Essai sur le rôle des tests d'hypothèse en sciences humaines

Rite propitiatoire ou pièce à conviction?

Roland Capel

Denis Monod

Jean-Pierre Müller

1. UTILISATION ABUSIVE DES STATISTIQUES EN SCIENCES HUMAINES : RÉACTION DE L'APA

Les psychologues qui ont écrit au moins un article scientifique au cours de leur carrière connaissent l'APA, *American Psychological Association*, organisation responsable, entre autres, des normes régissant l'écriture de toute publication scientifique en matière de psychologie. Les directives de l'APA s'appliquent aussi bien à la manière de rédiger une bibliographie, de présenter des résultats statistiques sous forme de chiffres et de tableaux, qu'au style propre de la communication scientifique. En bref, le but de l'APA est de créer un langage scientifique commun, dont la nature soit si possible « objective » et permette des confrontations et des échanges d'idées propices à un fructueux développement des disciplines.

En novembre 1995, lors de sa dernière réunion bisannuelle, une sous-commission de l'APA, 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 ». Cette

décision historique marque la reconnaissance officielle de problèmes dénoncés déjà dès les années 1950, liés à la compréhension, à l'interprétation et au mode de communication de certains résultats chiffrés. Précisément, c'est la pratique du test d'hypothèses nulles qui est en question et, plus généralement, c'est l'approche inférentielle elle-même qui semble devoir être rediscutée.

Pour qui s'intéresse de près aux rapports entre la théorie et l'application des techniques statistiques, particulièrement en psychologie, mais aussi en sociologie, biologie, médecine, géographie et bien d'autres disciplines, le problème soulevé par le BSA n'est pas vraiment nouveau. Les premières réflexions méthodologiques critiquant certains usages impropres des statistiques en psychologie comme dans d'autres domaines, apparaissent dès la fin de la seconde guerre mondiale. Une lecture attentive de cette littérature suggère que les problèmes liés au testage d'hypothèses nulles pourraient bien être apparus suite à la constitution, dans les années 1950, d'une théorie syncrétique, amalgame de théories originales d'inspiration différentes. Le présent article a pour but d'éclaircir cette question.

Contentons-nous pour l'instant de mentionner que dès les années 30, Sir R. A. Fisher, agrostatisticien ferru de mathématiques, formalise sa conception du raisonnement inférentiel, suivi quelques années plus tard par les mathématiciens J. Neyman et E. Pearson dont la théorie, très proche du point de vue des résultats, mais dans l'esprit très différente, va heurter violemment celle de leur illustre collègue, à tel point que le jeune professeur Neyman quittera l'Angleterre après une longue et stérile dispute. Ce conflit ne sera jamais réglé, si bien qu'en 1993, le méthodologue critique Gigerenzer déclare que la « logique hybride » héritée de ces deux théories concurrentes, constitue aujourd'hui le principal obstacle au développement de la science psychologique. Sans rechercher les causes profondes de telles pratiques, d'autres auteurs dénoncent l'utilisation rituelle, voire quasi religieuse de certaines techniques statistiques, en particulier le testage d'hypothèses nulles, dans les sciences humaines. Selon eux, les résultats sortis de l'encensoir

statistique n'auraient pas d'autre but que de satisfaire un rituel dont la finalité serait de « sacraliser » (Tukey, 1969) un résultat en lui permettant d'accéder au statut de vérité scientifique. Les titres des articles les plus virulents sont évocateurs : « *La religion statistique telle qu'elle est pratiquée dans les publications de médecine* » (Salsburg, 1985) ou « *Usage, sur-usage ou mauvais usage des tests statistiques en biologie et en écologie* » (Yoccozz, 1991). L'examen de la bibliographie relative à ce débat montre qu'il concerne au moins une cinquantaine d'auteurs et sur Internet même, une discussion sur le thème de l'« *over-reliance on significance* » a vu le jour et des échanges passionnés ont lieu à longueur d'année entre des universitaires de tous niveaux qui dénoncent des pratiques qu'ils jugent pernicieuses et ceux qui en prennent peu à peu conscience, non sans résistances. 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* »...

2. DES CONSTATATIONS ALARMANTES

Maintenant que le problème a été évoqué, analysons ses différents constituants dont les principaux, se résument dans la liste suivante, constituée à partir des points les plus cruciaux, régulièrement soulevés et discutés dans les travaux figurant dans la bibliographie ci-après :

Dans l'application de techniques statistiques :

- Les tests d'hypothèse sont trop souvent[1] mal compris et utilisés par les chercheurs de sciences humaines peu au fait des fondements théoriques des techniques qu'ils utilisent.

1. Pour être précis, le BSA vise toutes les techniques qui font 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.

- De ce fait, les résultats empiriques sont mal interprétés et légitiment des développements théoriques avec lesquels ils n'ont en fait aucune relation.
- L'utilisation très répandue de ces méthodes est révélatrice de l'existence d'une pratique qui relève parfois davantage de la croyance superstitieuse que de l'esprit hypothético-déductif.

Concernant la « culture » statistique :

- Tout chercheur qui se livre à une brève enquête auprès des usagers de statistiques arrive aisément à la conclusion que la plupart des étudiants sortent de l'université sans avoir réellement compris les tenants et aboutissants de la méthode inférentielle.
- Comme nous le verrons plus loin, des sommités respectées de la psychologie dite scientifique, tels Guilford, Nunnally, et Anastasi ont modifié le sens véritable des méthodes qu'ils utilisent, préconisent et enseignent sous une forme « pervertie » (Gigerenzer, 1982).

Aspects liés à la transmission du savoir :

- La mauvaise compréhension ainsi que l'usage inadéquat des méthodes inférentielles découlent d'une inconsistance interne, peut-être congénitale, imputable à l'origine conflictuelle des diverses théories (Goodman, 1993), dont la nature hybride constitue l'héritage « perverti ». Cette conception se perpétue en conservant ses incohérences miraculeusement intactes, d'enseignant à enseignant et de manuel en manuel, sans qu'aucune étude véritablement sérieuse sur l'origine de ces distorsions n'ait encore été entreprise. Mis à part la tentative de Gigerenzer (dont les idées seront exposées plus loin), on en est encore au stade de la dénonciation, et cet état dure depuis quarante ans.
- La déclaration du BSA révèle que le problème de l'« over-reliance on significance » s'est généralisé à toute la littérature psychologique : même les normes édictées par l'APA concernant la publication de résultats expérimentaux sont le fruit de conceptions erronées, et

certains auteurs n'hésitent pas à rejeter la responsabilité de la propagation de la « logique hybride » sur les éditeurs de revues et leurs rewiewers. À cet égard, la constitution, par l'APA elle-même, d'un organe de contrôle, marque un revirement salutaire.

Après ces accusations sévères qu'il conviendra de justifier, une remarque s'impose d'ores et déjà : la liste de critiques exposée ci-dessus ne constitue nullement une condamnation de la statistique inférentielle en soi. Celle-ci demeure, si elle est correctement appliquée et en attendant la généralisation d'alternatives sans doute plus coûteuses[2], un outils fort commode et économique, utile aussi bien à la prise de décision qu'au progrès de la connaissance. Il est important de comprendre que le problème ne réside pas dans la qualité propre de l'outil, mais dans l'usage discutable qui en est fait. La déclaration du BSA représente donc un espoir, non pas celui de voir condamnées une fois de plus « les statistiques » comme elles le sont communément par beaucoup de personnes qui n'ont pas consenti l'effort nécessaire à leur compréhension, mais bien celui de faire retrouver aux techniques inférentielles leur véritable utilité, dans le cadre restreint défini par leurs auteurs.

3. POUR DÉFINIR ET COMPRENDRE LE PROBLÈME : RETOUR AUX FONDEMENTS DE L'INFÉRENCE

Nous allons maintenant tenter de comprendre et décrire l'origine de ces pratiques perverties auxquelles certains chercheurs semblent parfois s'adonner. Pour ce faire, il est nécessaire de fixer le vocabulaire, ce qui obligera le lecteur profane à faire quelques pas en direction du coeur de la théorie de l'inférence statistique. Deux questions surgissent aussitôt : ce coeur théorique est-il unique, et peut-on le définir sans ambiguïté? Avouons d'emblée que les difficultés rencontrées lors de la rédaction des résumés qui vont suivre nous incitent à répondre par la négative. Par

2. Parmi les plus connues : les techniques du bootstrap et du jackknife (voir par exemple Efron & Tibshirani, 1993, ou Lebart & al., 1995, ou encore Capel, Müller & Monod, 1996).

ailleurs, le lecteur se rendra vite compte que toute tentative de « fixer » simplement le vocabulaire en l'espèce se heurte rapidement à l'obstacle de la polysémie des termes utilisés : plus l'on se montre exigeant dans la description des différentes théories, et plus se dévoile la variabilité du sens des mots qui les expriment. Par exemple, il est impossible de définir de manière unique (voir à ce propos : Bunge, 1981) la notion de *probabilité* pour exposer de façon cohérente les idées des trois auteurs exposées plus loin, et il en va de même avec les notions de *significativité*, de *confiance* et de bien d'autres encore.

Malgré ces difficultés, nous allons esquisser trois solutions au problème de l'inférence : celles proposées par Bayes, par Fisher, et celle proposée en collaboration par Neyman et Pearson. Ces théories sont celles que l'on retrouve le plus souvent dans les manuels de statistiques, quoique de manière implicite et déformée, comme nous allons le voir bientôt. Mais avant de nous lancer dans ce délicat exercice, nous ne saurions manquer de mettre en garde le lecteur contre le risque de conserver une image caricaturale, monolithique et dogmatique des idées que nous présentons ici le plus brièvement possible. Voici donc, au risque de fournir de nouvelles variantes à un éventail d'interprétations déjà très étendu, un rapide survol des doctrines inférentielles de trois maîtres reconnus de la statistique moderne.

Tout d'abord, rappelons 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[3], ou *inférence*, peut être défini comme la projection conjecturale et modalisée, sur une population *parente*, d'un savoir acquis sur l'un de ses échan-

3. A ce propos, Peters (1987) donne 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 subtratum de la perception.

tillons. 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 voeu pieux[4]. Il faut ajouter que si d'autres conditions peuvent encore être prises en compte – comme des contraintes de stratification, par exemple – 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, 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.

