In I. Drouet (éd) Le bayésianisme aujourd'hui: Fondements et pratiques. Paris: Editions Matériologiques, 193-219.

Pourquoi les méthodes bayésiennes sont-elles si peu utilisées dans les publications expérimentales? Les pratiques actuelles réexaminées à partir des conceptions des fondateurs de l'inférence statistique

Bruno LECOUTRE
ERIS, LMRS, UMR 6085, CNRS et Université de Rouen
Avenue de l'université, BP 12
76801 Saint-Etienne-du-Rouvray cedex
bruno.lecoutre@univ-rouen.fr
http://lmrs.univ-rouen.fr/Persopage/Lecoutre/Eris.html

Résumé

Les publications expérimentales font un usage abondant des procédures d'inférence statistique. En psychologie, par exemple, il n'est pas rare de trouver dans un même article plusieurs dizaines de tests de signification de l'hypothèse nulle. Suivant les recommandations d'un nombre croissant de revues, l'usage des intervalles de confiance fréquentistes se répand également. En revanche les méthodes bayésiennes ne sont guère utilisées. Nous examinons les raisons possibles de cet état de fait. Nous discutons notamment les conceptions (généralement ignorées) de Jeffreys vis-à-vis de l'utilisation des méthodes bayésiennes dans le cas particulier de l'analyse des données expérimentales et nous en tirons les leçons. Notre objectif premier est d'essayer de faciliter le dialogue entre les tenants de l'approche bayésienne et ceux qui analysent des données expérimentales.

Abstract

Statistical inference procedures are very extensively used in experimental publications. In psychology, for instance, it is not unusual to find in a single article several tens of null hypothesis significance tests. According to the guidelines of an increasing number of reviews, the use of frequentist confidence intervals is becoming more widely used. In contrast, Bayesian methods are hardly used. We consider the possible reasons for this state of affairs. We especially discuss the (largely ignored) Jeffreys conceptions about the use of Bayesian methods in the specific case of experimental data analysis, and we learn the lessons from them. Our primary goal is to attempt to facilitate dialogue between Bayesian proponents and experimental data analysts.

1. Introduction

En psychologie, il n'est pas rare que deux ou trois expériences soient présentées dans un même article. Chacune d'elles fait souvent intervenir un plan d'expérience complexe avec plusieurs facteurs expérimentaux. La technique la plus utilisée est l'analyse de variance (*ANOVA*). Outre les tests statistiques des effets principaux et des effets d'interaction, sont souvent présentés des tests complémentaires portant sur des effets partiels et/ou des effets conditionnels. Cela peut représenter plusieurs dizaines de tests de tests de signification de

l'hypothèse nulle d'absence d'effet. Suivant les recommandations d'un nombre croissant de revues, l'usage des intervalles de confiance fréquentistes, qui répond au besoin de l'estimation des effets, commence à se répandre. Ces intervalles sont utilisés à la place, ou le plus souvent en complément, des tests. Mais ils sont rarement commentés, ou alors les auteurs se contentent de dire si l'intervalle contient ou exclut la valeur zéro, le réduisant au rôle d'un test d'absence d'effet. En revanche les méthodes bayésiennes ne sont guère utilisées.

L'exemple de la psychologie est d'autant plus intéressant qu'elle est souvent apparue comme un domaine précurseur dans l'utilisation de l'inférence statistique. Ainsi, dès les années 1970, dans lesquelles on peut situer le *renouveau* de l'inférence bayésienne, ont été publiés des ouvrages bayésiens destinés aux psychologues (notamment Phillips, 1973 ; Novick et Jackson, 1974). En France, Henry Rouanet et Dominique Lépine ont proposé aux psychologues l'utilisation systématique de méthodes *fiducio-bayésiennes* (par exemple, Rouanet, Lépine et Pelnard-Considère, 1976), ouvrant une voie de recherche que nous avons poursuivie (*cf* Lecoutre et Poitevineau, 2014).

