous faisons intervenir dans ce débat, il découvrirait, comme Salsburg ou Carver, la surprenante réalité d'une pratique rituelle, véritable *religion statistique*. C'est ainsi qu'en très grand nombre, « les chercheurs s'engagent dans un rituel connu sous le nom de « chasse à la p-value[40] », que R. Abelson appelle aussi « la poursuite anxieuse de $p \leq .05$[41] ». Or, non seulement « [...] peu d'élus trouvent le salut[42] », mais l'adoption de cette attitude requiert de la part du chercheur le sacrifice de la compréhension réelle de la démarche, car il doit, nécessairement, à un moment ou à un autre, s'*en remettre* au statisticien. Cet acte de foi exige du chercheur un degré élevé de contrition, car ce n'est pas sans quelques réticences et sentiments de culpabilité qu'il accepte de laisser le grand-prêtre lui administrer ses sacrements, ses phrases

rituelles, ses formules imprégnées d'un ésotérisme complexe, ses dessins cryptiques, son sourire condescendant et ses remontrances paternalistes. Salsburg ironise à ce propos: « À ses risques et périls, on sollicite le prêtre (statisticien), lequel est rare et peu disponible, et fait en général de son mieux pour semer la confusion en posant des questions du genre « Pourquoi avez vous déjà réalisé l'expérience avant de me consulter ? » et il ne comprend pas du tout notre besoin de Rédemption [...][43] ».

C'est dans ce climat diffus de *culpabilité* que s'est installé ce que Gigerenzer appelle le *dogmatisme statistique* découlant de l'application de la statistique hybride héritée de Guilford: « Expérimenter, c'est fixer un seuil au niveau de signification, calculer les résultats, voir si le seuil est atteint, si oui publish, sinon perish[44] ». Selon Gigerenzer, le non-rejet de l'hypothèse nulle sera de plus en plus souvent interprété comme une mise en évidence d'un effet « nul », donc indigne de publication. Cette conclusion peut se justifier dans certains cas, mais on ne saurait nier que l'observation d'un effet nul peut, dans certains cas, également être d'un grand intérêt[45].

Dans sa pratique quotidienne de méthodologue, Gigerenzer déplore qu'il ne rencontre qu'une masse de travaux dont les hypothèses ne sont pas vraiment spécifiées, dans lesquels les seuils sont choisis par convention (.01 ou .05), généralement modifiés après l'expérience, pour être interprétés de manière fantaisiste, où l'ampleur des effets attendus n'est pas discutée, etc. Finalement, il qualifie de *comportement mécanique, obsessionnel et compulsif* ce qu'une grande partie de la communauté des chercheurs en sciences humaines considère comme une méthode scientifique « objective ».

Il arrive pourtant que certains chercheurs méritants s'insurgent contre ce rituel qu'ils jugent inacceptable et, suivant les conseils de flexibilité de Fisher, se fient davantage à leurs intuitions qu'aux injonctions dogmatiques.

▶▶▶

43. *Ibid.*

44. G. A. Barnard, J. C. Kiefer, L. M., Le Cam, & L. J. Savage, « Statistical inference », *in* D. G. Watts, (ed.), *The Future of Statistics*, New York, Academic Press, 1968, p. 147. (Cités par Gigerenzer, *op. cit.* p. 326)

45. Il suffit d'imaginer un test portant sur l'hypothèse: « H_0 = l'ESB ne se transmet pas à l'homme », quel soulagement de ne pouvoir la rejeter!

46. Cf. G. A. Barnard & alii., cités par Gigerenzer, *ibid.* p. 326.

47. On pourrait citer le cas (vécu) de ce chercheur venu nous consulter à propos des techniques statistiques qui pourraient être utilisées dans une recherche pour laquelle il espérait collecter des fonds. Nous avons tout d'abord voulu lui expliquer simplement notre point de vue, mais il nous interrompit *afin de nous supplier de lui dicter une dizaine de lignes incompréhensibles sans lesquelles sa recherche n'aurait aucune chance de trouver grâce aux yeux de ses supérieurs*, lesquels n'étaient par ailleurs pas utilisateurs de statistiques. Cette situation fait penser à l'histoire des habits neufs de l'empereur...

48. Rappelons que chez Fisher, la « signification » tire exclusivement son sens de la confrontation entre une hypothèse de travail (H_0) et des données expérimentales.

