Avec la prolifération des logiciels statistiques presse-bouton et leurs avalanches de p-values, une réflexion sur les tests statistiques est plus que jamais bienvenue. Souhaitons que le travail de J. Poitevineau (qu'on va lire ci-après) fasse des émules, et que la revue Mathématiques et Sciences humaines puisse se faire prochainement l'écho d'une réflexion critique à plusieurs voix sur les tests statistiques leur histoire, leur enseignement et leurs usages.

Henry Rouanet

L'USAGE DES TESTS STATISTIQUES PAR LES CHERCHEURS EN PSYCHOLOGIE ASPECTS NORMATIF, DESCRIPTIF ET PRESCRIPTIF¹

Jacques POITEVINEAU²

RÉSUMÉ — *La pratique des tests statistiques par les chercheurs en psychologie est abordée selon trois aspects. Du point de vue normatif les tests apparaissent inadaptés les principales critiques sont présentées. Du point de vue descriptif, l'examen des manuels statistiques, les ré-analyses d'articles publiés et les expériences auprès de chercheurs montrent l'existence de nombreux abus d'utilisation. Enfin, du point de vue prescriptif, des solutions de rechange sont envisagées, en particulier les méthodes bayésiennes qui apparaissent particulièrement prometteuses.*

MOTS CLÉS — Inférence statistique, Raisonnement inductif, Tests de signification, Intervalles de confiance, Méthodes bayésiennes, Méthodologie.

SUMMARY — *The use of significance tests by psychologists, from normative, descriptive and prescriptive viewpoints. At a normative level, the significance tests appear to be ill suited and the main criticisms are reported. At a descriptive level, both examination of statistical textbooks, re-analyses of published papers and experiments about the use of significance tests by psychologists clearly reveals many misuses. At a prescriptive level, alternative solutions are considered, especially the Bayesian methods, which appear to be especially attractive.*

KEY WORDS — Statistical inference, Inductive reasoning, Significance tests, Confidence intervals, Bayesian methods, Methodology.

---

¹ Article reçu le 29 octobre 2003, révisé le 3 décembre 2003, accepté le 5 décembre 2003

Cet article constitue un résumé d'une thèse intitulée « Méthodologie de l'analyse des données expérimentales étude de la pratique des tests statistiques chez les chercheurs en psychologie, approches normative, prescriptive et descriptive », réalisée sous la direction de B. Lecoutre et présentée à l'université de Rouen en 1998. Cette thèse est disponible sur demande auprès de l'auteur.

² LAM/LCPE, UMR7604 CNRS - Université Paris VI - Ministère de la Culture, 11 rue de Lourmel, 75015 Paris, poitevin@ccr.jussieu.fr.

1. INTRODUCTION

Les tests statistiques, sous la forme des tests de signification, ont envahi la plupart des domaines scientifiques. Un «résultat significatif»³ à l'un des seuils fatidiques 0.05 ou 0.01 fait maintenant incontestablement partie des normes en vigueur dans la communauté scientifique□il est même devenu quasiment obligatoire dès lors qu'il s'agit de publier dans une revue expérimentale. Le test de signification apparaît même souvent comme un label de «scientificité».

Un observateur extérieur pourrait donc croire que le test de signification est un outil parfaitement adapté à la méthodologie de la recherche expérimentale. Nous pensons qu'il n'en est rien et que l'apparente adaptation de la pratique du test de signification est illusoire, ce qui aboutit à une situation de fait pour le moins étrange, voire paradoxale. La recherche expérimentale peut en effet être rapprochée d'un jeu dans lequel seuls les résultats significatifs sont gagnants, alors que les résultats non significatifs sont en principe des constats d'ignorance, donc des échecs. Mais les règles de ce jeu sont inadaptées et par suite constamment transgressées, ce qui se traduit par d'innombrables abus d'interprétation et entraînent des distorsions considérables, notamment dans la conduite des expériences, dans la sélection des résultats publiés et dans la présentation de ceux-ci.

Nous étudierons ici la pratique des tests statistiques chez les chercheurs en psychologie en visant à montrer l'inadaptation fondamentale de l'usage des tests de signification traditionnels. Cette étude se structure selon trois approches complémentaires□ normative, descriptive et prescriptive. L'approche normative consistera essentiellement en un rappel des théories historiquement sous-jacentes aux tests de signification et en un exposé des principales critiques méthodologiques dont a fait l'objet l'usage de ces tests. L'approche descriptive étudiera la manière dont les tests de signification sont effectivement utilisés dans la pratique et permettra de mettre en évidence à la fois des distorsions et des mésusages. Enfin l'approche prescriptive, s'appuyant sur les aspects normatifs et descriptifs, visera une réflexion méthodologique sur le rôle des méthodes d'inférence statistique dans l'analyse des données expérimentales dont nous concluons que les méthodes bayésiennes devraient être recommandées comme la solution de rechange la plus prometteuse.

Au préalable, il est nécessaire de préciser qu'un test de signification traite d'une certaine *hypothèse statistique*, c'est-à-dire d'un énoncé spécifiant la loi de distribution d'une population statistique (souvent il s'agit d'un énoncé relatif à la valeur d'un ou plusieurs paramètres, par exemple que la moyenne d'une variable pour une population donnée est égale à telle valeur précise). Cette hypothèse statistique est dérivée de l'*hypothèse de recherche* à laquelle le chercheur s'intéresse en premier lieu, et qui concerne sa discipline, par des voies plus ou moins directes. S'il est évident que cette dérivation est cruciale pour la validité d'une inférence portant sur l'hypothèse de recherche, il est tout aussi évident que les théories des tests statistiques ne concernent que l'inférence portant sur l'hypothèse statistique et que la question de la pertinence de la dérivation qui y mène n'est pas du ressort de la statistique. Aussi, nous laisserons de côté cette question qui n'est en rien spécifique de l'utilisation du test de signification et se pose quelle que soit la méthode d'inférence utilisée.

---

³ Nous utiliserons quelques néologismes□ significativité (pour évoquer la propriété d'un résultat d'être «significatif» ou non), mésusage (pour signifier un usage abusif), prescriptif, réanalyse, répliquabilité.

2. ASPECTS NORMATIFS

La discussion de l'inadaptation d'une pratique nécessite d'abord la connaissance de l'outil utilisé et donc des théories qui lui sont sous-jacentes. Étant donné notre propos, ces théories seront abordées essentiellement selon leurs implications méthodologiques, et non selon un point de vue purement mathématique.

Les tests couramment utilisés reposent sur deux théories principales, celle de Fisher et celle de Neyman et Pearson, que l'on peut considérer comme rivales dans la mesure où elles sont logiquement incompatibles et où de vives controverses ont constamment opposé leurs auteurs.

2.1 LA THÉORIE DE FISHER

La théorie de Fisher s'inscrit dans la lignée des travaux de Yule et de Karl Pearson (c'est en 1900 que ce dernier présente le «test du χ^2 » [Pearson, 1900]). Fisher cristallise et développe les notions et les pratiques jusque-là en vigueur, tout en leur conférant le statut de méthode d'inférence incontournable. En 1925 paraît la première édition de *Statistical Methods for Research Workers* [Fisher, 1990(a)], puis en 1935 celle de *The Design of Experiments* [Fisher, 1990(b)] qui connaissent un succès considérable (14 éditions pour le premier ouvrage, 8 pour le second).

Fisher s'adresse principalement aux chercheurs scientifiques et voit le test comme un moyen d'apprendre à partir des données expérimentales [Fisher, 1990(b)].

Une seule hypothèse est mise à l'épreuve. Elle est appelée «l'hypothèse nulle», dans le sens de *to be nullified*, c'est-à-dire à réfuter, «lulle» ne signifiant donc pas nécessairement que la valeur testée pour le paramètre soit zéro, même si c'est un cas fréquent. En général l'hypothèse nulle sera la négation de l'hypothèse à laquelle le chercheur s'intéresse réellement [Fisher, 1990(b)].

Le résultat de la procédure de test est soit le rejet de «l'hypothèse nulle», soit la suspension du jugement, l'hypothèse nulle ne pouvant être acceptée [Fisher, 1990(b)]. Par conséquent, il n'y a qu'une seule possibilité d'erreur : rejeter l'hypothèse nulle alors qu'elle est vraie.

