L'Année psychologique, 2000, *100*, 683-713

REVUES CRITIQUES

Laboratoire de Mathématiques Raphaël-Salem,
CNRS UPRESA 6085,
Université de Rouen[1] *
LCPE, InaLF, FRE2173,
CNRS, Paris[2] **

ALLER AU-DELÀ DES TESTS DE SIGNIFICATION TRADITIONNELS : VERS DE NOUVELLES NORMES DE PUBLICATION

par Bruno LECOUTRE* et Jacques POITEVINEAU**

SUMMARY : Beyond traditional significance tests : Prime time for new publication norms.

There are good reasons to think that the role of usual null hypothesis significance testing in psychological research will be considerably reduced in the near future. Traditional statistical analysis results should be enhanced (« beyond simple p value statements ») to systematically include effect sizes and their interval estimates. Quite soon, these procedures could become new publication norms. *In this paper main abuses of significance tests and alternative available solutions are first reviewed. Among these solutions, both confidence interval* (frequentist) *methods and credibility interval* (fiducial Bayesian) *methods have been developed for assessing effect sizes, and especially for asserting the negligibility or the notability of effects. From a numerical example, these methods are illustrated for analysing contrasts between means in a complex experimental design. Both* raw *and* relative *(calibrated) effects are considered. The similarities and differences between the frequentist and Bayesian approaches, their correct interpretations, and their practical uses, are discussed.*

Key words : effect size, raw and relative effects, statistical inference, significance tests, confidence intervals, bayesian methods.

1. Mathématiques Site Colbert, 76821 Mont-Saint-Aignan Cedex. E-mail : bruno.lecoutre@univ-rouen.fr

2. 44, rue de l'Amiral-Mouchez, 75014 Paris. E-mail : jacques.poitevineau@ivry.cnrs.fr

INTRODUCTION

En dépit des critiques les plus sévères dont elle a toujours fait l'objet, l'utilisation des tests de signification était jusqu'à ce jour une quasi-obligation pour publier des résultats. Or, pour la première fois en psychologie, une prise de position *officielle* à l'encontre de l'usage actuel des tests de signification traditionnels se dessine. Elle émane du bureau des affaires scientifiques de l'*American Psychological Association* qui a chargé « un détachement spécial » *(Task Force)* d'étudier le rôle du test de signification dans la recherche en psychologie (APA, 1996). Un premier rapport aboutit à la conclusion que l'usage du test de signification ne doit pas être interdit, mais fait aussi expressément les recommandations suivantes, qui en modifient considérablement le statut : l'ouverture à d'autres méthodes d'analyse des résultats, entre autres les méthodes bayésiennes et les méthodes d'analyse des données graphiques et exploratoires ; le rapport systématique de la grandeur des effets observés et des intervalles de confiance correspondants ; la reconnaissance des études exploratoires bien formulées et bien conduites avec des traitements quantitatifs appropriés des résultats (en réaction contre les abus de la démarche hypothético-déductive) ; l'application du principe de parcimonie au choix des plans d'expérience et des analyses.

En ce qui concerne plus particulièrement la présentation habituelle des procédures d'inférence statistique, la recommandation est que « ... *enhanced characterization of the results of analyses (beyond p value statements) to include both direction and size of effects (*e.g., *mean difference, regression and correlation coefficients, odds-ratios, more complex effect size indicators) and their confidence intervals should be provided routinely as part of the presentation. These characterizations should be reported in the most interpretable metric (*e.g., *the expected unit change in the criterion for a unit change in the predictor, Cohen's* d*).* » Cette prise de position peut être considérée comme un événement, au sens où elle a rapidement suscité de nombreuses réactions et où son impact a rapidement dépassé le domaine de la psychologie (voir par ex. Hinkley, 1997).

Il y a donc de bonnes raisons de penser que ces recommandations pourraient devenir rapidement effectives dans les revues de psychologie, et qu'il faudra bientôt changer les habitudes de publication en présentant des procédures allant au-delà des tests de signification traditionnels[1]. Cette note

1. Depuis que le présent article a été accepté, les recommandations de la *Task Force* ont donné lieu à la publication d'un document détaillé (Wilkinson and Task Force on Statistical Inference, 1999). Ce document, ouvert à commentaires, a pour but explicite d'introduire dans le manuel de publication de l'APA de nouvelles directives normatives sur l'usage des méthodes statistiques dans les revues de psychologie. Le texte initial a été considérablement remanié, mais sans que cela remette en question nos commentaires sur la première version.

sera consacrée à une présentation générale des procédures d'inférence statistique qui devraient être utilisées *en plus* (ou *à la place*) des tests de signification usuels. Après avoir rappelé quelques principes généraux concernant la mesure de l'intensité des effets (absolus et relatifs), nous présenterons essentiellement ici les méthodes d'estimation par intervalle. Ces méthodes ont été développées, à la fois dans le cadre fréquentiste *(intervalle de confiance)* et dans le cadre bayésien *(intervalle de crédibilité)*. Ces deux approches, comme nous le verrons, fournissent des justifications et des interprétations différentes ; c'est pour marquer ces différences qu'on parle le plus souvent d'intervalle de crédibilité dans le cadre bayésien. Dans ce qui suit, quand nous utiliserons simplement « intervalle », cela renverra simultanément aux deux approches.

Nous nous limiterons ici à illustrer ces procédures dans le cas du traitement de données numériques par des techniques d'analyse de variance ; nous mentionnerons simplement ici que des solutions analogues existent pour les coefficients de corrélation (cf. Lecoutre, 1996*b* ; Lee, 1997) et pour les données catégorisées (cf. pour le cadre bayésien[1] Bernard, 1986, 1998 ; Lecoutre, Derzko et Grouin, 1995 ; Lecoutre et Charron, 2000). Nous rappellerons d'abord brièvement quelques abus d'utilisation des tests de signification usuels. Par test de signification usuel, nous entendrons ici le test qu'un effet est égal à zéro, auquel renvoie maintenant l'appellation consacrée « Null Hypothesis Significance Testing » utilisée dans la *Task Force*[2]. Puis nous passerons en revue des méthodes effectivement disponibles et acceptables, et nous en rappellerons l'interprétation correcte.

I. LES SOLUTIONS DE RECHANGE

1. Les abus d'utilisation des tests de signification

Le test de signification usuel ne dit rien quant à l'intensité, l'importance de l'effet parent (cf., par exemple, O'Brien et Shapiro, 1968 ; Rouanet, Lépine et Pelnard-Considère, 1976). C'est pour remédier à cette insuffisance méthodologique fondamentale que les chercheurs ont depuis longtemps commis deux abus principaux d'utilisation, qui peuvent en fait être considérés comme des « ajustements de jugement » (Bakan, 1966 ;

1. Dans le cadre fréquentiste, de nombreuses procédures d'intervalles de confiance ont été proposées pour l'analyse des tableaux de contingence, mais il n'existe pas à notre connaissance de synthèse facilement accessible.

2. Cette acception de *null hypothesis* correspond à l'usage courant, mais est restrictive. Il faut rappeler que pour Fisher, il s'agit de l'hypothèse à réfuter *(to be nullified)*, et non nécessairement, comme on le trouve parfois écrit, de l'hypothèse d'une valeur zéro pour le paramètre testé.

Phillips, 1973, p. 334) ou des « biais adaptatifs » (M.-P. Lecoutre, 1998) par rapport à une norme inadaptée.

Le premier abus est de confondre la significativité statistique avec la significativité scientifique ou substantielle. C'est 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. Cette erreur a été dénoncée très souvent, et depuis longtemps (voir, par exemple, Boring, 1919 ; Selvin, 1957 ; Kish, 1959 ; Bolles, 1962 ; Bakan, 1966 ; O'Brien et Shapiro, 1968 ; Gold, 1969 ; Morrison et Henkel, 1969 ; Winch et Campbell, 1969). D'une manière implicite, c'est contre elle que Reuchlin (1962, p. 370) met en garde le psychologue, lorsqu'il insiste sur le fait que c'est à celui-ci, et non au statisticien, de décider des hypothèses statistiques à tester ; c'est au psychologue de savoir si, du point de vue de la signification psychologique, il ne vaut pas mieux choisir pour hypothèse nulle qu'entre les moyennes de deux groupes la différence est inférieure à un point (pour une certaine échelle), plutôt qu'exactement égale à zéro.

Le second abus est de conclure à la véracité de l'hypothèse nulle en cas de résultat non significatif sur la seule base du risque de première espèce. Harcum (1990) donne des exemples d'acceptations « désinvoltes » d'hypothèses nulles dans des revues prestigieuses. Poitevineau (1998) passe en revue les articles publiés dans le *Journal of Abnormal Psychology* au cours de l'année 1994 et trouve dans environ la moitié des articles des conclusions telles que « il n'y a pas d'effet du facteur A » ou « il n'y a pas de différence entre les groupes ». Il montre que, même si « pas d'effet » est compris comme « effet faible ou négligeable », de telles conclusions sont généralement non fondées.

Les tests usuels sont en fait inadaptés à la nécessité de pouvoir mettre en évidence pour l'effet testé : soit une intensité, ou grandeur, faible ou *négligeable*, c'est-à-dire une valeur qui, si elle n'est pas strictement nulle, pourra être tenue pour suffisamment faible pour constituer une bonne approximation du zéro (au moins à un certain stade de la recherche) ; soit une intensité forte ou *notable*, c'est-à-dire, au contraire du cas précédent, importante, ou tout au moins impossible à négliger. Bien entendu, il peut se faire que l'intensité d'un effet ne soit ni négligeable, ni notable, c'est-à-dire qu'elle soit intermédiaire, moyenne. Cohen (1962, 1988) parle d'effet petit *(small)*, moyen *(medium)* ou grand *(large)*[1].

Cette incapacité des tests usuels à traiter le problème de l'intensité des effets fait notamment qu'ils sont inadaptés à la validation de modèles (Rouanet, 1967, 1986 ; Rouanet, Lépine et Holender, 1978) ; c'est ce que souligne encore récemment Bacher (1999), à propos des modèles structu-

1. Pour Cohen, un effet petit n'est pas nécessairement négligeable, ni nécessairement non négligeable d'ailleurs : c'est un effet difficile à déceler mais qui existe, alors que la notion de négligeabilité englobe celle d'un effet existant mais d'intensité inférieure à une certaine limite, aussi bien que celle d'un effet nul (inexistant).

raux. Même un auteur comme Frick (1996), qui défend l'utilité de ces tests dans certaines conditions, ne peut que partager ce point de vue. C'est d'ailleurs en niant l'intérêt d'étudier l'intensité des effets que l'un des plus ardents défenseurs des tests de signification usuels (Chow, 1988, 1996) justifie sa position.

Devant ces difficultés, il se trouve maintenant des partisans d'un bannissement pur et simple des tests de signification dans les publications ; Ceux-ci mettent en avant le « choc thérapeutique » que cela provoquerait (Shrout, 1997). Certains auteurs comme Hogben (1957) sont même allés plus loin et ont recommandé l'abandon de toute méthode d'inférence statistique. Cependant les méthodes d'inférence statistique sont souhaitables, car elles constituent un garde-fou indispensable pour éviter au chercheur (ou au lecteur d'une publication) de se laisser emporter par les interprétations spontanées pouvant conduire à des généralisations hâtives infondées.