Le bénéfice obtenu par la transformation d'un savoir à portée limitée en une conjecture dont l'extension est générale se paye par la nécessité de reconnaître, chez T. Bayes, une certaine part d'*a priori*, chez R. A. Fisher, un degré variable de confiance en une hypothèse, ou encore chez J. Neyman et E. Pearson, le risque de prendre une décision erronée. Toutes ces notions s'interprètent en termes de probabilité, quoique fort différemment selon l'auteur.

4. Née du sein des sciences de la terre (Fisher), la théorie des tests d'hypothèse a été progressivement acclimatée aux sciences humaines. Ce repiquage a sans doute créé des problèmes qui ne sont pas sans rapport avec ceux soulevés dans cet article.

4. LE TEST STATISTIQUE COMME SOLUTION AU PROBLÈME DE L'INFÉRENCE.

Deux siècles après les travaux de T. Bayes (1702-1761)[5], dont il sera question plus loin, K. Pearson (1857-1936), véritable « fondateur de la science statistique », selon les mots de Wilks (1941), a proposé, comme une solution formalisée au problème de l'inférence, sa théorie du « test d'ajustement » (*goodness of fit test*) qui s'apprête à fêter son premier centenaire. Ce test consiste en une comparaison entre distributions de fréquences théoriques et observées, 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[6] supposent en effet toutes 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 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.

5. Les illustres précurseurs : Pascal, Fermat, Bernouilli, de Moivre et les contemporains de Bayes, Laplace, Gauss, Poisson, etc. ont tous leurs mérites respectifs. Nous n'avons cité que ceux qui font encore école de nos jours. Le lecteur intéressé par l'histoire des statistiques trouvera de plus amples informations dans l'ouvrage de Peters (1987).

6. Appelée *test d'hypothèse* ou *test de signification* selon qu'on se réfère à un auteur ou à un autre.

L'expérience montre qu'il est vain 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. En effet, une compréhension claire de ce type de raisonnement nécessite des connaissances solides dans des domaines très divers, comme l'algèbre, l'analyse, la logique, le calcul des probabilités et la théorie des jeux. Aussi, dans le seul but d'introduire les mots clefs nécessaires à la compréhension des paragraphes qui suivront, et sans aucune prétention pédagogique, nous dirons, par exemple, qu'étant donné une hypothèse « nulle » d'indépendance supposée vraie dans une population, l'écart à l'indépendance observé dans un échantillon sera attribué, jusqu'à un certain point, au hasard de l'échantillonnage. Dans le cas où cet écart dépasse un certain *seuil critique*, on est amené à *rejeter* l'hypothèse nulle d'indépendance : les seuls aléas de l'échantillonnage n'étant plus déclarés suffisants, dans un tel cas, pour expliquer l'écart observé. 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[7] la probabilité conditionnelle de trouver un écart égal ou supérieur à celui mesuré dans une expérience particulière, entre ce que l'on observe sur un échantillon, et ce que l'on attend théoriquement sous l'hypothèse d'indépendance. Le seuil de signification (choisi par l'expérimentateur) définit 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 dire combien d'individus il faut tirer de la population pour garantir une qualité maximale du test.

7. On l'appelle aussi « p-value » d'après le langage de certains logiciels. Il va en être beaucoup question plus loin.

5. LE TEST DE SIGNIFICATION SELON FISHER : LA CONVICTION PAR L'ÉVIDENCE

Sir R. 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[8]. Remarquons, tout d'abord, qu'il n'a jamais parlé du « niveau de signification » d'un test en référence à la « p value », pas plus qu'il n'a utilisé l'expression « test d'hypothèse ». Par contre, il a effectivement qualifié sa démarche de « test de signification », et si la « p value » est 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. Certes, le choix d'un seuil pour la « p value » est inévitable, mais que cela reste implicite chez Fisher s'explique par le fait qu'il 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.

Fisher, en outre, n'envisage pas de définir une hypothèse alternative (H_1) à l'hypothèse nulle (H_0), le chercheur ne se voit donc contraint à aucune prise décision, ce qui ne l'expose pas au risque de choisir à tort un terme de l'alternative plutôt que l'autre. Les notions d'« erreur » de première et, *a fortiori,* de seconde espèce, n'existent donc pas chez Fisher, de même que celle de *puissance d'un test*.

8. Les affirmations attribuées à Fisher proviennent toutes de Goodman (1993), qui les a tirées de l'ouvrage principal (réédité) de l'auteur anglais (Fisher, 1973).

Fisher n'a jamais voulu entendre les arguments fréquentistes[9] invoquant des probabilités de décisions erronnées calculées sur la base d'expériences répétées[10]. A ce propos, Goodman (1993) 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'« evidence », basée sur une seule expérience, reflétant la crédibilité* post hoc *d'une hypothèse, expérience faite* » (p. 486). 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 : *evidence,* littéralement : *pièce à conviction*) en faveur de l'hypothèse nulle, la mesure de cette confiance étant donnée par la p-value[11].

Pour résumer la démarche « épistémique » de Fisher, on dira qu'elle offre au chercheur un cadre objectif propice à une réévaluation de la relation de confiance qu'il entretient avec ses hypothèses. Plus la p-value tirée d'une expérience est petite (c'est-à-dire plus l'écart entre le résultat observé et celui attendu sous H_0 est grand), moins il accordera de confiance à H_0; inversement, plus la p-value est grande (autrement dit, plus l'écart à l'indépendance est petit) et plus la confiance en H_0 se trouve renforcée. Pour Fisher, le test de signification ne devrait être utilisé que comme un argument de type *fiduciaire* (Reuchlin, 1977), et sa démarche n'a jamais cherché à cacher la portée limitée de ses résultats : si l'hypothèse nulle (X indépendant de Y) devait se trouver fortement mise en doute par l'expérience, cette dernière ne nous apprend rien sur la nature exacte de la relation existant entre les caractéristiques mesurées par X et Y, hormis qu'il est devenu difficile de croire à leur indépendance.

9. Et pourtant la plupart des méthodologues rangent sa théorie dans la famille fréquentiste... ce qui ne contribue pas à dissiper la confusion.

10. Fisher qualifie l'argumentation fréquentiste de « *puérile (...) provenant de mathématiciens sans contact personnel avec les sciences de la nature* ».

11. Les utilisateurs de *packages* statistiques savent bien que cet indice est très généreusement distribué par tous les logiciels statistiques modernes, au contraire du seuil exact de signification qui correspond à la valeur (score) de la variable de décision correspondant au centile 95, ou 99, ou autre). Sur ce point, la conception de ces logiciels incite donc plutôt à un raisonnement fishérien.

Pour ces raisons, la généralisation de résultats empiriques restera toujours du domaine de la *conjecture*. Dans l'optique de Fisher, 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 (*evidence*) en faveur ou contre une hypothèse donnée, mais elles ne peuvent pas, et ne prétendent pas, muer la conjecture en connaissance[12].

6. J. NEYMAN ET E. PEARSON : DE LA CONJECTURE À LA RÈGLE DE DÉCISION

Reprochant à Fisher son interprétation des probabilités, les mathématiciens J. Neyman et E. Pearson[13] proposent dès 1928, d'abandonner les notions gênantes d'« évidence » ou de « conviction », difficiles à définir et par trop subjectives à leurs yeux. Selon la théorie « fréquentiste » de ces auteurs, 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

12. Fisher : "*We may say that a nonsignificant result « confirms » but does not 'establish', the null hypothesis*". Plus prudents que le maître, Campbell & Stanley (1966) écrivent : « Fondamentalement, les résultats expérimentaux [provenant de tests inférentiels] ne prouvent ni ne confirment jamais aucune théorie, ils indiquent plutôt que celle-ci a été testée... et a échappé à l'infirmation ».

13. 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, Peters (1987) 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, le conflit s'envenima, si bien qu'en 1937 Neyman quitte l'Angleterre et ne reviendra en Europe que pour de brefs séjours.

par l'expérimentation. Ce que chercheurs, praticiens et profanes désirent selon eux savoir, c'est s'il convient d'avaler ou non un certain médicament, d'utiliser tel engrais ou tel autre, d'appliquer tel test ou inventaire, 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 du printemps 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 de décision *S* (qui serait, dans notre exemple, dérivée du nombre de décès humains imputables à un agent pathogène provenant d'animaux malades) est définie, et sa valeur, pour l'expérience particulière, 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 appelé *domaine de rejet* de H_0[14]. Dans ce dernier cas, c'est l'hypothèse alternative H_1 qui sera préférée à H_0 et finalement choisie. 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 anglais : *evidence*) en faveur de la vérité 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 α = 5%),* sait

14. Souvent appelée aussi : « région critique ».

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[15].

Le rejet de la théorie de Fisher est ainsi nettement marqué, 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* », 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* »... (Neyman et Pearson, 1933, cités par Goodman). 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 », car si *une* expérience a incité le chercheur à rejeter H_0, rien ne l'assure cependant qu'une autre expérience ne le convaincrait pas du contraire. 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.

On peut constater que ces deux stratégies se réclament de deux positions épistémologiques fort différentes : alors que Fisher tente d'accroître sa connaissance de la réalité sensible en mettant une hypothèse à l'épreuve des faits[16], la règle décisionnelle de Neyman et Pearson ne s'encombre

15. Pour Fisher, cette conception des probabilités restera toujours 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.

pas de préoccupations épistémologiques. Son apport se limite à une sorte de contournement de la connaissance : dès lors qu'elle incite, expérience faite, à *préférer* une hypothèse plutôt qu'une autre, elle ne force pas l'utilisateur à s'interroger sur le degré de vérité de l'un ou l'autre des modèle sous-jacents. En ce sens on pourrait affirmer, en paraphrasant Goodman (*ibid*), que l'oeuvre de Neyman et Pearson – dans la mesure où elle préconise une règle de pur *comportement* – représente « *une tentative de complet rejet du raisonnement inductif ».* Ce propos paraît toutefois excessif, car il ne rend pas grâce à des auteurs qui se trouvent par ailleurs être les créateurs, 150 ans après Bayes, d'une nouvelle théorie de l'estimation statistique basée sur la construction d'*intervalle de confiance*. Dans cette nouvelle optique, les notions de confiance, de probabilité et donc d'inférence reviennent en force, mais transformées au prix d'un glissement sémantique subtil. Ce que Neyman et Pearson appelleront la « confiance en une estimation » n'est d'aucune manière basée sur des considérations d'ordre subjectif, mais uniquement sur l'observation systématique de certaines occurrences d'un événement[17] en cas d'expérimentations répétées. Pour Neyman et Pearson, l'inférence redevient donc possible à la condition d'admettre leur définition fréquentiste de la probabilité.