Les données de l'échantillon, supposé provenir aléatoirement d'une population hypothétique, infinie et inconnue, étant recueillies, on calcule pour la statistique de test la probabilité d'obtenir un résultat au moins aussi extrême que celui observé en supposant vraie l'hypothèse nulle. Cette probabilité conditionnelle, habituellement notée p , est le niveau (ou seuil) observé de signification (*significance level*). Elle est, selon Fisher, caractéristique des données observées et indicatrice du degré de réfutation de l'hypothèse nulle auquel ces données conduisent [Fisher, 1990(a)]. Ce seuil de signification, caractéristique d'une expérience *unique*, ne doit pas être confondu avec un taux d'erreur que l'on obtiendrait par échantillonnage répété dans *une même* population, même s'ils peuvent coïncider dans certains cas [Fisher, 1990(c)]. Si p est jugé suffisamment faible, on considère qu'on a réussi à montrer la fausseté de l'hypothèse nulle, on la rejette et le résultat est déclaré *significatif*. Si p est trop élevé, on suspend le jugement et le résultat est déclaré *non significatif*. Pour juger de p , Fisher fait souvent référence dans ses premiers écrits à une valeur de 0.05 [Fisher, 1990(a)], mais il en vient, ultérieurement, à rejeter la notion d'un seuil absolu, intangible [Fisher, 1990(c)].

2.2 LA THÉORIE DE NEYMAN ET PEARSON

Leurs articles fondateurs paraissent en 1928 et 1933. Bien que s'inspirant des travaux de Fisher, leur approche participe davantage d'une théorie statistique de la décision et ils ne la présentent pas, au contraire de Fisher, comme le fondement d'une nouvelle logique inductive.

Pour eux il s'agit bien moins de pouvoir conclure sur la véracité d'une hypothèse que d'adopter, à son égard, un *comportement* rationnel selon un critère de contrôle à *long terme* des erreurs de décision [Neyman, Pearson, 1933(a)]. En envisageant une répétition à l'infini et dans des circonstances identiques, la procédure doit être telle qu'on contrôle le nombre de décisions erronées.

La population parente, dont l'échantillon est censé être tiré au hasard, est envisagée comme l'ensemble des répétitions à l'infini de l'expérience réalisée.

Comme chez Fisher il s'agit de mettre à l'épreuve une hypothèse particulière, qu'ils notent H_0 et appellent «*hypothèse testée*» (et non «*hypothèse nulle*», terme qui leur est étranger). Cette hypothèse est celle à laquelle on s'intéresse particulièrement, souvent celle qui semble la plus probable *a priori* [Neyman, Pearson, 1928, 1933(a)]. Par ailleurs, H_0 n'est plus suffisante pour construire le test, il convient de prendre également en compte les hypothèses alternatives admissibles (le plus souvent une seule) [Neyman, Pearson, 1928]. Autrement dit, il est vain de vouloir tester une hypothèse si on ne lui en reconnaît pas de concurrente.

La procédure de test doit permettre de décider entre soit rejeter H_0 , soit l'accepter. Il y a donc deux possibilités d'erreurs quand une décision est prise. L'erreur de première espèce ou de type I consiste à rejeter H_0 alors qu'elle est vraie ; la probabilité conditionnelle correspondante est appelée risque de première espèce, souvent noté α . L'erreur de deuxième espèce ou de type II consiste à accepter H_0 alors qu'une hypothèse alternative est vraie ; le risque de deuxième espèce est la probabilité conditionnelle correspondante, souvent notée β , et $1-\beta$ (probabilité de choisir H_1 alors qu'elle est vraie) est appelée la *puissance* du test (par rapport à H_1). Toutes choses égales par ailleurs, les risques α et β varient en sens inverse.

Le risque de première espèce α est choisi *a priori*. Selon Neyman et Pearson, il existe une dissymétrie fondamentale entre les deux risques d'erreurs [Neyman, Pearson, 1933(b)]. C'est le risque contre lequel il est le plus important de se prémunir, du point de vue de l'utilisateur, qui devra être posé comme étant, par définition, le risque de première espèce, et par conséquent l'hypothèse correspondante deviendra l'hypothèse testée [Neyman, 1950]. Quant au choix de la valeur précise à fixer pour α , les auteurs considèrent qu'il est du seul ressort du chercheur.

Les hypothèses possibles permettent de déterminer, étant donné le modèle d'échantillonnage, le sous-espace de l'espace des échantillons, appelé la région critique, qui conduira à rejeter H_0 si l'échantillon observé s'y situe. Si, au contraire, l'échantillon n'appartient pas à cette région, on acceptera H_0 . Cette région critique est calculée de façon que, α étant fixé, le risque de deuxième espèce β soit minimum.

Au terme de l'expérience, on applique la règle ainsi définie et on conclut selon la région à laquelle l'échantillon (en pratique, la statistique de test) appartient. La région critique est construite tout à fait indépendamment des données ; seules sont nécessaires les connaissances de α , N et des valeurs spécifiées par les hypothèses. Les données ne

servent que pour conclure, et peu importe à quel point précis de l'espace des échantillons elles correspondent, la seule question étant de savoir si ce point appartient à la région critique ou non.

Bien qu'il soit habituel de classer les deux approches précédentes dans le cadre d'une conception *fréquentiste* de la probabilité, c'est-à-dire où celle-ci est exclusivement définie comme la limite d'une fréquence, ce qualificatif s'applique surtout à la théorie de Neyman et Pearson. Rouanet parle de «*fréquentiste modéré*» pour Fisher et de «*fréquentistes radicaux*» pour Neyman et Pearson [Rouanet, 1991]⁴.

Le plus souvent les chercheurs, et même les statisticiens, ne distinguent pas clairement les deux théories mais utilisent divers amalgames. Ceux-ci ont souvent été dénoncés. Gigerenzer évoque à ce propos, en un sens péjoratif, la «*méthode hybride de l'inférence statistique*» [Gigerenzer, 1993]. Par exemple, les chercheurs se basent le plus souvent sur le seuil observé p (Fisher), mais font aussi appel, occasionnellement, à la notion de risque de deuxième espèce et/ou de puissance (Neyman et Pearson). Ou bien les deux risques, de première et deuxième espèces, sont mis en avant (Neyman et Pearson) mais le choix de H_0 se fait par négation de l'hypothèse d'intérêt, et sans considération de l'importance relative des risques d'erreur (approche fishérienne). Ces amalgames sont d'ailleurs révélés par la terminologie et les notations. Il est maintenant passé dans l'usage de parler «*l'hypothèse nulle*» (Fisher) et de la noter « H_0 » (Neyman et Pearson), alors même que ces auteurs envisageaient de façon opposée le rôle de cette hypothèse. Le seuil observé (Fisher) est parfois qualifié de «*risque*» (Neyman et Pearson).

2.3 CRITIQUES DE CES DEUX THÉORIES DES TESTS

Les tests de signification ont fait l'objet de nombreuses critiques méthodologiques qui mettent en cause l'adaptation de cet outil à l'usage auquel on le destine. Ces critiques sont apparues très tôt, en fait, dès leur «*naissance*» [Boring, 1919]. Dans le domaine de la psychologie et de la sociologie, elles se sont surtout multipliées dans les années soixante. Certaines s'appliquent seulement à l'une des théories, mais la plupart s'appliquent aux deux⁵. Nous n'évoquons ici que les principales en nous limitant aux références les plus anciennes (cf. [Lecoutre B., Lecoutre M.-P., Poitevineau, 2001] ou [Lecoutre B., Poitevineau, 2000] pour une bibliographie plus détaillée).

Répétition

Contrairement à ce à quoi renvoie la théorie de Neyman et Pearson, dans la recherche scientifique, même si le critère de reproductibilité est fondamental, une expérience n'est pratiquement jamais répétée à l'identique un très grand nombre de fois [Camilleri, 1962].

---

⁴ En fait, Fisher a une conception de la probabilité qui l'apparente beaucoup plus aux bayésiens qu'à Neyman et Pearson (voir, en particulier, ses travaux sur la «*probabilité fiduciaire*»). Ce n'est que parce qu'il considère qu'en général on ne possède pas d'information fiable extérieure aux données permettant de fixer raisonnablement les probabilités *a priori* qu'il rejette catégoriquement l'utilisation du théorème de Bayes comme outil pour obtenir les probabilités *a posteriori* (cf., par exemple, [Fisher, 1990(b)] p. 7, [Fisher, 1990(c)] p. 10).

⁵ Puisque c'est de la *pratique* des chercheurs qu'il est question, ne sont considérés ici que les tests *usuels* typiquement le test qu'un effet est égal à zéro, au moyen du t de Student, du F de l'analyse de variance, ou le test de l'indépendance ou de l'homogénéité de deux variables au moyen du χ^2 .

Arbitraire

Dans la théorie de Fisher, puisqu'il n'existe pas d'erreur de type II, donc pas de risque β à minimiser, aucune justification formelle n'existe quant au choix de la région de rejet de l'hypothèse nulle. Dans les deux théories, le choix de la valeur du seuil de référence, qui marque la frontière entre significatif et non significatif, est aussi éminemment arbitraire ou subjectif et introduire des fonctions de coûts des erreurs ne fait que déplacer le problème [Rozeboom, 1960].

Et encore, dans un même cadre théorique, pour un même seuil et à partir des mêmes données, deux chercheurs pourront ou non obtenir un résultat significatif selon qu'ils choisissent de mettre en œuvre un test bilatéral ou unilatéral.