Dès cette époque, Winkler (1974) s'interrogeait sur la réticence des psychologues à mettre en pratique les procédures inférentielles bayésiennes. La situation n'a guère changé et, dans l'article de Winkler, toujours d'actualité, on trouve que ceci « apparaît dû à une combinaison de facteurs incluant conviction philosophique, tradition, formation statistique, manque de disponibilité, difficultés de calcul, difficultés de présentation, et résistance perçue des rédacteurs de revues » (p. 129). Si nous laissons de côté le choix de l'approche philosophique, qui « n'est réellement pas aussi important que le fait que l'approche soit utilisée de manière cohérente, soigneuse et appropriée », aucun des arguments mentionnés n'est réellement convaincant.

Les bayésiens eux-mêmes ont certainement une part de responsabilité. Devant le front de défense uni des fréquentistes, qui se placent sous la bannière de « l'objectivité », les bayésiens présentent des lignes d'attaques dispersées. La diversité des approches pour le choix de la distribution *a priori* – non-informative, conjuguée, subjective… – devrait en principe répondre à des objectifs différents et complémentaires, tels que permettre d'exprimer l'apport propre des données, traduire des informations initiales « objectives », formaliser l'opinion d'experts du domaine, etc. La considération d'un ensemble de distributions *a posteriori* correspondant à différentes distributions *a priori* pertinentes est aussi une possibilité séduisante pour répondre au souci d'objectivité. Mais ceux qui analysent des données expérimentales sont souvent désemparés devant cette diversité, d'autant plus que les bayésiens leur apparaissent souvent ne pas être d'accord entre eux. Il en résulte que les méthodes bayésiennes sont souvent perçues comme trop compliquées à mettre en pratique et/ou trop *subjectives* pour être acceptées par la communauté scientifique. De plus, en insistant souvent sur l'approche décisionnelle, les bayésiens ne répondent pas nécessairement aux besoins des expérimentateurs. Soulignons bien le fait que nous nous restreignons ici à la situation de l'analyse des données en vue de *la communication des résultats à la communauté scientifique*. Cela ne préjuge donc pas de l'intérêt que pourrait avoir une approche décisionnelle, prenant éventuellement en compte des éléments subjectifs, dans d'autres situations.

Notre objectif premier est d'essayer de susciter (et si possible faciliter) le dialogue entre les tenants de l'approche bayésienne et ceux qui analysent des données expérimentales. Dans la mesure du possible, nous éviterons ici les aspects *normatifs* (voir pour cela le chapitre de Christian Robert) – quelle est d'un point de vue théorique la « meilleure » approche ? Nous n'aborderons pas non plus les aspects polémiques, concernant les controverses sur *l'usage* qui est fait des tests statistiques (« *the significance test controversy* »), celles-ci ayant fait l'objet

d'un nombre considérable de publications, ou encore celles portant sur des questions telles que « qu'est-ce qu'une procédure statistique objective ? »

Il convient d'abord de remarquer que la quasi-totalité des utilisateurs de l'inférence statistique, mais aussi la majorité des statisticiens, n'ont connaissance des procédures actuellement en usage qu'à travers les textes « modernes ». Il en résulte qu'ils ignorent souvent les débats qui ont opposés ceux qui sont présentés comme les grands fondateurs de la statistique, et en particulier qu'ils ne distinguent pas clairement les points de vue de Fisher d'une part et de Neyman et Pearson d'autre part. Pour ne prendre qu'un exemple, il est devenu si habituel de parler de « l'hypothèse nulle H_0 » que l'on est surpris d'apprendre que Fisher n'a jamais utilisé la notation H_0 et que Neyman et Pearson n'ont jamais parlé d'« hypothèse nulle. » Cette remarque pourrait sembler pointilleuse, mais elle révèle en fait que ces auteurs avaient des conceptions très différentes sur le rôle de l'inférence statistique dans la démarche scientifique. Il nous a donc paru utile de commencer par rappeler les définitions des tests statistiques données, non seulement par Fisher et Neyman-Pearson, mais aussi par Jeffreys pour ce qui concerne l'approche bayésienne.

