Significativité statistique et signification psychologique de la différence

Daniel Pasquier, psychologue
daniel.pasquier@libertysurf.fr
Avenir & Entreprise
http://www.avenirentreprise.tk/
Laboratoire PRIS, Rouen

Surely, God loves the .06 nearly as much as the .05." (Rosnow & Rosenthal, 1989)

Problématique :

Dans le cadre de la littérature traitant d'études interethniques ou transculturelles, les auteurs sont souvent amenés à prendre en considération des différences de scores obtenus à des tests ou à des questionnaires par des groupes de répondants appartenant à différents pays, à différentes races, à différentes cultures… Par exemple, Schwarzer & Jerusalem (1995) étudient les différences du sentiment d'auto-efficacité dans différents pays. Ces études deviennent susceptibles de prendre un caractère polémique lorsqu'elles concluent à la supériorité des uns sur les autres. L'exemple le plus marquant reste celui de la polémique qui a pu faire rage autour de la prétendue supériorité intellectuelle des américains blancs sur les américains noirs (Jensen, 1969).

Ces études quantitatives procèdent le plus souvent par des comparaisons de moyennes des données empiriques et on note une utilisation quasi systématique des tests d'inférence (test t de Student ; test F de Snedecor), dans la droite ligne des manuels de statistiques pour les sciences humaines (par exemple, Langouet et Porlier, 1981). Pourtant, ce type d'épreuves statistiques n'est absolument pas approprié à l'objectif recherché et il en résulte que la plupart du temps les auteurs tirent des conclusions peu ou prou abusives.

Dans ce papier, on reprendra quelques éléments de critiques de l'utilisation des tests d'inférence pour « montrer des différences significatives », et on tentera de répondre à une double question : pourquoi utilise-t-on ces tests, quelles sont les alternatives ?

1. Critique de l'utilisation des tests d'inférence :

Dans cette première partie, on passe en revue quelques points de critique de l'utilisation des tests d'inférence pour montrer des « différences significatives ». La critique de cette utilisation n'est pas nouvelle et on peut en trouver un développement quasi systématique depuis plusieurs décennies (par exemple Lecoutre, 1996). Insensible à ces critiques, le microcosme des chercheurs en sciences humaines a longtemps continué à appliquer des routines de manière aveugle. Les choses commencent à bouger timidement et on a vu dernièrement la *task force* de l'A.P.A. préconiser l'utilisation conjointe des intervalles de confiance et de l'importance de l'effet (*effect size)*.

Dans le présent cadre, on se contentera d'évoquer les approximations les plus fréquentes des applications des tests d'inférence ainsi que la dépendance de la conclusion de la taille de l'effectif. On verra ensuite ce qu'évaluent ces tests, leurs deux types d'erreur et ce qu'ils n'évaluent pas.

1.1 Conditions de mise en œuvre :

La mise ne œuvre des tests d'inférence les plus usités nécessitent quelques conditions préalables : les mesures doivent être effectuées au niveau d'échelles d'intervalles ; les données observées doivent être issues de populations parentes normales ; et enfin les variances observées doivent être homogènes c'est-à-dire que les distributions parentes des groupes comparés auront la même variance (principe d'homoscédasticité). Dans les faits, l'existence de ces conditions préalables n'est que très rarement vérifiée et on peut dire que ces épreuves sont la plupart du temps utilisées de manière extensive.

1.2 Importance de l'effectif :

Un second point de critique ne se réfère pas directement aux utilisateurs mais provient du fonctionnement même de ces tests : la significativité de la différence varie avec l'effectif. Pour simplifier, prenons l'exemple d'un test de Q.I. standardisé (QI moyen = 100, $\sigma = 15$). Imaginons que deux groupes de sujets, le groupe A et le groupe B ont passé ce test. Le groupe A obtient un QI moyen de 99 et un écart type de 15 points ; le groupe B obtient un QI moyen de 101 et un écart type de 15 points. On se pose la question de savoir si cette différence de 2 points est significative. Si l'effectif des deux groupes est de 10 répondants, la différence n'est pas

significative ; si les deux effectifs passent à 500, la différence devient significative et si les deux effectifs passent à 700 la différence devient très significative. De la même façon que pour les corrélations, pour rendre une différence significative, il suffit d'augmenter le nombre de répondants au test ou au questionnaire, ce qui somme toute peut ouvrir la porte à toutes les manipulations. On peut même aboutir à des propositions qui défient le bon sens. Par exemple, si on faisait passer le test de QI à deux groupes de un million de sujets, il suffirait d'un écart des moyennes de 0,05 point pour déclarer que l'un des deux groupes est plus intelligent que l'autre.