Degré de désaccord

Fisher considère que des hypothèses rejetées au même seuil observé p présentent un même degré d'évidence contre elles. Mais le seuil p , à lui seul, est insuffisant pour «mesurer» le degré de désaccord avec l'hypothèse testée. $p = 0.01$ n'évoque certainement pas la même chose pour un chercheur selon que l'effectif est de 10 ou de 1000. Et Lindley a montré que dans le cadre du test d'une hypothèse simple H_0 , il est toujours possible de trouver une taille d'échantillon N telle que H_0 soit rejetée à un seuil de 5 %, par exemple, et, simultanément, d'obtenir une probabilité *a posteriori* pour H_0 de 95 %, alors même que la probabilité (non nulle) *a priori* de H_0 est aussi faible qu'on veut [Lindley, 1957].

Événements inobservés

La définition du seuil observé p (ou du risque β) repose sur la prise en compte d'événements qui ne se sont *pas* produits puisqu'on calcule la probabilité d'observer un événement *au moins* aussi extrême que celui observé. Baser une inférence sur quelque chose qui ne s'est pas produit peut paraître pour le moins étrange, sinon non pertinent [Jeffreys, 1961].

Non naturel

Le raisonnement manque de naturel, il est contre-intuitif. On calcule une probabilité correspondant à l'événement observé, *conditionnellement à une hypothèse* sur le paramètre d'intérêt, $Pr(\text{Données}|\text{Hypothèse})$, alors qu'il semblerait plus naturel, à l'inverse, comme on le fait dans le cadre bayésien, de calculer une probabilité sur les valeurs possibles du paramètre, *conditionnellement à l'événement observé*, $Pr(\text{Hypothèse}|\text{Données})$ [Carver, 1978].

Décision ou jugement

Dans la recherche scientifique, le test d'une hypothèse ne correspond pas à un processus de décision où il faut choisir, de façon irréversible, entre deux actions. Une hypothèse scientifique n'est jamais définitivement acceptée ; elle peut toujours être remise en cause en fonction de découvertes ultérieures. C'est ce que Fisher, le premier, a toujours violemment objecté à Neyman et Pearson [Fisher, 1990(a)]. Cet argument a souvent été repris, mais certains auteurs, comme Bakan englobent l'approche fishérienne dans la critique, considérant que la méthode de Neyman et Pearson n'a fait que révéler ce qui était implicite chez Fisher [Bakan, 1966].

Biais de publication

Le fait qu'un résultat significatif soit un critère important dans la sélection des articles soumis à publication risque d'entraîner un biais, dans la mesure où un résultat non significatif a peu de chances d'être publié [Sterling, 1959].

Hypothèse inutile

Comme Tukey le remarque, une hypothèse nulle ponctuelle (attribuant au paramètre une valeur précise et non un intervalle) est pratiquement toujours fautive, ne serait-ce qu'à une quantité infinitésimale près □ "It is foolish to ask 'Are the effects of A and B different?' They are always different for some decimal place" [Tukey, 1991]. Aussi, pour autant qu'on y mette le prix en termes de taille de l'échantillon, tout résultat sera significatif et l'information apportée par le test est donc quasi-nulle [Berkson, 1942], [Meehl, 1967]. C'est ainsi que Carver qualifie l'hypothèse nulle d'«*homme de paille*» [Carver, 1978].

Intensité de l'effet

Le test ne dit rien quant à l'intensité de l'effet parent [O'Brien, Shapiro, 1968]. Un résultat significatif n'est qu'une indication de l'*existence* de l'effet supposé □ un résultat non significatif un constat d'ignorance [Rouanet, Lépine, Pelnard-Considère, 1976].

Paradoxe fondamental

Que le plus souvent l'hypothèse d'intérêt soit identifiée à l'hypothèse alternative conduit au paradoxe fondamental mis en avant par Berkson dès 1942 [Berkson, 1942]. Un résultat significatif n'est pas une information positive *en faveur* de l'hypothèse alternative, il n'est qu'une évidence *contre* l'hypothèse nulle ; alors que le but de l'inférence scientifique est au contraire d'apporter des éléments en faveur d'une hypothèse, d'une théorie.

D'une manière voisine, Carver évoque une corruption de la méthode scientifique [Carver, 1978]. Cette dernière consiste, selon lui, une fois les données recueillies, à examiner si celles-ci sont compatibles avec l'hypothèse de recherche, puis, quand elles ne le sont pas, à envisager d'autres hypothèses. Au contraire, avec le test de signification, l'hypothèse nulle est mise en avant et est généralement la seule testée, quelque intérêt scientifique qu'elle ait. En conséquence, l'hypothèse de recherche ne sera même pas examinée si le test est non significatif, alors même qu'elle pourrait présenter une bonne compatibilité avec les données.

3. ASPECTS DESCRIPTIFS

L'intérêt d'un outil se mesure finalement par la façon dont il est utilisé. Les tests de signification peuvent ainsi être appréhendés à travers les interprétations auxquelles ils donnent lieu, que ce soit dans des ouvrages statistiques à l'usage des chercheurs ou dans des articles scientifiques publiés. Le comportement des chercheurs dans des situations d'analyse statistique de données a également été étudié expérimentalement. Après avoir montré que les tests donnent lieu à beaucoup d'erreurs et d'abus d'interprétation, nous nous interrogerons finalement sur les raisons de la persistance de leur utilisation.

3.1 LES INTERPRÉTATIONS DES TESTS □ ERREURS ET ABUS

Non seulement les tests ont été sévèrement critiqués, mais encore ils donnent très fréquemment lieu à des erreurs d'interprétation, des abus (bien entendu, sont exclues ici les erreurs d'utilisation comme l'usage d'une procédure non appropriée). Quatre grands types d'erreurs peuvent être distingués.

- L'erreur la plus typique est de considérer la probabilité p (seuil observé) ou $\square$ comme une probabilité concernant l'hypothèse et non plus conditionnelle à celle-ci. Elle se présente sous diverses variantes □

La probabilité que l'hypothèse nulle soit vraie est p (ou $\square$)

La probabilité que l'hypothèse alternative soit vraie est $1-p$ (ou $1-\square$)

La probabilité que les résultats soient dus au seul hasard est p (ou $\square$)

Les deux dernières formes sont respectivement appelées “Valid Research Hypothesis Fantasy” et “Odds-Against-Chance Fantasy” par Carver [Carvert, 1978].

- Une deuxième erreur est de considérer $1-p$ (ou $1-\square$) comme la probabilité de reproduire le résultat observé (“Replicability or Reliability Fantasy”, dans les termes de Carver). Cette reproductibilité peut concerner soit la valeur même de l'effet, soit, plus couramment, seulement la significativité du résultat.
- La troisième erreur, très souvent dénoncée, est de confondre significativité statistique et significativité substantielle (*substantive significance*), c'est-à-dire considérer que plus un résultat est significatif, plus il est scientifiquement intéressant, et/ou que plus l'effet correspondant dans la population parente est grand (*substantive importance*) [Selvin, 1957].
- La quatrième erreur est de conclure à la véracité de l'hypothèse nulle en cas de résultat non significatif. Ceci n'est une erreur, à strictement parler, que dans le cadre fishérien (encore faut-il, dans le cadre neymanien, que la puissance soit suffisante). Par exemple, dans Bassock *et al.*⁶ on peut lire □ “Subject’s performance was not affected by differences in the size of the assigned and the receiving sets [$\square^2(1) \square \square .08$, n.s.], so we combined the results of subjects...”.

Harcum [1990] donne d'autres exemples d'acceptation « □ désinvolte □ de l'hypothèse nulle, dont certains ont été publiés dans des revues de renom comme le *Journal of Experimental Psychology*. Ainsi, Smith et Ramunas, sur la base de résultats non significatifs, soutiennent qu'ils ont éliminé des effets de champ visuel (alors même que seulement six sujets ont été examinés)⁷.

3.2 EXAMEN DE QUELQUES OUVRAGES STATISTIQUES DE RÉFÉRENCE

Si une pratique inadaptée peut venir de l'outil lui-même, elle peut aussi résulter de l'utilisation d'un mode d'emploi inadéquat. Il est donc intéressant d'étudier les manuels de statistique appliquée s'adressant directement au psychologue. La présentation des tests de signification devrait en principe y être une adaptation fidèle des théories.

Cet examen ne pouvant être exhaustif nous n'avons retenu que des ouvrages « □ incontournables □, dont le succès en librairie a été évident (à l'échelle des ouvrages

⁶ Bassok M., Wu L.-L., Olseth K. L., "Judging a book by its cover: Interpretative effects of content on problem-solving transfer", *Memory & Cognition* 23, 1995, p. 354-367.

