Former les étudiants et les chercheurs aux méthodes bayésiennes pour l'analyse des données expérimentales

Bruno Lecoutre

ERIS, Laboratoire de Mathématiques Raphaël Salem
UMR 6085 C.N.R.S. et Université de Rouen
Avenue de l'Université, BP 12, 76801 Saint-Etienne-du-Rouvray
bruno.lecoutre@univ-rouen.fr
Internet : http ://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris

Une version en langue anglaise de cet article, intitulée " *Training students and researchers in Bayesian methods for experimental data analysis*", est publiée dans *Journal of Data Science*, 2006, **4**, 2 [http ://www.sinica.edu.tw/ jds/JDS-7.html] (à paraître).

Résumé

Les tests de signification fréquentistes de l'hypothèse nulle (en anglais " *Null Hypothesis Significance Testing*" = NHST) font tellement partie des habitudes des scientifiques que l'on ne peut supprimer leur usage "en les jetant par la fenêtre". Face à cette situation, la stratégie proposée pour former les étudiants et les chercheurs aux méthodes d'inférence statistique pour l'analyse des données expérimentales repose sur une transition en douceur vers le paradigme bayésien. Les principes de base de cette stratégie sont les suivants. (1) Présenter les interprétations bayésiennes naturelles des tests de signification usuels pour attirer l'attention sur leurs insuffisances. (2) Créer en conséquence le besoin d'un changement dans la présentation et l'interprétation des résultats. (3) Finalement fournir aux utilisateurs la possibilité réelle de penser de manière rationnelle les problèmes d'inférence statistique et de se comporter d'une façon plus raisonnable. La conclusion est que l'enseignement de l'approche bayésienne dans le contexte de l'analyse des données expérimentales apparaît à la fois *désirable* et *faisable*. Cette faisabilité est illustrée pour les méthodes d'analyse de variance.

1 Introduction

La période actuelle est cruciale car on voit apparaître de nouvelles normes de publication pour la recherche expérimentale. Ainsi en psychologie la nécessité de changements dans la présentation des résultats expérimentaux a été récemment rendue officielle par l'American Psychological Association (Wilkinson *et al.*, 1999 ; American Psychological Association, 2001). Dans tous les domaines expérimentaux, et en particulier dans la recherche médicale, cette nécessité est de plus en plus mise en avant par les "éditeurs" des revues qui demandent aux auteurs de fournir de manière routinière des indicateurs de la taille des effets ("*effect size*") et leurs intervalles d'estimation ("*interval estimates*), en plus ou à la place des tests de signification traditionnels.

Cet article est divisé en quatre sections. (1) Je montre que le test de signification usuel est une méthode inappropriée pour l'analyse des données expérimentales, non pas parce qu'il est un modèle normatif incorrect, mais parce qu'il ne répond pas aux questions que pose la recherche scientifique. Je présente et critique les recommandations proposées par la "*Task Force*" de l'American Psychological Association pour surmonter l'inadéquation des tests. (2) Comme solution, je propose d'enseigner les méthodes bayésiennes comme une *thérapie* contre les mauvais usages et les abus d'utilisation des tests de signification. (3) La faisabilité de cet enseignement est illustrée dans le contexte des méthodes d'analyse de variance. (4) Ses avantages et difficultés sont discutés. En conclusion, former les étudiants et les chercheurs aux méthodes bayésiennes devrait devenir un défi motivant pour les formateurs en statistique.

2 Le contexte actuel de la recherche expérimentale

2.1 Le goulot d'étranglement des tests de signification

Dès les origines (Boring, 1919 ; Tyler, 1931 ; Berkson, 1938 ; etc.), le test de signification a fait l'objet de critiques intenses, tant sur des bases théoriques que sur des bases méthodologiques, pour ne pas mentionner la controverse aiguë qui opposa Fisher à Neyman et Pearson sur le fondations mêmes de l'inférence statistique. dans les années 1960, il y eut de plus en plus de critiques, en particulier dans les sciences du comportement et dans les sciences sociales (voir notamment Morrison & Henkel, 1970). L'inadéquation fondamentale du test de signification dans l'analyse des données expérimentales a été dénoncé par les scientifiques les plus éminents et les plus avertis (voir Poitevineau, 1998, 2004 ; Lecoutre, Lecoutre & Poitevineau, 2001).

Plusieurs études empiriques ont mis en avant l'existence largement répandue d'erreurs d'interprétation des tests chez les étudiants et le chercheurs en psychologie (Rosenthal & Gaito, 1963 ; Nelson, Rosenthal & Rosnow, 1986 ; Oakes, 1986 ; Zuckerman, Hodgins, Zuckerman & Rosenthal, 1993 ; Falk & Greenbaum, 1995 ; Mittag & Thompson, 2000 ; Gordon, 2001 ; Poitevineau & Lecoutre, 2001). Récemment, Haller et Krauss (2001) ont trouvé que la plupart des formateurs en méthodologie qui enseignent les statistiques aux étudiants en psychologie, y compris ceux qui travaillent dans le domaine de la statistique, font les mêmes erreurs d'interprétation que leurs étudiants. En outre, Lecoutre, Poitevineau et Lecoutre (2003) ont montré que des statisticiens de métier travaillant dans l'industrie pharmaceutique ne sont pas à l'abri des erreurs d'interprétation des tests, en particulier quand le test est non-significatif.

Si certains des résultats précédents peuvent être interprétés comme un manque de maîtrise de l'outil statistique, cette explication s'applique difficilement aux statisticiens de métier. Il est plus vraisemblable que ces résultats montrent que le test de signification ne répond pas aux questions que pose la recherche scientifique. Par suite les utilisateurs doivent effectuer une combinaison plus ou moins "naïve" des résultats du test avec d'autres informations. En d'autre mots, il doivent recourir à des ajustements pour adapter un outil inapproprié à leurs besoins réels, ce qui a été appelé "*judgmental adjustments*" (Bakan, 1966 ; Phillips, 1973, page 334) ou encore "*adaptative distorsions*" (M.-P. Lecoutre, 2000, page 74). Ainsi la confusion entre signification *statistique* et signification *scientifique* ("plus un résultat est significatif, plus il est scientifiquement intéressant,

et/ou plus l'effet vrai est grand") est une illustration d'un tel ajustement et peut être vu comme un *abus adaptatif*. L'utilisation impropre des résultats non-significatifs comme "preuve de l'hypothèse nulle" est encore plus illustrative ; en effet, face à un résultat non-significatif, les utilisateurs semblent n'avoir pas d'autre choix que de l'interpréter comme preuve de l'hypothèse nulle ou d'essayer de le justifier en invoquant une anomalie dans les conditions expérimentales ou dans l'échantillon. De même les interprétations "incorrectes" des seuils de signification observés ("*p-values*) comme des probabilités "inverses" (1-p est "la probabilité que l'hypothèse alternative soit vraie" ou est considéré comme "degré de certitude de la réplicabilité des résultats"), même par des utilisateurs avertis, révèlent des questions qui sont d'un intérêt primordial pour les utilisateurs. De telles interprétations suggèrent que les utilisateurs souhaitent réellement autre chose ("*really want to make a different kind of inference*", Robinson & Wainer, 2002, page 270). Plus encore, de nombreux chercheurs en psychologie disent explicitement qu'ils ne sont pas satisfaits des pratiques actuelles et apparaissent avoir une conscience réelle du goulot d'étranglement du test de signification (M.-P. Lecoutre, 2000). Ils utilisent celui-ci uniquement parce qu'ils ne connaissent pas d'autre méthode, mais ils expriment leur besoin pour des méthodes d'inférence qui répondraient mieux à leurs questions spécifiques. Dans ce contexte un consensus consiste à attendre de l'analyse statistique qu'elle exprime de manière objective "ce que les données ont à dire" indépendamment de toute information extérieure. Très peu de chercheurs disent effectivement qu'ils souhaitent intégrer des informations extérieures – et notamment le contexte théorique – dans l'analyse statistique de leurs données.

2.2 Le temps est venu de changements dans l'enseignement des méthodes d'inférence statistique

Ces résultats sont un encouragement pour les nombreuses tentatives récentes d'améliorer les pratiques usuelles pour analyser et rapporter les résultats expérimentaux. Nous pouvons espérer avec Kirk (2001, page 217)) que ces tentatives provoqueront des réactions en chaîne, et en particulier que "*teachers of statistics, methodology, and measurement courses will change their courses*" et que "*faculties will require students to learn the full arsenal of quantitative and qualitative statistical tools*". Nous ne pouvons pas accepter que les futurs utilisateurs des méthodes d'inférence statistique continuent à utiliser des procédures inappropriées "parce qu'ils ne connaissent pas d'autres méthodes".

Aussi le temps est-il venu de changements dans la conception de l'enseignement de l'inférence statistique, même dans les cours d'introduction pour les étudiants non statisticiens. Une opinion de plus en plus répandue est qu'il faut enseigner, en plus ou à la place des tests de signification, des procédures inférentielles qui permettent de surmonter les mauvais usages habituels de ces tests, tout en apportant une information pertinente sur la taille des effets. Pour cela les intervalles de confiance, les méthodes de vraisemblance, ou les méthodes bayésiennes sont manifestement appropriées (voir Goodman & Berlin, 1994 ; Nester, 1996 ; Rouanet, 1996). Aujourd'hui la tendance majoritaire est de proposer l'utilisation des intervalles de confiance. Les extraits suivants sont les recommandations de la *Task Force of the American Psychological Association* (Wilkinson *et al.* 1999) pour réviser la section statistique du manuel de l'*American Psychological Association* (les italiques sont ajoutées).