A travers cet exemple, on voit bien que l'importance de l'effectif comme déterminant d'un différence psychologique est parfaitement contre-intuitif et contre-productif : le nombre de sujets qui passent un test ne peut pas influer sur l'existence ou non de différences dans le fonctionnement cognitif de ces individus. Mais s'il n'existe pas de rapport entre le déterminant de la significativité de la différence et le fonctionnement psychologique qu'évaluent donc les tests d'inférence ?

1.3 Ce que les tests d'inférence évaluent et ce qu'ils n'évaluent pas

Langouet et Porlier (1981) résument ainsi la méthode pour comparer une moyenne observée à une norme :

« Le chercheur énoncera une *hypothèse expérimentale* (H_1). Elle peut selon les cas être une hypothèse de différence ou de non-différence.

Le statisticien énonce *toujours* une hypothèse de non-différence : nous n'avons pas apporté la preuve d'une différence ; les différences peuvent s'expliquer par le hasard de l'échantillonnage. C'est l'*hypothèse nulle* (H_0). »

Puis il y a calcul du test statistique (ici le t de Student) suivi de la lecture de la table au seuil P choisi pour N-1 degrés de liberté. Enfin le résultat de cette lecture est interprété de la manière suivante :

« Si |t| calculé > |t| *lu*, l'hypothèse nulle est rejetée. La norme M_0 n'appartient pas à l'intervalle de confiance de la moyenne observée. Au seuil considéré, on a apporté la preuve d'une différence entre *m* et M_0…

« Si |t| calculé < |t| *lu*, l'hypothèse nulle n'est pas rejetée. La norme M_0 appartient à l'intervalle de confiance de la moyenne observée. On n'a pas, au seuil considéré, apporté la

preuve d'une différence entre *m* et M_0. Les différences peuvent résulter du hasard de l'échantillonnage. La différence entre m et M_0 n'est pas significative.[1] »

Dans la comparaison des résultats de deux groupes de sujets, la valeur de la norme attendue est égale à 0. Les conclusions s'énoncent ainsi (cas des grands échantillons indépendants) :

« A P = .05 $z_{calculé} > z_{lu}$

Nous avons apporté la preuve que les deux échantillons n'appartiennent pas à la même population parente.

Nous pouvons affirmé que le groupe B (m' = 10,3) a mieux réussi l'épreuve que le groupe A (m = 9). L'hypothèse nulle est rejetée.

A P = .01 $z_{calculé} < z_{lu}$

Il n'est donc pas possible, avec une chance sur 100 de nous tromper, d'affirmer que les deux groupes n'appartiennent pas à la même population parente. Par conséquent, à ce seuil, l'hypothèse nulle ne peut pas être rejetée.

Dans ce cas, nous dirions que la différence entre les deux moyennes observées est *significative* (P = .05). Elle n'est pas *très significative* (P = .01). La différence n'est affirmée que si l'on prend un risque connu (ici 5 %)… ».

Cette longue citation est essentielle pour appréhender la logique de fonctionnement d'un test d'inférence statistique. L'idée de base en est qu'un échantillon de données est toujours extrait de façon aléatoire d'une population parente dont on ne sait rien par ailleurs. Si l'on pose ce principe comme axiome, **au mieux le test d'inférence indiquera le risque que cet échantillon aléatoire n'appartienne pas à cette population parente.** Un point, c'est tout. Cette procédure traite les données comme aléatoires, même après les avoir recueillies !