Pour traiter le problème de l'intensité des effets, l'usage systématique des estimations par intervalle proposé par la *Task Force* est effectivement la solution de rechange qui est de loin la plus souvent recommandée. On notera ici que l'étude de la puissance pour obtenir une conclusion sur l'importance d'un effet n'est pas retenue par la *Task Force*[1], sans doute parce qu'elle est maintenant désapprouvée par les statisticiens : la puissance peut être un guide utile pour planifier une expérience (choix des effectifs avant le recueil des observations), mais elle ne doit pas être utilisée pour interpréter les données (voir, par exemple, Schuirman, 1987 ; Goodman et Berlin, 1994).

2. La mesure de l'intensité des effets

La mesure de l'intensité des effets apparaît incontournable (voir en particulier : Yates, 1951 ; Nunnally, 1960 ; Cohen, 1962, 1988, 1990 ; Hays, 1963 ; Bakan, 1966 ; Vaughan et Corballis, 1969 ; Dwyer 1974 ; Craig, Eison et Metze, 1976 ; Cox, 1977 ; Carver, 1978 ; Guttman, 1983 ; Lecoutre, 1984, 1996*a* ; Harris, 1991 ; Rogers, Howard et Vessey, 1993 ; Rouanet, 1996 ; Schmidt, 1996), sans pour autant, bien sûr, assimiler l'intérêt d'un effet à sa grandeur (voir, par exemple : O'Grady, 1982 ; Rosenthal, 1990). Cette mesure est vue soit comme un prolongement, soit comme un remplacement de la procédure de test. Elle a été abordée de façons très différentes, et c'est une des raisons pour lesquelles on rencontre différents termes : intensité, taille *(size)*, ampleur *(magnitude)*, grandeur, importance, que nous traiterons ici comme équivalents[2]. Loin de s'opposer aux approches décrites dans les sections suivantes, elle en constitue au contraire un préalable.

1. Ceci malgré le fait que son plus ardent défenseur en psychologie, Cohen, soit un des cosignataires du rapport.

2. Mais on pourrait réserver le terme *importance* pour les aspects qualitatifs de la conclusion et les autres termes pour ses aspects quantitatifs.

La recommandation de la *Task Force* est l'utilisation routinière d'indicateurs relatifs de la grandeur de l'effet observé, tels que le *d* de Cohen, qui consiste à rapporter l'effet brut (par ex., une différence de moyennes) à l'écart type « d'erreur » qui lui est associé dans l'analyse de variance, ce qui permet d'obtenir un effet *standardisé* ou *calibré*[1]. On remarquera d'ailleurs que dans la littérature anglo-saxonne le terme *effect size* est presque toujours entendu comme *grandeur relative de l'effet*. Un tel indicateur est par définition indépendant de l'unité de mesure et présente ainsi l'intérêt de pouvoir comparer des effets portant sur des variables différentes (cf. Rouanet, Lépine et Pelnard-Considère, 1976).

Mais l'utilisation des indicateurs relatifs appelle un certain nombre de réserves. Ainsi, dans le cas du *d* de Cohen, l'écart type d'erreur apparaissant seul au dénominateur, l'effet relatif augmente dès que, ce qui est tout de même souhaitable, cet écart type d'erreur diminue, même si l'effet absolu reste très faible. Ces réserves sont encore accentuées en ce qui concerne l'utilisation des indicateurs *en part de variance expliquée*, notamment le coefficient de différenciation η^2 de K. Pearson et le coefficient ω^2 de Hays (1963), dont l'idée est de mesurer l'effet comme la proportion de variance qui lui est imputable par rapport à la variance totale. Pour ces indicateurs, un même facteur peut voir son importance augmenter d'une expérience à l'autre, simplement parce que la variabilité intragroupe est mieux contrôlée. D'autre part, pour une même variable dépendante, la nature des facteurs retenus dans le plan d'analyse ou contrôlés influence le résultat et la part de variance expliquée par tel facteur n'existe pas dans l'absolu (Oakes, 1986, p. 64). Par exemple, admettons que les facteurs A et B aient des effets additifs. On s'intéresse à l'effet de A, mais dans un cas on fait varier simultanément les deux facteurs alors que dans un second cas on opère à un niveau fixé du facteur B. Toutes choses égales par ailleurs, la variance totale sera plus grande dans le premier cas et le coefficient (η^2 ou ω^2) sera plus faible. Le résultat peut encore être fortement affecté par le choix des niveaux des facteurs (Levin, 1967), la fidélité des mesures (O'Grady, 1982). Le seul fait que ces coefficients puissent s'exprimer comme un pourcentage de variance est donc loin d'assurer leur comparabilité d'une étude à l'autre. Plus fondamentalement, Oakes (1986, p. 62-63) critique l'utilisation d'indicateurs relatifs car ils incitent le psychologue à « ne pas prendre au sérieux » les variables utilisées (et leurs unités), alors même que pour lui une tâche primordiale est justement de donner sens à ces variables. D'une manière générale, un indicateur relatif peut d'ailleurs ne pas donner une bonne image de l'importance *réelle* de l'effet (voir pour un exemple Rosenthal et Rubin, 1982). Ces remarques montrent l'utilité

1. Cette recommandation n'a cependant pas été reprise dans les nouvelles directives, qui privilégient l'utilisation d'un indicateur non standardisé : « If the units of measurement are meaningful on a practical level (*e.g.*, number of cigarettes smoked per day), then we usually prefer an unstandardized measure (regression coefficient or mean difference) to a standardized measure (*r* or *d*). »

(sinon la nécessité) de rapporter pour chaque effet observé un indicateur de l'effet brut observé (ce que soutiennent par exemple, Vaughan et Corballis, 1969), même dans le cas où on privilégie un indicateur relatif pour les analyses inférentielles.

Quel que soit l'indicateur utilisé, on peut se demander s'il est considéré seulement comme une description de l'effet observé ou comme une estimation d'un effet *vrai* (parent), auquel cas se pose le choix du « meilleur estimateur » (cf. Richardson, 1996). 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, on sait bien qu'un effet observé faible associé à un résultat non significatif est souvent perçu par le chercheur comme étant en faveur de l'absence d'effet vrai (M.-P. Lecoutre, 1998), alors qu'il n'est souvent qu'un constat d'ignorance. Il est donc nécessaire, comme nous l'avons dit en introduction, de fournir une estimation par intervalle pour l'effet vrai. En ce cas, la question de trouver la « meilleure » estimation ponctuelle apparaît généralement secondaire et peut par conséquent être évitée.

Enfin, dès le niveau *descriptif*, intervient le choix des critères pour juger de l'importance des effets, notamment des critères de négligeabilité/notabilité. Manifestement, ce choix dépendra des circonstances et des connaissances qu'on a du domaine, et contiendra une part d'arbitraire. C'est souvent l'importance respective des effets les uns par rapport aux autres qui sera un critère essentiel. Cependant, dans le cas d'effets relatifs, il est maintenant assez courant d'utiliser comme repères les conventions proposées par Cohen (1962, 1988), en y apportant éventuellement quelques modifications (pour plus de détails, cf. Corroyer et Rouanet, 1994). Une autre approche, illustrée par Haase, Waechter et Solomon (1982), est de fournir une base empirique de référence par la compilation d'un très grand nombre de résultats publiés, en l'occurrence les articles parus dans le *Journal of Counseling Psychology* de 1970 à 1979. Ils considèrent pour cela la distribution des 11 044 coefficients η^2 (part de variance expliquée par le facteur expérimental) calculés à partir des tests statistiques fournis dans les articles. Ils proposent cette distribution comme base de comparaison pour évaluer grossièrement de nouveaux résultats dans un domaine comparable à celui de la psychologie de *counseling*.

3. L'INTERVALLE DE CONFIANCE

3.1. *La méthode la plus souvent proposée*

L'intervalle de confiance, au sens *fréquentiste* (Neyman et Pearson), est incontestablement la méthode la plus souvent proposée pour pallier les insuffisances des tests usuels : voir, par exemple, Natrella (1960) ; Nun-

nally (1960) ; Rozeboom (1960) ; Grant (1962) ; LaForge (1967) ; Carver (1978) ; Oakes (1986) ; Casella et Berger (1987) ; Evans, Mills et Dawson (1988) ; Cohen (1994) ; Loftus et Masson (1994) ; Schmidt (1996). Il n'est maintenant plus rare de voir mentionnés des intervalles de confiance, en complément des tests, dans des journaux comme le *Journal of Abnormal Psychology*. Des journaux scientifiques, notamment dans le domaine médical, ont d'ailleurs déjà publié des éditoriaux préconisant l'utilisation systématique des intervalles de confiance : par exemple, Lutz et Nimmo (1977) ; Rothman (1978) ; Berry (1986) ; Evans, Mills et Dawson (1988) ; Braitman (1988, 1991) ; Loftus (1993) ; Falissard et Landais (1995). Il n'est donc pas surprenant que le rapport de la *Task Force* recommande également d'utiliser routinièrement des intervalles de confiance.

Le plus simple est d'utiliser les intervalles de confiance usuels, qui sont relativement familiers dans les cas élémentaires d'inférence sur une moyenne ou sur la différence de deux moyennes. Ainsi, dans ce dernier cas, on obtient comme cela est bien connu un intervalle symétrique centré sur la différence observée. Tout naturellement, cet intervalle est d'autant plus étroit que la précision expérimentale, qui dépend des variances et des effectifs, est plus grande ; il reflète donc directement et explicitement le rôle des effectifs. Il inclut en outre la procédure décisionnelle du test de signification usuel de l'hypothèse nulle selon laquelle la différence vraie δ est égale à 0 : ce test est significatif au seuil bilatéral α si et seulement si l'intervalle de confiance $1 - \alpha$ ne contient pas la valeur 0[1]. En revanche, il ne fournit pas d'indication sur la valeur du seuil observé p (autre que de situer p par rapport à α). Rapporter à la fois p et un intervalle de confiance suppose donc d'utiliser simultanément deux procédures d'inférence distinctes.

3.2. *Est-ce l'intervalle le mieux approprié ?*

Mais l'intervalle de confiance usuel n'est pas directement approprié à la problématique de l'importance de l'effet. En particulier, montrer qu'une différence est négligeable requiert un intervalle centré sur zéro (du type $[-x, +x]$ avec $x > 0$) et non sur la valeur particulière observée (soit encore un intervalle pour la valeur absolue de δ). La construction d'un tel intervalle, ou ce qui revient au même la construction d'un test de l'hypothèse nulle $H_0 : |\delta| > x$ (que l'on veut rejeter) contre $H_1 : |\delta| \leq x$ (où $x > 0$), est possible. Mais elle a une longue histoire, qui révèle bien des difficultés.

Ainsi Serlin et Lapsley (1985, 1993) traitent de la validité approximative des hypothèses (le principe du *good enough*) et proposent aux psychologues une procédure de test qui paraît s'imposer, dans la mesure où elle satisfait les critères formels habituels de choix des tests fréquentistes (tests uniformément plus puissants, tests invariants). Mais les auteurs semblent

1. Plus généralement, l'intervalle de confiance usuel est l'ensemble des valeurs δ_0 telles que le test de l'hypothèse nulle « $\delta = \delta_0$ » est non significatif au seuil bilatéral α.

ignorer qu'elle a été proposée par ailleurs à différentes reprises de manière indépendante (notamment Bondy, 1969 ; Anderson et Hauck, 1983 ; Fowler, 1984, 1985 ; Patel et Gupta, 1984 ; Rocke, 1984 ; Wellek et Michaelis, 1991), et qu'elle a toujours été abandonnée, en raison de propriétés indésirables qui la rendent inacceptable (cf. Schuirmann, 1987 ; Schervish, 1995, p. 252 ; Lecoutre, 1996*a*, p. 225).