Hypothesis tests. "It is hard to imagine a situation in which a dichotomous accept-

reject decision is better than reporting an actual p value or, better still, a confidence interval. *Never use the unfortunate expression 'accept the null hypothesis.' Always provide some effect-size estimate when reporting a* p *value*."

Interval estimates. "*Interval estimates should be given for any effect sizes involving principal outcomes*. Provide intervals for correlations and other coefficients of association or variation whenever possible."

Effect sizes. *Always present effect sizes for primary outcomes*. 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."

Power and sample size. "Provide information on sample size and the process that led to sample size decisions. *Document the effect sizes, sampling and measurement assumptions, as well as analytic procedures used in power calculations*. Because power computations are most meaningful when done before data are collected and examined, it is important to show how effect-size estimates have been derived from previous research and theory in order to dispel suspicions that they might have been taken from data used in the study or, even worse, constructed to justify a particular sample size."

2.3 De nouvelles difficultés

"*It would not be scientifically sound to justify a procedure by frequentist arguments and to interpret it in Bayesian terms*" (Rouanet, 2000, in Rouanet *et al.*, page 54).

Les intervalles de confiance pourraient devenir rapidement une norme impérative dans les publications expérimentales. Cependant, pour de nombreuses raisons dues à leur conception *fréquentiste*, les intervalles de confiance peuvent difficilement être considérés comme la méthode ultime. En effet la raison qui rend attrayants les intervalles de confiance résulte d'une incompréhension fondamentale. Comme c'est le cas pour les tests de signification, l'interprétation fréquentiste d'un intervalle de confiance 95% met en jeu une répétition infinie de la même expérience : à long terme 95% des intervalles de confiance calculés contiendront la "vraie valeur" du paramètre ; chaque intervalle isolé a une probabilité qui est soit **0** soit **1** de la contenir. Il est si étrange de traiter les données comme aléatoires même *après avoir recueilli les observations* que l'interprétation fréquentiste *orthodoxe* des intervalles de confiance n'a pas de sens pour la plupart des utilisateurs. C'est sans aucun doute l'interprétation naturelle (bayésienne) des intervalles de confiance dans les termes d'un "intervalle fixé ayant 95% de chances d'inclure la vraie valeur du paramètre" qui les rend attrayants.

Même les experts en statistique ne sont pas à l'abri de confusions *conceptuelles*. Ainsi, par exemple, Rosnow et Rosenthal (1996, page 336) prennent l'exemple d'une différence de deux moyennes observées $d = +0.266$ et considèrent l'intervalle $[0, +532]$ dont les bornes sont "l'hypothèse nulle" (0) et ce qu'ils appellent la valeur "contrenulle" ($2d = +0.532$) calculée comme la valeur symétrique de 0 par rapport à d. Ils interprètent cet intervalle particulier $[0, +532]$ comme "a 77% confidence interval" (étant donné un seuil observé unilatéral du test usuel de Student égal à 0.115, soit $0.77 = 1-2\times0.115$). Si nous observons un autre échantillon, la valeur contrenulle ainsi que le seuil observé seront différents, et clairement, pour un grand nombre d'échantillons, la proportion des intervalles "nulle-contrenulle" ($[-2d, 0]$ ou $[0, 2d]$ (suivant le signe de d) qui contiennent la vraie valeur de la différence δ ne sera pas 77%. A l'évidence, 0.77 est ici une probabilité qui dépend des

données, et il faut donc recourir à une justification bayésienne pour pouvoir l'interpréter (pour une distribution *a priori* non-informative, c'est précisément "la probabilité que δ soit comprise entre 0 et +0.532").

Par delà ces difficultés avec les intervalles de confiance fréquentistes, les propositions de la *Task Force* sont à la fois en partie techniquement redondantes et conceptuellement incohérentes. De même que pour le test de signification, cela reviendrait à enseigner un ensemble de recettes et de rituels (calcul de puissance, "*p-values*", intervalles de confiance...), sans apporter une réelle pensée statistique. En particulier, on peut redouter que les étudiants (et leurs enseignants) continuent de se focaliser sur la signification statistique du résultat (en se demandant seulement si l'intervalle de confiance inclut la valeur de l'hypothèse nulle) plutôt que sur toutes les implications des intervalles de confiance. Comme les auteurs de ces propositions le constatent eux-mêmes, il est probablement vrai que "*statistical methods should guide and discipline our thinking but should not determine it*." Mais il n'est pas moins vrai que ce serait "*folly of blindly adhering to a ritualized procedure*" (Kirk, 2001, page 207).

3 La solution bayésienne

Nous en venons alors naturellement à nous demander si le "choix bayésien" ne sera pas, tôt ou tard, incontournable (Lecoutre, Lecoutre & Poitevineau, 2001).

3.1 Qu'est ce que l'inférence bayésienne pour l'analyse des données expérimentales ?

"*But the primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested*" (Rozeboom, 1960).

Pour les statisticiens, le rôle des probabilités, et par conséquent le débat entre "fréquentistes" et "bayésiens", peut être exprimé dans les termes suivants (Lindley, 1993) : "whether the probabilities should only refer to data and be based on frequency or whether they should *also* apply to hypotheses and be regarded as measures of beliefs" (les italiques sont ajoutées). L'inférence bayésienne, qui est basée sur une définition opérationnelle plus générale et plus utile de la probabilité, peut traiter directement des problèmes que l'approche fréquentiste peut seulement traiter de manière indirecte en recourant à des artifices arbitraires.

La critique la plus souvent faite à l'approche bayésienne par les fréquentistes est la nécessité de probabilités *a priori*. Beaucoup de bayésiens mettent en avant une perspective subjective. Une conception extrémiste est celle de Savage (1954) qui affirmait son intention d'incorporer – non seulement des connaissances *a priori* – mais aussi des *opinions a priori* dans l'inférence scientifique. De plus, par leur insistance sur les aspects décisionnels de l'approche bayésienne, beaucoup d'auteurs ont rendu obscure la contribution de l'inférence bayésienne à l'analyse des données expérimentales et à la publication scientifique. Cela pourrait fournir les raisons pour lesquelles jusqu'à présent les scientifiques ont été peu disposés à utiliser en pratique les procédures d'inférence bayésienne pour analyser leurs données.

Sans diminuer les mérites du point de vue de la théorie de la décision, il faut reconnaître qu'il existe une autre approche, non moins bayésienne, qui a été développée par Jeffreys dans les années 1930 (Jeffreys, 1998/1939). A la suite de Laplace (1986/1825), cette approche a pour but d'assigner la probabilité *a priori* quand on ne connaît "rien" sur la valeur du paramètre. En pratique, ces probabilités *non-informatives* sont des distributions vagues qui, *a priori*, ne favorisent aucune valeur particulière. Par conséquent elles laissent les données parler d'elles-mêmes ("*speak for themselves*", Box & Tiao, 1973, page 2). Sous cette forme le paradigme bayésien fournit, sinon des méthodes objectives, du moins des méthodes *de référence* appropriées pour la présentation des résultats scientifiques. Cette approche de l'inférence bayésienne est maintenant reconnue comme un standard : "*We should indeed argue that noninformative prior Bayesian analysis is the single most powerful method of statistical analysis*" (Berger, 1985, page 90).

3.2 Méthodes bayésiennes de routine pour l'analyse des données expérimentales

Depuis près de 30 ans maintenant, nous avons travaillé avec d'autres collègues pour développer des méthodes bayésiennes "de routine" pour la plupart des situations familières rencontrées dans l'analyse des données expérimentales (*cf.* Rouanet & Lecoutre, 1983 ; Lecoutre, 1984 ; Lecoutre, Derzko & Grouin, 1995 ; Lecoutre, 1996 ; Lecoutre & Poitevineau, 1996 ; Lecoutre & Charron, 2000 ; Lecoutre & Poitevineau, 2000 ; Lecoutre & Derzko, 2001). Ces méthodes peuvent être utilisées et enseignées aussi facilement que les tests t, F ou *khi-deux*. Elles ouvrent de nouvelles voies prometteuses dans la méthodologie statistique (Rouanet *et al.*, 2000).

Nous avons tout particulièrement développé des méthodes non-informatives. Pour les promouvoir, il nous a paru important de leur donner un nom plus explicite que "standard", "non-informatives" ou "de référence". Nous proposons de les appeler *fiducio-bayésiennes* (B. Lecoutre, 2000). Ce nom délibérément provocateur par sa référence à l'inférence *fiduciaire* rend hommage au travail de Fisher sur l'inférence scientifique pour les chercheurs expérimentaux (Fisher, 1990/1925). Il indique leur spécificité et leur objectif de laisser l'analyse statistique exprimer *ce que les données ont à dire*, indépendamment de toute information extérieure.

Les méthodes fiducio-bayésiennes sont des propositions concrètes pour répondre aux insuffisances des procédures fréquentistes (voir Poitevineau, 2004). Elles ont été appliquées à de très nombreuses données réelles et ont toujours été bien acceptées par les revues expérimentales (voir par exemple Hoc & Leplat, 1983 ; Ciancia *et al.*, 1988 ; Lecoutre, 1992 ; Desperati & Stucchi, 1995 ; Hoc, 1996 ; Amorim & Stucchi, 1997 ; Amorim *et al.*, 1997 ; Clément & Richard, 1997 ; Amorim *et al.*, 1998 ; Amorim *et al.*, 2000 ; Lecoutre *et al.*, 2003, 2004 ; et de nombreux articles expérimentaux publiés en français).