Pour en revenir à la règle de décision mise au point par Neyman et Pearson, on peut se demander si celle-ci constitue réellement un cadre conceptuel plus « objectif » que celui de Fisher, comme 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 toute entière qui devient critique à ce point. Posée en termes humains, par exemple, la question du choix de

16. En bon agrobiologiste Fisher chercherait, pour reprendre l'exemple précédent, à infirmer ou accréditer la théorie de l'impossibilité de la transmission d'un éventuel germe pathogène de la vache malade à l'homme.

17. Par exemple : la valeur théorique d'un paramètre X se trouve dans un *intervalle de confiance* borné de nombres réels, construit sur la base d'une mesure empirique dudit paramètre. Si l'intervalle est construit selon les règles (défini par exemple avec une probabilité de 95%), l'affirmation « l'intervalle contient la valeur réelle du paramètre X » est vraie pour 95 expériences sur 100.

la valeur 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 du cheptel national »? Qui peut en toute bonne foi prétendre juger de ce délicat problème? Peut-on raisonner « objectivement » lorsqu'on joue, car il s'agit bien d'un *pari*, avec la vie de quelques citoyens inconnus, de notre soeur ou notre enfant, ou celle de la reine d'Angleterre, fût-elle l'unique victime reconnue (une sur quarante millions, c'est presque négligeable en termes fréquentistes...[18]) de la décision de ne pas tuer les « vaches folles »? 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 cadre apparemment strict et rassurant du test fréquentiste ne résiste pas aux considérations en termes de *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 domaine est crucial et marque les limites de leur compétence, car le choix du seuil « critique » (le bien nommé) n'est pas de leur ressort[19], mais bien de celui des commanditaires du test. 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 a 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. Le faire – et il est hélas demandé trop souvent au statisticien si une corrélation est « intéressante » ou « négligeable » – constitue selon nous, une preuve caractérisée de *perte du sens* dans le domaine de la recherche.

18. Que penser de l'avis d'un médecin français interrogé sur les ondes de la télévision, affirmant avec le plus grand sérieux : « Le risque de transmission de la maladie est nul... ou quasi nul ». Ce sont rarement les experts qui paient le prix des décisions erronnées qu'ils ont contribué à faire accepter.

19. Même si cette opération devait emprunter à la technique utilisée ses concepts et ses formalismes.

En résumé, la logique de Neyman et Pearson se présente comme une règle simple, applicable dans un cadre rigide et parfaitement formalisé, ayant l'avantage de permettre une décision *en bonne connaissance de cause*... Mais là réside le coeur du problème, car lorsque cette cause est particulièrement sensible, il est fort probable que *personne* ne connaisse le coût réel de l'erreur, et la confusion des compétences relativement à l'interprétation d'un effet ou du choix d'un seuil critique 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 »?

7. LA RÉPONSE DE BAYES : L'INFÉRENCE PAS À PAS

La comparaison entre les conceptions fishérienne et fréquentiste ne saurait être vraiment éclairante pour notre propos si on ne la situait pas dans la perspective plus générale de l'histoire du raisonnement inductif. Ce texte n'est évidemment pas le lieu d'une aussi monumentale entreprise[20], mais il est un auteur, précurseur de ceux dont il a été question, dont les idées n'ont pas manqué de les influencer. Fisher, aussi bien que Neyman et Pearson ont connu l'oeuvre 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. Quant à Neyman et Pearson, nous ne savons pas s'ils se réfèrent explicitement à Bayes; ce qui frappe cependant, c'est la violente opposition qu'ils manifestent à toute idée d'induction, par leur attachement à une définition purement *fréquentiste* – pragmatique – de la notion de probabilité, ce qui les place

20. Les ouvrages de Peters (1987) et de Boudot (1972) peuvent servir d'introduction à cette vaste problématique. Peters l'aborde sous une angle purement scientifique, alors que Boudot se centre plus exclusivement sur des aspects d'épistémologie formelle largement tirés des oeuvres de Wald, Popper et Carnap, entre autres.

aux antipodes de la conception bayesienne qui veut que cette notion de probabilité soit quasi *essentiellement* attachée aux phénomènes réels. Cette opposition radicale[21], de nature philosophique, entre les approches fréquentistes et baysiennes de la probabilité, a donné naissance à deux courants séparés, totalement cloisonnés de la théorie statistique.

Si la théorie de Bayes a été vigoureusement critiquée par le courant scientiste du XIXe siècle jusqu'à nos jours, c'est sans conteste 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. 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. Dans la logique bayesienne, 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 grandeurs 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 des intervalles de confiance. De plus, le mode et la moyenne peuvent constituer des estimateurs ponctuels et il est aussi possible de calculer des *tests de signification* sur une valeur fixée du

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 des « offenses à la science »...

paramètre. Le théorème de Bayes peut être utilisé en lieu et place des théories de Neyman et Pearson ou de Fisher : étant donné un nombre fini d'hypothèses, et les probabilités *a priori* de chacune d'entre elles, on peut en déduire la probabilité *a posteriori* de celles-ci. Les bayesiens justifient leurs conceptions par le fait qu'il existe une procédure formelle (la leur) pour résoudre les problèmes d'inférence, à la condition d'admettre que l'on puisse représenter toute incertitude par des probabilités. Leur second argument se réfère aux mathématiques : ils affirment être en mesure de définir une « cohérence » à l'aide d'axiomes naturels sur l'incertitude, et qu'alors seule la théorie bayesienne satisfait cette condition de cohérence[22].

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 bayesienne. Chacun d'eux doit décider si cette pièce est équilibrée ou non.

L'adepte de Fisher va d'emblée postuler que la pièce est équilibrée, puis mettre 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 truquée, 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, rien ne met 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 ?

22. On ne s'étonnera pas que les fréquentistes divers prétendent que cette notion n'a été définie que dans ce but.

Sachant 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 frontière de la région critique[23], décider de la sensibilité (puissance) du test et calculer le nombre de jets nécessaires à la mise en oeuvre 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, ils se calculent en termes de *risque délibérément calculé et consenti* : c'est – en principe – le coeur 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 bayesienne quant à elle, s'intéresse davantage à la nature des choses : le statisticien bayesien 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 truquée dans le sens d'une supériorité des occurrences de « Face » sur « Pile ». Il quantifiera cette conviction en définissant une densité de probabilités (en langage bayesien : *priors*) « avant expérience » privilégiant l'évènement « Face ». Une expérience aléatoire de 100 lancers, par exemple, lui donne un résultat de 55 jets « Face » et de 45 jets « Pile ».

23. Comme nous l'avons vu plus haut, ce choix est loin d'être aussi anodin 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 qui fixent machinalement, « par convention » leur seuil de rejet à 5 ou 1%, sans parler de ceux qui oublient de le mentionner... c'est avouer faire ─ par convention ─ bien peu de cas des ses hypothèses de travail... et de sa propre respectabilité.

Sur la base de cette observation, le théorème de Bayes lui permettra d'obtenir une nouvelle distribution de probabilités appelée *posteriors,* favorable à « Face ». Qu'aura ainsi appris notre bayesien? Le *mode* de cette nouvelle distribution lui fournit un estimateur de la fréquence de « Face », et il pourra alors donner la probabilité que cette fréquence soit plus grande que 0.5, par exemple.

À ceux qui leur reprochent que deux personnes ayant des *priors* différentes peuvent obtenir des conclusions contradictoires, les bayesiens répondent tout d'abord qu'il convient de choisir des *priors* « raisonnables » et, d'autre part, que si l'expérience est répétée, alors les distributions *a posteriori* deviennent de plus en plus semblables, quel que soit le choix des *priors*. Toutefois, malgré la valeur de ces réponses, la démarche bayesienne demeure essentiellement suspecte, comme en témoignent par exemple les deux dernières phrases de l'ouvrage que Alfred (1987) consacre 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* ». Fortement imprégnées de scientisme, les idées de Alfred montrent à quel point l'intervention explicite de la subjectivité heurte une conception de la science omnubilée par des critères d'objectivités importés des sciences dures (lire à ce propos : Berger et Berry, 1988). Ce fait, couplé à celui de la difficulté de trouver des logiciels performants, explique sans doute que les bayesiens restent encore peu nombreux, quand bien même leur approche semble plus naturelle que celle des statistiques qualifiées aujourd'hui de « standard ».

8. LES STATISTIQUES STANDARD : DES ORIGINES À LA CONCEPTION MODERNE

Mais 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? Gigerenzer (cf. *supra*) y répond à sa manière, usant aussi bien d'arguments statistiques que de métaphores psychologiques et religieuses. Selon cet auteur, la « logique hybride » 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*, malgré leurs aspects apparemment inconciliables, et ce dès les années cinquante.

Pour vérifier cette affirmation, le méthodologue américain C. Huberty (1993) a examiné le contenu de 28 manuels (textbooks) de statistique édités entre 1910 et 1992, dans le but d'analyser leur contenu, en relation avec les tests d'hypothèses. L'auteur s'intéresse en priorité à retrouver dans ces ouvrages les idées propres à Fisher et à Neyman et Pearson. Il ressort de ces travaux que jusqu'aux années trente, la statistique inférentielle s'appliquait surtout à estimer des paramètres et calculer des distributions de probabilités. Immédiatement avant les travaux de Fisher, les seuls tests connus étaient ceux de K. Pearson (*goodness of fit*); ceux-ci étaient surtout utilisés pour tester la normalité de distributions observées par le biais de l'indice « chi carré ». Dans la période des années 1930 à 1950, on voit apparaître 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 apparaissent : la logique fishérienne se trouve « enrichie » d'un seuil critique, et celle de Neyman et Pearson doit s'accommoder chez certains auteurs de considérations sur la « significativité » d'une p-value, alors que l'alternative H_0 / H_1 disparaît. Du coup, les erreurs de type I et II sont mal définies (Lindquist, 1940) et la notion de puissance passe au second plan pour disparaître quasiment des plans d'expérience en psychologie.