Cependant, des solutions beaucoup plus raisonnables ont été proposées et sont maintenant accessibles aux psychologues. Dans une courte note, Bartko (1991) fournit des références de biostatistique, domaine où l'on s'est préoccupé du problème depuis longtemps, dans le cadre des essais cliniques de bioéquivalence en pharmacologie. Rogers, Howard et Vessey (1993) présentent et recommandent aux psychologues une procédure d'usage courant dans ce domaine. Elle consiste à conclure $| \delta | < x$ si chacun des deux tests unilatéraux, $H_{01} : \delta = -x$ *vs* $H_{11} : \delta > -x$ et $H_{02} : \delta = x$ *vs* $H_{12} : \delta < x$, est significatif au seuil α. Cette procédure a été discutée sous l'appellation de *two one-sided tests*, notamment par Schuirmann (1987). On en déduit aisément une procédure très simple de construction d'un intervalle de confiance centré sur 0 (Deheuvels, 1984). Cette procédure, développée à l'origine pour la comparaison de deux moyennes, a été généralisée pour s'appliquer à toute comparaison de moyennes à un ou plusieurs degrés de liberté en analyse de variance et fournit donc une solution générale pour démontrer la négligeabilité d'un effet (Lecoutre et Derzko, 1999).

3.3. *La duplicité de l'utilisation de l'intervalle de confiance*

S'il n'est pas dans notre propos ici de polémiquer, il nous paraît cependant nécessaire de mettre en garde le lecteur contre la duplicité à laquelle risque de conduire l'utilisation de l'intervalle de confiance. Comme le test de signification, l'intervalle de confiance relève d'un cadre de justification *fréquentiste*. Dans ce cadre, les bornes observées de l'intervalle de confiance pour l'échantillon (unique) dont on dispose, de même que le seuil observé p, ne sont interprétables qu'en référence à l'ensemble de tous les intervalles qu'on aurait pu observer[1]. Supposons pour fixer les idées que dans une expérience l'on ait obtenu l'intervalle de confiance 0,95 [+ 1,58, + 2,64] pour une différence de moyenne δ. Formellement, les bornes de l'intervalle de confiance pour le paramètre δ sont des grandeurs aléatoires, qui varient d'un échantillon à un autre. L'interprétation (fréquentiste) *correcte* de l'intervalle de confiance 0,95 est alors la suivante : « 95 % des intervalles calculés sur l'ensemble des échantillons possibles contiennent la vraie

1. Mais que l'on n'a pas observés. Baser l'inférence sur les événements qui ne sont pas produits peut paraître pour le moins paradoxal, comme le souligne ironiquement Jeffreys à propos du seuil p (1961, p. 385) : « What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. »

valeur δ »[1]. Mais cet énoncé est *conditionnel à* δ : il ne dépend pas des observations et est déterminé avant leur recueil. Dans le cadre fréquentiste les valeurs possibles du paramètre *ne sont pas probabilisées*. L'événement « 1,58 < δ < + 2,64 », associé aux bornes observées dans une expérience particulière, est vrai ou faux (puisque δ est fixé), et nous ne pouvons pas lui attribuer de probabilité (sinon 1 ou 0). Il est donc *illégitime* d'écrire « Pr (+ 1,58 < δ < + 2,64) = 0,95 » ou d'énoncer que « il y a 95 % de chances que la différence inconnue δ soit comprise entre + 1,58 et + 2,64 ». Pourtant votre intuition profonde est peut-être qu'une telle interprétation est, soit correcte, soit peut-être incorrecte mais en tout cas souhaitable ; cette interprétation est en tout cas parfaitement naturelle (Kadane, 1995). Si tel est le cas vous devez sérieusement vous demander si vous n'êtes pas un bayésien « qui s'ignore » (Lecoutre, 1997). En effet, cette interprétation naturelle n'est légitime que dans le cadre bayésien.

Par ailleurs la conception fréquentiste de la confiance (et du seuil α du test) comme une proportion calculée sur un très grand nombre de répétitions identiques impose que la procédure particulière utilisée (de même que la confiance) soit choisie *indépendamment des données*, et non pas en fonction des résultats observés. En conséquence, dans ce cadre il est par exemple illégitime de rapporter, *au vu des données*, l'intervalle de confiance usuel quand l'effet observé est notable et un intervalle centré sur 0 quand cet effet est négligeable. Dans la pratique, l'utilisateur a donc le choix entre trois attitudes : 1 / conserver le cadre de justification fréquentiste de l'intervalle de confiance et se satisfaire alors de son interprétation et des contraintes qu'impose son usage *honnête* ; 2 / conserver ce cadre tout en adoptant l'interprétation bayésienne alors *erronée* ; 3 / 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 notera d'ailleurs que la même duplicité apparaît dans le cas du test de signification, pour lequel le seuil de signification est souvent interprété comme la probabilité « que l'hypothèse nulle soit vraie » alors qu'il est la probabilité *conditionnelle* « de rejeter [à tort] l'hypothèse nulle si cette hypothèse est vraie ». On peut donc penser que ce sont paradoxalement leurs interprétations bayésiennes sauvages qui rendent ces procédures populaires.

4. Les méthodes bayésiennes

Un certain nombre d'auteurs ont conseillé l'utilisation de méthodes bayésiennes (ou apparentées) en psychologie (cf. Rozeboom, 1960 ; Edwards, Lindman et Savage, 1963 ; Edwards, 1965 ; Bakan, 1966 ; Corn-

1. Ainsi, si l'on répétait l'expérience un nombre infini de fois, $100(1-\alpha)$ % des intervalles calculés contiendraient la vraie valeur du paramètre. Pour une expérience particulière, l'utilisateur affirme simplement 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 100α % des fois (voir, par ex., Neyman, 1952, p. 196, 209-210).

field, 1966 ; Wilson, Miller et Lower, 1967 ; Phillips, 1973 ; Novick et Jackson, 1974 ; Winkler, 1974 ; Rouanet, Lépine et Pelnard-Considère, 1976 ; Rouanet, Lépine et Holender, 1978 ; Hoc, 1983 ; Rouanet et Lecoutre, 1983 ; Lecoutre, 1984, 1985, 1996*a* ; Lewis, 1993 ; Rouanet, 1996 ; Pruzek, 1997 ; Rindskopf, 1997, 1998 ; Rouanet *et al.*, 1998-1991). Le rapport de la *Task Force*, sans recommander particulièrement ces méthodes, reconnaît leur usage : « It is the view of the task force that there are many ways of using statistical methods to help us understand the phenomena we are studying (*e.g. Bayesian methods*, graphical and exploratory data analysis methods, hypothesis testing strategies). We endorse a policy of *inclusiveness* that allows any procedure that appropriately sheds light on the phenomenon of interest to be included in the arsenal of the research scientist » [les italiques sont de nous].

Dans l'inférence bayésienne, au contraire de l'inférence fréquentiste, les valeurs possibles du paramètre (par exemple la différence de moyennes δ) sont explicitement probabilisées. À partir d'un état de connaissance *initial* formalisé par une distribution initiale *(a priori)*, et des données, la formule de Bayes permet d'obtenir une distribution finale *(a posteriori)* qui exprime directement l'incertitude sur le paramètre, *conditionnellement à l'échantillon particulier observé.* Cette distribution combine l'information initiale avec l'information apportée par les données. Nous pouvons en déduire la probabilité que le paramètre appartienne à une zone de valeurs spécifiée pour le paramètre, que l'on appelle habituellement intervalle de *crédibilité* pour le distinguer de l'intervalle de confiance. Par définition, nous attribuons à cet intervalle de crédibilité la probabilité considérée (par ex. 0,95) : étant donné les observations, nous avons une probabilité 0,95 (95 % de chances) que la proportion vraie soit comprise entre les bornes de l'intervalle. Le cadre bayésien, en fournissant une inférence conditionnelle à l'échantillon, permet de choisir le type d'intervalles au vu des données. Il permet donc de concevoir l'utilisation des procédures inductives (généralisantes) comme un prolongement le plus naturel possible des conclusions descriptives portant sur l'échantillon (Lecoutre, 1984 ; Rouanet, 1996).

De plus l'approche bayésienne non informative, dite aussi standard ou encore fiducio-bayésienne[1], selon la terminologie forgée par Rouanet, Lépine et Pelnard-Considère (1976) est une réponse directe (voir aussi l'importante note technique de l'article de Rozeboom, 1960) à l'accusation de subjectivité, qui est l'argument principal (par exemple : Wilson, Miller et Lower, 1967 ; Frick, 1996) utilisé à l'encontre de la méthode bayésienne. Cette approche fournit des procédures *objectives*, bien adaptées à la spécificité de l'analyse des données expérimentales, qui apparaissent de plus en

1. Cette appellation, en faisant explicitement référence à la méthode *fiduciaire* de Fisher (Fisher, 1948, 1990-1956), répare l'injustice dont certains critiques comme Cohen (1994) et Frick (1996) font preuve envers celui-ci. Elle reconnaît à Fisher son souci de fournir une solution exprimant l'apport propre des données et formulée en termes de probabilités sur le paramètre.

plus être une solution de rechange prometteuse. Ces procédures peuvent d'ailleurs être utilisées aussi facilement que les tests de signification ou les intervalles de confiance traditionnels, comme l'ont montré les travaux réalisés en France depuis maintenant plus de vingt ans (cf. Lépine et Rouanet, 1975 ; Lecoutre, 1978 ; Lecoutre et Lecoutre, 1979). Elles sont basées sur des définitions opérationnelles plus utiles (Goodman et Berlin, 1994), et sont au moins aussi objectives que les autres techniques statistiques (Berger, 1985, p. 110 ; voir aussi Bernard, 1996). Leur utilisation est d'ailleurs acceptée par les revues psychologiques : voir par exemple Hoc et Leplat (1983) ; Ciancia *et al.* (1988) ; Lecoutre (1992) ; Hoc (1996) ; Clément et Richard (1997) ; sans compter de très nombreux articles publiés en français, et en tout premier lieu dans *L'Année Psychologique.*

Dans d'autres domaines, et tout particulièrement dans celui des essais cliniques en médecine et en pharmacologie, les méthodes bayésiennes sont de plus en plus développées et connues. Des propositions méthodologiques voisines de celles que nous préconisons ont été avancées pour leur utilisation de routine, par exemple par Spiegelhalter, Freedman et Parmar (1994). Ces auteurs suggèrent également d'utiliser, en plus de l'approche non informative, des distributions *a priori* « sceptiques » ou « enthousiastes » pour éprouver la robustesse des conclusions vis-à-vis d'informations extérieures aux données (voir aussi Lecoutre, 1996*a*, chap. 3 ; Lecoutre *et al.*, 2000). Ils insistent sur le fait que la motivation pour utiliser la méthodologie bayésienne est davantage pratique qu'idéologique.