3.3 La désirabilité des méthodes bayésiennes

Manifestement, l'approche bayésienne offre plus de souplesse à l'analyse des données expérimentales. Pour illustrer ses avantages, je reprendrai l'exemple médical utilisé par Student (1908) dans son article originel sur le test t. Etant donné, pour chacun des n=10 patients, les heures de sommeil supplémentaires procurées par l'utilisation de chacun des deux somnifères ("*soporific* [1] *and* [2]"), Student utilisait son test t pour une inférence sur

la différence des moyennes entre les deux somnifères, en construisant une nouvelle série de données obtenue "en soustrayant 1 de 2". Les dix différences individuelles ainsi obtenues sont données dans le Tableau 1. Student calculait alors la moyenne +1.58 [d] et l'écart-type (non corrigé) standard 1.17 [soit $s = 1.23$, corrigé pour les degrés de liberté] de cette série, et concluait à partir de sa table de la "distribution t" : "*the probability is* .9985 *or the odds are about* 666 *to* 1 *than* 2 *is the better soporific*" (ce qui n'est pas une formulation fréquentiste *orthodoxe* !). En termes modernes, nous calculons la statistique de test t pour l'inférence sur une moyenne sous le modèle normal t=+1.58/(1.23/$\sqrt{10}$)=+4.06 et nous trouvons le seuil unilatéral $p = 0.0015$ (9 *dl*).

+1.2	+2.4	+1.3	+1.3	0	+1.0	+1.8	+0.8	+4.6	+1.4

Tableau 1 : Données de Student

Les points développés ci-après illustrent la désirabilité des méthodes bayésiennes qui apportent une solution de rechange aux recommendations de la *Task Force*.

"Hypothesis tests" : Interprétation fiducio-bayésienne des "p-values". L'inférence fiducio-bayésienne fournit des réinterprétations éclairantes des procédures fréquentistes, sous une forme intuitivement parlante et facilement interprétable, en utilisant le langage naturel des probabilités bayésiennes. Par exemple, le seuil unilatéral p du test t est exactement la probabilité fiducio-bayésienne que la vraie différence δ soit de signe opposé à celui de la différence observée. Pour les données de Student ($p = 0.0015$, unilatéral), il y a une probabilité *a posteriori* 0.15% que la différence soit négative et la probabilité complémentaire 99.85% qu'elle soit positive. Dans le cadre bayésien ces énoncés sont *statistiquement corrects*.

De plus l'interprétation fiducio-bayésienne du seuil met clairement en évidence les insuffisances méthodologiques des tests de signification. Il devient manifeste que le seuil observé p *en lui-même* ne dit rien sur la grandeur de δ. D'une part, un résultat, même "hautement significatif" (p "très petit"), permet seulement de conclure que δ a le même signe que la différence observée d. D'autre part, un résultat "non-significatif" n'est en toute rigueur qu'un constat d'ignorance, comme cela est illustré par l'interprétation fiducio-bayésienne $Pr(\delta < 0) = Pr(\delta > 0) = 1/2$ d'un test "parfaitement non-significatif" (soit $d = 0$).

"Interval estimates" : Interprétation fiducio-bayésienne de l'intervalle de confiance usuel. Un autre apport important est l'interprétation de l'intervalle de confiance usuel en termes naturels. Dans le cadre bayésien, cet intervalle est habituellement appelé *intervalle de crédibilité*, ce qui rend compte explicitement de la différence d'interprétation. Il devient correct de dire "il y a une probabilité (ou *garantie*) 95% que δ soit compris entre les limites fixées de l'intervalle" (conditionnellement aux données), soit pour l'exemple de Student entre +0.70 et +2.46 heures.

"Effect sizes" : Réponses bayésiennes directes. Au delà des réinterprétations des procédures fréquentistes usuelles, d'autres énoncés bayésiens fournissent des réponses directes aux questions sur la grandeur des effets. Nous pouvons calculer la probabilité que δ dépasse un temps de sommeil supplémentaire fixé, plus aisé à interpréter ; par exemple "il y a une probabilité 91.5% que δ dépasse une heure". Puisque l'unité de mesure est ici signifiante, il est aisé d'apprécier la signification pratique de la grandeur de δ. Pour résumer les résultats, on peut rapporter : "il y a une probabilité *a posteriori* 91.5% que

la différence soit positive et grande ($\delta > +1$), une probabilité 8.5% qu'elle soit positive mais limitée ($0 < \delta < +1$), et une probabilité 0.15% qu'elle soit négative". Un tel énoncé n'a pas d'équivalent fréquentiste.

La question de la réplication des observations. L'inférence bayésienne fournit une solution directe et très intuitive. Etant donné l'expérience réalisée, la distribution prédictive exprime notre état de connaissance sur des données futures. Par exemple, pour une unité expérimentale supplémentaire, "il y a une probabilité 87.4% que la différence soit positive et une probabilité 78.8% que la différence dépasse une demi-heure", et pour un échantillon futur de taille 10, "il y a une probabilité 99.1% que la différence soit positive et une probabilité 95.9% que la différence dépasse une demi-heure".

"Power and sample size" : La planification et la conduite bayésienne des expériences. "*An essential aspect of the process of evaluating design strategies is the ability to calculate predictive probabilities of potential results*." (Berry, 1991, page 81). Les procédures bayésiennes prédictives donnent aux utilisateurs une méthode très séduisante pour répondre à des questions essentielles telles que : "quelle devrait être la taille de l'expérience pour avoir des chances raisonnables de démontrer une conclusion donnée ?" ; "sur la base des données disponibles, quelles sont les chances que le résultat final conduise à la même conclusion, ou au contraire ne permette pas de conclure ?" Ces questions sont non conditionnelles en ce sens qu'elles nécessitent de considérer toutes les valeurs possibles des paramètres. Alors que les pratiques fréquentistes traditionnelles ne traitent pas ces questions, les probabilités prédictives leur apportent une réponse directe et naturelle.

En particulier, à partir d'une étude pilote, les probabilités prédictives portant sur les limites de crédibilité fournissent un résumé utile pour aider au choix de la taille de l'échantillon d'une expérience. On peut également prédire le résultat final pour l'ensemble des données, dans le cas où les données d'une étude pilote sont incluses dans l'analyse finale (Lecoutre, 2001). Les procédures prédictives peuvent aussi être utilisées pour aider à la décision d'interrompre une expérience si la probabilité prédictive apparaît insuffisante. Des références pertinentes sont Berry (1991), Lecoutre, Derzko et Grouin (1995), Joseph et Bélisle (1997), Dignam *et al.*. (1998), Johns et Andersen (1999), Lecoutre (2001), Lecoutre, Mabika et Derzko (2002).

Introduction de distributions *a priori* "informatives". Si l'usage de distributions non-informatives a un statut privilégié pour obtenir des énoncés "à usage public", d'autres techniques bayésiennes ont aussi un rôle important à jouer dans la recherche expérimentale. Elles sont idéalement adaptées pour combiner les informations de plusieurs études et par conséquent pour planifier une série d'expériences. Des utilisations réalistes de ces techniques ont été proposées. Quand une analyse fiducio-bayésienne suggère une conclusion donnée, différentes distributions *a priori* traduisant les résultats d'autres expériences ou des opinions subjectives d'individus particuliers, bien informés ("experts"), soit *sceptiques* soit *enthousiastes*, peuvent être utilisées pour éprouver la robustesse des conclusions (voir en particulier Spiegelhalter, Freedman & Parmar, 1994). En regard au besoin d'objectivité des scientifiques, on pourrait argumenter avec Dickey (1986, page 135) que "*an objective scientific report is a report of the whole prior-to-posterior mapping of a relevant range of prior probability distributions, keyed to meaningful uncertainty interpretations*".

3.4 La faisabilité des méthodes bayésiennes

Nous avons tout particulièrement développé les méthodes bayésiennes dans le cadre de l'analyse de variance, qui est d'une grande importance pour l'analyse des données expérimentales. Les recherches expérimentales mettent généralement en jeu des plans d'expérience complexes, en particulier des plans à mesures répétées. Des procédures bayésiennes ont été développées dans ce domaine, mais elles sont généralement considérées comme difficiles à mettre en œuvre et ne sont pas incluses dans les logiciels statistiques habituels. En conséquence la possibilité de les enseigner est encore largement problématique pour la plupart des enseignants de statistique.

Un moyen simple de prendre en compte la complexité des plans expérimentaux est d'utiliser *l'approche de l'analyse spécifique*. En bref, une analyse spécifique pour un effet particulier consiste à traiter seulement les données qui sont *pertinentes pour cet effet*. Le plus souvent, la structure du plan de ces données pertinentes est beaucoup plus simple que celle du plan d'origine, et le nombre de paramètres "parasites" mis en jeu dans l'inférence spécifique est réduit de façon drastique. En conséquence, dans le cadre bayésien, des procédures relativement *élémentaires* peuvent être appliquées et des distributions *a priori réalistes* peuvent être examinées. En outre, les conditions ("assomptions") nécessaires et suffisantes spécifiques à chaque inférence particulière sont explicitées. Quand ces conditions sont mises en doute, des procédures de rechange peuvent facilement être envisagées : par exemple on peut effectuer une transformation des données pertinentes, ou encore recourir à des solutions qui ne supposent pas l'égalité des variances, etc. Ainsi les avantages de l'approche de l'analyse spécifique sur l'approche conventionnelle du modèle général apparaissent décisifs à la fois pour la faisabilité et la compréhension des procédures.