On constate donc que l'occultation – par mélange – des idées originales a débuté très tôt, largement du temps des années professionnellement actives des pères fondateurs. Huberty s'est ensuite intéressé à suivre l'évolution des idées de cinq « auteurs-enseignants au long cours » de la statistique, de 1950 à la période contemporaine (1992). Cette recherche donne lieu à d'intéressantes observations : il découvre de purs adeptes de

Fisher, sans compromis (Ferguson, de 1959 à 1989); d'autres restituent les idées de Neyman et Pearson, mais en amputant leur théorie des points les plus malcommodes ou délicats, comme la définition d'hypothèses alternatives[24], le choix de la puissance du test et de la région critique. D'autres enfin, plus redoutables, font preuve de créativité et réinterprètent 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 l'ouvrage qu'il publie cette année-là, crée, définit et consacre les concepts qui essaimeront rapidement dans toutes les disciplines des sciences humaines pour fonder la logique actuelle – pervertie – dont l'APA cherche aujourd'hui à limiter les dégâts. Même le lecteur le plus indulgent reconnaîtra que Guilford mélange tout : pour ne prendre que quelques exemples : 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. Huberty en conclut que, d'une manière générale, Guilford confond les notions de « signification », « fidélité » et « importance » d'une statistique de test, et que de plus, il ne saisirait pas toujours leur sens statistique de « variables de décision ». En d'autres termes, et en suivant les conclusions de l'exégèse de Huberty, les conceptions infidèles de Guilford semblent permettre 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, « grande » ou « petite »*. Abelson (1995, p. 40) écrit à ce propos : « *Une confusion fréquente consiste à utiliser le niveau de signification comme un indicateur du mérite du résultat* ». 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

24. Etape essentielle puisqu'elle suppose une réflexion à propos de l'importance supposée d'un effet.

de son objectif original, orientant l'instrument vers des fins pour lesquelles il n'était pas le moins du monde destiné. Formulé et utilisé dans le cadre de la logique hybride, *le test statistique semble être devenu un outil magique*[25] *permettant de savoir si ce que l'on observe est intéressant (scientifique?) ou non*.

9. RÉVOLUTION MÉTHODOLOGIQUE ET CONSÉCRATION DE LA « LOGIQUE HYBRIDE »

Si l'on veut argumenter la thèse de l'utilisation magique des techniques inférentielles, il faut tenter de comprendre pourquoi et comment une telle théorie a pu naître, à qui elle profite, et quelles sont les conséquences de son application dans les divers domaines où elle est utilisée.

Selon Gigerenzer, la structure de la logique hybride peut s'expliquer par deux facteurs étroitement imbriqués, l'un psychologique et l'autre lié à ce qu'il appelle la « révolution inférentielle » de l'après-guerre aux USA. Rappelons brièvement le contexte : avant la seconde guerre mondiale, les hommes de science pratiquaient la méthode expérimentale ou l'observation systématique, Wundt, Fechner, Freud, et Piaget en fournissent les exemples les plus célèbres, pour ne citer que des psychologues. Mais cela ne signifie pas que les techniques inférentielles étaient inconnues, puisqu'une preuve de l'existence de Dieu par un raisonnement analogue à un test d'hypothèse a déjà été proposée[26] en 1710. Dans un ouvrage publié en 1897, Fechner aborde divers thèmes méthodologiques, dont quelques techniques inférentielles, mais celles-ci ne consti-

25. Magique, car intégré dans un rituel devenu incontournable : il semble que pour certains auteurs, hélas nombreux, tout résultat dépourvu d'un attribut cryptique du genre (p = .0023 ***) ne peut être que suspect ou « non-scientifique ». Nous reviendrons en détail sur cette question, mais remarquons qu'il y a beaucoup à dire à propos des rituels statistiques, l'anthropologie critique trouverait dans la recherche en science humaines un terrain fertile en observations propres à la sphère du « culte ». Salsburg (1985) a tenté une première approche dans ce sens.

26. Due à J. Arbuthnot (1710), cité par Gigerenzer.

tuent 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%.

Cette véritable révolution méthodologique, particulièrement évidente aux Etats-Unis[27], reflète une profonde mutation de la pratique expérimentale. Rappelons que dès la seconde moitié du XIX[e] siècle, (Wundt) et jusqu'aux travaux de Pavlov, 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 utilitaires, 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%[28]. 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

27. Gigerenzer appuie son analyse sur des articles exclusivement américains, mais le phénomène, encouragé par les directives de l'APA, s'est généralisé au monde entier. La contagion de l'Europe par la révolution inférentielle mériterait d'être étudiée de plus près, mais il semble que la France ait bien résisté, en partie grâce au rayonnement de quelques grands noms de l'analyse de données, tels Benzécri. À titre d'anecdote, citons une communication personnelle de Lebart qui estime, non sans ironie, que le besoin d'inférence est davantage développé chez les anglo-saxons que chez les français. On parle d'ailleurs toujours d'« *analyse de données à la française* » en désignant des techniques descriptives dans lesquelles il n'est pas question de tests. En Suisse, l'influence américaine semble prépondérante, mais une véritable recherche à ce propos mériterait d'être entreprise.

28. Ces analyses sont dues à Danziger, (1990), cité par Gigerenzer, (1993).

.01 car, selon lui, « *les résultats situés entre .05 et .01 prennent la place de résultats de meilleure qualité* » (1962).

Fisher, qui déclarait lui-même que le niveau de signification d'un test ne constituait pas un argument suffisant en matière de validité scientifique, n'aurait jamais cautionné de tels critères. Pas plus d'ailleurs que Neyman et Pearson dont la « règle de comportement inductif » est exempte, comme nous l'avons vu, de toute prétention épistémique.

Pour les psychologues de l'après-guerre, pressés par des contraintes utilitaristes et un climat de compétition du type « publish or perish » de plus en plus contraignant, il fallut rapidement remédier aux faiblesses de ces théories trop modestes en les dépouillant de leurs « *relents agricoles et de leur complexité mathématique* » (Gigerenzer, 1993, p. 323), afin de les unifier en une seule et unique *méthode efficace de production de faits scientifiques.* C'est donc sous une forme hybride, telle qu'elle fut enseignée, entre autres, par Guilford, 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.

10. Les conséquences de la trahison des pères : péché, culpabilité et nouveau dogme

Il est maintenant possible d'imaginer l'état d'esprit des futurs usagers de la *statistique syncrétique* inaugurée, entre autres, par Guilford. On peut supposer que, envieux du prestige dont jouissent largement les sciences exactes et leurs méthodes auprès de la population et du pouvoir, les chercheurs de sciences humaines, avides de « *findings* » ou de « *hard facts* » publiables dans les revues dont nous avons décrit ci-dessus les critères, crurent bon d'importer dans leur propre discipline au moins deux exigences qui leur ont paru absolument nécessaires à la légitimation de leur

activité :

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

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

Nous avons suffisamment montré qu'aucune des théories inférentielle originales ne peut, ni ne prétend atteindre ces objectifs[29]. L'oeuvre statistique de Guilford[30] a, par contre, cette ambition, si bien que les éléments constitutifs de sa théorie syncrétique apparaissent maintenant plus clairement, à la lumière des deux objectifs définis ci-dessus.

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 bayesien (posterior probability). Gigerenzer dissèque sans pitié la rhétorique de Guilford (p. 323), mais il n'est pas question d'entrer ici dans trop de détails. Nous nous contenterons de retenir 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, elle serait donc une mesure de sa véracité. Selon Gigerenzer : « *Entre les mains de Guilford, la p-value, qui spécifie : p (D / H_0),* i.e. *la probabilité d'observer nos données, étant donné H_0, devient miraculeusement p(H_0 / D), probabilité bayesienne* a posteriori

29. Faut-il relever que si ces caractéristiques étaient réellement des critères de qualité scientifique, les travaux de Piaget, Köhler, Pavlov, Skinner et de bien d'autres éminents scientifiques (pour ne pas citer Freud ou Jung) n'auraient jamais été pris au sérieux ?

30. Les critiques adressés à Guilford ne doivent pas être généralisées à son oeuvre de psychologue dont-il n'est pas question ici.

de l'hypothèse, étant donné nos données ». Cette conception va totalement à l'encontre des idées de Fisher qui a toujours refusé l'idée d'*inverse probability*[31].

Plus grave : chez d'autres auteurs, 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. Ainsi, pour Anastasi (1958) : « *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* ». Nunnally (1975), quant à lui, accorde cette valeur « prédictive » au niveau de signification associé au *seuil critique* de Neyman et Pearson : « *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* ». On peut montrer que ces idées sont fausses[32], mais il est piquant de constater que le cadre probabiliste dans lequel elles sont énoncées leur permet, théoriquement, d'être parfois vraies : il n'est en effet pas totalement exclu que deux expériences répétées donnent pratiquement la même p-value. Ces conceptions perverties relèvent-elles de la simple incompréhension ? Gigerenzer, qui n'accuse pas les psychologues d'incompétence, attribue cette déviance à une *distorsion volontaire* dictée par les « exigences de scientificité » évoquées plus haut.