II. EXEMPLE NUMÉRIQUE : MISE EN ŒUVRE DES PROCÉDURES

Nous considérerons un exemple typique d'analyse de variance à mesures répétées mettant en jeu plusieurs facteurs. Nous verrons, comment à partir d'une décomposition appropriée en comparaisons à un degré de liberté, il est possible de répondre de manière précise aux objectifs de l'expérience, avec des procédures simples à mettre en œuvre. D'un point de vue pratique, nous illustrerons les calculs d'intervalles, fréquentistes et fiducio-bayésiens, pour les effets bruts et pour les effets calibrés, ainsi que leurs différences d'interprétations. Nous verrons qu'il est immédiat de déduire les formules donnant ces intervalles à partir des tableaux de moyennes et des rapports F usuels de l'analyse de variance ; le calcul nécessite seulement l'utilisation d'une table de la distribution usuelle du *t* de Student pour les effets bruts et d'un programme d'une distribution que nous appelons *lambda-prime* pour les effets relatifs. Plus directement, le programme PAC (Lecoutre et Poitevineau, 1992) fournit tous les résultats, descriptifs et inférentiels[1].

1. Une version limitée, PAC Junior, qui permet d'obtenir tous les résultats présentés dans cet article, est disponible gratuitement sur internet à l'adresse http://www.cisia.com/cisia/logiciels/pac.htm ou sur demande aux auteurs.

1. Présentation de l'exemple

Considérons une expérience de décision lexicale à partir de la présentation visuelle sérielle rapide de phrases, réalisée par Aguilar et Denhière (1998). Les facteurs retenus pour l'analyse sont les suivants. Les sujets sont répartis aléatoirement en trois groupes de 30 sujets chacun, caractérisés par la *Durée de présentation* des mots de la phrase (d1 = 150 ms, d2 = 250 ms et d3 = 350 ms). Ils sont soumis à six occasions de mesure, définies par le croisement des deux facteurs *Nature* de la phrase à deux modalités (n1 = phrase dénotant une action et n2 = phrase dénotant un événement), et *Forme* de la phrase à trois modalités (f1 = explicite, f2 = prédictible et f3 = contrôle). Le plan d'analyse peut donc s'écrire : $S_{30} < D_3 > \times N_2 \times F_3$. La variable dépendante est la moyenne, pour chacune des occasions, des temps de décision lexicale à quatre phrases, exprimée en millisecondes. La question principale de cette expérience est de savoir si la forme de phrase prédictible conduit à un temps de décision plus court que la forme contrôle, ceci en fonction de la durée de présentation.

Le tableau I fournit les moyennes observées. L'analyse de variance traditionnelle, pour les sources de variation canoniquement associées aux trois facteurs du plan d'expérience, révèle des effets significatifs, au seuil 0,05, pour les trois effets principaux, ainsi que pour l'interaction N.D.

Tableau I. — *Moyennes observées*

Observed means

		f1	f2	f3	
d1	n1	893,93	910,47	923,17	909,19
	n2	889,93	918,93	947,53	918,80
		891,93	914,70	935,35	
d2	n1	802,33	847,33	862,43	837,37
	n2	812,23	839,47	881,67	844,46
		807,28	843,40	872,05	
d3	n1	754,63	803,13	859,70	805,82
	n2	763,20	833,83	891,17	829,40
		758,92	818,48	875,43	ms

2. Décomposition en comparaisons à un degré de liberté

La décomposition canonique des sources de variation (effets principaux et interaction) répond souvent imparfaitement aux questions posées dans l'expérience. La pratique d'analyses *complémentaires* – décompositions des effets à plusieurs degrés de liberté, « analyses en sous-plans », etc. – est

donc très répandue. On sait que des réserves méthodologiques sont souvent émises à l'égard de cette pratique ; il est cependant manifeste que ces réserves sont étroitement liées à l'usage exclusif des tests de signification, et traduisent une méfiance (d'ailleurs souvent justifiée) à l'égard de la « pêche aux résultats significatifs ». Clairement, prendre en compte explicitement l'intensité des effets avec l'usage d'intervalles prémunit contre des interprétations hâtives et autorise à largement minimiser ces réserves. Cela a pour conséquence une souplesse beaucoup plus grande dans le choix d'une décomposition appropriée, et devrait notamment permettre d'éluder le plus souvent le problème des comparaisons *a posteriori*.

Nous adopterons ici la décomposition suivante. Pour un facteur numérique tel que le facteur *Durée de présentation*, il est généralement approprié de considérer les composantes de la régression polynomiale ; nous retiendrons donc ici la composante linéaire (comparaison notée *lin* D) et sa résiduelle (*-lin* D). La structure du facteur *Forme* de la phrase, associée aux objectifs de l'expérience, conduit à comparer chacune des formes expérimentales (explicite et prédictible) à la forme contrôle, d'où les comparaisons notées f3, f1 et f3, f2. Pour la décomposition du croisement F × D, nous considérerons ces deux comparaisons conditionnellement à chacune des modalités de D, qui seront plus directement interprétables que les comparaisons d'interaction partielles.

Par suite chaque effet analysé sera à un degré de liberté. L'effet brut correspondra à une différence de moyennes (par ex., f3, f1 ou f3, f1/d1) ou à une différence de différences de moyennes (par ex., l'interaction n2, n1 . f3, f1) ou encore à la pente d'une droite de régression (*lin* D), c'est-à-dire de manière générale à un contraste entre moyennes. Suivant la convention usuelle en inférence statistique, nous utilisons des lettres grecques pour les paramètres et des lettres romanes pour les statistiques. Généralisant la notation utilisée précédemment pour une différence de moyennes, l'effet (brut) vrai du contraste sera noté δ et son effet observé sera noté d (qu'on ne confondra bien entendu pas avec le « d de Cohen »). De même les notations σ et s seront utilisées pour l'écart type qui sert à calibrer l'effet, d'ou les effets calibrés vrai δ/σ (qui est le d de Cohen) et observé d/s. D'une manière générale, on notera que d est proportionnel à la racine carrée du carré-moyen formant le numérateur du rapport F, alors que s est proportionnel à la racine carrée du dénominateur (cf. Lecoutre, 1996*a*, chap. 4).

3. Effets bruts : intervalle centré sur l'effet observé

Il est immédiat de passer du test usuel à cet intervalle. Dans la situation considérée ici, ce dernier peut être regardé comme un intervalle de confiance *(fréquentiste)*, comme un intervalle *fiduciaire*, ou comme un intervalle de crédibilité *fiducio-bayésien* (ou *bayésien standard*). Dans la suite nous l'appellerons simplement *intervalle*, laissant le lecteur libre de choisir son cadre de *justification* et d'*interprétation*.

À partir de l'effet observé d (que nous supposerons non nul) et du rapport F correspondant de l'analyse de variance (à 1 et q degrés de liberté), il est immédiat de déduire les bornes de l'intervalle à $100(1-\alpha)$ % pour l'effet vrai δ :

$$d - t_\alpha | d | / \sqrt{F} \quad \text{et} \quad d + t_\alpha | d | / \sqrt{F}$$

où t_α est la valeur critique bilatérale de la distribution de Student à q degrés de liberté : $\Pr(|t_q| > t_\alpha) = \alpha$. En d'autres termes, les limites de l'intervalle sont les $100(\alpha/2)$ % et $100(1-\alpha/2)$ % percentiles de la distribution de Student généralisée à q degrés de liberté, de centre (moyenne) d et d'échelle $e = | d | / \sqrt{F}$ (son écart type est $e\sqrt{q/(q-2)}$). Si q est élevé, il s'agit approximativement de la distribution normale de moyenne d et d'écart type e. C'est ici qu'apparaît clairement la distinction entre les cadres d'interprétation : dans le cadre fréquentiste, cette distribution, que nous notons $t_q(d, e^2)$, n'a qu'un statut technique d'intermédiaire de calcul et n'est pas interprétable ; dans le cadre fiducio-bayésien, c'est la distribution finale relative à l'effet vrai δ.

Il en résulte que, sans même utiliser de tables statistiques, il est immédiat, connaissant l'effet observé et le rapport F, de déduire au moins approximativement l'intervalle à 95 % en calculant :

$$[d - 2 | d | / \sqrt{F},\ d + 2 | d | / \sqrt{F}]$$

Pour plus de précision, on peut remplacer 2 par $1{,}96\sqrt{q/(q-2)}$.

Par exemple, pour la comparaison n2, n1, nous avons $d = +13{,}4259$ et $F = 6{,}1401$ (avec $q = 87$), d'où approximativement l'intervalle à 95 % [+ 2,6, + 24,3] ; l'approximation plus précise donne [+ 2,7, + 24,2], qui, avec une décimale coïncide avec l'intervalle exact.

4. Effets bruts : intervalles de notabilité et de négligibilité

Il est tout aussi immédiat de déterminer des intervalles directement appropriés pour des conclusions de notabilité et de négligeabilité d'un effet[1]. Pour montrer qu'un effet est notable, on calculera un intervalle unilatéral à $100(1-\alpha)$ % ; pour cela, il suffit de retenir celle des deux limites de l'intervalle bilatéral usuel à $100(1-2\alpha)$ % qui est appropriée. Par exemple, pour la comparaison n2, n1, l'intervalle unilatéral à 95 % [+ 4,4, + ∞] est déduit de l'intervalle bilatéral à 90 % pour δ [+ 4,4, + 22,4]. Les solutions fréquentistes et fiducio-bayésiennes coïncident encore.

Pour montrer qu'un effet est négligeable, on calculera un intervalle centré sur 0. Mais dans ce cas, les solutions fréquentistes et fiducio-

1. Dans le cas où l'on veut montrer qu'un effet est « moyen », on pourra utiliser l'intervalle usuel. Il est également possible de construire un intervalle centré sur une valeur spécifiée autre que 0 (cf. Lecoutre et Derzko, 1999).

bayésiennes divergent. Dans le cadre fréquentiste, il suffit de considérer la plus grande en valeur absolue des deux limites de l'intervalle bilatéral usuel à $100(1 - 2\alpha)$ % (et non $100(1 - \alpha)$ %). Ainsi, par exemple, si l'intervalle centré sur d à 90 % pour δ est [– 13,5, + 4,5] (ce qui est le cas de la comparaison n2, n1.*-lin* D), nous en déduisons immédiatement l'intervalle de confiance (fréquentiste) centré sur 0 à 95 % (et non 90 %) [– 13,5, + 13,5][1]. Dans le cadre fiducio-bayésien, l'intervalle centré sur 0 de crédibilité 95 % se déduit de la distribution finale, soit dans l'exemple précédent $t_{87}(-11{,}73, 10{,}39^2)$, d'où l'intervalle [– 16,8, + 16,8].

Dans ce cas, même si les deux solutions sont proches, elles apparaissent irréconciliables d'un point de vue théorique. Pour le fréquentiste, la solution bayésienne ne peut être retenue car elle donne un intervalle trop large (qui contient toujours l'intervalle fréquentiste), et pour le bayésien, l'intervalle de confiance fréquentiste a une trop faible probabilité (inférieure à $1 - \alpha$).

5. Effets calibrés : intervalles

Conceptuellement, les résultats précédents se généralisent directement au cas d'un effet calibré. Les limites de l'intervalle usuel (bilatéral) pour δ/σ sont les $100(\alpha/2)$ % et $100(1 - \alpha/2)$ % percentiles d'une nouvelle distribution qui est déterminée par d/s et par F. Nous l'appelons *lambda-prime* et nous la notons $\Lambda'_q(d/s, b^2)$, où $b = |\, d/s \,|/\sqrt{F}$ (Lecoutre, 1996*a*, p. 55). Seulement cette distribution est plus complexe (asymétrique notamment) et nécessite de recourir à un programme informatique[2]. Comme pour l'effet brut, dans le cadre fréquentiste cette distribution n'a qu'un statut technique d'intermédiaire de calcul et n'est pas interprétable ; dans le cadre fiducio-bayésien, c'est la distribution finale relative à l'effet calibré vrai δ/σ.