Des justifications plus complètes sont données dans Rouanet et Lecoutre (1983) (voir aussi Lecoutre, 1984 et Rouanet, 1996). Il faut souligner que l'intérêt de l'approche de l'analyse spécifique pour l'analyse de variance est souvent implicitement reconnu. Ainsi Hand et Taylor (1987) suggèrent de dériver systématiquement les données pertinentes avant d'utiliser les logiciels statistiques habituels. Dans un contexte plus particulier Jones et Kenward (1989) développent "*a simple and robust analysis for two-group dual designs*" (page 160) qui est typiquement une analyse spécifique.

Trois avantages décisifs de l'approche de l'analyse spécifique peuvent être mis en avant. (1) Toutes les procédures d'analyse de variance traditionnelles peuvent être dérivées comme une extension directe des procédures de base élémentaires utilisées en statistique descriptive (moyennes, écarts-types) et en statistique inférentielle (tests t de Student). (2) Les plans complexes mettant en jeu plusieurs facteurs sont facilement pris en compte ; en particulier les conditions de validité précises pour chaque inférence sont explicitées et rendues compréhensibles. (3) Les procédures bayésiennes deviennent faciles à mettre en œuvre.

Des programmes informatiques basés sur l'approche de l'analyse spécifique ont été développés (Lecoutre et Poitevineau, 1992 ; Lecoutre, 1996). Ils incluent à la fois les pratiques fréquentistes traditionnelles (tests de signification, intervalles de confiance) et des procédures bayésiennes (non-informatives et utilisant des distributions *a priori* conjuguées). Ces procédures sont applicables à des plans expérimentaux généraux (en particulier, les plans à mesures répétées), équilibrés ou non équilibrés, avec des données univariées ou multivariées, et des covariables.

D'autres logiciels destinés à enseigner ou à s'initier à l'inférence bayésienne élémentaire

sont *First Bayes* (O'Hagan, 1996) et un ensemble de macros pour *Minitab* (Albert, 1996).

J'ai limité ici ma présentation au cadre de l'analyse de variance, mais des procédures similaires sont aussi disponibles pour les inférences sur des proportions (Lecoutre, Derzko & Grouin, 1995 ; Bernard, 2000 ; Lecoutre & Charron, 2000).

4 la formation des étudiants et des chercheurs aux méthodes bayésiennes

"*It is their straightforward, natural approach to inference that makes them* [Bayesian methods] *so attractive*" (Schmitt, 1969, preface)

En 1976 Jaynes écrivait "*As a teacher, I therefore feel that to continue the time honoured practice – still in effect in many schools – of teaching pure orthodox statistics to students, with only a passing sneer at Bayes and Laplace, is to perpetuate a tragic error which has already wasted thousands of man-years of our fines mathematical talent in pursuit of false goals. If this talent had been directed toward understanding Laplace's contributions and learning how to use them properly, statistical practice would be far more advanced than it is.*" (Jaynes, 1976, page 256). Ce serait une folie de perpétuer cette erreur ! Depuis plus de 25 ans maintenant, avec mes collègues nous avons progressivement introduit des méthodes bayésiennes dans des cours et des stages pour des auditoires de différentes formations, tout particulièrement en psychologie. Notre expérience d'enseignement et de conseils statistiques nous a montré que ces méthodes sont beaucoup plus intuitives et plus proches de la pensée des scientifiques que les procédures fréquentistes. Aussi sommes nous complètement en désaccord avec Moore (1997) qui affirmait "*Bayesian reasoning is considerably more difficult to assimilate than the reasoning of standard inference*".

4.1 Notre stratégie d'enseignement

Du fait que les publications expérimentales sont remplies de tests de signification, les étudiants et les chercheurs sont (et seront encore dans le futur) constamment confrontés à leur utilisation. Les tests font tellement partie des habitudes des scientifiques que l'on ne peut supprimer leur usage "en les jetant par la fenêtre", même si je suis entièrement d'accord avec Rozeboom (1997, page 335) qu'ils ont "*surely the most bone-headedly misguided procedure ever institutionalised in the rote training of science students*". Cette réalité ne peut être ignorée, et c'est un défi pour les enseignants de statistique d'introduire l'inférence bayésienne sans écarter, ni les tests de signification usuels ni les "*guidelines*" qui visent à les remplacer par des intervalles de confiance. Aussi je défends la position que la seule stratégie efficace est *une transition en douceur vers le paradigme bayésien* (voir Lecoutre, Lecoutre & Poitevineau, 2001).

La stratégie d'enseignement que je propose est d'introduire les méthodes bayésiennes de la manière suivante. (1) Présenter les *interprétations fiducio-bayésiennes* naturelles des tests de signification usuels pour attirer l'attention sur leurs insuffisances. (2) Créer en conséquence le besoin d'un changement dans la présentation et l'interprétation des résultats. (3) Finalement fournir aux utilisateurs la possibilité réelle de penser de manière rationnelle les problèmes d'inférence statistique et de se comporter d'une façon plus raisonnable.

A partir d'une utilisation interactive de nos programmes informatiques, il suffit d'un ensemble très réduit de notions préliminaires pour introduire les procédures de base de l'analyse de variance ("ANOVA"), c'est-à-dire les inférences sur les effets à un degré de liberté dans les plans d'expérience complexes. La possibilité d'appliquer les méthodes bayésiennes dans le contexte de plans réalistes est essentielle pour motiver les étudiants et les chercheurs. On peut concentrer l'attention sur les principes de base et la signification pratique des procédures. En conséquence, les principes de techniques plus avancées peuvent être plus facilement compris, indépendamment de leur difficulté mathématique.

4.2 Premier exemple : student data

Il est révélateur de remarquer que l'exemple historique de Student présenté dans la Section 3.3 était une application typique de l'approche de l'analyse spécifique. Les données de base étaient pour chacun des n=10 patients la différence entre les heures de sommeil supplémentaires obtenues par l'usage de d'un somnifère ("hyoscyamine hydobromide"), les heures de sommeil étant mesurées sans médicament et après traitement avec soit (1) "dextro hyoscyamine hydobromide" soit (2) "laevo hyoscyamine hydobromide" (on notera que c'étaient elles-mêmes déjà des données dérivées). L'analyse de Student est un exemple typique d'inférence spécifique : elle ne met en jeu que l'inférence élémentaire sur la moyenne d'une distribution normale.

De la même manière, nous pouvons appliquer aux données du Tableau 1 l'inférence bayésienne élémentaire sur la moyenne, avec seulement deux paramètres, la différence moyenne de la population δ et l'écart-type σ. En utilisant la distribution *a priori* non-informative usuelle, la distribution *a posteriori* (fiducio-bayésienne) de δ est une distribution t *généralisée*. Elle est centrée sur la différence moyenne observée $d = +1.58$ et a pour facteur d'échelle $e = s/\sqrt{n} = 0.39$. Cette distribution a le même nombre de degrés de liberté q=9 que le test t.

Elle s'écrit $\delta \sim d + et_q$, ou encore $\delta \sim t_q(d, e^2)$ – soit ici $\delta \sim t_9(+1.58, 0.39^2)$ – par analogie avec la distribution normale (il faut noter que cette distribution ne doit pas être confondue avec la distribution t *noncentrée*, familière aux utilisateurs de l'analyse de la puissance). Le facteur d'échelle e est le dénominateur de la statistique de test t usuelle ($t = d/e$) (en supposant $d \neq 0$). Par conséquent, la distribution fiducio-bayésienne de δ peut être directement dérivée de t=+4.06. Ce résultat met en évidence la propriété fondamentale de la statistique de test t d'être un estimateur de la précision expérimentale, *conditionnellement à la valeur observée* d. Plus précisément, $(d/t)^2$ estime la variance d'erreur d'échantillonnage de d.

Le recours aux ordinateurs résout les problèmes techniques soulevés par l'usage des distributions bayésiennes. Il donne aux étudiants un outil attrayant et intuitif pour comprendre l'impact des effectifs d'échantillons, des données et des distributions *a priori*. La distribution *a posteriori* peut être explorée visuellement. L'interprétation fiducio-bayésienne des tests de signification usuels est explicitée. Les limites de crédibilité pour une probabilité (ou garantie) donnée, ou inversement la probabilité d'un intervalle donné peuvent être calculées.

Un aspect important de l'inférence statistique est de faire des prédictions. Dans ce cas encore l'inférence bayésienne fournit une solution directe et très intuitive. Par exemple, que pouvons-nous dire de la valeur de la différence d' que nous observerions pour de nouvelles données ? La distribution prédictive pour d' dans un échantillon futur d'effectif n'

est naturellement plus dispersée que la distribution de δ relative à la population (c'est d'autant plus vrai que l'effectif du nouvel échantillon est plus petit). Ainsi la distribution fiducio-bayésienne (*a posteriori*) prédictive pour d', étant donné la valeur d observée dans les données disponibles, est encore une distribution t généralisée (naturellement centrée sur d), $d' \sim t_q(d, e^2 + e'^2)$, où $e' = s/\sqrt{n'}$. En fait, l'incertitude sur δ conditionnellement aux données disponibles (reflétée par e^2) s'ajoute à l'incertitude sur les résultats de l'échantillon futur quand δ est connue (reflétée par e'^2). A partir des données de Student, la distribution prédictive est $d' \sim t_9(+1.58, 1.29^2)$ pour une unité expérimentale future ($n' = 1$) et $d' \sim t_9(+1.58, 0.68^2)$ pour une réplique avec le même effectif ($e' = e$).