49. L'économiste McCloskey appelle le niveau de signification « *The gold standard* ». (D. N. McCloskey, « The insignificance of statistical significance », *Scientific American*, vol. 272, n°4, 1995, pp. 20-21)

50. Elazar J. Pedhazur, *Multiple regression in behavioral research (2nd ed.)*, New York, Holt, Rinehart & Winston, 1982, p. 24. Mis à part dans Carver et Meehl, déjà mentionnés, on trouve de semblables critiques également dans les articles (américains) suivants: Denton E. Morrison et Ramon E. Henkel (eds), *The Significance Test Controversy*, Chicago, Aldine, 1970. William D. Schafer, « Interpreting statistical significance and non-significance », *Journal of Experimental Education*, vol. 61, n°4, 1993, pp. 383-387; Bruce Thompson, « The use of statistical significance test in research: bootstrap and other alternatives », *Journal of Experimental Education*, vol. 61, pp. 334-349. Et pour citer des auteurs français, voir aussi, Daniel Corroyer & Henry Rouannet, « Note méthodologique sur l'importance des effets et ses indicateurs dans l'analyse des données », *L'Année psychologique*, vol. 94, 1994, pp. 607-624.

Mais Savage et ses collaborateurs considèrent leur destin avec pessimisme : « [Ces personnes] font le meilleur, inspirées par leur bon instinct, mais elles pensent néanmoins qu'elles vivent dans le péché[46] ».

Ces quelques remarques, parmi bien d'autres que le lecteur trouvera dans la littérature citée en référence, suggèrent que *les procédures statistiques telles que tests et techniques complexes ont acquis une fonction d'*imprimatur, *comme si le recours à celles-ci garantissait, à lui seul, la scientificité des résultats obtenus.*[47]

Des significations de la signification

Nous avons comparé plus haut la p-value à la pierre philosophale de la recherche quantitative contemporaine, son pouvoir (supposé) étant de permettre la transmutation de simples résultats d'observations en « findings » scientifiques dignes de publication. Il reste encore à évoquer le rôle de la formule magique qui réalise effectivement cet étonnant prodige.

Remarquons tout d'abord que par la grâce de la p-value, les chiffres issus des expérimentations de sciences humaines acquièrent (ou n'acquièrent pas) le sacrement de la *signification*[48]. Or, toute la question est de savoir en quoi le « signifié » ainsi dévoilé peut satisfaire notre désir de connaissance. Si, dans le sens courant, un résultat est dit « significatif » *s'il se prête à l'interprétation* (Robert), pour un statisticien, en revanche, cela voudra dire qu'un tel résultat n'est probablement pas dû au seul hasard de l'échantillonnage. Cependant, et nous touchons ici au nœud du problème de l'usage abusif des théories inférentielles, le sens de ce mot a subi une dérive parallèle à la logique à laquelle il est associé. En effet, le raisonnement hybride ne se borne plus à reconnaître à la p-value sa fonction originelle d'indicateur de la possibilité d'une interprétation, il lui ajoute indûment celle de *clé de cette interprétation*. Dans cette optique, la « signification » d'une p-value demeure bien l'attribut de ce qui est susceptible d'être interprété, mais en tant qu'indicateur de cette signification, elle se trouve du même coup promue au rang d'étalon d'évaluation[49]. Dans le cadre de la logique hybride, une p-value « significative » constitue *par elle-même* un quantificateur de l'importance d'un effet mesuré sur un échantillon, effet dont elle ne devrait être que l'indice révélateur de la probable non-nullité, sans plus. Cette conception est qualifiée sans détours de « *fantaisie statistique* »[50] par un nombre impressionnant de statisticiens.

Nous sommes ici au cœur du problème de l'utilisation abusive de statistiques en sciences humaines : les conséquences en paraissent pour le moins inquiétantes. Car, ayant ainsi élucidé le nouveau rôle donné par ses adeptes à la p-value, il devient possible de comprendre le sens des étranges critères de publication des principales revues scientifiques. Ces critères ne peuvent en effet se justifier que par la croyance qu'en plus d'être un indicateur de l'existence *probable* d'un effet d'une variable sur une autre (ce qui est vrai), la p-value serait encore un indicateur de l'importance de cet effet (ce qui est faux), et même un indicateur de *la qualité et de l'intérêt scientifique du résultat* (ce qui est proprement délirant).

Tout porte donc à croire que, tombé entre les mains des psychologues, et particulièrement de certains psychométriciens, le test statistique s'est mué en une sorte d'échelle « scientométrique », fournissant, à l'instar de n'importe quel autre test métrique, des mesures absolues et commodes (lecture facile, échelle sur 100), jouissant de toutes les apparences de l'objectivité découlant de l'usage de chiffres et de formules compliquées, permettant, ni plus ni moins, d'évaluer la qualité scientifique d'une recherche. Et c'est ainsi que la p-value fut considérée dès la fin de la seconde guerre mondiale comme l'augure de deux qualités propres à toute mesure psycho-

logique jugée «fiable», à savoir sa *fidélité* (reproductibilité), ainsi que sa *validité* (valeur «réelle» de l'influence d'une grandeur mesurable sur une autre).