Les conséquences psychologiques de cette « profanation » ne sont pas sans intérêt. Jugeant que l'héritage des pères a été transmis de manière pour le moins tendancieuse – *la connaissance comme conviction* de Fisher se muant, dans la logique hybride, en *connaissance comme vérité* – Gigerenzer fait l'hypothèse que les artisans de cette hérésie n'ont pas enfreint la Loi sans quelques remords... Les modalités de cette trahison

31. Mais on ne saurait toutefois nier que $p(H_o / D)$ est liée à la p-value $p(D / H_o)$ par la formule de Bayes.

32. Selon Bakan (1966), Oakes (1986) et bien d'autres, cités par Gigerenzer, 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* (Abelson, 1995).

sont d'ailleurs plus complexes car, outre la théorie de Fisher, la « théorie hybride » a également intégré, comme nous l'avons vu, la structure rigide de la mécanique décisionnelle de Neyman et Pearson, auteurs dont les idées sont 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 apparemment indestructible[33]. La faute d'une conception aussi monstrueuse va nécessairement retomber sur ses auteurs, désormais habités par des « *Sentiments de malhonnêteté et de culpabilité pour avoir violé les règles* ». Des mécanismes de défense ne tardent pas à s'installer, si bien que : « *La logique hybride tente de résoudre le conflit de ses parents en leur déniant le statut de parents* » (p. 326). À l'appui de cette thèse, Gigerenzer recense trente manuels de statistique modernes et relève 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. Le déni ne s'arrête pas là : à celui des parents, s'ajoute encore le *déni du conflit* entre parents (p. 327) : le seul auteur (Hays, 1963) qui, dans son manuel, les mentionne nommément, présente la théorie de Neyman et Pearson comme un simple progrès par rapport à celle de Fisher. Dans aucun de ces trente ouvrages, il n'est fait mention que la théorie statistique a toujours été un lieu fertile en controverses. Bien au contraire, la culpabilité liée au « meurtre des pères ennemis » a contribué à créer un climat de dogmatisme extrêmement contraignant : inspiré par la nécessité de se conformer aux prétendues normes de « scientificité », le respect absolu du rituel statistique est devenu de nos jours la voie obligée menant au ciel de la publication.

Et 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* » (Salsburg, 1985), que Abelson (1995) appelle aussi « *la poursuite anxieuse de pð.05* ». Or

33. On est surpris de l'inanité des efforts des nombreux statisticiens qui se sont efforcés, dès sa naissance, de l'éradiquer.

il se trouve non seulement que « *peu d'élus trouvent le salut* » (p. 220), mais que cette pratique exige de lourds sacrifices car il s'agit, ni plus ni moins, de renoncer à la satisfaction (qui paraît à première vue assez légitime de la part d'un scientifique...) de simplement comprendre ce qu'il fait. Car, essentiellement syncrétique et incohérente, la « logique hybride » de cette théorie demeure 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 ne manquerait pas de se heurter tôt ou tard aux inconsistances de l'argumentation fallacieuse qu'elle propose. Comme nous allons le voir, il découvrirait au contraire, comme Salsburg ou Carver (1978), la surprenante réalité d'une pratique rituelle, véritable *religion statistique*.

Le chercheur qui n'a pas pu (ou voulu) se livrer à cette critique pourra se contenter de jouir des bienfaits de la méthode : être publié et accéder à des postes élevés. Mais, nous l'avons vu, cette attitude requiert le sacrifice de la compréhension réelle de la démarche, et il devra nécessairement, à un moment ou à un autre, *s'en remettre* – au sens théologique du terme – au statisticien[34]. Cet acte de foi exige de lui 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 [...]* » (1985, p. 220) .

34. Une autre curiosité du monde des statistiques : le marché des statisticiens de recherche se porte bien, notre pays en importe volontiers des Etats-Unis, mais paradoxalement, du moins en Suisse, le métier de statisticien n'existe pas. Personne (ou presque, à force de se l'entendre dire) ne répond « je suis statisticien » lorsqu'on lui demande sa profession. Par contre, dans toute équipe de recherche, on sait désigner « le statisticien ».

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 (du genre : « *il faut fixer d'abord un seuil*, » etc...). Mais Savage (*In:* Barnard & *al.*, 1968) considère leur destin avec pessimisme : « *[Ces personnes] font le meilleur, inspirées par leur bon instinct, mais elles pensent néanmoins qu'elles vivent dans le péché* ». Et c'est dans ce climat diffus de *culpabilité* que s'est installé ce que Gigerenzer appelle le *dogmatisme statistique*.

Ce nouveau dogme, découlant de l'application de la statistique hybride héritée de Guilford, impose une discipline implacable : « *Expérimenter, c'est fixer un seuil au niveau de signification, calculer les résultats, voir si le seuil est atteint, si oui* publish*, sinon* perish[35] ». 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 (Abelson, 1995) que l'observation d'un effet nul peut également être de grand intérêt[36]. De plus, les tests inférentiels ne sont pas du tout limités à la postulation d'effets nuls, il est parfaitement possible de définir H_0 comme « la moyenne du groupe 1 est trois fois plus élevée que celle du groupe 2 », par exemple. Mais cette pratique requiert des hypothèses de travail précises et des plans expérimentaux rigoureux, dans lesquels la puissance du test et le nombre d'individus nécessaires à la mise en évidence d'un effet déterminé, à un seuil choisi, doivent être minutieusement calculés auparavant.

Gigerenzer ne constate pas ce comportement en sciences humaines, il ne rencontre au contraire 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

35. Savage, *op. cit.* p. 326.

36. 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.

être interprétés de manière fantaisiste, etc. Finalement, il qualifie de *comportement mécanique, obsessionnel et compulsif* (p. 327) ce qu'une grande partie de la communauté des chercheurs en psychologie considère comme une méthode scientifique « objective ».

Un tel comportement est sans doute encouragé par l'usage de logiciels statistiques proposant des techniques de plus en plus complexes, mais dont l'emploi se fait de plus en plus simple. Des centaines de tests simultanés peuvent maintenant être effectués en quelques secondes, notamment des analyses multivariées dont la réalisation était impensable voici encore trente ans. On peut supposer qu'un tel « environnement » peut parfois provoquer un certain relâchement dans la rigueur des plans expérimentaux : il est déjà bien assez astreignant d'apprendre à maîtriser la machine, et le but est considéré comme atteint lorsqu'elle est dressée à produire des p-values si possible « significatives ».

En guise d'exemple, Gigerenzer décrit le comportement d'un de ses étudiants qui, ayant calculé deux moyennes empiriques rigoureusement égales, se servit tout de même de son ordinateur, au grand désespoir de son maître, qui commente : « *[...] pour les tester, comme si le simple fait de dire qu'elles sont égales n'était pas suffisamment objectif* » (p. 328). 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*[37] *ont acquis une fonction d'imprimatur, comme si le recours à celles-ci garantissait, À LUI SEUL, la scientificité des résultats obtenus.*

37. On pourrait citer le cas (vécu) de ce chercheur venu consulter à propos des techniques statistiques qui pourraient être utilisées dans une recherche pour laquelle il espèrait collecter des fonds. Nous avons tout d'abord voulu lui expliquer simplement notre point de vue, mais il nous interrompit a*fin 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'étant par ailleurs pas utilisateurs de statistiques. Cette situation fait penser à l'histoire des habits neufs de l'empereur...

11. LES DÉRIVES DE LA SIGNIFICATION : LA « PERTE DU SENS »

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 de nos expériences acquièrent (ou n'acquièrent pas) le *sacrement* (Tuckey, 1969) de la signification, or toute la question est de savoir en quoi le « sens » ainsi dévoilé peut satisfaire notre désir de connaissance. Selon McCloskey (1995), l'introduction de l'usage du mot *signification* en statistique serait due au marquis de Laplace; pour sa part, Fisher s'en servait couramment comme s'il faisait appel à une notion bien connue. Le sens statistique du mot « signification » est donc chez lui clair et bien délimité : il découle directement, et tire exclusivement son sens de la confrontation entre une hypothèse de travail (H_0) et des données expérimentales. Le dictionnaire (Robert) enseigne qu'un résultat est significatif *s'il se prête à l'interprétation* ou, en termes statistiques, qu'il n'est probablement pas dû au seul hasard de l'échantillonnage. Cependant, et nous touchons ici au noeud 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 contente plus de donner à la p-value sa fonction originelle d'indicateur de la possibilité d'une interprétation, mais lui ajoute indûment celle de *support 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'une évaluation[38]. Dans le

38. Dans un article initulé « The insignifiance of statistical significance », McCloskey (1995) appelle le niveau de signification : « *The gold standard...* ».

cadre de la logique hybride, une p-value « significative » constitue *pour 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 non-nullité, *sans plus* (Cardinet, 1985). Cette attitude est qualifiée sans détours de « *fantaisie statistique* » par Pedhazur (1982, p. 24) et condamnée par un nombre impressionnant d'auteurs (Morrison, 1970; Carver, 1978; Meehl, 1978; Schafer, 1993; Thompson; 1993), pour ne citer que les auteurs américains les plus virulents, sans oublier les européens : Corroyer (1994), Cardinet (1985), Pochon (1991), dans un registre moins polémique.

Ayant ainsi élucidé le nouveau rôle donné par les psychologues à la p-value, il devient possible d'intégrer dans cette compréhension le sens des étranges critères de publication des principales revues scientifiques. Ces critères ne peuvent en effet se justifier que par la croyance (erronée) que l'indicateur supposé de la force d'un effet *est aussi l'indicateur de la qualité scientifique de ce résultat*. Ainsi compris et utilisé, le test statistique a perdu sa qualité originelle *d'indicateur d'existence probable d'un effet*, pour être miraculeusement métamorphosé en *instrument de mesure* du degré de l'intérêt scientifique d'une expérience. Le sens des critères de publications en vigueur devient donc limpide, et tout porte à croire que, tombé entre les mains des psychologues, et particulièrement de certains psychométriciens (Guilford, etc...), 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.

Une conséquence remarquable de cette mutation de la fonction du test statistique sur l'interprétation de la notion de variable de décision est que le *domaine de rejet* cher à Neyman et Pearson se mue en domaine du « scientifique » (comprendre : du publiable), alors que le domaine complémentaire, favorable à la conservation de l'hypothèse nulle, est voué à

recueillir des résultats banals et non scientifiques, provenant d'auteurs condamnés à l'anonymat. Cette conception aberrante du test d'hypothèse a donc transformé la notion mathématique de *variable de décision* en une entité curieuse[39] que nous proposons de nommer *variable de publication*, concept qui ne manquerait pas de surprendre dans leur retraite les vénérables auteurs des techniques originales.