4.3 Deuxième exemple : Expérience de temps de réaction

Comme illustration d'un plan d'expérience plus complexe, considérons l'exemple suivant, dérivé de Holender et Bertelson (1975). Dans une expérience psychologique, le sujet doit réagir à un signal. Le plan met en jeu deux facteurs répétés croisés : le facteur A (fréquence du signal) à deux modalités ($a1$: fréquent et $a2$: rare), et le facteur B (durée de la période préparatoire), à deux modalités ($b1$: courte et $b2$: longue). L'hypothèse de recherche principale est que l'effet d'interaction entre les facteurs A et B est nul (ou quasi nul) (*modèle additif*). Les $n = 12$ sujets sont divisés en trois groupes de quatre sujets chacun. Les données traitées ici et présentées dans le Tableau 2 sont les temps de réaction au signal en millisecondes (moyennés sur les essais). Ils ont été précédemment analysés en détail avec des méthodes bayésiennes dans Rouanet et Lecoutre (1983), Rouanet (1996) et Lecoutre et Derzko (2001). Je me concentrerai ici sur les aspects techniques de l'appproche de l'analyse spécifique pour des sources de variations à un degré de liberté mais cette approche peut être facilement généralisée pour des sources à plusieurs degrés de liberté.

Dans le cas présent les données de base consistent en trois "groupes" et quatre "occasions" de mesure. Puisque A et B sont tous deux des facteurs à deux modalités, leur interaction peut être représentée par un seul contraste entre les quatre conditions. Considérons le contraste de coefficient s $[w_o]_{o \in O} = [+1 - 1 - 1 + 1]$. Les coefficients $[w_o]$ sont appelés *coefficients de dérivation sur les occasions*. Les données dérivées pertinentes pour l'interaction sont les douze effets d'interaction individuels rapportés dans le Tableau 2. Elles constituent un plan simple (équilibré) en groupes indépendants et l'effet d'interaction est simplement la moyenne globale δ. Cette moyenne est donnée par les *coefficients de dérivation sur les groupes* $[v_g]_{g \in G} = [1/3\ 1/3\ 1/3]$.

Un résultat à valeur générale pour un effet à un dl est qu'il peut être testé en utilisant la statistique de test $t = d/e = -2.08$, où $e = bs = 9.61$ est précisément le facteur d'échelle de la distribution fiducio-bayésienne. La constante b dépend des coefficients de dérivation sur les groupes v_g et des effectifs des groupes f_g (ici $f_{g1} = f_{g2} = f_{g3} = 4$) : $b^2 = \sum(v_g^2/f_g) = 1/12 = 0.289^2$. La variance *intra-groupe* $s^2 = 33.28^2$ est la moyenne des variances des groupes pondérées par leurs nombres de degrés de liberté $f_g - 1$. Dans le cas d'effectifs inégaux, nous pourrions considérer soit la moyenne équipondérée soit la moyenne pondérée, obtenues respectivement pour les coefficients $[v_g]_{g \in G} = [1/3\ 1/3\ 1/3]$ (*équipondération*) et $[v_g]_{g \in G} = [f_{g1}/12\ f_{g2}/12\ f_{g3}/12]$ (*pondération par les effectifs*).

Les résultats généraux suivants assurent le lien avec les procédures "ANOVA" traditionnelles. Les deux carrés-moyens du rapport F usuel, $F = MS_{A.B}/MS_{S(G).A.B} = 0.047$, sont respectivement proportionnels à d^2 et s^2 : $MS_{A.B} = (d/(ab))^2 = 13.02$ et $MS_{S(G).A.B} =$

						données individuelles dérivées	
groupe	sujet	$a1b1$	$a2b1$	$a1b2$	$a2b2$	effet d'interaction	moyenne
g1	1	387	435	416	473	+9	427.75
	2	321	336	343	368	+10	342.00
	3	333	362	358	390	+3	360.75
	4	344	430	352	393	-45	379.75
	moyenne					$d_{g1} = -5.75$ ms	$d_{g1} = 377.56$ ms
						$s_{g1} = 26.35$ ms	$s_{g1} = 36.84$ ms
g2	5	368	432	432	504	+8	434.00
	6	357	367	394	411	+7	382.25
	7	336	346	340	421	+71	360.75
	8	387	454	438	496	-9	443.75
	moyenne					$d_{g2} = +19.25$ ms	$d_{g2} = 405.19$ ms
						$s_{g2} = 35.37$ ms	$s_{g2} = 40.08$ ms
g3	9	345	408	417	479	-1	412.25
	10	358	389	372	407	+4	381.50
	11	317	375	341	392	-7	356.25
	12	386	510	464	513	-75	468.25
	moyenne					$d_{g3} = -19.75$ ms	$d_{g3} = 404.56$ ms
						$s_{g3} = 37.11$ ms	$s_{g3} = 48.24$ ms
	moyenne	353.3	403.7	388.9	437.3	$d = -2.08$ ms	$d = 395.71$ ms
						$s = 33.28$ ms	$s = 41.99$ ms

Tableau 2 : Expérience de temps de réaction : données de base et données pertinentes pour l'interaction et pour la comparaison des groupes

$(s/a)^2 = 276.84$. La constante a dépend seulement des coefficients de dérivation sur les occasions w_o : $a^2 = \sum w_o^2 = 4$. Toutes ces formules sont explicitées dans nos programmes informatiques. Avec ces notations, toutes les procédures inférentielles (fréquentistes et bayésiennes) sont simplement calquées sur l'inférence pour une moyenne sous le modèle normal.

Toute source de variation à un *dl* peut être analysée de la même manière. Supposons par exemple que le groupe $g3$ soit un groupe *contrôle* ; nous pouvons alors planifier de décomposer les effets mettant en jeu le facteur G suivant les deux contrastes suivants : $g2, g1$ (qui oppose $g2$ et $g1$) et $g3, g1\_g2$ (qui oppose $g3$ d'une part à $g1$ et $g2$ d'autre part. L'analyse spécifique de ces deux contrastes repose sur les données pertinentes constituées des douze moyennes individuelles rapportées dans le Tableau 2. Les coefficients de dérivation sur les occasions sont $[w_o] = [1/4\ 1/4\ 1/4\ 1/4]$ $(a^2 = 1/4)$ et nous considérons pour les données dérivées les deux contrastes (orthogonaux) entre les groupes de coefficients respectifs $[-1\ +1\ 0]$ $(b^2 = 1/2)$ et $[-1/2\ -1/2\ +1]$ $(b^2 = 3/8)$. A partir des données pertinentes pour l'interaction, nous pouvons encore analyser les interactions entre $A.B$ et ces deux contrastes. Le Tableau 3 fournit un résumé des analyses spécifiques de toutes les sources de variations.

Entre sujets	$[w_o]$	$[v_g]$	a	b	d	s	$e{=}bs$
$g2, g1$	$^1/_4$ $^1/_4$ $^1/_4$ $^1/_4$	-1 0 +1	0.5	0.7071	+27.63	41.99	29.69
$g3, g1\_g2$	$^1/_4$ $^1/_4$ $^1/_4$ $^1/_4$	$^{-1}/_2$ $^{-1}/_2$ +1	0.5	0.6124	+13.19	41.99	25.71
Intra sujets	$[w_o]$	$[v_g]$	a	b	d	s	$e{=}bs$
$a2, a1$	$^{-1}/_2$ $^{+1}/_2$ $^{-1}/_2$ $^{+1}/_2$	$^1/_3$ $^1/_3$ $^1/_3$	1	0.2887	+49.38	22.26	6.42
$a2, a1.g2, g1$	$^{-1}/_2$ $^{+1}/_2$ $^{-1}/_2$ $^{+1}/_2$	-1 0 +1	1	0.7071	+5.75	22.26	15.74
$a2, a1.g3, g1\_g2$	$^{-1}/_2$ $^{+1}/_2$ $^{-1}/_2$ $^{+1}/_2$	$^{-1}/_2$ $^{-1}/_2$ +1	1	0.6124	+14.63	22.26	13.63
$b2, b1$	$^{-1}/_2$ $^{+1}/_2$ $^{-1}/_2$ $^{+1}/_2$	$^1/_3$ $^1/_3$ $^1/_3$	1	0.2887	+34.63	20.80	6.01
$b2, b1.g2, g1$	$^{-1}/_2$ $^{-1}/_2$ $^{+1}/_2$ $^{+1}/_2$	-1 0 +1	1	0.7071	+30.50	20.80	14.71
$b2, b1.g3, g1\_g2$	$^{-1}/_2$ $^{-1}/_2$ $^{+1}/_2$ $^{+1}/_2$	$^{-1}/_2$ $^{-1}/_2$ +1	1	0.6124	+3.75	20.80	12.74
$A.B$	+1 -1 -1 +1	$^1/_3$ $^1/_3$ $^1/_3$	2	0.2887	−2.08	33.28	9.61
$A.B.g2, g1$	+1 -1 -1 +1	-1 0 +1	2	0.7071	+25.00	33.28	23.53
$A.B.g3, g1\_g2$	+1 -1 -1 +1	$^{-1}/_2$ $^{-1}/_2$ +1	2	0.6124	−26.50	33.28	20.38

Tableau 3 : Expérience de temps de réaction : Tableau résumé des analyses spécifiques

5 Un défi pour les formateurs en statistique

Former les étudiants et les chercheurs aux méthodes bayésiennes devrait constituer un défi motivant pour les formateurs en statistique. Il est souvent affirmé que les méthodes bayésiennes nécessitent de nouveaux concepts probabilistes, en particulier la définition bayésienne de la probabilité, les probabilités conditionnelles et la formule de Bayes. Mais, puisque la plupart des gens utilisent des énoncés de "probabilité inverse" pour interpréter les tests de signification et les intervalles de confiance, ces notions sont déjà – au moins implicitement – mises en jeu dans les méthodes fréquentistes. Ce qui est simplement nécessaire pour enseigner l'approche bayésienne est un changement très naturel de mise en avant de ces concepts, montrant qu'ils peuvent être utilisés de façon cohérente et appropriée dans l'analyse statistique.