Il faut souligner que le risque de décision du rejet de H_0 est choisi de façon parfaitement arbitraire par le chercheur, et on voit bien dans la dernière phrase citée que, selon ce risque choisi, la différence sera déclarée significative ou non significative. C'est uniquement par tradition que les risques de 5% ou de 1% sont le plus souvent considérés, sans autre argument que cette tradition, éventuellement modulée par la prise en compte des erreurs dites de type I ou de type II : « Intuitivement, il semblerait qu'on ait intérêt à abaisser le seuil de rejet de l'hypothèse

[1] Dans ce passage, N est égal au nombre de données recueillies et prises en compte dans les calculs ; *m* est la moyenne observée qu'on compare à la norme M_0 afin de savoir si l'échantillon appartient bien à la population parente ou non à un risque choisi par le chercheur.

nulle, de façon à n'avancer que des hypothèses très fiables. L'inconvénient est que, ce faisant, on augmente les chances de commettre une autre erreur, celle de ne pas rejeter l'hypothèse nulle alors qu'elle est fausse.

On appelle ces deux types d'erreurs respectivement :

* erreur de type I (ou de première espèce) = rejet de l'hypothèse nulle alors qu'elle est vraie ;

* erreur de type II (ou de seconde espèce) = acceptation de l'hypothèse nulle alors qu'elle est fausse.

Ces deux erreurs sont antagonistes : abaisser l'une augmente immédiatement l'autre, et la décision que doit prendre le chercheur est un compromis adapté à la situation. » (Sironi, xxx site) Par exemple, à partir du calcul de la puissance d'un test, c'est-à-dire de la probabilité de ne pas commettre d'erreur de type II, on peut déterminer *a priori* l'effectif de l'échantillon à constituer. Concrètement, on observe une fois de plus que le chercheur dispose d'un panel de moyens lui permettant de conclure à une différence, essentiellement en jouant sur l'effectif de l'échantillon : « Plutôt que de véritables "erreurs", les nombreux abus d'interprétation des tests peuvent être regardés comme des *biais adaptatifs*, que les chercheurs développent spontanément pour tenter d'ajuster les tests à leurs besoins réels. » (Lecoutre, 2005)

Au niveau de l'interprétation des résultats d'un test statistique, on note dans la citation en début de section le glissement de sens qui intervient dans l'énoncé des conclusions. En effet, on passe sans autre forme de procès de l'appartenance à une population parente à la preuve d'une différence. Les deux évènements ne sont pas de même nature. Dire de deux échantillons qu'ils se situent plutôt aux marges d'une population parente virtuelle ne constitue pas une preuve qu'un des deux groupes présente des résultats meilleurs que ceux de l'autre groupe et inversement dire de ces échantillons qu'il se situent dans la moyenne des échantillons de cette population parente n'annule pas la différence éventuellement constatée. Ce n'est pas la fréquence d'occurrence supposée d'un événement[2] ou d'un non-événement qui peut décider ni de son existence, ni de son importance !

Le risque est pris par rapport à une population parente supposée, dont on ne connaît rien et par rapport à une hypothèse somme toute peu crédible : comment argumenter l'existence d'une non-différence ? La question de l'appartenance ou non à une population parente relève du défi au

[2] D'où l'expression d'inférence fréquentiste ; on parle également de procédures d'échantillonnage.

bon sens une fois appliquée aux situations concrètes. Par exemple, on compare des résultats scolaires d'un groupe de garçons et d'un groupe de filles de même âge et d'une même classe. Si le test est considéré comme significatif, on dira que les deux groupes n'appartiennent pas à la même population parente et si le test n'est pas considéré comme significatif, on dira qu'on ne montre pas que les deux groupes n'appartiennent pas à la même population parente. En somme, c'est le résultat du test qui détermine *a posteriori* l'appartenance de l'échantillon à une population parente virtuelle. Ici, c'est la conclusion du test statistique qui déterminerait le sexe ou l'indétermination du sexe des élèves !

En résumé, comme le souligne Reuchlin (1998), « Les critiques ou réserves que l'on vient de lire portent seulement sur l'inférence statistique. Dans ce domaine, elles portent essentiellement… sur l'ambiguïté qui s'attache à l'identification concrète de la population de référence, fondement même de la démarche inductive ; sur le caractère arbitraire des seuils et des effectifs en fonction desquels on décidera cependant qu'un effet est significatif ou non, sur le caractère invraisemblable de l'hypothèse nulle ». Si, au mieux les tests d'inférence ne renseignent que sur l'appartenance ou non à une population parente - information qui peut être, à la limite, intéressante de considérer si l'on souhaite généraliser les conclusions d'une étude -, quelles informations utiles ne donnent-ils pas ?