Le sens de la perversion opérée sur la technique du test d'hypothèse par les psychométriciens nous semble maintenant bien éclairci : grand producteurs de tests et de mesures, les jeunes chercheurs de cette discipline en pleine expansion avaient prioritairement besoin d'une garantie d'objectivité pour le produit de leurs recherches. Quoi de plus naturel pour eux en effet, (déformation professionnelle?) que de chercher à construire une échelle numérique d'une telle assurance? Le niveau de signification fishérien qui, rappelons-le, s'utilise comme une échelle de la confiance qu'un chercheur peut accorder à une hypothèse en fonction de ses observations, fut pressenti pour jouer ce rôle. Il fut interprété *en une autre mesure de confiance*, celle que l'on peut accorder à des résultats étant donné une hypothèse. D'origine mathématique, complexe, obscure – car suffisamment mal interprétée pour le rester – la p-value fut considérée dès la fin de la seconde guerre mondiale comme l'augure de deux qualités propres à toute mesure psychologique jugée « fiable » : la *fidélité* (reproductibilité) et la *validité* (valeur « réelle » de l'influence d'une variable sur une autre). Et c'est ainsi que le test d'hypothèse devint *en soi* un instrument permettant la mesure de la légitimité et de la qualité des productions psychométriques dont le nombre allait dès lors croître de manière exponentielle.

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

39. Cet objet devrait sans doute trouver une place de choix dans la description du rituel institutionnel régissant le mode de production moderne des faits scientifiques, tel qu'il est défini par Bourdieu (1982).

graphie, etc.

12. QUELQUES EXEMPLES

Le problème central de l'interpétation abusive de la technique du test d'hypothèse est que cet « instrument » ne mesure pas ce qu'il est censé mesurer, et que la p-value ne signifie pas ce qu'on veut lui faire signifier, si bien que les formules magiques trop souvent rencontrées commeþ : « 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 contre-sens. Si l'on veut justifier cette 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.

Outre le sens très général donné par le Robert, (« significatif = qui se prête à l'interprétation »), nous avons montré que « *significatif* » est aussi utilisé comme un indicateur de l'ampleur d'un effet (cf. Guilford), voire de l'*importance* scientifique de cet effet. Ces confusions sont extrêmement courantes et peuvent être relevées dans pratiquement n'importe quel journal ou revue faisant intervenir des statistiques à fin de comparaison. On en trouve des illustrations jusque dans certaines formules publicitaires : lorsqu'un fabricant de médicaments affirme, par exemple, que son produit « améliore significativement le bien-être du malade ».[40]

Comment comprendre la formule rituelle, si fréquente dans les publications scientifiques : « *la différence entre les deux moyennes est significative...* », sans autre spécification? Si l'on tient compte de ce qui

40. Peut-on croire que ce mot est utilisé dans le sens très restreint utilisé en statistique? Certainement pas, car on conçoit mal un malade se contenter d'une modification non nulle de son état, sans prendre en considération *l'ampleur* de cette amélioration. Dans ce cas particulier, « significatif » est employé dans un sens évaluatif : le médicament est sensé avoir un effet puissant, plus « efficace » que les produits concurrents, et il est naturel d'en déduire que l'attribut « significatif » constitue dans ce cas un *argument de vente*.

a été dit plus haut, son auteur peut vouloir dire l'une au moins des trois 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 pragmatique).

Remarquons tout d'abord que ces trois 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 – pensons aux conséquences d'un réchauffement minime de l'atmosphère sur le volume des calottes polaires, donc sur la sécurité des populations vivant près du niveau de la mer. Il se peut aussi qu'un résultat soit très intéressant, « grand » en termes d'effet, et statistiquement non significatif[41] ; toutes les combinaisons sont possibles sans qu'il soit possible de présumer de l'importance pratique et scientifique d'un résultat sur la seule base de son ampleur et de son niveau de signification.

Il faut insister sur le cas d'un effet « insignifiant mais significatif » (sans intérêt pratique mais susceptible d'être interprété comme n'étant pas dû au seul hasard). On ne peut ici ignorer l'effet que la taille de l'effectif d'un échantillon peut avoir sur le niveau de signification : plus l'échantillon est grand et plus la « significativité » d'un résultat augmente (*i.e.* plus la p-value baisse). Au-delà de certaines tailles, les p-values sont inférieures aux seuils convenus pour la raison qu'il n'existe pas d'effets rigoureusement nuls tels que ceux postulés par une H_0 usuelle. Par conséquent, si l'on croit que la p-value est un indicateur de l'importance scientifique d'un effet, on devrait en conclure que les grands échantillons révèlent toujours des effets plus forts (et donc plus intéressants...) que

41. McCloskey (1995) 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 au sujets du groupe témoin.

les petits. Une conséquence miraculeuse de cette interprétation fallacieuse de la p-value 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 (*effect size*) ou l'intérêt des effets observés.[42]

Pour répondre à la question posée tout à l'heure : que déduire d'une différence déclarée « significative »? Au vu de ce qui précède, lorsque le nombre de sujets n'est pas donné, ni le seuil α, la formule « la différence est significative » a, au mieux, un sens sous-déterminé, si elle n'est pas simple poudre aux yeux. Le lecteur charitable, ou respectueux de l'autorité scientifique, peut sans doute admettre que la différence observée est grande, ou scientifiquement intéressante, ou non pas due au seul hasard de l'échantillonnage, mais rien ne lui permet d'en juger lui-même dans les conditions décrites. Le sérieux et la pertinence de ce commentaire ne peuvent en ce cas être garantis que par la seule renommée de l'auteur.

Dans le but de prévenir ce défaut d'information, l'APA exige de ses auteurs le report méticuleux des effectifs, de la valeur de l'effet (*effect size*), des seuils et des niveaux exacts de signification. Le risque de mauvaise interprétation a-t-il pour autant disparu? On peut craindre que non, car la logique hybride s'accommode fort bien de cette apparence de rigueur, comme nous allons le voir dans l'exemple suivant.

Afin d'assurer l'actualité de cette discussion, nous avons « tiré au hasard » un périodique scientifique américain consacré à la psychologie[43] sur le présentoir de notre bibliothèque universitaire. La problématique de l'article est fort complexe et se réfère à des concepts psycho-sociologiques spécialisés, un résumé serait long et inutile pour

42. Certes, il est vrai que les grand échantillons donnent des estimations plus précises, donc des comparaisons plus fiables, mais le nombre de sujets (et la qualité des mesures) est seul garant de cette précision, le détour par la p-value est donc inutile, au contraire de la définition d'un intervalle de confiance qui lui, est informatif.

43. Le nom des auteurs importe peu ici, notre intention n'étant pas polémique. Toutefois nous tenons la référence à disposition de tout lecteur qui nous suspecterait d'avoir trahi ou déformé la pensée des auteurs.

notre propos. La méthode utilisée fait recours principalement à l'analyse de variance simple (ANOVA) et au calcul de corrélations entre différentes mesures de comportements sociaux; nous en reportons une table en remplaçant le nom des variables par des lettres, pour plus de commodité :

TABLEAU 1.

Corrélations entre deux jeux de variables

	X_1	X_2	X_3
Y_1	–.52 ****	–.37*	.90*****
Y_2	–.48**	.50***	.02
Y_3	.82*****	–.29	–.29

Note :

*p < .02, df=40.

**p < .001, df=40.

***p < .0007, df=40.

****p < .0003, df=40.

*****p < .0001, df=40.

Une telle présentation permet effectivement de connaître le nombre de cas étudiés (df+2=42) et la précision des indices statistiques y est exemplaire. Mais que doit comprendre le lecteur? Pour répondre à cette question, il faut d'abord identifier les bases statistiques qui fondent l'argumentation des auteurs. Font-ils référence à des conceptions fishériennes, neyman-pearsoniennes, bayesiennes, ou s'inspirent-ils de la conception syncrétique que Gigerenzer a qualifiée d'*hybride*?

L'analyse du texte révèle que les auteurs exposent bien leurs hypothèses de travail, mais ne les traduisent jamais clairement en hypothèses nulles ou alternatives. La puissance du test reste donc inconnue et aucune interprétation psychologique de l'ampleur attendue d'un effet (d'une variable sur une autre) n'est proposée. En ce qui concerne le seuil critique utilisé dans les tests, le tableau ci-dessus n'en mentionne aucun, alors qu'il spé-

cifie des variations infinitésimales de p. L'existence de seuils est néanmoins implicite puisque le texte juge « significatifs » tous les résultats associés à des niveaux de signification inférieurs à .05. Par ailleurs on observe qu'aucune règle de décision n'est suivie en ce qui concerne l'éventuel rejet d'une hypothèse ou d'une autre. De plus, les variations de la valeur de p sont de toute évidence interprétées comme révélatrices des variations de la force d'un effet, puisque les auteurs déclarent « significatif » l'effet d'une variable X_1 sur une variable Y (ils précisent *F = 4.2, p<.05*), alors que l'effet (mesuré sur le même échantillon) d'une autre variable X_2 sur la même variable Y est déclaré marginal[44] avec l'argument : *F = 3.3 , p<.07* . Comme on voit, tous les ingrédients du micmac logique dénoncé par Gigerenzer semblent être réunis. Tentons maintenant d'en analyser les différents constituants.

L'usage du terme « significatif » renvoie nécessairement à la notion de seuil, implicite chez nos auteurs, comme à la règle décisionnelle définie par Neyman et Pearson. Faut-il pour cette raison considérer leur approche comme inspirée par la théorie de ces derniers? Rappelons que pour ceux-ci le niveau exact de signification (p-value) n'a pas d'intérêt en lui-même, puisque toute p-value inférieure au seuil α, fixé d'avance, signe le rejet de H_0, le risque d'erreur de première espèce restant toujours le même, égal à α. Or, comme nous l'avons vu, nos auteurs ne définissent pas d'hypothèses alternatives, ne fixent pas de seuil, et ne tranchent donc jamais entre deux hypothèses; enfin, ils précisent avec force détails les différentes valeurs de p. Il s'ensuit que nos auteurs ne s'inspirent certainement pas des idées de Neyman et Pearson.