5.1 Un changement naturel dans la présentation des concepts probabilistes

"[Bayesian analysis provides] *direct probability statements – which are what most people wrongly assume they are getting from conventional statistics*" (Grunkemeier & Payne, 2002, page 1901)

Une étude empirique récente (Albert, 2003) montre que les étudiants dans les cours d'introduction à la statistique font généralement des confusions entre les différentes conceptions de la probabilité. Clairement, l'enseignement des tests de signification et des intervalles de confiance ne peut qu'ajouter à la confusion, puisque ces méthodes sont justifiées par des arguments fréquentistes et généralement interprétées (de façon injustifiée) en termes bayésiens. Ironiquement ces interprétations *hérétiques* sont encouragées par la duplicité de la plupart des formateurs en statistique qui les tolèrent et même les utilisent. Par exemple, Pagano (1990, page 288) décrit un intervalle de confiance 95% comme un intervalle "*such that the probability is* 0.95 *that the interval contains the population value*". D'autres auteurs affirment que l'interprétation fréquentiste "correcte" qu'ils défendent peut s'exprimer comme "*we can be* 95% *confident that the population mean is between*

114.06 *and* 119.94" (Kirk, 1982, page 43), "95% *confident that* θ *is below* $B(X)$" (Steiger & Fouladi, 1997, page 230) ou encore "*we may claim* 95% *confidence that the population value of multiple* R^2 *is no lower than* .0266" (Smithson, 2001, page 614). Il est difficile d'imaginer que les étudiants ou les scientifiques puissent comprendre que "*confident*" renvoie ici à une conception fréquentiste de la probabilité ! Ainsi, dans un papier récent, Schweder et Hjort (2002) donnent la définition suivante de la probabilité, particulièrement révélatrice : "*we will distinguish between probability as frequency, termed probability, and probability as information/uncertainty, termed confidence*". Après de nombreuses tentatives pour enseigner l'interprétation "correcte" des procédures fréquentistes, je suis entièrement d'accord avec Freeman (1993) que dans ces tentatives "*we are fighting a losing battle*".

En ce qui concerne la probabilité conditionnelle et la formule de Bayes, l'enseignement traditionnel des procédures fréquentistes est également une source de confusions. Cela est tout particulièrement mis en évidence par le fait que même des chercheurs avertis confondent fréquemment "la probabilité [*conditionnelle*] de faire une erreur de Type I si l'hypothèse nulle est vraie" et "la probabilité *marginale* de faire une erreur de Type I". Ainsi Azar (1999) écrit : "[a significant result] *indicates that the chances of the finding being random is only 5 percent or less*" ; cet énoncé a été commenté ultérieurement par Bakeman (1999) comme "*a misunderstanding that generations of instructors of statistics clearly have failed to eradicate*". Cela est sans doute dû au fait que la plupart des présentations fréquentistes ne mettent pas ou peu l'accent sur les probabilités conditionnelles. Par exemple, les livres de statistique usuels parlent de "la probabilité de faire une erreur de Type I [Type 2]" en omettant la condition "étant donné H_0 [H_1]" (voir par exemple Kirk, 1982, pages 36-37). Je considère avec Berry (1997) que les probabilités conditionnelles sont intuitives pour beaucoup de gens. Ainsi la formule de Bayes est facilement comprise si elle est introduite à partir de tableaux de contingence avec des probabilités interprétées comme des fréquences, de sorte que les probabilités *a priori* peuvent être supposées connues exactement (voir Box & Tiao, 1973, page 12).

Des difficultés considérables sont dues à l'utilisation mystérieuse et irréaliste faite de la distribution d'échantillonnage pour justifier les tests de signification et les intervalles de confiance. Des questions souvent posées par les étudiants nous montrent combien cette utilisation est contre-intuitive : "pourquoi doit-on calculer la probabilité d'échantillons qui n'ont pas été observés ?" ; "pourquoi considérer la probabilité de résultats qui sont plus extrêmes que le résultat observé ?" ; etc. On ne rencontre pas de telles difficultés avec l'inférence bayésienne : la distribution *a posteriori*, étant conditionnelle aux données, ne met en jeu que la probabilité d'échantillonnage des données ("en mains") que l'on a effectivement observées, par l'intermédiaire de la fonction de vraisemblance qui transcrit la distribution d'échantillonnage dans l'"ordre naturel".

5.2 L'approche bayésienne fournit des outils pour surmonter les difficultés usuelles

"*I stopped teaching frequentist methods when I decided that they could not be learned*" (Berry, 1997).

Il est difficile de trouver des justifications *intuitives* des procédures fréquentistes autres que bayésiennes. Au contraire, avec l'approche bayésienne, on peut donner des justifications et des interprétations intuitives des procédures, de sorte que le niveau des justifications mathématiques peut être aisément adapté aux connaissances des étudiants. Ainsi

on peut argumenter avec Albert (1995, 1997) et Berry (1997) que l'inférence bayésienne élémentaire peut être effectivement enseignée à des étudiants novices et que ces étudiants tirent un réel bénéfice d'un tel enseignement. Plus encore, la mise en œuvre pratique des procédures bayésiennes conduit à une compréhension empirique des concepts de la probabilité, et ce d'autant plus que l'on s'aide de l'ordinateur.

Notre expérience des méthodes bayésiennes est qu'elles permettent aux étudiants de surmonter les difficultés usuelles rencontrées avec l'approche fréquentiste. Bien entendu, la liste qui suit n'est pas exhaustive et des études empiriques seraient les bienvenues pour confirmer nos conclusions.

Il peut être difficile pour les étudiants de distinguer un paramètre, tel que la moyenne d'une population, de la statistique moyenne observée calculée à partir d'un échantillon. Les deux notions de distribution *a posteriori* et de distribution prédictive de données futures, conditionnellement aux données disponibles, sont des outils utiles pour donner aux étudiants une compréhension de cette distinction essentielle. De plus, on peut utiliser la distribution prédictive pour obtenir, comme cas limites : (1) la distribution d'échantillonnage d'une statistique quand la distribution *a priori* tend vers une distribution ponctuelle ("paramètre connu") ; (2) la distribution *a posteriori* quand l'effectif des données futures tend vers l'infini (le paramètre peut être regardé comme la statistique observée dans un échantillon futur d'effectif très grand).

En outre, les notions de distributions *a posteriori* et prédictive, en étant des outils fondamentaux pour une meilleure compréhension des fluctuations d'échantillonnage, permettent aux étudiants de prendre conscience des conceptions erronées relatives à la réplique des expériences. En effet beaucoup de gens surestiment la probabilité de retrouver un résultat significatif (Tversky & Kahneman, 1971 ; Lecoutre & Rouanet, 1993). On recontre des conceptions erronées similaires avec les intervalles de confiance. Une étude empirique (Cumming *et al.*, 2004) suggère que de nombreux chercheurs avertis ("*leading researchers*") en psychologie, neurosciences du comportement et médecine "*hold the confidence level misconception that a* 95% *CI will on average capture* 95% *of replication means*" (page 299), sous-estimant les fluctuations des répliques.

Une difficulté importante avec la logique du test de signification est qu'il nécessite que l'hypothèse que l'on veut démontrer soit l'hypothèse alternative. A l'évidence cet artifice peut être complètement évité avec l'approche bayésienne, qui fournit des réponses directes aux questions auxquelles on s'intéresse : "quelle est la probabilité que la différence entre deux moyennes soit grande ?" ; "quelle est la probabilité que la différence (en valeur absolue) soit petite ?" ; "sur la base des données partielles qui ne permettent pas de conclusion, quelles sont les chances que le résultat final soit concluant ?" ; etc.

L'utilisation des interprétations fiducio-bayésiennes des tests de signification et des intervalles de confiance dans le langage naturel des probabilités sur les effets inconnus vient très naturellement aux étudiants. En retour les mauvais usages des tests de signification apparaissent être plus clairement compris. En particulier les étudiants deviennent très vite alertés sur le fait que les résultats non-significatifs ne peuvent pas être interprétés comme "preuve de l'absence d'effet". Je suis entièrement d'accord avec Berry (1997) qui conclut ironiquement que les étudiants exposés uniquement à une approche bayésienne en viennent à mieux comprendre les concepts fréquentistes d'intervalles de confiance et de "*p-value*" que les étudiants exposés uniquement à une approche fréquentiste.