Il s'agira ensuite d'examiner les pratiques actuelles dans l'analyse statistique des données expérimentales, notamment à partir les *normes* officielles qui existent pour ces pratiques. A titre d'illustration, nous décrirons d'une part les lignes directrices (« *guidelines* ») imposées aux statisticiens de l'industrie pharmaceutique pour l'analyse des essais cliniques en vue de la mise sur le marché de nouveaux médicaments, et d'autre part les directives données pour la majorité des revues de psychologie. Nous reviendrons ensuite sur les raisons qui ont conduit les auteurs précédents à des définitions différentes des tests statistiques. Nous présenterons d'abord leurs conceptions de l'utilisation de l'inférence statistique dans la démarche expérimentale – avec notamment la distinction entre décision et estimation, – puis nous examinerons le statut qu'ils attribuaient aux probabilités bayésiennes. Nous reviendrons enfin plus particulièrement sur les conceptions (généralement ignorées) de Jeffreys vis-à-vis de l'utilisation des méthodes bayésiennes dans le cas particulier de l'analyse des données expérimentales.

Notre exposé sera très loin d'être exhaustif et nous espérons que nos choix – qui reflètent inévitablement de façon plus ou moins implicite nos propres conceptions – ne paraîtront pas subjectifs. Nous nous efforcerons pour cela d'adopter une présentation la plus descriptive et la moins polémique possible, nous référant aux textes mêmes des auteurs.

Les aspects prescriptifs – quelles procédures bayésiennes sont appropriées pour l'analyse et la présentation des données expérimentales ?– ne sont pas l'objet du présent texte. En conclusion, nous aborderons très brièvement notre point de vue. Bien entendu la question reste largement ouverte.

2. Trois définitions des tests statistiques

2.1 Le test de signification de Fisher

Sir Ronald Fisher (1890-1962), généticien et statisticien britannique, est reconnu pour avoir révolutionné la statistique. Ses trois livres, réédités dans un seul volume en 1990 (Fisher, 1990a, 1990b, 1990c), étaient avant tout destinés aux chercheurs et ont eu un succès considérable, avec chacun plusieurs éditions.

Dans le *test de signification* de Fisher, une seule hypothèse, appelée « l'hypothèse nulle » est considérée. C'est une hypothèse que l'on cherche à réfuter (« *to be nullified* »), souvent, mais pas nécessairement, l'hypothèse que le paramètre a une valeur nulle. On considère une statistique de test appropriée dont la distribution d'échantillonnage, quand cette l'hypothèse

est vraie, est exactement connue. Celle-ci donne la probabilité que la statistique de test « excède par chance » la valeur observée dans l'échantillon, *si l'hypothèse nulle est vraie*. C'est le seuil (ou niveau) de signification, aujourd'hui appelé « *p-value* ». Le résultat expérimental est jugé *significatif* – l'hypothèse nulle est réfutée – quand p est jugé suffisamment petit. En pratique, on peut considérer la probabilité que la statistique de test excède la valeur observée dans une direction ou dans l'autre relativement à la valeur fixée par l'hypothèse nulle – test bilatéral (« *two-sided* ») – ou la probabilité qu'elle l'excède dans une direction donnée – test orienté ou unilatéral (« *one-sided* »).

2.2 Le test d'hypothèses de Neyman-Pearson

Jerzy Neyman (1894-1981), mathématicien et statisticien d'origine polonaise, et Egon Pearson (1895-1980, fils de Karl Pearson), statisticien britannique, ont publié leurs articles de base en 1928 et 1933. Leur objectif était de donner des règles de *comportement rationnel*, pour prendre des *décisions statistiques* sur des hypothèses. Cette collaboration a conduit plus tard Neyman à formuler sa méthode *d'intervalles de confiance* avec la même perspective.

Neyman et Pearson ont rejeté la conception de Fisher d'une seule hypothèse et mis en avant la nécessité d'hypothèses alternatives. Pour cela, on considère des hypothèses mutuellement exclusives, généralement deux, notées H_0 et H_1 (mais il peut y avoir plusieurs hypothèses alternatives), H_0 étant appelée l'hypothèse à tester. Le *test d'hypothèses* de Neyman-Pearson est une règle de décision basée sur la division de l'espace des échantillons en deux régions: une région critique pour laquelle on rejette H_0 et une région d'acceptation (complémentaire) pour laquelle on accepte H_0 (il peut y avoir aussi une région d'indécision, mais ceci est rarement considéré en pratique). Le rôle du test est de minimiser « sur le long terme » la proportion de décisions erronées. Il y a deux types d'erreurs: Type I, rejeter H_0 quand elle est vraie dont la probabilité est notée α et Type II, accepter H_0 quand H_1 est vraie dont la probabilité est notée β. Pour *l'hypothèse alternative* H_1, la *puissance* d'une région critique est la probabilité de rejeter l'hypothèse testée H_0 quand H_1 est vraie, sous la condition que α *est fixée* ; cette probabilité est donc égale à $1-\beta$. Plus généralement la puissance est une fonction du paramètre. Le, maintenant célèbre, lemme de Neyman-Pearson fournit, au moins sous certaines conditions (test d'hypothèses ponctuelles), un moyen de trouver une *meilleure région critique*, ce qui définit en un sens un test optimal (« uniformément plus puissant »).