Par exemple, pour la comparaison n2, n1, nous avons d/s =+ 0,2612 et F = 6,1401 (avec $q = 87$), d'où l'intervalle à 95 % [+ 0,05, + 0,47]. On détermine l'intervalle unilatéral et l'intervalle centré sur 0 en procédant comme pour les effets bruts.

1. En fait cette procédure peut être raffinée de la manière suivante : dans le cas où l'intervalle bilatéral usuel à $100(1 - 2\alpha)$ % ne contient pas zéro, c'est-à-dire où le test usuel (de l'hypothèse nulle $\delta = 0$) est significatif au seuil unilatéral α, on remplace par 0 la limite inférieure si $d > 0$ ou la limite supérieure si $d < 0$ (Berger et Hsu, 1996, p. 294). Par exemple, pour la comparaison n2, n1 ($p = 0{,}015$, bilatéral), l'intervalle usuel à 90 % [+ 4,4, + 22,4] ne contient pas 0 ; on obtiendrait « l'intervalle de négligeabilité » à 95 % [0, + 22,4], qui prend explicitement en compte le résultat significatif du test t usuel.

2. Les percentiles de la distribution *lambda-prime* peuvent être obtenus à partir de la distribution, plus facilement disponible, du t *non centré* (qui intervient dans les calculs de puissance du test de Student). On a en effet la relation, écrite symboliquement, $\Pr(\Lambda'_q(d/s,\ b^2) < x) = \Pr(t'_q(x, b^2) > d/s)$ (Lecoutre, 1999).

6. Effets calibrés : interprétation pour une comparaison intra-sujet

L'interprétation de l'effet calibré dans le cas d'une comparaison *intra-sujet* mérite une attention particulière. Le plus simple est de considérer cette comparaison conditionnellement à chaque groupe de sujets. Par exemple, pour la comparaison f3, f1/d3, on obtient l'intervalle de notabilité à 95 % pour $\delta/\sigma[+ 1{,}32, + \infty]$. Dans ce cas δ et σ sont respectivement la moyenne et l'écart type de la distribution parente des 30 effets individuels du groupe d3 (350 ms). Si on suppose que celle-ci est une distribution normale $N(\delta, \sigma^2)$, la proportion de ses valeurs plus grandes que x est un paramètre π_x, déterminé par δ et σ, soit formellement :

$$\pi_x = \Pr(N(\delta, \sigma^2) > x) = \Pr(N(0,1) < (\delta - x)/\sigma).$$

En conséquence un énoncé sur δ/σ ($x = 0$) nous renseigne certes sur l'importance relative de l'effet moyen, mais renvoie plus fondamentalement à un énoncé sur π_0. Ainsi, dans l'exemple précédent, l'inégalité $\delta/\sigma > + 1{,}32$ signifie que la proportion π_0 des valeurs *positives* dans la distribution parente est supérieure à 90,7 % ($\pi_0 > 0{,}907$). On dispose en outre pour une comparaison intra-sujet de solutions directes et élégantes qui permettent de prendre en compte simultanément l'importance de l'effet et les différences individuelles. Par exemple, pour une conclusion d'effet notable, il s'agit de montrer que la proportion π_x des effets parents supérieurs à x (limite jugée notable) est au moins égale à une proportion P (par ex. 80 %), ce qui renvoie à une inférence sur $(\delta - x)/\sigma$, et non δ/σ[1] (cf. Lecoutre, 1996*a*, p. 52-57) ; pour une conclusion d'effet négligeable, il s'agit de montrer que la proportion $\pi_{[-x, x]}$ des effets parents compris entre $-x$ et x (limite jugée négligeable) est au moins égale à une proportion P[2]. Mais les procédures correspondantes, bien que disponibles, sont encore peu courantes et sortiraient du cadre de cet article.

7. Résultats et commentaires

À titre d'illustration, nous considérerons, d'une part, les intervalles bilatéraux usuels à 95 % présentés dans le tableau II, et, d'autre part, les intervalles de notabilité (qui correspondent ici aux résultats significa-

1. Du moins dans le cas d'une comparaison conditionnelle à un groupe de sujets. Pour une comparaison portant sur plusieurs groupes g, le paramètre π_x est une moyenne des proportions des différents groupes et dépend des effets δ_g et des variances de ces groupes ; l'inférence est donc beaucoup plus complexe.

2. Cette question est actuellement étudiée dans le cadre de l'*équivalence individuelle* en pharmacologie : cf. Schall et Luus, 1993 ; Lecoutre, 1995 ; Schall, 1995 ; on pourra également se référer à Rouanet, Lépine et Holender, 1978.

TABLEAU II. — *Intervalles bilatéraux usuels à 95 % pour les effets bruts (δ) et les effets calibrés (δ/σ)*

Usual two-tailed 95 % intervals for raw effects (δ) and calibrated effects (δ/σ)

Comparaison	$\mid t \mid = \sqrt{F}$ [q dl]	p	d
lin D	4,0538 [58]	0,0002	– 0,4819
-lin D	1,1291 [87]	0,26	– 11,7340
n2, n1	2,4779 [87]	0,015	+ 13,4259
f3, f1	10,326 [87]	< 0,0001	+ 74,9000
f3, f1/d1	3,4229 [29]	0,002	+ 43,4167
f3, f1/d2	4,8753 [29]	< 0,0001	+ 64,7667
f3, f1/d3	9,9856 [29]	< 0,0001	+ 116,517
f3, f2	4,9355 [87]	< 0,0001	+ 35,4167
f3, f2/d1	2,1765 [29]	0,038	+ 20,6500
f3, f2/d2	2,1335 [29]	0,041	+ 28,6500
f3, f2/d3	4,0984 [29]	0,0003	+ 56,9500
n2, n1 . f3, f1	1,6133 [87]	0,11	+ 20,2000
n2, n1 . f3, f2	1,0911 [87]	0,28	+ 14,5889

	Intervalle bilatéral pour δ	d/s	Intervalle bilatéral pour δ/σ
lin D	[– 0,72, – 0,24]	– 0,0052	[– 0,0079, – 0,0025]
-lin D	[– 32,4, + 8,9]	– 0,119	[– 0,33, + 0,09]
n2, n1	[+ 2,7, + 24,2]	+ 0,261	[0,05, + 0,47]
f3, f1	[+ 60,5, + 89,3]	+ 1,089	[+ 0,82, + 1,35]
f3, f1/d1	[+ 17,5, + 69,4]	+ 0,625	[+ 0,23, + 1,01]
f3, f1/d2	[+ 37,6, + 91,9]	+ 0,890	[+ 0,46, + 1,31]
f3, f1/d3	[+ 92,7, + 140,4]	+ 1,823	[+ 1,23, + 2,41]
f3, f2	[+ 21,2, + 49,7]	+ 0,520	[+ 0,30, + 0,74]
f3, f2/d1	[+ 1,2, + 40,1]	+ 0,397	[+ 0,02, + 0,77]
f3, f2/d2	[+ 1,2, + 56,1]	+ 0,390	[+ 0,01, + 0,76]
f3, f2/d3	[+ 28,5, + 85,4]	+ 0,748	[+ 0,34, + 1,15]
n2, n1 . f3, f1	[– 4,7, + 45,1]	+ 0,170	[– 0,04, + 0,38]
n2, n1 . f3, f2	[– 12,0, + 41,2]	+ 0,115	[– 0,09, + 0,32]

tifs)[1] et de négligeabilité (qui correspondent ici aux résultats non significatifs) présentés dans le tableau III, ceci à la fois pour les effets bruts et pour les effets calibrés des principales comparaisons à un degré de liberté de la décomposition retenue. Rappelons encore que seul le cadre bayésien permet de choisir le type d'intervalles au vu des données.

TABLEAU III. — *Intervalles de notabilité et de négligeabilité à 95 % pour les effets bruts (δ) et les effets calibrés (δ/σ)*

95 % notability intervals and negligibility intervals for raw effects (δ) and calibrated effects (δ/σ)

Comparaison	p	Intervalle de notabilité pour δ	Intervalle de notabilité pour δ/σ
lin D	0,0002	$]-\infty, -0{,}28]$	$]-\infty, -0{,}003]$
n2, n1	0,015	$[+4{,}4, +\infty[$	$[+0{,}08, +\infty[$
f3, f1	< 0,0001	$[+62{,}8, +\infty[$	$[+0{,}87, +\infty[$
f3, f1/d1	0,002	$[+21{,}9, +\infty[$	$[+0{,}29, +\infty[$
f3, f1/d2	< 0,0001	$[+42{,}2, +\infty[$	$[+0{,}53, +\infty[$
f3, f1/d3	< 0,0001	$[+96{,}7, +\infty[$	$[+1{,}32, +\infty[$
f3, f2	< 0,0001	$[+23{,}5, +\infty[$	$[+0{,}33, +\infty[$
f3, f2/d1	0,038	$[+4{,}5, +\infty[$	$[+0{,}08, +\infty[$
f3, f2/d2	0,041	$[+5{,}8, +\infty[$	$[+0{,}07, +\infty[$
f3, f2/d3	0,0003	$[+33{,}3, +\infty[$	$[+0{,}40, +\infty[$
		Intervalle de négligeabilité* pour δ	Intervalle de négligeabilité* pour δ/σ
n2, n1 . f3, f1	0,11	$[-41{,}0, +41{,}0]$	$[-0{,}34, +0{,}34]$
n2, n1 . f3, f2	0,28	$[-36{,}8, +36{,}8]$	$[-0{,}29, +0{,}29]$
-*lin* D	0,26	$[-29{,}0, +29{,}0]$	$[-0{,}29, +0{,}29]$

* Les limites des intervalles de négligeabilité fréquentistes et bayésiens sont différentes (l'intervalle fréquentiste est toujours contenu dans l'intervalle bayésien), mais dans les exemples considérés ici coïncident cependant pour les décimales indiquées.

* *Frequentist and Bayesian negligibility intervals have unequal limits (the frequentist interval is always included within the Bayesian interval). Yet in the present examples they coincide for the given decimal places.*

1. On gardera à l'esprit qu'une conclusion de négligeabilité peut dans certain cas être obtenue alors que le test est significatif : ceci peut se produire si la différence observée est négligeable et si la précision expérimentale est très grande.

7.1. *Commentaires sur les effets bruts*

Nous nous conterons ici de commenter brièvement les principaux résultats. Les conclusions inférentielles sur les effets *vrais* δ, basées sur les intervalles à 95 % précédents, sont indiquées par un énoncé entre crochets, par exemple [$\delta < -0{,}28$ ms] ; rappelons encore que seul le cadre bayésien permet d'interpréter la confiance 95 % comme une *probabilité* relative à δ (on parle aussi dans ce cadre de *garantie*).