On ne s'étonnera pas que la p-value, dotée de ces nouvelles vertus miraculeuses, ne soit pas restée longtemps l'apanage des psychologues: elle fut rapidement adoptée et exploitée dans d'autres domaines de recherche comme la sociologie, l'épidémiologie, la médecine, l'économie, la géographie, etc.

Pour en revenir à la question posée en introduction, on aura compris que dans l'usage des tests inférentiels, l'*abus* dénoncé par Dacunha-Castelle réside dans le fait que ces techniques sont utilisées improprement, puisque le niveau de signification (p-value) ne signifie pas ce qu'on veut lui faire signifier, ne mesure pas ce qu'on veut lui faire mesurer et ne dit donc pas ce qu'on veut lui faire dire. Si bien que les formules magiques trop souvent rencontrées comme: «la différence trouvée est significative ($p = 0.003$)», ou «la corrélation vaut 0.34 (***)», constituent dans certains contextes des conclusions vides de sens, si ce n'est de véritables contresens.

Pour justifier cette dernière affirmation, il faut tout d'abord chercher à comprendre en quoi le caractère polysémique du terme «*significatif*» peut biaiser gravement l'interprétation de résultats numériques. Nous avons montré qu'en plus du sens très général (cf. Robert) dans lequel le terme «*significatif*» est généralement utilisé, il pouvait aussi être considéré par certains comme un indicateur de l'ampleur d'un phénomène observé (cf. Guilford), voire de l'intérêt scientifique d'une recherche, et même de la qualité de celle-ci (cf. Melton). Ces confusions sont extrêmement courantes et peuvent être relevées dans pratiquement n'importe quelle revue faisant intervenir des statistiques.

Comment comprendre alors la formule rituelle, si fréquente dans les publications

▶▶▶

51. McCloskey (Cf. *op. cit. supra*) rapporte le cas devenu célèbre de l'expérimentation de l'effet de l'aspirine sur l'occurrence d'infarctus: l'expérience fut stoppée et considérée comme parfaitement convaincante bien avant que les standards de signification statistique ne fussent atteints. Les effets du traitement étaient si évidents dès les premiers contrôles qu'il eût été immoral de continuer à donner des placebos aux sujets du groupe témoin.

52. C'est pourtant ce que suppose H_0, l'hypothèse nulle.

53. Cf. D. McCloskey, *op. cit.*

scientifiques: «la différence entre les deux moyennes est significative... », sans autre spécification? Si l'on tient compte de ce qui précède, son auteur peut vouloir dire l'une au moins des quatre choses suivantes: 1) que la différence observée est *susceptible d'être interprétée* comme n'étant pas due au seul hasard de l'échantillonnage (sens statistique strict); ou 2) que la différence est importante, c'est à dire *grande* (sens évaluatif); ou 3) que la différence est *intéressante* pour son domaine (sens épistémique); ou, finalement 4) que le phénomène observé est issu d'une recherche sérieuse, donc *digne de publication* (sens pragmatique).

Remarquons tout d'abord que ces quatre acceptions sont *indépendantes*: un résultat peut être digne d'interprétation au sens statistique, *petit* du point de vue de l'ampleur de l'effet mesuré, mais *important* si l'on tient compte de ses retombées pratiques, et ceci, quelle que soit la qualité de la recherche effectuée. Il se peut aussi qu'un résultat intermédiaire soit suffisamment intéressant et impressionnant du point de vue de l'effet qu'il met en lumière, pour qu'il soit inutile d'établir sa significativité statistique[51]. Quoiqu'il en soit, il est impossible de présumer de l'importance pratique et scientifique d'un résultat de recherche en se fiant uniquement à son ampleur et à son niveau de signification; il faut pour cela impérativement tenir compte encore de la taille de l'échantillon et de la nature du domaine étudié. Sur ce point, les compétences du statisticien et celles de l'expert du domaine doivent se compléter.