Quand on s'intéresse à une différence, il serait pertinent de s'interroger sur l'importance de cette différence : l'effet qu'elle traduit est-il important, notable, ou encore négligeable ? Il est également essentiel de prendre en compte la question de la signification psychologique de la différence : par quel processus psychologique la différence se crée-t-elle ? ou encore, à quelles différences de fonctionnement mental renvoient des différences de scores sur une variable ?

Les tests d'inférence ne répondent donc pas de façon adaptée à l'étude statistique des différences observées. Il devient alors légitime de s'interroger sur les raisons de leur utilisation.

2. Et pourtant ils sont utilisés : l'ethnicisation du rapport au savoir

Pour Reuchlin (1998), la place exorbitante prise par les tests d'inférence alors que leurs critiques, jamais réfutées, s'étendent sur plusieurs décennies relève du mystère : « Il y a là un problème étonnant dont l'élucidation pourrait justifier la préparation d'une thèse sur la psychologie du chercheur et la sociologie de la recherche. Comment en est-on arrivé là ? ».

Reuchlin évoque plusieurs types de raisons comme la relation implicite entre le

statisticien et le chercheur, chacun pouvant penser que l'autre détient la justification. Il évoque aussi la peur de l'expert d'un comité de lecture de s'engager : « …il s'engage moins en faisant remarquer qu'une épreuve n'atteint pas le seuil de .05 qu'il le ferait en soutenant par exemple que la recherche qu'il lit, malgré ses éventuelles faiblesses, pourrait ouvrir la voie à une nouvelle façon d'envisager le problème dont elle traite, ce qui justifie sa publication ».

D'autre part, cet auteur remet en cause une conception de la recherche qui privilégie des plans d'expérience basés sur une relation cause-effet ponctuelle plutôt que des approches structurelles et systémiques. Plus largement, d'un point de vue épistémologique, il se demande si on n'assiste pas là « …à un véritable dévoiement de la notion d'objectivité ». La routine formelle des tests de signification protègerait le chercheur de tout risque de subjectivité et « …l'intérêt des résultats paraît ne pouvoir être évalué objectivement que par la lecture d'une table numérique ».

De fait, une démarche scientifique nécessite une prise de risque de la part du chercheur, un engagement, une implication forte dans une démarche tâtonnante, incertaine dont la pertinence des résultats ne peut émerger qu'au fil des expériences réinsérées dans des contextes scientifiques élargis.

A ces explications, Lecoutre (2005) ajoute la nécessité pour le chercheur d'utiliser des tests d'inférence pour optimiser ses chances de publication et de reconnaissance par ses pairs : « La conduite d'une majorité de chercheurs apparaît dictée par un résultat de test de signification usuel significatif qui est souvent le seul critère retenu.

Les arguments invoqués dans ce cas sont le plus souvent d'ordre conjoncturel : un résultat significatif fait incontestablement partie des normes de présentation des travaux dans la communauté scientifique actuelle des chercheurs expérimentalistes, qui se disent souvent contraints de "se plier" à cet usage ».

Mais qu'est-ce qui pousse ainsi les chercheurs à se plier à un tel usage ? C'est à ce niveau que nous proposerons une explication relevant d'une forme de processus d'ethnicisation du rapport au savoir. D'une façon générale, l'ethnicisation est un « … processus de construction de frontières et de désignation… », processus produit par le jeu des rapports sociaux et des relations sociales historiquement constitués et transformés (De Rudder et al., 2000). Bien évidemment, ces frontières ne concernent pas à proprement parlé l'origine ethnique[3], mais plutôt le sentiment de

[3] « L'ethnie … désigne un groupement humain qui se réclame d'une même origine, possède un nom… et une tradition culturelle commune. » (Ferréol et al., 2003).

l'appartenance à une culture scientifique commune. Notre proposition relève plus de l'ordre du symbolique que de la simple ethnologie car comme le souligne l'auteur, « La classe sociale, elle-même, est susceptible de racisation (les ouvriers du XIX^{e} siècle, héréditairement stupides, immoraux, etc.) ou d'ethnicisation ».