Un objectif essentiel de l'enseignement de la statistique est de préparer les étudiants à lire des publications expérimentales. Pour les raisons exposées plus haut, avec l'approche

bayésienne les étudiants sont très bien préparés à une lecture intelligente et critique. En fait, l'approche bayésienne est mieux adaptée que l'approche fréquentiste à la manière usuelle de présenter les résultats expérimentaux, du fait que cette dernière met rarement en jeu de manière explicite les concepts de base du raisonnement du test de signification (hypothèse nulle, seuil $\alpha \ldots$).

Examiner interactivement différentes distributions *a priori* et comparer les distributions *a posteriori* correspondantes avec la solution fiducio-bayésienne permet aux étudiants de se forger une intuition correcte et une compréhension sur les rôles relatifs des effectifs, des données et de l'information extérieure. Faire varier les effectifs respectifs des données disponibles et des données futures et examiner les distributions prédictives est également utile pour donner aux étudiants une compréhension intuitive du rôle des effectifs.

5.3 Quelques difficultés éventuelles avec l'approche bayésienne

Les difficultés de l'approche bayésienne les plus souvent dénoncées portent sur l'explicitation de la distribution *a priori*. Berry (1997) met en avant le fait que les distributions bayésiennes *a priori* et *a posteriori* sont subjectives et il oblige les étudiants à estimer leurs probabilités *a priori*, tout en reconnaissant les difficultés de cette tâche ("ils n'aiment pas ça"). Pour le moins, le rôle de la probabilité subjective devrait être clarifié (D'Agostini, 1999).

Cependant, dans la mesure où c'est l'analyse des données expérimentales qui nous préoccupe, je ne pense pas que ce soit une bonne stratégie d'attirer l'attention des étudiants (ou des chercheurs) sur une approche qui ne répond pas à leurs attentes (voir Section 2.1). C'est pourquoi nous évitons toujours – au moins *dans un premier temps* – le problème de l'estimation d'une distribution *a priori* "subjective" et nous axons notre enseignement sur les procédures fiducio-bayésiennes. Une fois que les étudiants sont devenus familiarisés avec leur utilisation et leur interprétation, il y a des façons motivantes d'introduire dans un deuxième temps les distributions *a priori* "informatives". En particulier, les étudiants sont généralement séduits pas l'idée d'explorer l'impact de distributions *a priori handicapantes* ("sceptiques") et d'examiner si les données constituent un contre-poids suffisant. Des distributions *a priori* qui expriment les résultats d'expériences antérieures sont également généralement bien acceptées. Finalement, on peut montrer comment l'explicitation des opinions *a priori* d'"experts" du domaine peut être utile dans certaines études, mais il faut insister sur le fait que cela nécessite des techniques appropriées (pour un exemple dans le domaine des essais cliniques, voir Tan *et al.*, 2003).

D'autres difficultés peuvent être dues à des confusions avec les interprétations fréquentistes. Par exemple, certains étudiants concluent de manière erronée de la distribution *a posteriori* que la différence *observée* – et non la différence parente – est grande, ce qui peut être dû à une confusion avec le raisonnement du test de signification (un résultat est significatif si la différence observée est "en un sens" grande). Une possibilité serait de ne pas enseigner les méthodes fréquentistes (Berry, 1997). Cependant, dans le contexte actuel, ce serait difficilement une attitude réaliste. Une solution pour affronter ce problème est d'utiliser l'approche de l'inférence combinatoire (ou ensembliste) proposée par Rouanet et Bert (2002) (voir aussi Rouanet, Bernard & Lecoutre, 1986 ; Rouanet, Bernard & Le Roux, 1990). En bref, cette approche consiste à écarter le caractère "aléatoire" du concept d'échantillon et à remplacer les formulations probabilistes par des formulations en termes de "proportions d'échantillons". La motivation pour l'enseignement est de permettre aux

étudiants d'apprendre les aspects *calculatoires* des procédures d'inférence fréquentistes sans être concernés prématurément par les difficultés conceptuelles des concepts probabilistes. En conséquence, les formulations probabilistes – et en particulier l'*interprétation* des procédures fréquentistes – sont réservés à l'approche bayésienne, minimisant ainsi les sources de confusion potentielles.

6 Conclusion

"*It could be argued that since most physicians use statement A* [the probability the true mean value is in the interval is 95%] *to describe 'confidence' intervals, what they really want are 'probability' intervals. Since to get them they must use Bayesian methods, then they are really Bayesians at heart !*" (Grunkemeier & Payne, 2002, page 1904)

De nos jours des méthodes bayésiennes de routine pour les situations familières d'analyse des données expérimentales sont faciles à mettre en œuvre. Elles satisfont les *desiderata* des scientifiques et sont mieux en accord avec leurs interprétations spontanées des données que les procédures fréquentistes. Par suite elles peuvent être enseignées à des étudiants et des chercheurs non-statisticiens sous une forme intuitive et motivante. L'utilisation des interprétations fiducio-bayésiennes (basées sur des distributions *a priori* non-informatives) des tests de signification et des intervalles de confiance dans le langage naturel des probabilités sur les effets inconnus vient très spontanément aux étudiants. En retour l'approche bayésienne évite les difficultés usuelles rencontrées avec les procédures fréquentistes, et en particulier les mauvais usages et les abus communs des tests de signification sont plus clairement compris. L'attention des utilisateurs peut être focalisée sur des stratégies plus appropriées telles que la considération de la signification pratique des résultats et la réplication des expériences.

Références

Albert, J. (1995). Teaching Inference about Proportions Using Bayes and Discrete Models. *Journal of Statistics Education 3*(3). Retrieved July 2, 2003, from http ://www.amstat.org/publications/jse/v3n3/albert.html.

Albert, J. (1996). *Bayesian Computation Using Minitab*. Wadsworth Publishing Company, Belmont.

Albert, J. (1997). Teaching Bayes' rule : a data-oriented approach. *The American Statistician* **51**, 247-253.

Albert, J. (2003). College students' conceptions of probability. *The American Statistician* **57**, 37-45.

American Psychological Association (2001). *Publication Manual of the American Psychological Association* (5th edition). Author, Washington, DC.

Amorim, M.A., Glasauer, S., Corpinot, K. & Berthoz, A. (1997). Updating an object's orientation and location during nonvisual navigation : A comparison between two processing modes. *Perception and Psychophysics* **59**, 404-418.

Amorim, M.-A., Loomis, J.M. & Fukusima, S.S. (1998). Reproduction of object shape is more accurate without the continued availability of visual information. *Perception* **27**, 69-86.

Amorim, M.-A. & Stucchi, N. (1997). Viewer- and object-centered mental explorations of an imagined environment are not equivalent. *Cognitive Brain Research* **5**, 229-239.

Amorim, M.-A., Trumbore, B. & Chogyen, P. L. (2000). Cognitive repositioning inside a desktop VE : The constraints introduced by first- versus third-person imagery and mental representation richness. *Presence : Teleoperators and Virtual Environments* **9**,165-186.

Azar, B. (1999). APA statistics task force prepares to release recommendations for public comment. *APA Monitor Online* **30**(5). Retrieved July 2, 2003, from http ://www.apa.org/monitor/may99/task.html.

Bakeman, R. (1999). Statistical matters (letter). *APA Monitor Online* **30**(7). Retrieved July 2, 2003, from http ://www.apa.org/monitor/julaug99/letters.html.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Verlag, New York.

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

Bernard, J.-M. (2000). Bayesian inference for categorized data. *In* H. Rouanet, J.-M. Bernard, M.-C. Bert, B. Lecoutre, M.-P. Lecoutre & B. Le Roux, *New Ways in Statistical Methodology : From Significance Tests to Bayesian Inference* (2nd edition), 159-226, Peter Lang, Bern, SW.

Berry, D. A. (1991). Experimental design for drug development : a Bayesian approach. *Journal of Biopharmaceutical Statististics* **1**, 81-101.

Berry, D. A. (1997). Teaching elementary Bayesian statistics with real applications in science. *The American Statistician* **51**, 241-246.

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

Box, G. E. P. & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison Wesley, Reading, MA.

Ciancia, F., Maitte, M., Honoré, J., Lecoutre, B. & Coquery, J.-M. (1988). Orientation of attention and sensory gating : 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.

Cumming, G., Williams, J. & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics* **3**, 299-311.

D'Agostini, G. (1999). Teaching statistics in the physics curriculum. Unifying and clarifying role of subjective probability**.** *American Journal of Physics*, 67, 1260-1268.

Desperati, C. & Stucchi, N. (1995). The role of eye-movements. Experimental Brain Research **105**, 254-260.

Dickey J. M. (1986). Discussion of Racine, A., Grieve, A. P., Flühler, H. & Smith, A. F. M., Bayesian methods in practice : Experiences in the pharmaceutical industry. *Applied Statistics* **35**, 93-150.

Dignam, J., Bryant, J., Wieand, H. S., Fisher, B.&d Wolmark, N. (1998). Early stopping of a clinical trial when there is evidence of no treatment benefit : protocol B-14 of the National Surgical Adjuvant Breast and Bowel Project. *Controlled Clinical Trials* **19**, 575-588.

Falk, R. & Greenbaum, C. W. (1995). Significance tests die hard. The amazing persistence of a probabilistic misconception. *Theory & Psychology* **5**, 75-98.

Fisher, R.A. (1990/1925). *Statistical Methods for Research Workers*. Oliver and Boyd, London. (Reprint, 14th edition, *in* Fisher, 1990).

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

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.

Gordon, H. R. D. (2001). American Vocational Education Research Association members' perceptions of statistical significance tests and other statistical controversies. *Journal of Vocational Education Research* **26**(2).