La plupart des intervalles de confiance, au sens de Neyman, utilisés en pratique peuvent être obtenus par *inversion* d'un test d'hypothèses : intuitivement, si on teste chaque valeur possible du paramètre, l'intervalle est l'ensemble des valeurs qui ne sont pas rejetées par ce test. Un intervalle de « niveau de confiance » 95% (par exemple) doit satisfaire la probabilité fréquentiste d'échantillonnage suivante : pour toute valeur fixée du paramètre, sur le long terme 95% au moins des intervalles calculés contiennent cette valeur. Si la probabilité excède 95% l'intervalle est dit *conservateur*. En pratique on utilise aussi des intervalles *approchés*, dont la probabilité d'échantillonnage fluctue autour de 95%.

2.3 Le test de signification bayésien de Jeffreys

Sir Harold Jeffreys (1891-1989) fut un géophysicien britannique réputé. Son principal livre statistique, d'abord publié en 1939, est considéré comme la première tentative de développer une théorie formelle de l'inférence statistique basée sur l'approche bayésienne. Deux éditions augmentées furent publiées en 1948 et 1961, celle-ci ayant été réimprimée en 1967 avec quelques corrections (Jeffreys, 1967).

Dans la suite des travaux pionniers de Bayes et Laplace, Jeffreys a cherché à développer des méthodes bayésiennes « objectives », applicables quand on ne connaît rien sur la valeur

du paramètre (« pas d'information initialement »): « nous recherchons principalement une théorie qui peut être utilisée dans la première étape d'un sujet » (Jeffreys, 1967, p. 252). Les probabilités bayésiennes *a priori* sont utilisées pour cet objectif : « Si nous n'avons *pas d'information pertinente* sur la valeur réelle d'un paramètre, la probabilité doit être choisie de manière à exprimer le fait que nous n'avons aucune information » (Jeffreys, 1967, p. 118, italiques ajoutées). Il a développé pour cela la *règle de Jeffreys* (ainsi appelée maintenant), sur laquelle nous ne nous étendrons pas. L'*a priori* de Jeffreys qui en résulte est usuellement appelé non-informatif (le plus fréquemment), objectif, de référence, ou par défaut.

Pour Jeffreys la fonction du test de signification est de comparer une *valeur* [précise] *suggérée* d'un « nouveau paramètre » (nous y reviendrons), souvent zéro, avec l'ensemble des autres valeurs possibles. Pour cela il considérait deux hypothèses complémentaires, qu'il notait q et q' : « q, que le paramètre a cette valeur suggérée, et q', qu'il a une autre valeur qui est *à déterminer à partir des observations* (Jeffreys, 1967, p. 246, italiques ajoutées). Comme Fisher, il utilisait les termes hypothèse nulle. Dans le but de « dire que nous n'avons pas d'information initialement », il semblait évident à Jeffreys que les deux hypothèses sont initialement également probables, donc une probabilité ½ pour l'hypothèse nulle, et une probabilité ½ à répartir sur l'ensemble des autres valeurs possibles selon une distribution *a priori* ne favorisant aucune valeur particulière du paramètre.