D'autre part, ce processus de prise en compte et de pointage de la différence entre *Alter* et *Ego*, n'est pas neutre dans la mesure où il est « … en même temps et indissolublement un processus de classement sur une échelle qui ordonne des statuts sociaux, économiques, politiques… », le groupe dominant s'identifiant moins comme un groupe ethnique que comme un groupe de référence universelle (De Rudder, 1995).

Et en matière de recherche, si on admet qu'à l'heure actuelle le groupe qui se pose comme référence universelle se situe aux Etats-Unis, alors on comprend mieux pourquoi les chercheurs qui souhaitent parvenir à une reconnaissance internationale se glissent dans le moule des normes anglo-saxonnes. La rédaction des articles se fait en langue anglaise pour publication dans des revues nord-américaines, les normes bibliographiques utilisées seront celles préconisées par l'A.P.A. et les tests d'inférence sont quasi-systématiquement mis en œuvre : « [...] Le chercheur qui présente un résultat significatif, tel le vainqueur d'une épreuve sportive, fait souvent l'objet de suspicion et doit satisfaire à un contrôle avant que son résultat soit homologué (publié). C'est le rôle des éditeurs et rapporteurs des revues aux réserves desquels l'expérimentateur est souvent confronté. Malheureusement, la norme est si bien établie que ces réserves portent presque exclusivement sur la validité des tests (A-t-on utilisé le bon test ? Les conditions d'application sont-elles satisfaisantes ? Etc.) et non sur leur pertinence (Le test répond-il vraiment à la question posée ?) » (Poitevineau, 1998).

Le chercheur qui veut réussir par le biais de la publication de ses articles doit se conformer aux normes du groupe dominant la communauté scientifique internationale. Ce mouvement d'allégeance (Gangloff, 1997) du chercheur en mal de réussite, lui permettant d'appartenir au *Nous* du monde académique officiel, ne supporte pas d'exception même si la qualité scientifique de son travail doit en pâtir et l'amener à la pratique de routines inadaptées au problème qu'il se pose.

Et plus subtilement encore, en adoptant ces routines il peut également s'imprégner de l'idéologie nord-américaine néo-libérale sociologiquement, économiquement, militairement, politiquement dominante et tellement destructrice des environnements naturels, culturels et

humains. Il peut se laisser aller plus ou moins consciemment à justifier cette domination impérialiste en apportant la preuve illusoire de la supériorité des *W.A.P.*[4] sur les autres groupes humains.

Les exemples ne manquent pas des dégâts sociaux et humains justifiés par la naturalisation des différences et le cautionnement pseudo-scientifique de toutes les formes de racismes et de discriminations (colonisation, esclavagisme, génocides, ethnocides, holocauste…), et moins spectaculaires, des litanies des justifications des positions sociales par l'idéologie des dons (Léon, 1980 ; Pasquier, 1992), ou pis encore par l'appartenance à une race donnée[5], certains auteurs n'hésitant pas à naturaliser la courbe de Gauss (Herrnstein & Murray, 1994).

Pourtant, les alternatives existent, mieux appropriées à la description des différences.

3. Les alternatives

Parmi les alternatives aux tests de signification fréquentistes, on peut citer les intervalles de confiance basés sur la fidélité des épreuves, la prise en compte de la grandeur de l'effet, l'approche bayésienne, la prise en compte de réseaux nomologiques dans le cadre d'une démarche scientifique structurale… Dans le cadre restreint de cet article, nous évoquerons les trois derniers points.