⁷ Smith M.C., Ramunas S., "Elimination of visual field effects by use of a single report technique: Evidence for order-of-report artifact", *Journal of Experimental Psychology* 87, 1971, p. 23-28.

universitaires) (cf. [Faverge, 1975], [Hays, 1973], [Kirk, 1982], [Reuchlin, 1976], [Siegel, 1956], [Winer, 1971]).

À la lecture, il apparaît de nettes distorsions entre les deux théories des tests de signification et leur présentation dans ces six ouvrages. Des ambiguïtés, sinon des abus, apparaissent le plus souvent dans les interprétations des résultats de plus ces interprétations révèlent des divergences entre la présentation des principes et l'exposé de la pratique à partir des exemples.

Sur le seul critère de la présentation des tests, l'approche de Neyman et Pearson l'emporte sur celle de Fisher qui n'est jamais présentée en tant que telle et n'apparaît que comme élément d'un amalgame des deux théories. Les six livres se répartissent en

- un neyman-pearsonien pur (Hays), qui se distingue nettement par sa rigueur, sa conformité à la théorie statistique,
- deux neyman-pearsoniens plus «approximatifs» (Winer et Kirk),
- et trois «hybrides» (Siegel, Faverge et Reuchlin).

3.3 RÉ-ANALYSES STATISTIQUES D'ARTICLES PUBLIÉS

L'objectif premier des études de ce type est d'ordre méthodologique : mettre en évidence certains problèmes soulevés par l'usage des tests de signification, et en tirer éventuellement les conséquences pour une meilleure pratique.

Ce type d'étude est inauguré en 1962 par Cohen dont le but est de déterminer si la puissance des tests utilisés par les psychologues est suffisante pour que l'hypothèse nulle ait de bonnes chances d'être rejetée quand elle est fautive (il s'inscrit explicitement dans la lignée de Neyman et Pearson) [Cohen, 1962]. À cette fin, il met au point une grille pour la classification des effets et analyse les articles parus en 1960 dans le *Journal of Abnormal and Social Psychology*. Chacun des 70 articles retenus est alors caractérisé par trois puissances moyennes des tests relatifs aux hypothèses, puissances calculées en supposant l'effet vrai «faible», «moyen» ou «fort». En moyenne, Cohen trouve, pour ces puissances moyennes 0.20 (effet supposé faible), 0.50 (effet supposé moyen) et 0.83 (effet supposé fort). Sauf à supposer un effet vrai fort, en moyenne la puissance est donc peu élevée. Dans ces conditions un chercheur a assez peu de chances de rejeter l'hypothèse nulle alors même qu'elle est fautive, et cependant presque tous ces articles font état de résultats significatifs. Plutôt que supposer des effets parents forts, Cohen pense qu'il existe un biais dans la sélection des articles acceptés pour publication.

Depuis, ce type d'étude a souvent été répété (cf. par exemple [Clark-Carter, 1997]). Les résultats et les conclusions sont similaires à ceux de l'article de 1962, ce qui montre l'absence d'évolution des comportements des chercheurs.

Haase *et al.* ont opté pour une approche différente puisque pour permettre de juger de l'intensité d'un effet, ils proposent de fournir une base empirique par la compilation d'un très grand nombre de résultats [Haase, Waechter, Solomon, 1982]. Ils ont ainsi recensé les grandeurs des effets, exprimées par le coefficient f^2 (part de variance «expliquée» par le facteur expérimental), calculées à partir des tests statistiques rapportés dans les articles parus dans le *Journal of Counseling Psychology* de 1970 à 1979. La moyenne obtenue pour ces coefficients est de 0.1589, valeur plutôt faible.

Nous avons nous-même procédé à une ré-analyse d'articles parus dans le *Journal of Abnormal Psychology* en 1994 en privilégiant l'aspect descriptif et des analyses

fiducio-bayésiennes [Poitevineau, 1999]. Il en résulte que la moyenne des effets apparaît plus importante si l'on utilise un critère pour lequel la calibration se fait par rapport à la variabilité inter-sujets plutôt que par rapport à la variance totale. Cette étude peut donc apparaître comme une illustration de la critique de Rosenthal et Rubin sur le «biais» des indicateurs en part de variance [Rosenthal, Rubin, 1982]. Cependant, d'un point de vue inférentiel, les effectifs sont en général trop faibles pour aller au-delà de la simple significativité. Par ailleurs, l'abus consistant à conclure à une absence d'effet en cas de résultat non significatif s'avère particulièrement fréquent.

3.4 QUELQUES RÉSULTATS EXPÉRIMENTAUX

Les expériences menées auprès de chercheurs peuvent être subdivisées en deux groupes.

Un premier groupe comprend des études par questionnaires qui portent directement sur la signification des probabilités associées aux tests de signification et qui font directement appel à des connaissances «vivresques» [Oakes, 1986], [Freeman, 1993], [Falk, Greenbaum, 1995]. Ici il ne s'agit que de confirmer l'existence des erreurs d'interprétation de ces probabilités par rapport à la référence normative et de révéler leur grande fréquence.

Un second groupe comprend des études plus proches de situations réelles, où il est davantage fait appel à des jugements «spontanés», pour lesquels le sujet n'a pas forcément à sa disposition une réponse stéréotypée. Tversky et Kahneman ont été les instigateurs des études sur les représentations des chercheurs dans diverses situations d'inférence statistique [Tversky, Kahneman, 1971]. Toutefois leur perspective reste relativement normative, puisqu'il s'agit avant tout de répertorier des distorsions, des *biais*, par rapport à une référence normative (le cadre bayésien). Sans doute le plus connu est celui de «représentativité», c'est-à-dire la surestimation par les chercheurs de la représentativité d'un échantillon. Au contraire, M.-P. Lecoutre adopte une approche beaucoup plus descriptive avec pour finalité la formalisation de modèles descriptifs du fonctionnement cognitif développés dans ce type de situations. Elle s'intéresse particulièrement à des situations «conflituelles», où l'on fournit des informations qui semblent contradictoires [Lecoutre M.-P., 1982, 1991], [Lecoutre M.-P., Rouanet, 1993]. Elle constate alors que la plupart des chercheurs fondent essentiellement leurs conclusions sur les résultats des tests mais que, s'ils concluent unanimement, et indépendamment des effets observés et des attentes, à l'existence d'un effet en cas de résultat significatif, leurs attitudes sont très diversifiées en cas de résultat non significatif. On notera que les statisticiens se conduisent globalement d'une manière voisine de celle des psychologues, et qu'ils ne sont pas, malgré leur formation, à l'abri des abus d'interprétation des tests, tout particulièrement de celui consistant à voir dans un résultat non significatif, surtout quand l'effet observé est faible, la démonstration d'une absence d'effet [Lecoutre M.-P., Poitevineau, Lecoutre B., 2003].

Dans ce même contexte, une expérience intéressante est celle de Rosenthal et Gaito car elle fut la première portant sur la manière dont les chercheurs en psychologie expérimentale interprètent les seuils de signification associés à un test [Rosenthal, Gaito, 1963]. Face à une série de 14 seuils de signification (de 0.001 à 0.90), supposés associés à des résultats expérimentaux issus tour à tour d'un échantillon de taille 10 ou 100, les sujets étaient invités à exprimer leur «degré de conviction dans les résultats expérimentaux comme une fonction des valeurs p correspondantes». Il apparaît que les courbes reliant la confiance au seuil p sont des fonctions monotones décroissantes de type exponentiel, celle correspondant au plus grand effectif étant toujours la plus haute

(plus de confiance). Les auteurs concluent aussi, en s'appuyant sur un test significatif, à un effet de chute brutale de confiance juste après la valeur $p = 0.05$. Oakes a tiré argument de ce résultat pour conclure, à l'appui de son hypothèse de «significativité», que la «chute» de confiance indiquait que des résultats expérimentaux hypothétiques prennent soudainement statut de réalité dès lors qu'ils sont associés à un seuil observé juste inférieur à 0.05 [Oakes, 1986]. Cependant nous avons pu montrer, en répliquant l'expérience et en procédant à des analyses individuelles que cet effet de chute n'existe sensiblement que chez une minorité de chercheurs. Au contraire la majorité des chercheurs ont tendance à porter un jugement nuancé en fonction du seuil observé, et non à fonctionner en tout ou rien en opposant systématiquement significatif et non significatif [Poitevineau, Lecoutre B., 2001].

Une méthode qui donne lieu à tant d'erreurs dans son application, même chez des usagers avertis, pose, pour le moins, le problème de son adéquation aux besoins des chercheurs. Pourtant, les critiques des tests n'ont guère influé sur la pratique des chercheurs en psychologie, le titre de l'article de Falk et Greenbaum est d'ailleurs particulièrement explicite : «Significance tests die hard» [Falk, Greenbaum, 1995]. On est donc conduit à rechercher les raisons de ces erreurs et de leur persistance.