Jeffreys a proposé de mesurer l'évidence contre l'hypothèse nulle par le rapport des quotes (*odds ratios*) *a* posteriori aux quotes *a priori*, qu'il notait K. De nos jours K est appelé le *facteur de Bayes*. En pratique, Jeffreys (1967, p. 432) a proposé de graduer le degré d'évidence. Une valeur $K > 1$ (degré 0) est à l'appui de l'hypothèse nulle (« *supported* »). Autrement l'évidence contre q est qualifiée selon 5 degrés :

(1) $1 > K > 10^{-1/2}$ (0.3162) ne vaut pas plus qu'une simple mention (« *a bare mention* »
(2) $10^{-1/2} > K > 0.1$, est appréciable (« *substantial* »)
(3) $0.1 > K > 10^{-3/2}$ 0.0316), est forte (« *srong* »)
(4) $10^{-3/2} > \mathrm{K} > 0.01$, est très forte (« very *srong* »)
(5) $0.01 > K$, est déterminante (« decisive »)

2.4 En conclusion de la section

Sans surprise, c'est l'approche décisionnelle des tests d'hypothèses de Neyman-Pearson, beaucoup plus formelle que celle du test de signification de Fisher, qui est devenue la théorie fréquentiste « officielle » de la statistique mathématique. Mais l'apport de Fisher reste omniprésent dans les publications expérimentales (au moins), avec l'utilisation de la *p-value* et l'usage de sa terminologie : hypothèse nulle, significatif *vs* non-significatif, etc. Pourtant la conception de Fisher n'est pas compatible avec l'approche décisionnelle de Neyman-Pearson. En particulier, la *p-value* ne joue aucun rôle dans leur test d'hypothèses et, en toute rigueur, ne devrait donc pas être considérée par ses utilisateurs.

Quand ils présentent le test bayésien d'une hypothèse nulle ponctuelle, les statisticiens bayésiens actuels ne manquent pas de souligner le fait que, pour les tailles d'échantillon usuels, quand l'hypothèse nulle d'une valeur précise est rejetée par un test fréquentiste, la probabilité *a posteriori* bayésienne de l'hypothèse nulle est généralement nettement plus grande que la *p-value*. Pour Berger (2003, p. 3), cela démontre que la « l'interprétation erronée trop commune des *p-values* comme probabilités d'erreur résulte très souvent en une surestimation considérable de l'évidence contre H_0. » Cela traduirait une différence radicale (« *dramatic* ») entre l'approche bayésienne de Jeffreys et l'approche de Fisher. Mais on peut aussi se demander si des probabilités aussi radicalement différentes qu'une probabilité d'échantillonnage conditionnelle au paramètre et une probabilité *a posteriori* conditionnelle

aux données sont ici directement comparables, et si elles ne doivent pas plutôt être jugées avec des critères différents.

3. Les utilisateurs face à la norme

Un élément nouveau à prendre en compte, à notre époque où tout devient de plus en plus réglementé, est qu'ont été édictées des normes pour la mise en œuvre et la présentation des procédures d'inférence statistique. Une particularité des psychologues est qu'ils procèdent souvent eux-mêmes aux analyses statistiques, à l'aide des logiciels dont ils disposent. Il est donc intéressant de comparer leurs pratiques à celles qui concernent le développement de nouveaux médicaments par l'industrie pharmaceutique. En effet, dans ce domaine, les analyses statistiques sont effectuées par des statisticiens professionnels. Comme en psychologie, l'usage des tests statistiques a été largement débattu mais ils restent la procédure la plus répandue.

3.1 La réglementation statistique pour les essais cliniques

Le développement de nouveaux médicaments par l'industrie pharmaceutique fait apparaître des questions statistiques spécifiques, dont les réponses doivent obéir à un cadre réglementaire. Ainsi la Conférence Internationale d'Harmonisation (*International Conference on Harmonization of Technical Requirements for Registration of Pharmaceutical for Human Use* – ICH), qui réunit les autorités réglementaires et l'industrie pharmaceutique de l'Europe, du Japon et des Etats-Unis, a rédigé des « lignes directrices » (*guidelines*), désignées ci-après par « l'ICH E9 », qui fournissent les principes statistiques à suivre dans un essai clinique.