3.1 Grandeur de l'effet: d de Cohen

Cohen (1988) a proposé un indicateur de l'importance d'une différence qui se calcule comme $d = (M_1 - M_2) / \sigma$. Il propose les limites suivantes : de 0,20 à 0,50 pour un petit effet ; de 0,50 à 0,80 pour un effet moyen ; 0,80 ou plus pour un effet important. Dans l'exemple présenté au début de cet article, la différence brute de 2 points de QI se traduit par $d = 0,13$. Pour prétendre à un petit effet, il faudrait au moins 3 points d'écart entre les moyennes des groupes. Cet indicateur descriptif présente l'avantage d'orienter le chercheur vers cette question de la grandeur de l'effet : à partir de quelle différence peut-on conclure à l'existence d'un effet qui présente une signification psychologique. Les limites proposées par Cohen n'offre pas un caractère coercitif ;

4 Blanc, anglo-saxon, protestant.

5 « C'est ainsi que les noirs américains ont un Q.I. inférieur en moyenne de 15 points par rapport au reste de la population, ce qui est considérable, puisque la moyenne générale est de 100. Et l'on sait, depuis les travaux fondamentaux d'Arthur Jensen et d'autres, que cet écart de 15 points est essentiellement imputable à des différences génétiques. Cette conclusion n'a rien de surprenant, dès lors que le Q.I. a un coefficient d'héritabilité de 80 %, au niveau individuel. » (De Lesquen, 1996)

seulement indicatives, elles demandent à être redéfinies au cas par cas. L'approche bayésienne accentue cette nécessité de poser la question de la signification psychologique de la différence.

3.2 Approche bayésienne

Dans cette approche, on probabilise les valeurs possibles d'une différence dans une population parente. Pour les aspects théoriques et les procédures de calcul, on se reportera à l'ouvrage de Lecoutre (1996).

Par rapport à la pratique des tests fréquentistes, la pratique bayésienne impose au chercheur le choix de l'écart brut entre les moyennes - par exemple lorsque les objets quantifiés ont d'emblée une signification -, ou de la grandeur de l'effet dont il veut tester la probabilité. Il peut se contenter de prendre en considération des critères psychométriques comme un effet seuil (on teste par exemple la probabilité d'un petit effet au sens de Cohen), ou bien l'erreur de mesure dérivée de la fidélité d'un test, ou encore l'écart inter-classe de scores standardisés…

Plus stimulante, cette approche offre au chercheur l'opportunité de se poser la question de déterminer *a priori* l'importance de l'effet qui reflète une véritable différenciation du processus mental. Dans l'exemple fictif d'une différence de δ points de QI, il doit argumenter le choix de la valeur de δ qu'il considère comme l'effet d'une réelle différence de fonctionnement cognitif. Il doit alors choisir entre une différence de degré comme par exemple la vitesse de traitement de l'information, et une différence de nature comme la mise en œuvre de schèmes de résolution plus ou moins bien adaptés à la résolution du problème posé, ou encore une différence culturelle liée à la spécificité du contenu des items ou à l'attitude d'un groupe dans des situations d'évaluation… Sur ce dernier point on peut citer entre mille autres exemples les difficultés de l'anthropologue des sensations à se décentrer de son modèle pour comprendre le fonctionnement sensoriel d'individus d'autres cultures : «…les sens sont construits par la culture tout autant que par la biologie. Un exemple typique est le fait que le nombre de sens reconnus change selon les cultures. » ou encore si nous avons pour habitude de séparer le fonctionnement des sens, en d'autres cultures il y a synesthésie : « Les Dogons du Mali soutiennent que l'odeur et le son sont tous deux apportés par des vibrations et croient ainsi qu'ils peuvent « entendre » des odeurs » (Howes, 2003). La question de fond n'est plus le rejet de H_0 mais celle de l'interprétation d'une différence, de sa signification psychologique avant même le calcul statistique.

Sur ce dernier plan, le comportement des deux types de tests d'inférence diffèrent et peuvent ne pas conduire aux mêmes conclusions. Le tableau 1 présente une simulation de l'influence de la taille de l'effectif. Rappelons (colonnes 1 et 2 du tableau) qu'à différence égale, le rejet de H_0 est fonction de la grandeur de l'échantillon. Le test bayésien[6] fonctionne différemment (colonnes 3 à 5 du tableau). A grandeur égale de l'effet (d = 0,13), l'augmentation de l'effectif va dans le sens d'une levée de l'incertitude et confirme dans notre exemple la probabilité de l'existence d'un effet négligeable.