Durée de présentation. — Les temps de décision moyens observés diminuent avec la durée de présentation : 914,0 ms (150 ms), 940,9 (250 ms) et 817,6 ms (350 ms). L'analyse de la composante linéaire *lin* D (qui correspond ici à la comparaison des deux durées extrêmes) montre que cette diminution, rapportée à un ms d'augmentation de la durée de présentation (soit un effet observé $d = (817{,}6\text{-}914{,}0)/200 = -0{,}48$ ms par ms) est importante comme le montre l'intervalle de notabilité [$\delta < -0{,}28$ ms par ms]. La comparaison résiduelle -*lin* D (qui correspond ici à la comparaison de la durée intermédiaire aux deux autres durées extrêmes) indique que l'hypothèse d'une diminution linéaire du temps de décision en fonction du temps de présentation apparaît ici être une approximation acceptable, l'écart entre les moyennes observées et les moyennes théoriques étant relativement limité comme le montre l'intervalle de négligeabilité [$|\delta| < 29{,}0$ ms].

Nature de la phrase. — Le temps de décision observé pour une phrase dénotant un événement (864,2 ms) est plus long pour une phrase dénotant une action (850,8 ms). Mais la différence (comparaison n2, n1), bien que significative, est relativement limitée par rapport aux effets des autres facteurs, comme le montre l'intervalle bilatéral [$+2{,}7$ ms $< \delta < +24{,}2$ ms]. On remarquera ici l'équivalence entre les propriétés « le test F de l'hypothèse nulle $\delta = 0$ est significatif au seuil (bilatéral) α » et « l'intervalle bilatéral à $1-\alpha$ pour δ ne contient pas la valeur 0 ». Mais l'intervalle est manifestement plus informatif que ce seul énoncé.

Forme de la phrase. — Les temps de décision moyens observés pour les trois formes sont respectivement 819,4 ms (explicite), 858,9 ms (prédictible) et 894,3 ms (contrôle). Le temps moyen de décision à la forme contrôle est notablement plus grand que celui de la forme explicite (comparaison f3, f1) [$\delta > 62{,}8$ ms]. Il est également plus grand que celui de la forme prédictible (comparaison f3, f2), mais la différence apparaît moindre [$+21{,}2 < \delta < +49{,}7$]. Ceci exprime le fait que la différence entre explicite et prédictible est voisine de celle entre prédictible et contrôle.

Interactions. — La différence observée entre la forme contrôle et la forme explicite augmente avec la durée de présentation (respectivement 43,4 ms, 64,8 ms et 116,5 ms). Cette différence est notablement plus grande pour la durée de 350 ms (comparaison f3, f1/d3 [$\delta > +96{,}7$]) que pour la durée de 150 ms (comparaison f3, f1/d1 [$+17{,}5 < \delta < +69{,}4$]). La différence observée entre la forme contrôle et la forme prédictible augmente également, mais de façon plus limitée avec la durée de présentation (respectivement 20,6 ms, 28,6 ms et 56,9 ms).

7.2. *Commentaires sur les effets calibrés*

L'utilisation des effets calibrés pour juger de l'importance des effets conduit essentiellement aux mêmes conclusions. Par exemple, si nous adoptons les critères de Cohen (1988) pour juger de l'importance d'une différence de moyennes, les valeurs repères de δ/σ sont 0,20 (effet faible), 0,50 (effet moyen) et 0,80 (effet fort). Dans ce cas, l'effet de la nature de la phrase (n2, n1) apparaît négligeable *ou* moyen [$0{,}05 < \delta/\sigma < + 0{,}47$], la précision expérimentale reflétée par la largeur de l'intervalle étant insuffisante pour trancher ; au contraire, la différence entre la forme contrôle et la forme explicite (f3, f1) est notable [$\delta/\sigma > 0{,}87$ ms].

Cependant l'utilisation de critères conventionnels doit être effectuée avec prudence, en tenant compte du type de la comparaison. Ainsi, dans le cas d'une comparaison telle que *lin* D, pour apprécier l'importance de l'effet calibré, il faut tenir compte du fait que celui-ci est rapporté à un temps unité (1 ms), ce qui explique les valeurs en apparence faibles pour l'effet calibré observé ($d/s = - 0{,}0052$) et les limites correspondantes. Mais la situation est différente selon que la comparaison est inter-sujet (comme ici) ou intra-sujet. Dans le premier cas seul l'effet brut est rapporté à l'unité, alors que dans le second cas l'écart type serait lui aussi rapporté au temps unité (d'où un effet calibré 200 fois plus grand).

Surtout on gardera à l'esprit le fait que, pour une comparaison intra-sujet (telle que n2, n1), l'effet calibré renvoie fondamentalement aux différences individuelles (ce qui ne diminue en rien son intérêt), et non seulement à l'importance de l'effet moyen.

CONCLUSION

Il ne peut y avoir de psychologie quantitative si l'on se limite à un catalogue d'effets significatifs ou non significatifs et s'il ne se constitue pas un corpus de résultats eux-mêmes quantitatifs. L'utilisation systématique d'intervalles, qui à la suite de l'intervention de l'*American Psychological Association* pourrait rapidement devenir une *norme* de publication, apparaît être un progrès méthodologique indéniable par rapport à la situation actuelle. En particulier, la variabilité des mesures étant exprimée de façon manifeste, cet usage devrait réduire l'abus consistant à inférer une absence d'effet sur la seule base d'un résultat non significatif.

En pratique, il est le plus souvent suffisant de considérer des comparaisons à un degré de liberté[1], à partir desquelles on peut répondre de manière précise aux objectifs de l'expérience. Dans ce cas, les intervalles bilatéraux,

1. Les solutions, fréquentistes et bayésiennes, présentées ici sont également disponibles pour les effets à plusieurs degrés de liberté (Lecoutre et Poitevineau, 1992 ; Lecoutre et Derzko, 1999).

centrés (du moins approximativement pour l'effet calibré) sur l'effet observé, sont très simples à calculer. De plus on peut aussi facilement obtenir des intervalles directement appropriés pour montrer qu'un effet est notable (intervalle unilatéral), ou au contraire est négligeable (intervalle centré sur 0). Nous donnerons pour terminer quelques brèves indications générales qui peuvent servir de guide pour la présentation des procédures. Ces indications tiennent compte, outre des conventions actuelles et des recommandations de la *Task Force*, des pratiques qui semblent se dégager, à partir des articles déjà publiés dans des revues expérimentales.

Effets bruts et/ou effets calibrés

Il paraît utile de fournir au moins les valeurs observées pour les deux types d'effets, et si l'on ne rapporte que les effets bruts, de fournir également les écarts types associés. Pour ne pas multiplier les inférences, il peut sembler raisonnable d'en privilégier un pour le calcul des intervalles. Si la recommandation actuelle de la *Task Force* est de considérer l'effet calibré (*d* de Cohen), on gardera à l'esprit le fait que la calibration ne permet pas d'assurer, à elle seule, la comparabilité des résultats entre variables, entre expériences.

Choix de la confiance ou garantie

La tradition favorise indéniablement la convention de 95 %, qui ne saurait cependant avoir un statut de standard incontournable, étant donné son caractère arbitraire. Il s'agit en fait de trouver un équilibre entre le souci de fournir un énoncé le plus assuré possible et celui d'obtenir un intervalle informatif (le plus court possible). Une confiance ou garantie de 0,90 pourra apparaître souvent comme un bon compromis[1]. Il faut souligner le fait que dans le cadre fréquentiste, pour assurer la cohérence de l'interprétation de α (dans le cas d'un intervalle de confiance $1 - \alpha$) comme un taux d'erreur à long terme, il est nécessaire que α soit choisi indépendamment des données, par exemple en le spécifiant avant leur recueil.

Types d'intervalles

On utilisera bien entendu, selon la conclusion recherchée (effet faible, moyen ou limité, fort...), l'intervalle le mieux adapté. Mais ceci n'est cependant pas sans poser problème dans le cadre fréquentiste. En effet, comme pour la confiance, l'interprétation repose sur la condition que le type d'intervalle est choisi indépendamment des données, et non pas en fonction des résultats observés. Le cadre bayésien, au contraire, permet de choisir les types d'intervalles et la garantie appropriés en fonction des données ; il

1. Ce compromis est souvent raisonnable compte tenu des effectifs habituellement utilisés en psychologie.

apporte donc incontestablement beaucoup plus de souplesse à l'analyse et permet de rechercher la plus grande cohérence possible dans les conclusions. Ceci sera particulièrement apprécié dans le cas d'une étude exploratoire, où on pourra rapporter différents énoncés qui serviront de référence aux études ultérieures.

Critères pour juger de l'importance

Même si des conventions comme celles proposées par Cohen sont d'usage assez courant pour juger des effets calibrés, et peuvent d'ailleurs être une référence utile, il est clair que le jugement critique du chercheur devra toujours s'exercer. C'est le plus souvent l'importance respective des effets, dans l'expérience analysée (référence interne), et/ou dans d'autres expériences (référence externe), qui fournit la base de l'interprétation. Mais il n'existe sans doute pas de critères absolus, et au contraire les critères doivent dépendre du domaine de la recherche, de ses objectifs (notamment de son caractère fondamental ou appliqué), ainsi que du type d'effets (inter-sujets ou intra-sujets). De plus les critères devraient évoluer au fur et à mesure que progresse la connaissance.

Cadre fréquentiste ou cadre bayésien

La question du choix entre le cadre bayésien et cadre fréquentiste, aussi ancienne que l'inférence statistique et qui divise les statisticiens, dépasse notre présent propos. Dans le domaine expérimental, il est manifeste que la tradition favorise l'approche fréquentiste, autoritairement parée des vertus d'objectivité et souvent acceptée sans réserve. À l'opposé, il y apparaît encore souvent provocateur d'utiliser l'approche bayésienne (même non informative), en dépit du fait que le rapport de force en sa faveur a considérablement évolué ces dernières années chez les statisticiens.

Pour l'utilisateur la distinction peut cependant sembler relativement minime, dans la mesure où, comme nous l'avons vu, il peut obtenir des résultats voisins et même souvent identiques dans les deux cadres. Mais, d'une part, il faut garder à l'esprit le fait que l'usage honnête des intervalles de confiance fréquentistes, outre la nécessité de renoncer à l'interprétation naturelle en termes probabilistes, impose aussi la contrainte impérieuse de choisir le type d'intervalles et la confiance *indépendamment* des données. Et, d'autre part, il faut aussi être conscient du fait que la « liberté » autorisée dans le cadre bayésien est considérée par les tenants de l'interprétation fréquentiste comme la porte ouverte à toutes les dérives.

Comment présenter les procédures

La conception des procédures comme un prolongement, technique et conceptuel, des tests de signification, permet de les intégrer sans difficulté. De même que pour les tests, des variantes de détail pourront tenir compte

des habitudes des revues, mais il suffit en général de compléter les résultats traditionnellement fournis (statistique de test, degrés de liberté, seuil observé) par l'effet observé et par l'intervalle correspondant, par exemple : « $d = 43{,}42$, $F(1,29) = 11{,}72$, $p < 0{,}001$, $17{,}5 < \delta < 69{,}4$ »[1] ou « $F(1,87) = 57{,}83$, $p < 0{,}001$, $d/s = 0{,}52$, $\delta/\sigma > 0{,}33$ ». En préalable, il convient bien entendu de préciser la procédure utilisée (effet brut et/ou effet calibré, ainsi que le type d'intervalles si on ne s'en tient pas aux intervalles de confiance bilatéraux usuels) et éventuellement d'expliciter le choix de l'approche bayésienne (« intervalle de confiance » renvoyant implicitement à l'approche fréquentiste). Il faut également indiquer la confiance ou garantie associée aux intervalles. Si celle-ci varie d'un énoncé à l'autre (ce qui ne devrait concerner que le cadre bayésien), l'intervalle sera suivi de la mention correspondante, par exemple « $\gamma = 0{,}99$ ». Enfin, dans le cas où l'on ne rapporte que l'effet brut, il est utile d'indiquer également l'écart type associé.