Les essais de supériorité. La plus grande partie des directives de l'ICH E9 concernent les *essais de supériorité*, c'est-à-dire des études expérimentales *comparatives* destinées à démontrer qu'un traitement, typiquement un nouveau médicament, est supérieur à un autre, typiquement un médicament de référence, dans une indication médicale donnée. La procédure préconisée utilise l'approche des tests d'hypothèses de Neyman-Pearson, basée sur la puissance (*power-based*). L'hypothèse testée, dite « privilégiée », est l'absence de différence entre les deux traitements. L'hypothèse alternative, dite « de travail », est l'existence d'une différence ayant une valeur spécifiée à l'avance. Le choix de cette valeur doit être justifié, soit par un jugement portant sur l'effet minimal pertinent au sens clinique, soit par un jugement sur l'effet attendu du nouveau traitement, la valeur étant plus grande dans cette deuxième éventualité. On utilise habituellement un test bilatéral et la probabilité de rejeter à tort l'hypothèse d'absence de différence si elle est vraie (erreur de type I) est généralement fixée de manière conventionnelle à α=5% (éventuellement moins). Avant de commencer l'expérimentation, on détermine le nombre de sujets nécessaire pour que la probabilité de rejeter cette même hypothèse, lorsque l'hypothèse de travail alternative est vraie, soit suffisamment grande : au moins 0.80 ou 0.90 (soit respectivement β=20% et β=10% pour l'erreur de type II).

L'ICH E9 s'inscrit donc prioritairement dans la perspective décisionnelle de l'approche de Neyman-Pearson. Ceci est en accord avec le fait que le processus d'approbation des médicaments a, par tradition, un caractère décisionnel : acceptation/rejet. En conséquence l'utilisation des tests d'hypothèses est toujours très prégnante dans l'industrie pharmaceutique. Mais on peut objecter que cette pratique aboutit au paradoxe bien connu suivant. (1) Si le test n'est pas assez puissant, on risque de ne pas pouvoir démontrer la supériorité du nouveau médicament ; (2) s'il est trop puissant, on risque de conclure à sa supériorité alors que la

différence vraie avec le médicament de référence est en fait triviale. La nécessité de procéder à l'estimation de la différence est manifeste.

L'ICH E9 prend en compte cette objection et reconnaît explicitement l'insuffisance d'une décision en tout ou rien. Il est ainsi recommandé de rapporter la valeur exacte du seuil observé des tests (la « *p-value* »), plutôt que de faire référence exclusivement à une valeur critique α. De plus, il est suggéré que les valeurs de *p* peuvent être utilisées pour évaluer les différences : « Le calcul des *p-values* est parfois utile, soit comme une aide pour évaluer une différence spécifique d'intérêt, soit comme un indice de signalisation appliqué à un grand nombre de variables de sécurité ou de tolérabilité pour mettre en avant des différences méritant davantage d'attention (page 1935). » Cette suggestion pose cependant un réel problème, puisque la valeur de *p* dépend des effectifs et n'est donc qu'un indicateur très indirect de la différence vraie. La procédure, plus satisfaisante, également recommandée est de rapporter, outre la valeur de *p*, une estimation de la différence (la « taille de l'effet ») et un intervalle de confiance.

Les essais d'équivalence et de non infériorité. L'ICH E9 traite différemment le cas des *essais d'équivalence*. Pour pouvoir conclure à l'équivalence de deux médicaments, il s'agit de démontrer que ceux-ci diffèrent au plus d'une « marge de petitesse », c'est-à-dire d'une quantité *cliniquement négligeable* justifiée scientifiquement. De nombreux statisticiens ont objecté que, dans cette situation, le test d'absence de différence n'est pas satisfaisant, même si on prend en compte *a posteriori* ce qu'aurait été sa puissance si les médicaments n'étaient pas équivalents (au sens précédent). Ce raisonnement indirect ne garantit pas nécessairement l'équivalence dans le cas où le test ne permet pas de rejeter l'hypothèse d'absence de différence. Selon l'ICH E9, ce test ne doit donc pas être utilisé, et il est explicitement demandé d'utiliser une procédure spécifique pour cette situation, de préférence un intervalle de confiance.

C'est également l'utilisation exclusive d'un intervalle de confiance qui est recommandée pour les *essais de non infériorité*, dans lesquels il s'agit de démontrer qu'un nouveau médicament n'est pas inférieur à un médicament de référence de plus d'une marge admissible.