N	**P**	**petit effet (d > 0,20)**	**effet moyen (d > 0,50)**	**effet important (d > 0,80)**
10	0,76	?	?	NON
500	0,03*	?	NON	NON
700	0,01**	?	NON	NON
1 000 000	**	NON	NON	NON

N : effectif. P : valeur de la probabilité fréquentiste. * : significatif à P < 0,05. ** : très significatif à P < 0,01. petit effet (d > 0,20) : probabilité bayésienne > 90% d'avoir un petit effet. effet moyen (d > 0,50) : probabilité bayésienne > 90% d'avoir un effet moyen. effet important (d > 0,80) **:** probabilité bayésienne > 90% d'avoir un petit effet.

Tableau 1 : comparaison de l'influence des effectifs dans les inférences fréquentiste et bayésienne.

Pour obtenir un résultat significatif, ce n'est donc pas la grandeur de l'effectif qu'il convient d'augmenter, mais bel et bien l'écart des moyennes. Le tableau 2 illustre la divergence des conclusions auxquelles on peut aboutir. A effectifs égaux ($n_1 = n_2 = 500$), l'augmentation de l'écart des différences conduit toujours au rejet de H_0. *A contrario*, pour parvenir à l'existence d'un effet important, il faut que la différence des moyennes soit suffisante.

m_1 ; m_2	**δ**	**d**	**p**	**petit effet (d > 0,20)**	**effet moyen (d > 0,50)**	**effet important (d > 0,80)**
104 ; 99	5	0,33	**	OUI	NON	NON
108 ; 99	9	0,60	**	OUI	OUI	NON
113 ; 99	14	0,93	**	OUI	OUI	OUI

m_1 ; m_2 : moyennes des deux groupes. δ : différence des moyennes. **d** : d de Cohen.

Tableau 2 : comparaison de l'influence de l'écart des différences dans les inférences fréquentiste et bayésienne.

[6] Calculs effectués avec le gratuiciel *LeBayésien* de Lecoutre & Poitevineau (1996).

Ici, il faut une différence de 14 points, soit pratiquement un écart type d'écart pour conclure à un effet notable important, ce qui ne froisse pas le bon sens. De ce fait, la manipulation des conclusions par les effectifs devient impossible. Même si on peut reprocher à cette approche le recours à une population parente virtuelle, on ne peut lui reprocher de parvenir à des conclusions contre-intuitives.

Au-delà de la statistique enfin, on évoquera brièvement la question de l'alternative aux approches cause-effet ponctuelles.

3.3 Réseau nomologique et démarche scientifique structurale

Claude Bernard a écrit en 1865 « …quand on réunit des éléments physiologiques, on voit apparaître des propriétés qui n'étaient pas appréciables dans les éléments séparés … Leur union exprime plus que l'addition de leurs propriétés séparées. »

Cette citation invite à dépasser les modèles d'analyse des associations simples et par conséquence les méthodes statistiques qui les accompagnent. Comme le souligne Reuchlin (2003), « Il semble bien que le fonctionnement du vivant non seulement ignore les principes de simplicité et d'économie mais même en prend le contre-pied. L'interdépendance des processus, la redondance de plusieurs processus vicariants capables d'assumer la même fonction, la mise en jeu de niveaux de contrôle hiérarchisés témoignent, entre autres faits, que la "cause" d'un comportement n'est pas à rechercher dans une variable ou quelques variables que l'expérience et l'analyse statistique pourraient dissocier et dont elles seraient en mesure d'évaluer indépendamment la présence et le poids. Elle réside plus vraisemblablement dans le fonctionnement d'une certaine structure. »

En d'autres termes, l'étude des différences nécessiterait l'inscription de chaque variable dans un réseau nomologique avec d'autres variables, les variations apportées à l'une d'entre elles se répercutant sur le système d'ensemble. Ce type d'approche structurale fait appel à d'autres statistiques, essentiellement descriptives et basées sur les équations structurelles (Roussel et al., 2002), l'hypothèse nulle étant rejetée, une fois pour toutes.

Pour conclure :

L'utilisation des tests d'inférence a été maintes fois critiquée et ces critiques n'ont jamais été réfutées. Leur utilisation perdure contre vents et marées et cette aberration peut s'expliquer, entre autres facteurs, par le comportement d'allégeance des chercheurs à l'égard des normes édictées par la communauté scientifique mondialement dominante. Cette allégeance opère dans un processus d'ethnicisation d'un rapport au savoir qui trace la frontière symbolique, mais aux effets d'exclusion, de non-prise en considération et de non-reconnaissance redoutables et redoutés, entre chercheurs allégeants et chercheurs rebelles.