Si l'on souhaite présenter une synthèse des résultats, le tableau d'analyse de variance, souvent délaissé de nos jours, peut être facilement complété et retrouver ainsi un intérêt qui pourrait conduire à sa réhabilitation.

Pour conclure, nous rejoindrons beaucoup de ceux qui ont critiqué les tests de signification en soulignant le fait qu'aucune procédure statistique ne dispensera jamais le chercheur d'un effort de réflexion sur ses données. L'usage de nouvelles procédures d'inférence statistique devra fournir un outil, une aide, mieux appropriés à cette réflexion.

RÉSUMÉ

Il y a de bonnes raisons de penser que le rôle des tests de signification usuels dans la recherche en psychologie sera considérablement réduit dans un proche avenir. Les résultats des analyses statistiques traditionnelles devraient être systématiquement complétés (« au-delà des seuls seuils observés p *») pour inclure systématiquement la présentation d'indicateurs de la grandeur des effets et leurs estimations par intervalles. Ces procédures pourraient rapidement devenir de nouvelles* normes *de publication. Dans cet article, nous passons d'abord en revue les principaux abus des tests de signification et les solutions de rechange proposées. Parmi celles-ci, des méthodes d'intervalle de confiance* (fréquentistes) *et des méthodes d'intervalles de crédibilité* (fiducio-bayésiens) *permettent d'estimer l'importance réelle des effets, et en particulier d'apprécier leur caractère négligeable ou notable. À partir d'un exemple numérique, nous illustrons ces méthodes pour l'analyse de contrastes entre moyennes dans un plan d'expérience complexe, en considérant à la fois les effets* bruts *et*

1. L'intervalle pourrait également être donné sous la forme traditionnelle [17,5, 69,4], mais celle-ci est moins explicite, et surtout peu commode pour les intervalles unilatéraux. Bien entendu l'écriture probabiliste $\Pr(17{,}5 < \delta < 69{,}4) = 0{,}95$ ne peut être utilisée que dans le cadre bayésien.

les effets relatifs ***(calibrés). Nous discutons les similitudes et les différences des approches fréquentistes et bayésiennes, leur interprétation correcte et leur utilisation pratique.***

Mots-clés : grandeur de l'effet, effets bruts et relatifs, inférence statistique, tests de signification, intervalles de confiance, méthodes bayésiennes.

BIBLIOGRAPHIE

Aguilar N., Denhière G. — (1999) La production des inférences sur la conséquence probable des actions et des événements : influence du temps de présentation et de l'intervalle interstimulus, *L'Année Psychologique* (soumis).

American Psychological Association, Board of Scientific Affairs — (1996) Task force on statistical inference initial report (draft), disponible sur Internet à l'adresse : http://www.apa.org/science/tfsi.html

Anderson S., Hauck W. W. — (1983) A new procedure for testing equivalence in comparative bioavailability and other clinical trials, *Communications in Statistics, Theory and Methods, 12*, 2663-2692.

Bacher F. — (1999) L'utilisation des modèles dans l'analyse des structures de covariance, *L'Année Psychologique, 99* (1), 99-122.

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

Bartko J. J. — (1991) Proving the null hypothesis, *American Psychologist, 46*, 1089.

Berger G. O. — (1985) *Statistical decision theory and Bayesian analysis*, New York, Springer Verlag.

Berger R. L., Hsu J. C. — (1996) Bioequivalence trials, intersection-union tests and equivalence confidence sets, *Statistical Science, 11*, 283-319.

Bernard J.-M. — (1986) Méthodes d'inférence bayésienne sur des fréquences, *Informatique et Sciences Humaines, 68-69*, 89-133.

Bernard, J.-M. — (1996) Bayesian interpretation of frequentist procedures for a Bernoulli process, *The American Statistician, 50*, 7-13.

Bernard J.-M. — (1998) Bayesian inference on frequencies, *in* H. Rouanet, J.-M. Bernard, M.-C. Bert, B. Lecoutre, M.-P. Lecoutre et B. Le Roux (Edit.), *New ways in statistical methodology : From significance tests to Bayesian inference*, Berne, Peter Lang, 159-226.

Berry G. — (1986) Statistical significance and confidence intervals [Editorial], *The Medical Journal of Australia, 144*, 618-619.

Bolles R. — (1962) The difference between statistical hypotheses and scientific hypotheses, *Psychological Reports, 11*, 639-645.

Bondy W. A. — (1969) A test of an experimental hypothesis of negligible difference between means, *The American Statistician, 23*, 28-30.

Boring E. G. — (1919) Mathematical versus scientific significance, *Psychological Bulletin, 16*, 335-338.

Braitman L. E. — (1988) Confidence intervals extract clinically useful information from data [Editorial], *Annals of Internal Medicine, 108*, 296-298.

Braitman L. E. — (1991) Confidence intervals assess both clinical significance and statistical significance [Editorial], *Annals of Internal Medicine, 114*, 515-517.

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

Casella G., Berger L. — (1987) Reconciling Bayesian and frequentist evidence in the one-sided testing problem (with discussion), *Journal of the American Statistical Association*, *82*, 106-111.

Chow S. L. — (1988) Significance test or effect size ? *Psychological Bulletin*, *103*, 105-110.

Chow S. L. — (1996) *Statistical significance : Rationale, validity, and utility*, Londres, Sage.

Ciancia F., Maitte M., Honoré J., Lecoutre B., Coquery J.-M. — (1988) Orientation of attention and sensory gatting : An evoked potential and RT study in cat, *Experimental Neurology*, *100*, 274-287.

Clément E., Richard J.-F. — (1997) Knowledge of domain effects in problem representation : The case of tower of Hanoi isomorphs, *Thinking and Reasoning*, *3*, 133-157.

Cohen J. — (1962) The statistical power of abnormal-social psychological research : A review, *Journal of Abnormal and Social Psychology*, *65*, 145-153.

Cohen J. — (1988) *Statistical power analysis for the behavioral sciences* (1re éd. en 1969), Hillsdale (NJ), Erlbaum.

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

Cohen J. — (1994) The earth is round ($p < .05$), *American Psychologist*, *49*, 997-1003, with replies in *American Psychologist*, 1995, *50*, 1098-1103.

Cornfield J. — (1966) Sequential trials, sequential analysis and the likelihood principle, *The American Statistician*, *20*, 18-23.

Corroyer D., Rouanet H. — (1994) Sur l'importance des effets et des indicateurs dans l'analyse statistique des données, *L'Année Psychologique*, *94*, 607-624.

Cox D. R. — (1977) The role of significance tests, *Scandinavian Journal of Statistics*, *4*, 49-70.

Craig J. R., Eison C. L., Metze L. P. — (1976) Significance tests and their interpretation : An example utilizing published research and ω^2, *Bulletin of the Psychonomic Society*, *7*, 280-282.

Deheuvels, P. — (1984) How to analyze bio-equivalence studies ? The right use of confidence intervals, *Journal of Organizational Behaviour and Statistics*, *1*, 1-15.

Dwyer J. H. — (1974) Analysis of variance and the magnitude of effects : A general approach, *Psychological Bulletin*, *81*, 731-737.

Edwards W. — (1965) Tactical note on the relation between scientific and statistical hypotheses, *Psychological Bulletin*, *63*, 400-402.

Edwards W., Lindman H., Savage L. J. — (1963) Bayesian statistical inference for psychological research, *Psychological Review*, *70*, 193-242.

Evans S. J. W, Mills P., Dawson J. — (1988) The end of the p-value ?, *British Heart Journal*, *60*, 177-180.

Falissard B., Landais P. — (1995) Les statistiques en médecine : et s'il était temps de prendre un peu de distance ?, *Médecine Thérapeutique*, *1*, 775-781.

Fisher R. A. — (1948) Conclusions fiduciaires, *Annales de l'Institut Henri-Poincaré*, *10*, 191-213.

Fisher R. A. — (1990-1956) *Statistical methods and scientific inference*, Réimpression de la 3e édition de 1973, *in* J. H. Bennett (Edit.), *Statistical methods, experimental design, and scientific inference*, Oxford, Oxford University Press (1re éd. en 1956, Londres, Oliver & Boyd).

Fowler R. L. — (1984) Approximating probability levels for testing null hypotheses with noncentral F distributions, *Educational and Psychological Measurement*, *44*, 275-281.

Fowler R. L. — (1985) Testing for substantive significance in applied research by specifying nonzero effect nullhypotheses, *Journal of Applied Statistics*, *70*, 215-218.

Frick R. W. — (1996) The appropriate use of null hypothesis testing, *Psychological Methods*, *1*, 379-390.

Gold D. — (1969) Statistical tests and substantive significance, *The American Sociologist*, *4*, 42-46.

Goodman S. N., Berlin J. A. — (1994) The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results, *Annals of Internal Medicine*, *121*, 200-206.

Grant D. A. — (1962) Testing the null hypothesis and the strategy and tactics of investigating theoretical models, *Psychological Review*, *69*, 54-61.

Guttman L. — (1983) What is not what in statistics ?, *The Statistician*, *26*, 81-107.

Haase R. F., Waechter D. M., Solomon G. S. — (1982) How significant is a significant difference ? Average effect size of research in counseling psychology, *Journal of Counseling Psychology*, *29*, 58-65.

Harcum E. R. — (1990) Methodological versus empirical literature : Two views on casual acceptance of the null hypothesis, *American Psychologist*, *45*, 404-405.

Harris M. J. — (1991) Significance tests are not enough : The role of effect size estimation in theory corroboration, *Theory and Psychology*, *1*, 337-360, 375-382.

Hays W. L. — (1963) *Statistics for psychologists*, New York, Holt, Rinehart & Winston.

Hinkley D. — (1997) Comment (discussion de l'article de J. O. Berger, B. Boukai et Y. Wang, Unified frequentist-Bayesian testing of a precise hypothesis), *Statistical Science*, *12*, 155-156.

Hoc J.-M. — (1983) *L'Analyse planifiée des données en psychologie*, Paris, PUF.

Hoc J.-M. — (1996) Operator expertise and verbal reports on temporal data, *Ergonomics*, *39*, 811-825.

Hoc J.-M., Leplat J. — (1983) Evaluation of different modalities of verbalization in a sorting task, *International Journal of Man-Machine Studies*, *18*, 283-306.

Hogben L. — (1957) *Statistical theory : The relationship of probability, credibility, and error. An examination of the contemporary crisis in statistical theory from a behaviourist viewpoint*, New York, W. W. Norton.

Jeffreys H. — (1961) *Theory of probability*, Oxford, Clarendon (3e éd., 1re éd. en 1939).

Kadane J. B. — (1995) Prime time for Bayes, *Controlled Clinical Trials*, *16*, 313-318.

Kish L. — (1959) Some statistical problems in research design, *American Sociological Review*, *24*, 328-338.

LaForge R. — (1967) Confidence intervals or test of significance in scientific research ?, *Psychological Bulletin*, *68*, 446-447.

Lecoutre B. — (1978) Note sur le calcul de la distribution fiduciaire pour une inférence sur un contraste entre moyennes, *Cahiers de Psychologie*, *21*, 279-282.