3.5 POURQUOI LES TESTS DE SIGNIFICATION SONT-ILS TOUJOURS EN USAGE ?

Plusieurs raisons ont été avancées pour expliquer la popularité des tests.

- Les tests confèrent aux conclusions une impression d'objectivité [Bakan, 1966]. Sans aucun doute ceci est une raison importante, l'objectivité étant une préoccupation fondamentale des chercheurs scientifiques.
- L'appareillage mathématique des tests fournit une apparence de scientificité [Carver, 1978]. La rigueur des mathématiques et l'*aura* dont elles jouissent sont censés rejaillir sur l'ensemble de la recherche, assurant sa validité.
- L'approche de Fisher offre une ressemblance avec les idées de Popper [Popper, 1959]. Pour ce dernier, la démarcation entre énoncés scientifiques et non scientifiques est réalisée sur la base du caractère réfutable, ou non, de ces énoncés : une hypothèse scientifique est une hypothèse qui peut être empiriquement contredite ; en revanche, jamais elle ne pourra être vérifiée (au sens de «logiquement prouvée»). On retrouve la même asymétrie entre réfutation et vérification que chez Fisher. La théorie des tests de Fisher, comme le pense Oakes, a ainsi pu bénéficier du succès certain des idées de Popper [Oakes, 1986]. Encore récemment Reuchlin invoque ces idées et leur consacre un paragraphe dans son ouvrage de 1992 [Reuchlin, 1992].
- En semblant fournir une procédure automatique, les tests dispensent d'une réflexion supplémentaire et assurent un confort certain à leurs utilisateurs [Bakan, 1966], [Falk, Greenbaum, 1995].
- Les tests déchargent le chercheur de la tâche d'interprétation. Carver évoque le fait particulier que bien souvent le chercheur se trouve confronté à des unités de mesure qu'il ne sait interpréter (il prend l'exemple d'une échelle de «Qualité d'enseignement» : que représente une différence de 10 points sur cette échelle?) [Carver, 1978]. Dans ces conditions, le test, avec sa possibilité de déclarer «significatif» un effet, est vu comme une solution, déchargeant le chercheur de la tâche d'interprétation, comme si la significativité statistique se suffisait à elle-même (on rejoint la confusion entre significativité statistique et significativité substantielle).
- Les tests semblent répondre exactement à l'attente des chercheurs, à propos de la probabilité des hypothèses, de la répliquabilité des résultats, de l'importance des effets,

en raison justement des abus dont ils font l'objet. Selon beaucoup d'auteurs, il faut voir là la principale raison de la popularité des tests [Schmidt, 1996]. Ce sont les abus mêmes qui conduiraient à une illusion d'adaptation. Et ces abus seraient favorisés par le manque de naturel du raisonnement sur lequel est basé le test, ce qui est une difficulté à comprendre la portée exacte du test.

Ensuite, plusieurs hypothèses peuvent être avancées pour expliquer la persistance de cette popularité, mais sans doute aucune n'est à elle seule suffisante.

- On peut penser que les chercheurs, à part ceux travaillant dans le domaine ou particulièrement intéressés par ces questions, ne lisent pas, en général, les articles méthodologiques.
- Le nombre considérable de domaines de l'activité scientifique touchés par les tests et le temps passé depuis l'apparition des premières critiques peuvent en eux-mêmes apparaître comme des arguments en faveur des pratiques actuelles qui irait imaginer qu'une méthode aussi universelle et qui a survécu à de telles attaques depuis si longtemps ne soit pas appropriée (c'est l'argument de Frick [1996]).
- Les problèmes soulevés par les critiques sont fondamentaux et les résistances des chercheurs à modifier leur pratique sont en conséquence. C'est ce qu'exprime Bakan sous une forme imagée en considérant le test comme un des « fils de la tapisserie culturalo-scientifique » de la psychologie, profondément emmêlé avec les autres qui tirer dessus (reconnaître son inadéquation) risquerait de démanteler l'ensemble de la tapisserie ; aussi les psychologues préféreraient-ils effectuer des ajustements (pourvoir le test de qualités qu'il n'a pas) [Bakan, 1966].
- Le fait que le test statistique soit devenu un critère très important, une norme pour la publication joue certainement un grand rôle dans la perpétuation des attitudes. Winkler relève la grande uniformité des plans d'expérience et des analyses statistiques afférentes dans les articles publiés dans les revues relevant de l'*American Psychological Association* [Winkler, 1974]. Aussi un chercheur, même convaincu du peu d'intérêt d'un test, hésitera à ne pas en produire, craignant un refus de publication [Lecoutre M.-P., 1991].
- Oakes invoque la soumission des chercheurs à l'autorité des statisticiens [Oakes, 1986]. Mais cela ne tient pas compte du fait que ces derniers sont eux-mêmes divisés quant aux théories statistiques.
- Un facteur probablement décisif, bien qu'assez peu mentionné en général, est le rôle de l'enseignement des statistiques en psychologie. D'une part, cet enseignement est muet sur les controverses parmi les statisticiens [Gigerenzer, 1993], et d'autre part, malgré toutes les critiques, cet enseignement n'a pratiquement pas évolué, tout au moins jusqu'à un passé récent [Schmidt, 1996].
- Gigerenzer propose une métaphore psychanalytique dans laquelle les idées bayésiennes jouent le rôle du ça, la théorie de Neyman et Pearson le rôle du surmoi et celle de Fisher le rôle du moi [Gigerenzer, 1993]. Il existerait un conflit entre le ça et le surmoi, d'où anxiété, et il en résulterait un aveuglement dogmatique aboutissant à la négation du problème, ce qui expliquerait que les choses ne changent pas, ou si peu. Mais cet argument a été écarté par Falk et Greenbaum [1995] qui ont fort justement remarqué que les psychologues ne connaissent pas d'angoisse à ce propos. Il semble effectivement que l'angoisse des psychologues ne porte pas sur la question « Dois-je faire un test », mais seulement sur la question « Ai-je fait le bon test ».
- Les tests font souvent l'objet d'une véritable religion et sont pratiqués comme un rituel [Salsburg, 1985]. Guttman a d'ailleurs dénoncé à ce propos les « adorateurs des étoiles » [Guttman, 1979, 1983], et Rouanet [1991(a)] évoque, ironiquement, une

- «respectabilité hôtelière». Il n'est pas facile de s'opposer à une religion, surtout quand elle est si répandue et les arguments logiques ont peu de prise en ce domaine.
- Finalement, en raison même de leurs défauts, les tests, dans leur usage actuel, sont un allié précieux du chercheur : un effet significatif, un «succès», donc une possibilité de publier, peut s'obtenir avec un effort intellectuel réduit en recueillant suffisamment de données. D'où une réticence certaine à abandonner un outil aussi pratique.

Il est nécessaire d'ajouter que les logiciels statistiques de grande diffusion ne font que contribuer à la popularité des tests dont ils automatisent l'usage. Ces logiciels ont de plus en plus tendance à se substituer aux manuels de statistiques dans un rôle prescriptif qui peut se retrouver à différents niveaux :

- i) dans le manuel d'utilisation qui, par-delà sa fonction de mode d'emploi du logiciel, peut se présenter comme un véritable manuel de statistique ;
- ii) dans l'aide en ligne qui peut être une simple reprise du manuel ou apporter d'autres informations ;
- iii) dans les sorties présentées, où les résultats numériques sont de plus en plus souvent agrémentés de commentaires ou de conclusions «en clair». On peut à l'évidence craindre que ce rôle, en raison de simplifications souvent abusives (nécessité commerciale de diffusion auprès du plus grand nombre d'utilisateurs oblige), ne fassent que renforcer les abus existants.

4. ASPECTS PRESCRIPTIFS

Les prescripteurs (auteurs de manuels, d'articles méthodologiques, de logiciels, conseillers statistiques, enseignants...) ont un rôle important à jouer, qui consiste à préciser l'usage des tests de signification ou à le contre-indiquer et recommander d'autres méthodes.

Relativement peu d'auteurs, comparé au nombre d'auteurs critiques, ont défendu les tests de signification dans le contexte de la recherche scientifique [Sullivan, 2000]. Quant aux auteurs qui ont critiqué l'usage des tests de signification, la plupart ont proposé des solutions de rechange. Nous passerons brièvement en revue les principales et examinerons leurs implications méthodologiques.

4.1 LES SOLUTIONS DE RECHANGE PRÉCONISÉES

La mesure de la grandeur de l'effet

Une suggestion très fréquente est d'étudier l'intensité des effets (par exemple [Nunnally, 1960], [Lecoutre B., 1984], [Schmidt, 1996]). Cette étude est vue, soit comme un prolongement, soit comme un remplacement de la procédure de test, et elle a été abordée de façons très différentes. Loin de s'opposer aux approches décrites dans les sections suivantes, elle en constitue au contraire le plus souvent un préalable. Elle a donné lieu à la proposition de plusieurs statistiques pour mesurer la grandeur de l'effet.