Cette allégeance sans conscience conduit certains à utiliser, éventuellement « malgré eux », des routines statistiques inappropriées à la nature du problème posé, ce qui en soi n'a guère d'importance. Cette attitude de soumission à la loi du plus fort[7] peut prendre une tournure contraire à toute déontologie quand des conclusions posées abusivement servent à disqualifier et à discriminer socialement, voire à modifier ou détruire physiquement - tentation toujours présente de l'eugénisme -, les groupes fragilisés comme les femmes, les ouvriers ou les migrants.

L'usage de ces routines aux effets potentiellement pervers lorsqu'elles s'appliquent à l'étude des différences entre les êtres humains doit être abandonné au profit d'approches expérimentales et de modèles statistiques mieux adaptés à ce type d'études.

[7] On ne peut s'empêcher de penser ici aux travaux de Milgram (1974) relatifs à la soumission librement consentie.

Bibliographie :

Bernard, C. (1865). Introduction à la médecine expérimentale. Paris : Baillière.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

De Lesquen, H. (1996). *André Langaney, prix Lyssenko en 1996 pour sa contribution à l'étude des races humaines.* http://www.clubdelhorloge.fr/lyssenko_1996langaney.php

De Rudder, V. (1995). Ethnicisation. In P-J Simon (Ed.). *Pluriel recherches. Vocabulaire historique et critique des relations inter-ethniques. Cahier n°3.* Paris : L'Harmattan.

De Rudder V., Poiret C. & Vourc'h F. (2000). *L'inégalité raciste. L'universalité républicaine à l'épreuve.* Paris : P.U.F.

Ferréol G. & Jucquois G., (Eds) (2003). *Dictionnaire de l'altérité et des relations interculturelles.* Paris : Armand Colin.

Gangloff, B. (1997). Les implications théoriques d'un choix d'items : de la norme d'internalité à la norme d'allégeance. *Pratiques psychologiques, 2*, 99-106

Herrnstein, R-J. & Murray, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life.* New York: The Free Press.

Jensen, A. (1969). How Much Can We Boost I.Q. and Scholastic Achievement. *Harvard Educational Review, 39*, 1-123.

Howes D. (2003). Evaluation sensorielle et diversité culturelle. *Psychologie française, 48-4,*117-125.

Langouet, G. & et Porlier, J-C. (1981). *Mesure et statistique en milieu éducatif.* Paris : E.S.F.

Lecoutre, B. (1996). *Traitement statistique des données expérimentales.* Montreuil : C.I.S.I.A.

Lecoutre, B. & Poitevineau, J. (1996). *LeBayésien*. Montreuil : C.I.S.I.A.

Lecoutre, B.(2005) *1. Comportement des utilisateurs des méthodes statistiques.* http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris.htm

Léon, A. (1980). *Introduction à l'histoire des faits éducatifs.* Paris : P.U.F.

Milgram, S. (1974). *Soumission à l'autorité*. Paris : Calmann-Lévy.

Pasquier, D. (1992). *Agir pour la réussite scolaire.* Paris : Hachette.

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

Reuchlin, M. (1998). Signification statistique et signification scientifique. *La revue française de psychiatrie et de psychologie médicale,18,* 9-104.

Reuchlin, M. (2003). Contributions à l'histoire des méthodes statistiques employées en psychologie. 2. Les plans d'expérience et l'analyse de variance : Ronald Aymler Fisher (1890-1962). *Psychologie et Histoire*, 2003, vol. 4, 31-60.

Roussel, P., Durrieu, F., Campoy, E. & El Akremi, A. (2002). *Méthodes d'équations structurelles : recherche et applications en gestion.* Paris : Economica.

Schwarzer, R., & Jerusalem, M. (1995). Generalized Self-Efficacy scale. In J. Weinman, S. Wright, & M. Johnston, *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35-37). Windsor, UK: NFER-NELSON.