Grunkemeier, G. L. & Payne, N. (2002). Bayesian analysis : A new statistical paradigm for new technology. *The Annals of Thoracic Surgery* **74**, 1901-1908.

Haller, H. & Krauss, S. (2002). Misinterpretations of significance : A problem students share with their teachers ? *Methods of Psychological Research* **7**(1). Retrieved July 2, 2003, from http ://www.mpr-online.de.

Hand, D. J.& Taylor, C. (1987). Multivariate Analysis of Variances and Repeated Measures : A practical Approach for Behavioural Scientists, Chapman and Hall, London.

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.

Holender, D. & Bertelson, P. (1975). Selective preparation and time uncertainty. *Acta Psychologica* **39**, 193-203.

Jaynes, E. T. (1976). Confidence intervals vs Bayesian intervals. *In* W.L. Harper & C.A. Hooker (Eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. 2*, 175-257, D. Reidel, Dordrecht, Netherlands.

Jeffreys, H. (1998). *Theory of Probability* (3rd edition). Clarendon, Oxford (1st edition : 1939).

Johns, D. & Andersen, J.S. (1999). Use of prédictive probabilities in phase II and phase III clinical trials. *Journal of Biopharmaceutical Statistics* **9**, 67-79.

Jones, B. & Kenward, M. G. (1989). *Design and Analysis of Cross-over Trials*. Chapman and Hall, London.

Joseph, L. & Bélisle, P. (1997). Bayesian sample size determination for normal means and differences between normal means. *The Statistician* **46**, 209-226.

Kirk, R. E. (1982). *Experimental Design. Procedures for the Behavioral Sciences*. Brooks /Cole, Pacific Grove, CA.

Kirk, R. E. (2001). Promoting good statistical practices : Some suggestions. *Educational and Psychological Measurement* **61**, 213-218.

Laplace, P.-S. (1986/1825). *Essai Philosophique sur les Probabilités*. Christian Bourgois, Paris (English translation : *A Philosophical Essay on Probability*, 1952, Dover, New York).

Lecoutre, B. (1984). L'Analyse Bayésienne des Comparaisons [The Bayesian Analysis of Comparisons]. Presses Universitaires de Lille, Lille.

Lecoutre, B. (1996). *Traitement statistique des données expérimentales : des pratiques traditionnelles aux pratiques bayésiennes* (avec programmes Windows par B. Lecoutre & J. Poitevineau, disponibles gratuitement sur Internet à l'addresse http ://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris ou contacter bruno.lecoutre@univ-rouen.fr). DECISIA Editions, Paris, FR.

Lecoutre, B. (2000). From significance tests to fiducial Bayesian inference. *In* H. Rouanet, J.-M. Bernard, M.-C. Bert, B. Lecoutre, M.-P. Lecoutre & B. Le Roux, *New ways in statistical methodology : From significance tests to Bayesian inference* (2nd edition), 123-157, Peter Lang, Bern, SW.

Lecoutre B. (2001). Bayesian prédictive procedure for designing and monitoring experiments. *In Bayesian Methods with Applications to Science, Policy and Official Statistics*, 301-310, Office for Official Publications of the European Communities, Luxembourg.

Lecoutre, B. & Charron, C. (2000). Bayesian procedures for prediction analysis of implication hypotheses in 2×2 contingency tables. *Journal of Educational and Behavioral Statistics* **25**, 185-201.

Lecoutre, B. & Derzko, G. (2001). Asserting the smallness of effects in ANOVA. *Methods of Psychological Research* **6**(1), 1-32. Retrieved July 2, 2003, from http ://www.mpr-online.de.

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

Lecoutre, B., Lecoutre, M.-P. & Poitevineau, J. (2001). Uses, abuses and misuses of significance tests in the scientific community : won't the Bayesian choice be unavoidable ? *International Statistical Review* **69**, 399-418.

Lecoutre, B., Mabika, B. & Derzko, G. (2002). Assessment and monitoring in clinical trials when survival curves have distinct shapes in two groups : a Bayesian approach with Weibull modeling illustrated. *Statistics in Medicine* **21**, 663-674.

Lecoutre, B. & Poitevineau, J. (1992). PAC (*Programme d'Analyse des Comparaisons*) : *Guide d'utilisation et manuel de référence*. CISIA-CERESTA, Montreuil, France.

Lecoutre, B. & Poitevineau, J. (2000). Aller au delà des tests de signification traditionnels : vers de nouvelles normes de publication [Beyond traditional significance tests : Prime time for new publication norms]. *L'Année Psychologique* **100**, 683-713.

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

Lecoutre, M.-P. (2000). And... What about the researcher's point of view. In H. Rouanet, J.-M. Bernard, M.-C. Bert, B. Lecoutre, M.-P. Lecoutre & B. Le Roux, *New ways in statistical methodology : from significance tests to Bayesian inference* (2nd edition), 65-95, Peter Lang, Bern, SW.

Lecoutre, M.-P., Poitevineau, J. & Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. *International Journal of Psychology* **38**, 37-45.

Lecoutre, M.-P., Clément, E. & Lecoutre, B. (2004). Failure to construct and transfer correct representations across probability problems. *Psychological Reports* **94**, 151-162.

Lecoutre, M.-P. & Rouanet H. (1993). Predictive judgments in situations of statistical analysis. *Organizational Behavior and Human Decision Processes* **54**, 45-56.

Lee, P. (1997). *Bayesian Statistics : An Introduction* (2nd edition). Oxford University Press, Oxford.

Lindley, D. V. (1993). The analysis of experimental data : The appreciation of tea and wine. *Teaching Statistics* **15**, 22-25.

Mittag, K. C. & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher* **29**, 14-20.

Moore, D.S. (1997). Bayes for Beginners ? Some Pedagogical Questions. In S. Panchapakesan & N. Balakrishnan (eds.), *Advances in Statistical Decision Theory*, 3-17, Birkhäuser.

Morrison, D. E. & Henkel, R. E. (Eds.) (1970). *The Significance Test Controversy – A Reader*. Butterwoths, London.

Nelson, N., Rosenthal, R. & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist* **41**, 1299-1301.

Nester, M. R. (1996). An applied statistician's creed. *Applied Statistics* **45**, 401-410.

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

O'Hagan, A. (1996). *First Bayes* [Teaching package for elementary Bayesian Statistics]. Retrieved July 2, 2003, from http ://www.shef.ac.uk/∼st1ao/1b.html.

Pagano, R. R. (1990). *Understanding statistics in the behavioral sciences* (3rd edition). West, St. Paul, MN.

Phillips, L. D. (1973). *Bayesian Statistics for Social Scientists*. Nelson, London.

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

Poitevineau, J. & Lecoutre, B. (2001). The interpretation of significance levels by psychological researchers : The .05-cliff effect may be overstated.*Psychonomic Bulletin and Review* **8**, 847850.

Poitevineau, J. (2004). L'usage des tests statistiques par les chercheurs en psychologie : Aspects normatif, descriptif et prescriptif. *Mathématiques et Sciences Humaines* **167**, 5-25.

Robinson, D. H. & Wainer, H. (2002). On the past and future of Null Hypothesis Significance Testing. *Journal of Wildlife Management* **66**, 263-271.

Rosenthal, R. & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology***55**, 33-38.

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

Rouanet, H. (2000). Statistical practice revisited. *In* H. Rouanet, J.-M. Bernard, M.-C. Bert, B. Lecoutre, M.-P. Lecoutre & B. Le Roux, *New ways in statistical methodology : From significance tests to Bayesian inference* (2nd edition), 29-64, Peter Lang, Bern, SW.

Rouanet, H., Bernard, J.-M. & Lecoutre, B. (1986). Non-probabilistic statistical inference : A set theoretic approach. *The American Statistician* **40**, 60-65.

Rouanet, H., Bernard, J.-M. & Leroux, B. (1990). *Statistique en Sciences Humaines : Analyse Inductive des Données*. Dunod, Paris.

Rouanet, H. & Bert, M.-C. (2000). Introduction to combinatorial inference. *In* H. Rouanet, J.-M. Bernard, M.-C. Bert, B. Lecoutre, M.-P. Lecoutre & B. Le Roux, *New ways in statistical methodology : From significance tests to Bayesian inference* (2nd edition), 97-122, Peter Lang, Bern, SW.

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.

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

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. *In* L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), *What If There Were No Significance Tests ?* 335-392, Erlbaum, Hillsdale, NJ.

Savage, L. (1954). *The Foundations of Statistical Inference.* John Wiley & Sons, New York.

Schmitt, S. A. (1969). *Measuring Uncertainty : An Elementary Introduction to Bayesian Statistics.* Addison Wesley, Reading, MA.

Schweder, T. & Hjort, N. L. (2002). Confidence and likelihood. *Scandinavian Journal of Statistics* **29**, 309-332.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters : The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement* **61**, 605-632.

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

Steiger, J. H. & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. *In* L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What If There Were No Significance Tests ?* 221-257, Erlbaum, Hillsdale, NJ.

Student (1908). The probable error of a mean. *Biometrika* **6**, 1-25.

Tversky, A. & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin* **76**, 237-251.

Tyler, R. (1931). What is statistical significance ? *Educational Research Bulletin* **10**, 118-142.

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.

Zuckerman, M., Hodgins, H., Zuckerman, A. & Rosenthal, R. (1993). Contemporary issues in the analysis of data : A survey of 551 psychologists. *Psychological Science* **4**, 49-53