La description de la grandeur des effets observés est certainement une étape indispensable, mais en ce qui concerne l'inférence, la simple estimation ponctuelle est incontestablement insuffisante. Ajouter une telle estimation au test de signification usuel d'un effet nul est certes un progrès, mais suggère fortement une généralisation qui reste impressionniste et présente des dangers réels. En particulier, comme l'ont montré

les expériences auprès des utilisateurs, un effet observé faible associé à un résultat non significatif est souvent perçu comme un cas «*en* faveur de l'absence d'effet vrai», alors qu'il n'est souvent qu'un constat d'ignorance.

L'étude de la puissance

Son défenseur le plus ardent en psychologie est Cohen [1969]. La puissance n'est, bien entendu, pas envisagée contre la théorie de Neyman et Pearson, puisqu'elle en est le cœur, mais elle est invoquée pour pallier les insuffisances de la théorie de Fisher au regard des problèmes de sensibilité.

Mais s'il est indéniable que prendre en compte la puissance du test utilisé peut effectivement constituer un garde-fou pour éviter des conclusions injustifiées d'absence d'effet en cas de résultat non significatif, dès que l'on tente d'utiliser la puissance pour interpréter les données, les critiques développées contre la théorie neyman-pearsonienne s'appliquent. Par ailleurs, Schuirmann a montré, dans le cadre du «*problème de l'équivalence*» en pharmacologie, que cette méthode de la puissance possède des propriétés formelles indésirables qui font qu'elle n'est pas recommandable [Schuirmann, 1987].

L'intervalle de confiance

L'intervalle de confiance, au sens de Neyman et Pearson, est de très loin la solution la plus souvent mentionnée [Thompson, 2002]. On ne trouve guère que Cornfield [1966], Johnstone [1988], Frick [1995] ou Kadane [1995] pour englober l'intervalle de confiance dans la même critique que le test. Il est symptomatique qu'apparaissent maintenant dans des journaux scientifiques des éditoriaux mettant en cause la pratique des tests et/ou prônant un développement de l'utilisation des intervalles de confiance [Fidler, Thompson, 2001].

Rappelons qu'il existe une correspondance directe entre l'intervalle de confiance et le test [Lehmann, 1959] : un intervalle de confiance $1-\alpha$ est généralement construit comme l'ensemble des valeurs qui, si elles étaient choisies pour hypothèse nulle, conduiraient à un résultat de test non significatif au seuil (bilatéral) α .

En tant que méthode d'estimation de l'effet, il présente déjà l'avantage de recentrer la discussion sur celui-ci. En tant qu'intervalle, par rapport au test usuel d'une hypothèse ponctuelle, il invite à prendre en considération la possibilité d'autres valeurs que celle choisie pour hypothèse nulle. Il apporte aussi, par sa largeur, une information directe sur la variabilité en jeu, sur la précision des résultats.

Cependant, l'intervalle de confiance pose, lui aussi, bien des problèmes. D'un point de vue pratique, en dehors des situations élémentaires, les intervalles de confiance ne sont pas toujours facilement disponibles. À ce propos, Rouanet évoque la «*trouerie des conseillers statistiques*» qui recommandent d'utiliser davantage de méthodes qu'ils savent non forcément disponibles [Rouanet, 1991(b)]. Enfin, peut-être plus encore que les tests traditionnels, les intervalles de confiance donnent lieu à des erreurs ou abus d'interprétation. L'interprétation correcte, dans le cadre fréquentiste, d'un intervalle de confiance $1-\alpha$ est que si l'on répétait l'expérience un nombre infini de fois, $100(1-\alpha)\%$ des intervalles calculés contiendraient la vraie valeur du paramètre. Le paramètre d'intérêt n'est pas probabilisé, ce sont les bornes de l'intervalle qui varient d'une répétition à l'autre. Pour une expérience particulière, l'utilisateur affirme que l'intervalle calculé contient effectivement le paramètre, ou décide d'agir comme si c'était le cas ;

ainsi, sur un très grand nombre de répétitions, cet utilisateur ne se trompera que $100(1-\alpha)\%$ parfois [Neyman, 1952]. Or, il est très courant de trouver des énoncés du type *La probabilité que la vraie valeur du paramètre soit comprise entre x et y (les bornes particulières calculées à partir des données observées) est $1-\alpha$* . Ce qui est une interprétation erronée dans le cadre fréquentiste. Si cet abus est si fréquent, même chez des utilisateurs avertis ou chez des statisticiens [Lecoutre B., 1997], c'est tout simplement qu'il relève d'une interprétation *bayésienne* parfaitement naturelle, sinon justifiée [Kadane, 1995].

Dans la pratique, l'utilisateur a donc le choix entre trois attitudes

- i) conserver le cadre de justification fréquentiste de l'intervalle de confiance et se satisfaire de l'interprétation «correcte» ;
- ii) conserver ce cadre tout en adoptant l'interprétation bayésienne alors «erronée» ;
- iii) adopter explicitement le cadre de justification bayésien. Comme nous l'avons dit, tout montre que la majorité des utilisateurs adoptent actuellement la deuxième attitude. On peut donc penser que ce sont paradoxalement leurs interprétations bayésiennes sauvages qui rendent ces procédures populaires.

Les méthodes de vraisemblance

Un certain nombre d'auteurs ont prôné l'utilisation de la fonction de vraisemblance, et particulièrement du rapport de vraisemblance (ainsi [Rozeboom, 1960], [Edwards, Lindman, Savage, 1963], [Cornfield, 1966], [Winkler, 1974], [Oakes, 1986]). Dans le cas simple où deux hypothèses ponctuelles H_0 et H_1 sont en cause, cela présente l'avantage de ne faire intervenir ni probabilités *a priori*, ni éléments non observés, mais dans le cas d'hypothèses composées, cela nécessite l'introduction des probabilités *a priori*.

Les méthodes bayésiennes

La proposition précédente est souvent associée aux méthodes bayésiennes qui sont d'ailleurs souvent préconisées par les mêmes auteurs (voir aussi [Hoc, 1983], [Lecoutre B., 1984], [Lecoutre B., Lecoutre M.-P., Poitevineau, 2001], [Rouanet, Bernard, Bert, Lecoutre B., Lecoutre M.-P., Leroux, 2000], [Rouanet, Lépine, Pelnard-Considère, 1976]).

La méthode bayésienne consiste à calculer, au moyen du théorème de Bayes, la distribution *a posteriori* pour le paramètre auquel on s'intéresse, à partir des données observées (et d'un modèle d'échantillonnage associé) et d'une distribution de probabilités *a priori* sur le paramètre. Elle suppose donc, en général, de considérer la probabilité comme une mesure de l'incertitude relative à la valeur du paramètre inconnu. Portant directement sur l'effet d'intérêt (la valeur du paramètre retenu), elle est beaucoup plus intuitive et correspond beaucoup plus aux demandes des chercheurs que les méthodes fréquentistes [Kadane, 1995] elle fournit ainsi une solution directe aux questions de «négligeabilité» ou de notabilité de la grandeur de l'effet [Lecoutre B., 1984], [Rouanet, 1996].

La nécessité de spécifier des probabilités *a priori* a été, et est encore, beaucoup critiquée comme introduisant un élément subjectif. Mais, d'une part, l'approche non-informative, dite aussi standard ou encore fiducio-bayésienne selon la terminologie forgée par Rouanet *et al.* [1976], qui consiste à se placer dans un état d'ignorance *a priori* sur la valeur du paramètre, peut être vue comme une réponse à cette critique, et

d'autre part les théories fréquentistes des tests ne sont pas moins subjectives [Bernard, 1996].

Par ailleurs, il est possible de justifier certaines erreurs d'interprétation des tests et des intervalles de confiance en passant du cadre théorique traditionnel au cadre bayésien (plus particulièrement au cadre bayésien standard). Il est même possible de considérer que la plupart des chercheurs se comportent comme des “fiducio-bayésiens” naturels ou naïfs [Lecoutre B., 1991], [Lecoutre M.-P., 1991].

Si l'on s'en tient à l'intérêt méthodologique, seules les méthodes d'intervalle de confiance et, sans doute plus encore les méthodes bayésiennes paraissent devoir s'imposer comme véritables «challengers» des tests traditionnels. Dans cette perspective, un important effort de développement des méthodes bayésiennes standard, «non informative» a été entrepris en France depuis de nombreuses années pour couvrir les cas les plus usuels d'analyse des données expérimentales [Lecoutre B, 1984], [Lecoutre B, 1996], [Lecoutre B, Derzko, 2001], [Rouanet, Bernard, Bert, Lecoutre B, Lecoutre M.-P., 2000], [Rouanet, Lépine, Pelnard-Considère, 1976]. Des logiciels appropriés sont maintenant disponibles et ces méthodes peuvent être utilisées aussi facilement que les tests usuels [Lecoutre B, 1996], [Lecoutre B, Poitevineau, 1992].