Peut-on alors invoquer le recours à une logique fisherienne? Nous savons que Fisher défendait effectivement l'emploi d'une « échelle de conviction » en faveur ou contre une hypothèse nulle (H_o). Mais peut-on interpréter de cette manière le système d'étoiles qui préfixent dans le tableau les différentes mentions de p? Le texte ne permet aucune illusion à ce sujet, car il interprète systématiquement les différentes valeurs de p

44. « *The test revealed a marginal effect* (...) »

comme différents degrés de l'importance des effets mesurés, et non comme des repères indiquant le degré de confiance que l'on peut accorder à une hypothèse. Il est donc clair que les auteurs assimilent la *mesure de la probabilité* d'observer, dans un échantillon, l'effet d'une variable sur une autre – sous l'hypothèse qu'un tel effet n'existe pas dans la population –, à la *mesure de l'intensité* de l'effet observé sur cet unique échantillon. Une telle confusion, couramment rencontrée et régulièrement dénoncée depuis des décennies, revient à tenir pour équivalentes les propositions « l'effet de X sur Y est faible » et « X a peut-être un effet sur Y ». Ce procédé, qui centre l'attention sur la force de l'effet de X sur Y, préjuge ni plus ni moins de cela même qu'on aimerait pouvoir montrer, à savoir l'existence d'un tel effet .

Il y aurait certes lieu de s'inquiéter des conséquences funestes de cette confusion sur le sens des résultats d'une telle recherche, mais il faut toutefois reconnaître que, si les indices probabilistes sont calculés sur le même échantillon, il est vrai qu'un effet déclaré « fort » sera plus « significatif » qu'un effet considéré comme « faible », et dans ce cas, mais dans ce cas seulement, la comparaison des p-values reflète effectivement l'ordre d'importance des effets. Par contre, c'est au niveau du sens pragmatique du terme « significatif », *le seul qui intéresse vraiment le praticien ou le lecteur non spécialiste* intéressé par la problématique, que le sens induit par ce type d'argument peut être gravement biaisé. Car c'est en général dans la partie finale (discussion) des articles relatant les résultats d'une recherche que leur interprétation peut devenir cruciale : c'est alors que la confusion entre « fort », « important » et « significatif », ou entre « faible », « peu important », et « non significatif » n'est plus tolérable, car si comme le montre notre exemple, le résultat associé à une p-value de .07 est déclaré « non significatif » et, *sur la seule base de l'examen de la p-value*, interprété comme « faible » ou « marginal », *comment un lecteur non spécialiste en statistique pourrait-il ne pas en déduire que la relation en question est* insignifiante *et donc sans intérêt dans son domaine*?

La perversion la plus grave de la pratique hybride apparaît maintenant au

grand jour : jouant sur la polysémie du terme « significatif », elle donne à la p-value un rôle d'index d'importance scientifique, et crée ainsi l'illusion d'une échelle absolue que le lecteur praticien intéressé devrait utiliser pour juger de l'importance et de la validité des résultats qui lui sont présentés. Or, la pseudo-échelle qui lui est présentée n'est pas absolue puisqu'elle ne dépend pas seulement de la force de l'effet mesuré, mais aussi de la taille de l'échantillon. En d'autres termes, et en reprenant notre exemple, un lecteur non statisticien mais intéressé par les problèmes soulevés par les résultats dont il est question croira (!) que tel effet (traduit par un « F » égal à 3.3 par exemple[45]) mesuré sur 42 individus, est sans importance, sous-entendu qu'il n'est pas « significatif », alors que s'il avait été mesuré sur, disons 420 individus, il eût été déclaré « très significatif »... et donc interprété par lui comme très intéressant pour sa pratique.

Afin de convaincre le lecteur de l'importance cruciale de ces problèmes d'interprétation, et afin de ne pas lui laisser l'impression d'une critique d'arrière-garde fondamentaliste, prenons un exemple plus concret. La revue *The 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

45. Les valeurs « F » issues de l'analyse de variance sont en général très difficiles à interpréter en ce qui concerne l'importance de l'effet mesuré (car l'indice « F » n'est pas calibré, c'est à dire rapporté à une référence interne aux données). Lorsque cette mesure s'exprime sous forme de corrélation (indice standardisé), l'interprétation est plus aisée et même un non-spécialiste statisticien est capable, en principe, de distinguer une corrélation « élevée » d'une corrélation « très significative ». À propos de la calibration des statistiques de test et du problème fondamental de l'importance des effets mesurés, voir Corroyer & Rouannet (1994, en particulier p. 610).

verser les salaires. L'auteur de l'article (un professeur d'économie, McCloskey, 1995) 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 », et aurait sans doute mis tout le monde d'accord. L'article ne dit pas ce que le congrès a finalement décidé, mais les possibles conséquences sociales de la confusion des sens du terme « significatif » peuvent être dans ce cas mesurées en unités bien réelles de qualité de vie pour des milliers de personnes disposant de revenus très faibles.

13. CONCLUSION

Cet exposé ne constitue en aucune manière un réquisitoire contre les statistiques *en général* et leur fonction dans la recherche en sciences humaines. Notre discussion a surtout porté sur le crédit et la place que l'on peut accorder aux techniques inférentielles dans le processus de production de savoirs. De nombreux critiques proposent le rejet pur et simple de ces techniques du champ de la recherche en sciences humaines, tout en reconnaissant qu'une telle révolution méthodologique – une véritable conversion pour un grand nombre de chercheurs – n'a que très peu de chances de se réaliser. L'extraordinaire résistance du monde de la recherche à cette réforme a trouvé quelques explications partielles, la plus fréquemment invoquée étant celle de la crainte de ne plus être publié. Mais ce sentiment ne peut expliquer seul l'« *addiction* » aux techniques inférentielles (Schmidt, 1996), et la métaphore religieuse inaugurée par Salsburg, Gigerenzer et d'autres auteurs permet peut-être d'explorer une voie plus éclairante. Le constat surprenant de l'attachement inconditionnel, apparemment imperméables aux critiques les plus brillantes, féroces et définitives, des chercheurs à des pratiques jugées aujourd'hui dépassées par l'APA elle-même, ne doit-il pas nous amener à nous interroger sur le rôle social, culturel et psychologique des techniques incriminées ?

Entre croire et connaître, il faut maintenant tenter de situer l'objectif réel de l'usage singulier des techniques inférentielles. La réponse varie évidemment en fonction de *qui s'en sert* et surtout de *qui en interprète les résultats*. Remarquons qu'en ce qui concerne les utilisateurs, les chercheurs en sciences humaines sont rarement statisticiens et, parallèlement, on note que peu de statisticiens se livrent à la recherche appliquée. Il y a pourtant des exceptions célèbres : Spearman, Thurstone ainsi que d'autres grands psychologues-statisticiens ont amplement démontré leur polyvalence en créant les méthodes multivariées qui s'utilisent très largement de nos jours[46]. Dans cette situation idéale, lorsqu'une seule personne (ou équipe) crée ses propres outils, et interprète les résultats que ceux-ci ont permis d'obtenir, les problèmes de compréhension et de cohérence de l'argumentation ne se posent en principe pas. De tels cas de figure ne sont, hélas, pas courants, particulièrement en sciences humaines où les disciplines formelles, mathématiques et statistiques, ne sont de loin pas les plus investies. Rares sont les étudiants de psychologie, par exemple, qui s'orientent vers la recherche. La plupart d'entre eux – dont une grande majorité est de sexe féminin – aspirent à des activités cliniques à forte composante relationnelle, privilégiant le contact humain et les méthodes d'investigation qualitatives[47]. De ce fait, la majorité des étudiants issus de cette filière manquent certainement de compétences en statistique, c'est du reste un constat quasi unanime parmi les enseignants de ces branches (voir par exemple Antonietti, 1996). Or, les rares licenciés en sciences humaines qui désirent s'orienter vers la recherche découvrent assez rapidement que leur seule curiosité, leur imagination et leur éventuelle intuition psychologique ne suffisent en général pas à satisfaire les critères de publication des revues prestigieuses. Il suffit pour s'en convaincre de consulter les revues spécialisées de psychologie sociale, scolaire ou celles consacrées à l'orientation de carrière, et d'y mesurer le niveau des exigences statistiques et informatiques exigé. C'est donc souvent sans enthousiasme que les jeu-

46. Ce dont les théoriciens-statisticiens ne se souviennent pas très bien lorsqu'ils tentent d'imposer leur langage à des techniques qui ne sont pas de leur conception.

47. Voir à ce propos Dupont & al. 1992, entre autres.

nes chercheurs acceptent de subir le joug des incontournables tableaux, graphiques, techniques multivariées et tests statistiques indispensables à la bonne tenue de toute publication scientifique respectable. Désireux de progresser néanmoins dans cette direction, l'apprenti chercheur fera alliance, de plus ou moins bonne grâce, avec un « statisticien » ou toute personne investie de cette compétence, dans le but d'écrire des articles reconnus comme scientifiques par des *rewiewers* ayant par ailleurs adopté les mêmes habitudes. De tels processus d'auto-légitimation circulaire de compétences « déléguées » se rencontrent sans doute dans d'autres domaines que les sciences humaines, mais en ce qui concerne ces dernières, l'usage qu'elles font de techniques dans lesquelles les utilisateurs ne sont pas réellement compétents pose la question cruciale de « qui légitime quoi? comment? et pour qui? ».

D'après notre expérience (cf. note 37), et à lire la plupart des revues scientifiques de sciences humaines, il fait peu de doute que les dépositaires de la *parole qui sauve* sont les statisticiens. Or, comme nous l'avons vu, le chercheur psychologue (ou sociologue, ou autre) n'est que rarement spécialiste en matière de statistiques, et son drame réside en ce qu'il doit impérativement légitimer son travail en usant de techniques et de formulations complexes, étrangères à sa formation et à ses intérêts réels. Le recours au « prêtre statisticien » évoqué plus haut ne résout pas le problème : au contraire, il consacre une incompétence technique et, plus grave, encourage le transfert de l'expertise en ce qui concerne, d'une part l'interprétation des résultats et, d'autre part, l'importance qu'il convient de leur accorder. Si, en effet, et comme on le constate trop souvent, on laisse au statisticien le soin de juger de l'importance de l'effet de telle dimension psychologique (par exemple) sur telle autre, alors à quoi bon former des chercheurs psychologues, puisque les statisticiens semblent, mystérieusement et pour ainsi dire « par nature », investis de ces compétences, sans qu'aucune formation ne leur ait été nécessaire dans les divers domaines où leur discipline est appliquée?