Lecoutre B. — (1984) *L'analyse bayésienne des comparaisons*, Lille, Presses Universitaires de Lille.

Lecoutre B. — (1985) How to derive Bayes-fiducial conclusions from usual significance tests, *Cahiers de Psychologie Cognitive*, *5*, 553-563.

Lecoutre B. — (1995) *Bayesian procedures for asserting individual bioequivalence*, conférence invitée, Meeting of the International Society for Clinical Biostatistics, Barcelone, juillet 1995.

Lecoutre B. — (1996*a*) *Traitement statistique des données expérimentales : des pratiques traditionnelles aux pratiques bayésiennes*, avec programmes Windows® par B. Lecoutre et J. Poitevineau (disponibles sur Internet à l'adresse : http://epeire.univ-rouen.fr/labos/eris/pac.html), Saint-Mandé, CISIA.

Lecoutre B. — (1996*b*) Au-delà du test de signification ou l'inférence statistique sans tables (à la suite d'Alain Morineau), *La Revue de Modulad*, *17*, 98-100.

Lecoutre B. — (1997) Et si vous étiez un bayésien « qui s'ignore » ?, *La Revue de Modulad*, *18*, 81-87.

Lecoutre B. — (1999) Two useful distributions for Bayesian predictive procedures under normal models, *Journal of Statistical Planning and Inference*, *77*, 93-105.

Lecoutre B., Charron C. — (2000) Bayesian procedures for prediction analysis of implication hypothesis in 2 × 2 contingency tables, *Journal of Educational and Behavioral Statistics*, *25* (sous presse).

Lecoutre M.-P. — (1992) Cognitive models and problem spaces in « purely random » situations, *Educational Studies in Mathematics*, *23*, 557-568.

Lecoutre M.-P. — (1998) And... what about the researcher's point of view ? *In* H. Rouanet, J.-M. Bernard, M.-C. Bert, B. Lecoutre, M.-P. Lecoutre et Le Roux (Edit.), *New ways in statistical methodology : From significance tests to Bayesian inference*, Berne, Peter Lang, 65-95.

Lecoutre B., Derzko G. — (1999) *Asserting the smallness of effects in ANOVA* (soumis pour publication).

Lecoutre B., Lecoutre M.-P. — (1979) À propos d'une expérience d'apprentissage perceptif incident : quelques aspects de la démarche d'analyse des données et méthodes fiduciaires, *Psychologie Française*, *24*, 269-276.

Lecoutre B., Poitevineau J. — (1992) *PAC® (Programme d'Analyse des Comparaisons) : guide d'utilisation et manuel de référence*, Saint-Mandé, CISIA.

Lecoutre B., Derzko G., Grouin J.-M. — (1995) Bayesian predictive approach for inference about proportions, *Statistics in Medicine*, *14*, 1057-1063.

Lecoutre B., Poitevineau J., Derzko G., Grouin J.-M. — (2000) Désirabilité et faisabilité des méthodes bayésiennes en analyse de variance : application à des plans d'expérience complexes utilisés dans les essais cliniques, *in* I. Albert et B. Asselain (Édit.), *Biométrie et méthodes bayésiennes*, *14*, 1-23.

Lee P. — (1997) *Bayesian statistics : An introduction* (2e éd.), Oxford, Oxford University Press.

Lépine D., Rouanet H. — (1975) Introduction aux méthodes fiduciaires : inférence sur un contraste entre moyennes, *Cahiers de Psychologie*, *18*, 193-218.

Levin J. R. — (1967) Misinterpreting the significance of « explained variation », *American Psychologist*, *22*, 675-676.

Lewis C. — (1993) Bayesian methods for the analysis of variance, *in* G. Keren et C. Lewis (Edit.), *A Handbook for data analysis in the behavioral sciences*, vol. 2 : *Statistical Issues*, Hillsdale (NJ), Erlbaum, 233-256.

Loftus G. R. — (1993) Editorial comment, *Memory and Cognition*, *21*, 1-3.

Loftus G. R., Masson M. E. — (1994) Using confidence intervals in within-subject designs, *Psychonomic Bulletin and Review*, *1*, 476-490.

Lutz W., Nimmo I. A. — (1977) The inadequacy of statistical significance (Editorial), *European Journal of Clinical Investigation*, *7*, 77-78.

Morrison D. E., Henkel R. E. — (1969) Significance tests reconsidered, *The American Sociologist*, *4*, 131-140.

Natrella M. G. — (1960) The relation between confidence intervals and tests of significance, *The American Statistician*, *14*, 20-22.

Neyman J. — (1952) *Lectures and Conferences on Mathematical Statistics and Probability* (2e éd.), Washington, Graduate School US Department of Agriculture.

Novick M. R., Jackson P. H. — (1974) *Statistical methods for educational and psychological research*, NewYork, McGraw-Hill.

Nunnally J. C. — (1960) The place of statistics in psychology, *Educational and Psychological Measurement*, *20*, 641-650.

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

O'Brien T. C., Shapiro B. J. — (1968) Statistical significance. What ?, *Mathematics Teacher*, *61*, 673-676.

O'Grady K. E. — (1982) Measures of explained variance : Cautions and limitations, *Psychological Bulletin*, *92*, 766-767.

Patel H. I., Gupta G. D. — (1984) A problem of equivalence in clinical trials, *Biometrical Journal*, *33*, 1225-1230.

Phillips L. D. — (1973) *Bayesian statistics for social scientists*, Londres, Nelson.

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.

Pruzek R. M. (1997) — An introduction to Bayesian inference and its applications, *in* L. L. Harlow, S. A. Mulaik et J. H. Steiger (Edit.), *What if there were no significance tests ?*, Hillsdale (NJ), Erlbaum, 287-318.

Reuchlin M. — (1962) *Les méthodes quantitatives en psychologie*, Paris, PUF.

Richardson J. T. E. — (1996) Measures of effect size, *Behavior Research Methods, Instruments, and Computers*, *28*, 12-22.

Rindskopf D. — (1997) Testing « small », not null, hypotheses : Classical and Bayesian approaches, *in* L. L. Harlow, S. A. Mulaik et J. H. Steiger (Edit.), *What if there were no significance tests*, Hillsdale (NJ), Erlbaum, 319-332.

Rindskopf D. — (1998) Null-hypothesis tests are not completely stupid, but Bayesian statistics are better, *Behavioral and Brain Sciences*, *21*, 215-216.

Rocke D. M. — (1984) On testing for bioequivalence, *Biometrics*, *40*, 220-225.

Rogers J. L., Howard K. I., Vessey J. — (1993) Using significance tests to evaluate equivalence between two experimental groups, *Psychological Bulletin*, *113*, 553-565.

Rosenthal R. — (1990) How are we doing in soft psychology, *American Psychologist*, *45*, 775-777.

Rosenthal R., Rubin B. D. — (1982) A simple, general purpose display of magnitude of experimental effects, *Journal of Educational Psychology*, *74*, 166-169.

Rothman K. J. — (1978) A show of confidence [Editorial], *The New England Journal of Medicine*, *299*, 1362-1363.

Rouanet H. — (1967) *Les modèles stochastiques d'apprentissage*, Paris, Gauthier-Villars.

Rouanet H. — (1986) Modèles en tout genre et pratiques statisticiennes, *Comportements*, *4*, 113-124.

Rouanet H. — (1996) Bayesian procedures for assessing importance of effects, *Psychological Bulletin*, *119*, 149-158.

Rouanet H., Lecoutre B. — (1983) Specific inference in ANOVA : From significance tests to Bayesian procedures, *British Journal of Mathematical and Statistical Psychology*, *36*, 252-268.

Rouanet H., Lépine D., Holender D. — (1978) Model acceptability and the use of Bayes-fiducial methods for validating models, *in* J. Requin (Edit.), *Attention and Performance VII*, Hillsdale (NJ), Erlbaum, 687-701.

Rouanet H., Lépine D., Pelnard-Considère J. — (1976) Bayes-fiducial procedures as practical substitutes for misplaced significance testing : An application to educational data, *in* D. N. M. De Gruijter et L. J. T. Van Der Kamp (Edit.), *Advances in psychological and educational measurement*, New York, Wiley, 33-50.

Rouanet H., Bernard J.-M., Bert M.-C., Lecoutre B., Lecoutre M.-P., Le Roux B. — (1998-1991) *New ways in statistical methodology : From significance tests to Bayesian inference* (première édition en français intitulée *L'inférence statistique dans la démarche du chercheur*, 1991), Berne, Peter Lang.

Rozeboom W. W. — (1960) The fallacy of the hypothetico-deductive significance test, *Psychological Bulletin*, *57*, 416-428.

Schervish M. J. — (1995) *Theory of statistics*, New York, Springer Verlag.

Schall R. — (1995) Assessment of individual and population bioequivalence using the probability that bioavailabilities are similar, *Biometrics*, *51*, 615-626.

Schall R., Luus H. G. — (1993) On population and individual equivalence, *Statistics in Medicine*, *12*, 1109-1124.

Schmidt F. L. — (1996) Statistical significance testing and cumulative knowledge in psychology : Implications for training of researchers, *Psychological Methods*, *1*, 115-129.

Schuirmann D. J. — (1987) 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*, 657-680.

Selvin H. C. — (1957) A Critique of tests of significance in survey research, *American Sociological Review*, *22*, 519-527.

Serlin, R. C., Lapsley, D. K. — (1985) Rationality in psychological research : The good-enough principle, *American Psychologist*, *40*, 77-83.

Serlin R. C., Lapsley D. K. — (1993) Rational appraisal of psychological research and the good-enough principle, *in* G. Keren et C. Lewis (Edit.), *A Handbook for data analysis in the behavioral sciences*, vol. 1 : *Methodological issues*, Hillsdale (NJ), Erlbaum, 219-228.

Shrout P. E. — (1997) Should significance tests be banned ? Introduction to a special section exploring the pros and cons, *Psychological Science*, *8*, 1-2.

Spiegelhalter D. J., Freedman L. S., Parmar M. K. B. — (1994) Bayesian approaches to randomized trials (with discussion), *Journal of the Royal Statistical Society*, *A*, *157*, 357-416.

Vaughan G. M., Corballis M. C. — (1969) Beyond tests of significance : Estimating strength of effects in selected ANOVA designs, *Psychological Bulletin*, *72*, 204-213.

Wellek S., Michaelis J. — (1991) Elements of significance testing with equivalence problems, *Methods of Information in Medicine*, *30*, 194-198.

Wilkinson L. and Task Force on Statistical Inference, APA Board of Scientific Affairs — (1999) Statistical methods in psychology journals : Guidelines and explanations, *American Psychologist*, *54*, 594-604.

Wilson W. R., Miller H. L., Lower J. S. — (1967) Much ado about the null hypothesis, *Psychological Bulletin*, *67*, 188-196.

Winch, R. F., Campbell, D. T. — (1969) Proof ? No. Evidence ? Yes. The significance of tests of significance, *The American Sociologist*, *4*, 140-143.

Winkler R. L. — (1974) Statistical analysis : Theory versus practice, *in* C.-A. S. Staël Von Holstein (Edit.), *The concept of probability in psychological experiments*, Dordrecht, D. Reidel Publishing Company, 127-140.

Yates F. — (1951) The influence of statistical methods for research workers on the development of the science of statistics, *Journal of the American Statistical Association*, *46*, 19-34.

(Accepté le 11 mai 1999.)