5. CONCLUSION

Aucune des deux théories fréquentistes des tests de Fisher et de Neyman et Pearson n'apparaissant appropriée à l'usage, il n'est pas surprenant que la pratique actuellement dominante en psychologie consiste à en utiliser un amalgame. Mais un tel amalgame ne peut être justifié d'un point de vue formel. C'est donc à une véritable distorsion de l'outil statistique qu'on assiste dans la pratique. Cette distorsion est encouragée par les manuels de statistique appliquée qui constituent, directement ou indirectement, les références de base pour les chercheurs en psychologie. De plus, la plupart de ces manuels comportent des abus d'interprétation explicites, particulièrement dans les exemples d'application, et fournissent donc un mode d'emploi trompeur. Il en résulte bon nombre de mésusages du test de signification, les chercheurs tendant à lui faire dire plus que ne le permet la théorie.

Ces mésusages ressortent différemment selon les situations, comme le montrent les ré-analyses d'articles publiés et les travaux expérimentaux. Dans le cadre d'une publication, les chercheurs se comportent essentiellement en «tout ou rien» en ce qu'ils s'en tiennent à la dichotomie test significatif/non significatif et commentent peu les résultats des tests, d'où des conclusions d'ordre décisionnel. En revanche, les expériences montrent que les jugements de la plupart d'entre eux sont nuancés mais aussi que la confusion entre significativité statistique et significativité substantielle, éludée dans les articles publiés, réapparaît. Dans les deux cas, il se confirme qu'un résultat non significatif est très souvent abusivement interprété comme la démonstration d'une absence d'effet, ceci même par des statisticiens, ce qui laisse penser que le remède aux abus d'interprétation ne peut être recherché simplement par une formation plus poussée des chercheurs aux tests statistiques.

Les abus d'interprétation ne sont pas de simples erreurs de compréhension des théories statistiques, mais ils vont dans le sens d'une correction intuitive des défauts méthodologiques des tests certains auteurs, comme Bakan [1966], Phillips [1973] ou Lecoutre *et al.* [2003] parlent d'ailleurs d'ajustements ou de biais adaptatifs. Par ailleurs,

la pratique des chercheurs en psychologie est socialement adaptée à une norme qui, surmontant toutes les critiques, a fait l'objet d'un véritable consensus social impliquant tous les acteurs de la recherche scientifique – prescripteurs, auteurs et censeurs. Ce consensus est révélé par la stéréotypie, mais aussi la pauvreté, de la présentation des résultats statistiques dans les revues expérimentales – il est sans nul doute l'une des raisons essentielles de la longévité de l'usage des tests.

Si la situation a peu évolué jusqu'à présent, il peut en aller autrement dans un avenir proche. En effet, l'American Psychological Association [2001] a dernièrement reconsidéré l'inférence statistique et modifié ses normes de publication, [Wilkinson, 1999]. Les tests traditionnels ne sont pas abandonnés, mais l'ouverture à d'autres méthodes d'analyse des résultats est clairement affichée, notamment aux méthodes d'intervalles de confiance, et aussi aux méthodes bayésiennes. Par ailleurs, l'examen des publications actuelles, tant en statistique mathématique qu'en statistique appliquée, révèle une évolution considérable vers les méthodes bayésiennes. L'utilisation de ces méthodes devrait donc devenir un défi motivant pour la communauté scientifique [Lecoutre B, Lecoutre M.-P., 2001].

Remerciements. Nous remercions Marie-Paule et Bruno Lecoutre pour leur lecture attentive et leurs remarques.

BIBLIOGRAPHIE

AMERICAN PSYCHOLOGICAL ASSOCIATION, *Publication Manual of the American Psychological Association* (5th ed.), Washington, DC, Author, 2001.

BAKAN D., "The test of significance in psychological research", *Psychological Bulletin* 66, 1966, p. 423-437.

BERKSON J., "Some difficulties of interpretation encountered in the application of the chi-square test", *Journal of the American Statistical Association* 33, 1938, p. 526-542.

BERKSON J., "Tests of significance considered as evidence", *Journal of the American Statistical Association* 37, 1942, p. 325-335.

BERNARD J.-M., "Bayesian interpretation of frequentist procedures for a Bernoulli process", *The American Statistician* 50, 1996, p. 7-13.

BORING E.G., "Mathematical versus scientific significance", *Psychological Bulletin* 16, 1919, p. 335-338.

CAMILLERI S. F., "Theory, probability, and induction in social research", *American Sociological Review* 27, 1962, p. 170-178.

CARVER R. P., "The case against statistical significance testing", *Harvard Educational Review* 48, 1978, p. 378-399.

CLARK-CARTER D., "The account taken of statistical power in research published in the *British Journal of Psychology*", *British Journal of Psychology* 88, 1997, p. 71-83.

COHEN J., "The statistical power of abnormal-social psychological research: A Review", *Journal of Abnormal and Social Psychology* 65, 1962, p. 145-153.

COHEN J., *Statistical Power Analysis for the Behavioral Sciences*, New York, Academic Press, 1969.

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

CORNFIELD J., "Sequential trials, sequential analysis and the likelihood principle", *The American Statistician* 20, 1966, p. 18-23.

EDWARDS W., LINDMAN H., SAVAGE L. J., "Bayesian statistical inference for psychological research", *Psychological Review* 70, 1963, p. 193-242.

FALK R., GREENBAUM C. W., "Significance tests die hard. The amazing persistence of a probabilistic misconception", *Theory & Psychology* 5, 1995, p. 75-98.

FAVERGE J.-M., *Méthodes statistiques en psychologie appliquée*, (3 vol.), (7^e édition) Paris, Presses Universitaires de France, 1975, (1^e édition, 1950)

FIDLER F., THOMPSON B., "Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes", *Educational and Psychological Measurement* 61, 2001, p. 575-604.

FISHER R. A., *Statistical Methods for Research Workers*, (réimp. 14^e édition de 1970), J. H. Bennett (ed.), *Statistical Methods, Experimental Design, and Scientific Inference*, Oxford, Oxford University Press, 1990(a), (1^e édition, 1925, London, Oliver and Boyd.)

FISHER R. A., *The Design of Experiments*, London, (réimp. 8^e édition de 1966), J. H. Bennett (ed.), *Statistical Methods, Experimental Design, and Scientific Inference*, Oxford, Oxford University Press, 1990(b), (1^e éd., 1935, London, Oliver and Boyd).

FISHER R. A., *Statistical Methods and Scientific Inference*, (réimp. 3^e édition de 1973), J. H. Bennett (ed.), *Statistical Methods, Experimental Design, and Scientific Inference*, Oxford, Oxford University Press, 1990(c), (1^e édition, 1956, London, Oliver and Boyd).

FREEMAN P. R., "The role of *p*-values in analysing trial results", *Statistics in Medicine* 12, 1993, p. 1443-1452.

FRICK R. W., "A problem with confidence intervals", *American Psychologist* 50, 1995, p. 1002-1003.

FRICK R. W., "The appropriate use of null hypothesis testing", *Psychological Methods* 1, 1996, p. 379-390.

GIGERENZER G., "The superego, the ego, and the id in statistical reasoning", G. Keren, C. Lewis (eds), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, Hillsdale (N-J), Lawrence Erlbaum associates, 1993.

GUTTMAN L., "Cyril Burt and the careless star worshippers", *L'Echo des Messaches* 9, (1979), p. 7-8.

GUTTMAN L., "What is not what in statistics?", *The Statistician* 26, 1983, p. 81-107.

HAASE R. F., WAECHTER D. M., SOLOMON G. S., "How significant is a significant difference? Average effect size of research in counseling psychology", *Journal of Counseling Psychology* 29, 1982, p. 58-65.

HARCUM E. R., "Methodological versus empirical literature: Two views on casual acceptance of the null hypothesis", *American Psychologist* 45, 1990, p. 404-405.

HAYS W. L., *Statistics for the Social Sciences*, 2^e édition, New York, Holt, Rinehart & Winston, 1973.

HOC J.-M., *L'analyse planifiée des données en psychologie*, Paris, Presses Universitaires de France, 1983.

JEFFREYS H., *Theory of Probability*, 3^e édition, Oxford, Clarendon, 1961, (1^e édition, 1939.)