On voit ainsi que, malgré l'intérêt potentiel de plus en plus souvent reconnu, d'une démarche bayésienne, c'est l'inférence fréquentiste traditionnelle qui reste l'approche dominante et qui est imposée par les règles de l'ICH E9. Un document de la Food and Drug Administration (2010) constitue bien un début de reconnaissance des méthodes bayésiennes dans un cadre réglementaire, mais il reste une exception. En outre ce document restreint leur usage aux situations où l'on dispose de « bonnes informations », un concept qui apparaît difficile à définir et justifier en pratique.

3.2 Les normes de publication en psychologie

Dans la deuxième moitié des années 1990, le bureau des affaires scientifiques de l'*American Psychological Association* (APA) a chargé une « *Task Force* » d'étudier le rôle du test de signification dans la recherche en psychologie. A l'issue de longs débats contradictoires, pouvant laisser espérer une remise en question des pratiques, la conclusion peut apparaître décevante dans la mesure où elle n'a fait qu'encourager la poursuite de ces pratiques. Ainsi, même si la présentation est moins détaillée et moins formelle, les extraits suivants des recommandations faites (Wilkinson & APA Task Force on Statistical Inference, 1999) apparaissent en plein accord avec l'ICH E9.

« ***Tests d'hypothèses***. Il est difficile d'imaginer une situation dans laquelle une décision dichotomique accepter-rejeter est meilleure que de rapporter une *p* value exacte ou, mieux encore, un intervalle de confiance.

Puissance et taille d'échantillon. Fournir des informations sur la taille de l'échantilllon et sur le processus qui a conduit à la décision de cette taille.

Grandeur de l'effet **(**« ***effect size*** »***)***. Toujours fournir un estimateur de la grandeur de l'effet quand on rapporte une *p* value.

Intervalle d'estimation. Des intervalles d'estimation devraient être donnés pour toute grandeur d'effet concernant les résultats principaux. »

L'inférence bayésienne est bien mentionnée dans le rapport de la *Tak Force*, mais elle est restreinte à deux situations très particulières, comme une possibilité (parmi d'autres) : d'une part l'usage de méthodes bayésiennes empiriques dans le cas de comparaisons multiples, et d'autre part l'usage de distribution *a posteriori* dans le cadre du « modèle causal de Rubin » pour l'inférence causale.

Les recommandations concernant les méthodes fréquentistes ont été reprises sous une forme voisine dans la 6ème édition du « manuel de publication » de l'APA (American Psychological Association, 2010, pp. 33-35), qui sert de référence pour un grand nombre de revues. Mais les méthodes bayésiennes y sont ignorées, à l'exception des termes et définitions « BIC Critère d'information bayésien » et « p_{rep} La probabilité [prédictive bayésienne] qu'une réplique donne un résultat de même signe que le résultat initial » (voir plus loin), qui apparaissent de manière anecdotique dans la liste des « abréviations et symboles statistiques ».

3.3 Un cercle vicieux

Les principales lignes directrices sont donc les mêmes dans les deux domaines considérés. La norme qui en résulte constitue en fait un cercle vicieux dont il sera très difficile de sortir : d'une part cette norme impose, ou au moins favorise très fortement, l'usage des procédures d'inférence fréquentistes, mais d'autre part elle ne fait qu'officialiser les pratiques existantes et donc les renforcer. Une critique communément faite à ces pratiques est qu'elles n'ont pas de réelle justification, théorique ou méthodologique, ce qui entraîne de nombreux problèmes. Elles correspondent à ce que Gigerenzer a appelé une « logique hybride » de l'inférence statistique : « C'est un méli-mélo incohérent de certaines des idées de Fisher d'un côté et de certaines des idées de Neyman et E.S. Pearson de l'autre » (Gigerenzer, 1993, p. 314).

La norme en vigueur est effectivement un amalgame de différentes procédures plus ou moins compatibles :

– la conception décisionnelle du test d'hypothèses de Neyman-Pearson, avec notamment les notions d'hypothèse alternative et de puissance ;

– la conception du test de signification de Fisher avec l'usage du seuil observé, exclu par l'approche décisionnelle ;

– l'utilisation d'intervalles de confiance, lesquels dans l'ICH E9 sont en outre utilisés soit comme procédure complémentaire dans les essais de supériorité, soit comme procédure principale dans les essais d'équivalence ou de non infériorité.