Il faut insister sur le cas d'un effet «insignifiant mais significatif», c'est à dire sans intérêt pratique mais susceptible d'être interprété comme n'étant pas dû au seul hasard. On ne peut en effet ignorer l'influence que la taille d'un échantillon a sur le niveau de signification: deux résultats numériquement égaux peuvent être, à un même seuil, l'un *significatif* et l'autre *non significatif*, selon que le premier a été calculé sur un grand échantillon, et l'autre sur un petit. De plus, au-delà de certaines tailles d'échantillons, les p-values obtenues sont nécessairement inférieures aux seuils convenus, quels qu'ils soient, donc *significatives*, pour la simple raison que dans une population *réelle*, l'influence (au sens statistique) d'une caractéristique mesurable sur une autre ne peut être absolument nulle[52]. On voit immédiatement qu'une conséquence pernicieuse de ce qui précède est que les recherches à budget élevé, ou disposant des moyens d'interroger un grand nombre de sujets, atteindront immanquablement les standards actuels de publication, quels que soient l'ampleur ou l'intérêt des résultats.

Pour convaincre le lecteur de l'importance cruciale de ces problèmes d'interprétation, et ne pas lui laisser l'impression d'une critique d'arrière-garde fondamentaliste, prenons un exemple. La revue *Scientific American* a publié une discussion concernant le problème de l'effet éventuel qu'aurait une augmentation du salaire minimum sur le taux de chômage. Avant de voter cette loi, le Congrès américain attendait l'avis des experts, qui, en l'occurrence, ne s'accordaient pas. Certains d'entre eux affirmaient que cette mesure pouvait être introduite sans tarder, car son effet sur le taux de chômage était jugé statistiquement «non significatif». D'autres prétendaient au contraire que cet argument était sans valeur, car même un effet statistiquement faible aurait des conséquences catastrophiques sur le niveau de vie de nombreux ménages, certaines petites entreprises n'étant plus en mesure de verser les salaires. L'auteur de l'article[53] se scandalise fort légitimement que l'on puisse ainsi confondre la significativité statistique et l'importance sociale d'un effet. À la lumière de ce qui vient d'être dit plus haut, on ne saurait que l'approuver, sachant qu'une autre étude menée sur un échantillon plus grand aurait sans doute

abouti à la mise en évidence d'un effet «significatif», voire «hautement significatif» qui aurait sans doute mis tout le monde d'accord. L'article ne dit pas ce que le Congrès a finalement décidé, mais les conséquences sociales de la confusion des sens du terme «significatif» peuvent être, le cas échéant, mesurées en unités bien réelles de qualité de vie pour des milliers de personnes disposant de revenus très faibles.

Conclusion

Conscients des nombreux problèmes causés par l'usage qu'ils jugent fantaisiste des tests inférentiels, de nombreux statisticiens et méthodologues ont proposé le rejet pur et simple de ces techniques du champ de la recherche en sciences humaines (tout en proposant de leur substituer d'autres procédures plus adéquates: calcul d'intervalles de confiance, *bootstrap, jackknife*[54], etc.). Cette initiative a rencontré, et rencontre toujours, une extraordinaire résistance de la part du monde de la recherche, à tel point que cette *diabolique persévérance* a motivé quelques tentatives d'explication. Par exemple, celle invoquant la crainte qu'éprouveraient la plupart des chercheurs de ne plus être reconnus par leurs pairs, et de voir ainsi leurs travaux interdits de publication. Connaissant les critères de certains comités d'édition, comme ceux de Melton (cf. *supra*), on ne peut que les comprendre. Cependant, cette seule raison ne suffit pas pour expliquer l'incroyable «*addiction*»[55] de la communauté des chercheurs aux techniques inférentielles. Certes, les explications «psychanalytiques» et «religieuses» de Gigerenzer et Salsburg permettent sans doute d'apporter quelques lumières (et humour) sur ce point, mais quelle que soit la valeur de ces interprétations, le constat surprenant de l'attachement inconditionnel des chercheurs à des modes de raisonnement pour le moins critiquables devra nécessairement déboucher sur une réflexion plus approfondie, analysant le rôle social, culturel et psychologique des pratiques incriminées.

L'étude critique du paradigme «tests inférentiels» en sciences humaines constitue dès lors à nos yeux l'objectif prioritaire d'une recherche pluridisciplinaire dont les retombées sur les décisions politiques, sociales et économiques du siècle à venir seraient, si elle devait être un jour menée à bien, non pas «significatives», mais sans aucun doute capitales.

54. À propos de la technique du *Bootstrap*, voir B. Efron & R. J. Tibshirani, *An introduction to the bootstrap*, New York, Chapman & Hall, 1993. Une illustration de la méthode *Jackknife* se trouve dans Roland Capel, Jean-Pierre Müller & Denis Monod, «Modèles discriminants et classification: l'apport de la méthode "jackknife" à la stabilité des résultats», *Revue Suisse de Psychologie*, vol. 55, n°4, 1996.

55. «*Why are researchers addicted to significance testing?*» (F. Schmidt, *op. cit.* p. 124.)