- JOHNSTONE D., "Comments on Oakes on the foundations of statistical inference in the social and behavioral sciences: The market for statistical significance", *Psychological Reports* 63, 1988, p. 319-331.
- KADANE J. B., "Prime time for Bayes", *Controlled Clinical Trials* 16, 1995, p. 313-318.
- KIRK R. E., *Experimental Design*, 2^e édition, Belmont, Brook-Cole, 1982.
- LECOUTRE B., *L'analyse bayésienne des comparaisons*, Lille, Presses Universitaires de Lille, 1984.
- LECOUTRE B., «Du test de signification à l'inférence fiducio-bayésienne», H. Rouanet, M.-P. Lecoutre, M.-C. Bert, B. Lecoutre, J.-M. Bernard, B. Leroux (eds), *L'inférence statistique dans la démarche du chercheur*, Berne, Peter Lang, 1991.
- LECOUTRE B., *Traitement Statistique des Données Expérimentales : Des Pratiques Traditionnelles aux Pratiques Bayésiennes - Avec programmes Windows® par B. Lecoutre et J. Poitevineau*, Montreuil, CISIA-CERES, 1996.
- LECOUTRE B., «Et si vous étiez un bayésien 'qui s'ignore'», *La Revue de Modulad* 18, 1997, p. 81-87.
- LECOUTRE B., DERZKO, G., "Asserting the smallness of effects in ANOVA", *Methods of Psychological Research* 6, 2001, p. 1-32, [<http://www.mpr-online.de>].
- LECOUTRE B., LECOUTRE M.-P., POITEVINEAU J., "Uses, abuses and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable?", *International Statistical Review* 69, 2001, p. 399-417.
- LECOUTRE B., POITEVINEAU J., "PAC (Programme d'Analyse des Comparaisons) : Guide d'utilisation et manuel de référence", Montreuil, CISIA-CERESTA, 1992.
- LECOUTRE B., POITEVINEAU J., «Aller au-delà des tests de signification usuels vers de nouvelles normes de publication», *l'Année Psychologique* 100, 2000, p. 683-713.
- LECOUTRE M.-P., «Comportements des chercheurs dans des situations conflictuelles d'analyse de données expérimentales», *Psychologie Française* 27, 1982, p. 1-8.
- LECOUTRE M.-P., «Et le point de vue des chercheurs? Quelques éléments de réflexion», H. Rouanet, M.-P. Lecoutre, M.-C. Bert, B. Lecoutre, J.-M. Bernard, B. Leroux (eds), *L'inférence statistique dans la démarche du chercheur*, Berne, Peter Lang, 1991.
- LECOUTRE M.-P., POITEVINEAU J., LECOUTRE B., "Even statisticians are not immune from misinterpretations of Null Hypothesis Significance Tests", *International Journal of Psychology* 38, 2003, p. 37-45.
- LECOUTRE M.-P., ROUANET, H., "Predictive judgements in situations of statistical analysis", *Organizational Behavior and Human Decision Processes* 54, 1993, p. 45-56.
- LEHMANN E. L., *Testing Statistical Hypotheses*, New York, Wiley, 1959.
- LINDLEY D. V., "A statistical paradox", *Biometrika* 44, 1957, p. 187-192.
- MEEHL P. E., "Theory testing in psychology and physics: A methodological paradox", *Philosophy of Science* 34, 1967, p. 103-115.
- MORRISON D. E., HENKEL R. E. (eds), *The Significance Test Controversy*, London, Butterworths, 1970.
- NEYMAN J., *First Course in Probability and Statistics*, New York, Holt, 1950.
- NEYMAN J., *Lectures and Conferences on Mathematical Statistics and Probability*, (2^e édition), Washington, Graduate School U.S. Department of Agriculture, 1952.

NEYMAN J., PEARSON E. S., "On the use and interpretation of certain test criteria for purposes of statistical inference. Part I", *Biometrika* 20A, 1928, p. 175-240.

NEYMAN J., PEARSON E. S., "On the problem of the most efficient tests of statistical hypotheses", *Philosophical Transactions of the Royal Society of London, Series A* 231, 1933, p. 289-337.

NEYMAN J., PEARSON E. S., "The testing of statistical hypotheses in relation to probabilities *a priori*", *Proceedings of the Cambridge Philosophical Society* 29, 1933, p. 492-510.

NUNNALLY J. C., "The place of statistics in psychology", *Educational and Psychological Measurement* 20, 1960, p. 641-650.

OAKES M., *Statistical Inference: A Commentary for the Social and Behavioural Sciences*, New York, Wiley, 1986.

O'BRIEN T. C., SHAPIRO B. J., "Statistical significance, What?", *Mathematics Teacher* 61, 1968, p. 673-676.

PEARSON, K., "On the criterion that a given system of deviations form the probable in the case of correlated systems of variables is such that it can reasonably be supposed to have arisen from random sampling", *Philosophical Magazine, Series V*, 1, 1900, p. 157-175.

PHILLIPS L. D., *Bayesian Statistics for Social Scientists*, London, Nelson, 1973.

POITEVINEAU J., «Pratiques des tests statistiques en psychologie cognitive : L'exemple d'une année d'un journal», *Cahiers du LCPE* 3, 1999, p. 67-87.

POITEVINEAU J., LECOUTRE B., "Interpretation of significance levels by psychological researchers: The .05-cliff effect may be overstated", *Psychonomic Bulletin & Review* 8, 2001, p. 847-850.

POPPER K. R., *La logique de la découverte scientifique*, Paris, Payot, 1973, (réimp. 1984, 1^e édition originale, en allemand, 1939 ; 1^e traduction en anglais, 1959).

PRENTICE D. A., MILLER D. T., "When small effects are impressive", *Psychological Bulletin* 112, 1992, p. 160-164.

REUCHLIN M., *Précis de statistique*, Paris, Presses Universitaires de France, 1976.

REUCHLIN M., *Introduction à la recherche en psychologie*, Paris, Nathan, 1992.

ROSENTHAL R., GAITO J., "The interpretation of levels of significance by psychological researchers", *Journal of Psychology* 55, 1963, p. 33-38.

ROSENTHAL R., RUBIN D. B., "A simple, general purpose display of magnitude of experimental effect", *Journal of Educational Psychology* 74, 1982, p. 166-169.

ROUANET H., «Les pratiques statisticiennes en question», H. Rouanet, M.-P. Lecoutre, M.-C. Bert, B. Lecoutre, J.-M. Bernard, B. Leroux (eds), *L'inférence statistique dans la démarche du chercheur*, Berne, Peter Lang, 1991(a).

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

ROUANET H., "Bayesian methods for assessing importance of effects", *Psychological Bulletin* 119, 1996, p. 149-158.

ROUANET H., BERNARD J.-M., BERT M.-C., LECOUTRE B., LECOUTRE M.-P., LEROUX B. (eds), *New Ways in Statistical Methodology: From Significance Tests to Bayesian Inference*, (2nd edition), Berne, Peter Lang, 2000.

- ROUANET H., LÉPINE D., PELNARD-CONSIDÈRE J., "Bayes-fiducial procedures as practical substitutes for misplaced significance testing: An application to educational data", D. N. M. De Gy, L. J. T. Van Der Kamp (eds), *Advances in Psychological and Educational Measurement*, New York, Wiley, 1976.
- ROZEBOOM W. W., "The fallacy of the null hypothesis significance test", *Psychological Bulletin* 57, 1960, p. 416-428.
- SALSBURG D. S., "The religion of statistics as practiced in medical journals", *The American Statistician* 39, 1985, p. 220-223.
- SCHMIDT F. L., "Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers", *Psychological Methods* 1, 1996, p. 115-129.
- SCHUIRMANN D. J., "A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability", *Journal of Pharmacokinetics and Biopharmaceutics* 15, 1987, p. 657-680.
- SELVIN H. C., "A critique of tests of significance in survey research", *American Sociological Review* 22, 1957, p. 519-527.
- SIEGEL S., *Nonparametric Statistics for the Behavioral Sciences*, New York, McGraw-Hill, 1956.
- STERLING T. D., "Publication decisions and their possible effects on inferences drawn from tests of significance, or vice versa", *Journal of the American Statistical Association* 54, 1959, p. 30-34.
- SULLIVAN J. R., "A review of post-1994 literature on whether statistical significance test should be banned", *Paper presented at the annual meeting of the Southwest Educational Research Association*, Dallas, 2000.
- THOMPSON B., "What future quantitative social science research could look like: Confidence intervals for effect sizes", *Educational Researcher* 31, 2002 p. 24-31.
- TUKEY J. W., "The Philosophy of Multiple Comparisons", *Statistical Science* 6, 1991, p. 100-116.
- TVERSKY A., KAHNEMAN D., "Belief in the law of small numbers", *Psychological Bulletin* 76, 1971, p. 105-110.
- WILKINSON L., and Task Force on Statistical Inference, APA Board on Scientific Affairs, "Statistical methods in psychology journals: Guidelines and explanations", *American Psychologist* 54(8), 1999, p. 594-604.
- WINER B. J., *Statistical Principles in Experimental Designs*, (2^e édition) New York, McGraw-Hill, 1971, (1^e édition, 1962).
- WINKLER R. L., "Statistical analysis: Theory versus practice", C.-A. S. Stael Von Holstein (ed.), *The Concept of Probability in Psychological Experiments*, Dordrecht, Reidel, 1974.