Pour en venir aux consommateurs du discours et des résultats produits par la recherche, on doit évoquer le « praticien », en principe usager des

résultats de la recherche, et aussi le « profane », Monsieur Tout-le-Monde, baîlleur de fonds de l'industrie de production de savoirs académiques. On contestera difficilement que ces personnes ne maîtrisent pas, ou ignorent tout des techniques statistiques mises en oeuvre dans les travaux qu'elles consultent ou qu'elles financent très indirectement. Ce problème de communication existant entre le chercheur et son lecteur relève presque du tragique : n'est-il pas pathétique de voir en effet le chercheur en sciences humaines s'efforcer de légitimer ses résultats aux yeux de ses pairs, des praticiens et du grand public, en recourant à des techniques qu'il ne maîtrise que partiellement, et provoquer en retour méfiance et incompréhension pour son langage « ésotérique » et son « superbe éloignement des réalités pratiques » ... et se trouver reconnu comme un scientifique pour la raison même qu'on ne le comprend pas? À la fois signe de légitimité et cause de rejet, l'argumentation statistique ne peut donc pas, en l'état, satisfaire simultanément les exigences de la recherche en sciences humaines *et* les attentes du praticien.

Il est maintenant possible de mesurer les conséquences de la situation ambiguë du chercheur qui puise ses données dans le monde réel (le « terrain »), et se sert, pour les traiter, de méthodes provenant du monde abstrait des statistiques (les techniques inférentielles). Le fruit de son travail devrait en principe être accessible aux deux sources qui ont permis son élaboration : les résultats devraient « parler » aussi bien au praticien qu'au statisticien. Mais comment se peut-il que le chercheur, qui n'est en général ni l'un l'autre, assume et prenne en charge un langage qui soit commun à des personnes de formations si différentes?

L'usage généralisé de techniques et de modes de penser hybrides apparaît dès lors comme une fatalité : faut-il s'étonner que dans un univers conceptuel aussi étendu et hétérogène, leur usage ait pris racine si facilement et prospéré sans rencontrer de réelles résistances, bien au contraire. Ces pratiques ne proposent-t-elles pas un *consensus interprétatif* commode, et en même temps, pour celui qui s'en sert, des perspectives professionnelles intéressantes? Doit-on vraiment blâmer les chercheurs de

ne pas vouloir tuer la poule aux oeufs d'or, comme le proposent un nombre croissant de critiques ?

Chacun répondra à ces questions selon ses propres convictions ou intérêts. Remarquons toutefois que le *statu quo* expose les chercheurs trop insouciants de ces problèmes à devoir, tôt ou tard, reconnaître que : soit leur mode d'argumentation, supposé légitimer la nature scientifique de leurs travaux, n'est pas consistant, soit qu'ils ne le comprennent pas vraiment, soit qu'ils ne sont pas véritablement animés par le désir de faire progresser la science. En tout état de cause, une compréhension *réelle* du rôle du raisonnement inductif dans les sciences humaines qui soit commune au statisticien, au chercheur et au praticien n'existe manifestement pas, et en particulier, comme nous l'avons vu, en ce qui concerne l'usage des tests d'hypothèse. Notre voeu est d'avoir montré qu'un « savoir » ainsi légitimé ne peut pas être considéré comme du ressort de la seule rationalité[48], mais qu'il procède aussi, et pour une part non négligeable, peut être profondément ancrée dans la nature humaine, du *besoin de croire* aux vertus propitiatoires d'un rituel nécessairement mystérieux.

BIBLIOGRAPHIE

Note: les ouvrages ou travaux précédés d'un * ne sont pas cités dans l'article, ils constituent des références supplémentaires en rapport avec le sujet.

Abelson, R. (1995). *Statistics as principled argument.* Hillsdale, NJ : Lawrence Erlbaum.

Alfred, B. M. (1987). *Elements of statistics for the life and social sciences.* New York, N. Y. : Springer Verlag.

Anastasi, A. (1958). *Differential psychology (3rd Ed.).* New York : Macmillan.

Antonietti, J.-P. (1996). Au diable les statistiques ? *Psychoscope, 17*, 15-17.

48. D'autant plus qu'il ne saurait en être autrement d'un savoir « inféré ». Que l'on se rappelle le truisme de Hume : « Toute induction ou généralisation ne saurait se justifier par la seule logique ».

Arbuthnot, J. (1710). An argument for Divine Providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions of the Royal Society, 27*, 186-190.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423-437.

Barnard, G. A., Kiefer, J. C., Le Cam, L. M., & Savage, L. J. (1968). Statistical inference. In Watts, D. G. (Ed.), *The Future of Statistics,* (p. 147), New York : Academic Press.

Bayes, T (1763). An essay towards solving a problem in the doctrine of chance. *Philosophical Transactions of the Royal Society, 53*, 370-418.

Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist, 76*, 159-165.

Bourdieu, P. (1982). Les rites commes actes d'institution. *Actes de la Recherche en Sciences Sociales, 43,* 58-63.

Boudot, M. (1972). *Logique inductive et probabilité.* Paris : Armand Colin.

* Brewer, J. K. (1985). Behavioral statistics textbooks : source of myths and misconceptions? *Journal of Educational Statistics, 10*, 252-268.

Bunge, M. (1981). Four concepts of probability. *Applied Mathématical Modelling, 5 (5)*, 306-312.

Campbel, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago : Rand McNally.

Capel, R, Müller, J-P., & Monod, D. (à paraître). Modèles discriminants et classification : l'apport de la méthode « jackknife » à la stabilité des résultats. *Revue Suisse de Psychologie.*

Cardinet, J. (1985). *Du test de signification au coefficient d'assurance.* Neuchâtel *:* Institut Romand de Recherches et de Documentation Pédagogique.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48 (3)*, 378-399.

* Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Corroyer, D., & Rouannet, H. (1994). Note méthodologique sur l'importance des effets et ses indicateurs dans l'analyse des données. *L'année psychologique, 94*, 607-624.

Danziger, K. (1990). *Constructing the subject.* Cambridge : Cambridge University Press.

* Devereux, G. (1967). *De l'angoisse à la méthode.* Paris : Mouton.

Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. *In : Monographs on statistics and applied probability – vol 57 – ed*. Chapman & Hall.

* Fisher, R. A. (1935). *The design of experiments (8th Ed. 1966).* Edinburgh : Oliver & Boyd.

Gigerenzer, G. (1993). 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* (pp. 311-339). Hillsdale : Lawrence Erlbaum.

Goodmann, S. N. (1993). P values, hypothesis tests, and likelihood : implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology, 137*, (5) 485-495.

Guilford, J. P. (1942). *Fundamental statistics in psychology and education (6th. Ed. 1978)*. New York : McGraw-Hill.

Hays, W. L. (1963, 1973, 1981, 1988). *Statistics (for psychologists)*. New York : Holt, Rinehart & Winston.

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) 317-333.

* Lebart, L., Morineau, A., & Piron, M. (1995). *Statistique exploratoire multidimensionnelle*. Paris : Dunod.

Lindquist, E. F. (1938). *Statistical analysis in educational research.* Boston : Houghton Mifflin.

McCloskey, D. N. (1995). The insignificance of statistical significance. *Scientific American. 272* (4) p. 20-21.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks : sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology, 64*, 553-557.

Morrison, D. E., & Henckel, R. E. (Eds.). (1970). *The Significance Test Controversy.* Chicago : Aldine.

Müller, J-P., & Capel, R. (1995). *Statistiques : retour à quelques notions élémentaires.* Texte d'une communication, Séminaire de Statistiques en Sciences Humaines (SSSH), Université de Lausanne.

* Neyman, J. (1957). Inductive behavior as a basic concept of philosophy of science. *International Statistical Review, 25*, 7-22.

* Nunnaly, J. C. (1975). *Introduction to statistics for psychology and education (6th Ed. 1978).* New York : McGraw-Hill.

Oakes, M. (1986). *Statistical inference : a commentary for the social and behavioral sciences.* New York : Wiley.

* Pearson, E. S. (1962). Some thoughts on statistical inference. *Annals of mathematical statistics, 33*, 394-403.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research (2nd Ed.)*. New York, N. Y. : Holt, Rinehart & Winston.

Peters, W. S. (1987). *Counting for something*. New York, N. Y : Springer.

Pochon, L-O. (1991). Statistiques et sciences humaines, notes de travail. *Dossiers de psychologie, 38*. Université de Neuchâtel.

Publication Manual of the American Psychological Association. (4th Ed., 1994). Washington : American Psychological Association.

Reuchlin, M. (1977). Epreuves d'hypothèses nulles et inférence fiduciaire en psychologie. *Journal de Psychologie, 3*, 277-292.

* Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

* Rozenboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57*, 416-428.

Salsburg, D. S. (1985). The religion of statistics as practiced in medical journals. *American Statistician, 39 (3)*, 220-223.

Schafer, W. D. (1993). Interpreting statistical significance and non-significance. *Journal of Experimental Education, 61 (4)*, 383-387.

* Shaver, J. P. (1993). What statistical significance is, and what it is not. *Journal of experimental education, 61 (4)*, 293-316.

* Thomas, D. H. (1978). The awfull truth about statistics in archeology. *American Antiquity, 43*, 231-244.

Thompson, B. (1989). Statistical significance, result importance, and result generalisability : three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development, 22*, 2-6.

Thompson, B. (1993). The use of statistical significance test in research : bootstrap and other alternatives. *Journal of Experimental Education, 61*, 334-349.

Tuckey, J. W. (1969). Analysing data : sanctification or detective work? *American Psychologist*, 24, 83-91.

Wilks, S.S. (1941). Karl Pearson : Founder of the science of statistics. *The Scientific Monthly, 53*, 249-253.

Yoccozz, N. G. (1991). Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America, 72 (2)*, 106-111.