Pour ne prendre qu'un exemple indiscutable, le test utilisé dans les essais de supériorité n'apparaît pas cohérent avec l'approche utilisée pour les essais d'équivalence et de non infériorité. D'une part, on utilise généralement un test bilatéral qui, en toute rigueur, ne permet pas de se prononcer sur la direction de l'effet, alors que la question posée est manifestement orientée. D'autre part, le test de l'absence de différence, quand l'hypothèse privilégiée est rejetée, permet seulement de conclure que la différence est non nulle. C'est

l'hypothèse d'un effet minimal pertinent au sens clinique – et non l'hypothèse d'absence d'effet – qui devrait être l'hypothèse privilégiée, testée.

3.4 On n'échappe pas au fréquentisme

Dans le domaine des « innovations » des psychologues, mentionnons la proposition de Killeen (2005) de présenter le résultat du test pour comparer deux moyennes (ou plus généralement d'un test portant sur un contraste entre moyennes) sous la forme de la probabilité prédictive, conditionnellement aux données de l'expérience réalisée, de retrouver un effet de même signe dans une réplique de cette expérience (p_{rep}). D'un point de vue formel p_{rep} peut être dérivée par un argument bayésien supposant une distribution *a priori* non-informative. D'un point de vue pratique, elle peut être déduite directement de la *p-value*, et en un sens lui est donc équivalente. Mais elle a une autre interprétation, puisque p_{rep} est une expression *bayésienne prédictive* du résultat statistique de l'expérience. A la suite de cette proposition, la revue *Psychological Science* ?, et plus généralement les revues de l'*Association for Psychological Science*, ont recommandé son usage : « les auteurs sont encouragés à utiliser prep plutôt que les p values ». Même si cela peut paraître anecdotique, pour la première fois une probabilité bayésienne a été rapportée de manière routinière dans des revues de psychologie. On aurait donc pu espérer que cela ouvrît la voie à des débats constructifs sur le rôle de l'inférence bayésienne (cf Lecoutre, Poitevineau et Lecoutre, 2010). On a seulement eu des critiques systématiques qui ont conduit à l'abandon de la recommandation. La plupart des auteurs de ces critiques n'ont pas admis, ou n'ont pas compris, la justification bayésienne. Ainsi, Iverson, Lee et Wagenmakers (voir Lecoutre et Killeen, 2010) ont confondu la probabilité prédictive de réplication, conditionnelle aux données observées avec la probabilité conjointe d'observer le même signe dans deux expériences futures, conditionnelle au paramètre (probabilité fréquentiste d'échantillonnage).

3.5 L'interprétation bayésienne naïve des procédures fréquentistes

Même si le fréquentisme règne sur les publications des résultats expérimentaux, les interprétations spontanées de ses procédures (seuils de signification, intervalles de confiance), même par des utilisateurs « avertis », sont le plus souvent en termes de probabilités sur les paramètres, qui leur apparaissent comme les probabilités *naturelles*: celles qui vont « du connu vers l'inconnu ». Ainsi, beaucoup d'utilisateurs interprètent incorrectement la « *p-value* » comme une probabilité bayésienne « inverse » – 1-*p* est la probabilité que l'hypothèse alternative est vraie – ou encore comme une probabilité bayésienne prédictive : 1-*p* est la probabilité de retrouver un résultat significatif dans une réplique de l'expérience. De même ils interprètent le niveau de confiance comme une probabilité bayésienne. Plus encore, ces interprétations « hérétiques » sont tolérées, ou même utilisées par ceux qui leur inculquent les principes de ces procédures, enseignants, auteurs d'ouvrages statistiques, etc.

On pourrait multiplier les exemples. Pour n'en prendre qu'un, dans un ouvrage d'introduction à la statistique dans une collection destinée au grand public, dont l'objectif est de permettre au lecteur d' « accéder aux intuitions profondes du domaine », on trouve l'interprétation bayésienne suivante de l'intervalle de confiance (ou « fourchette ») pour une proportion: « si dans un sondage de taille 1000, on trouve *P* [la proportion observée] = 0.613, la proportion π_1 à estimer a une probabilité 0.95 de se trouver dans la fourchette: [0.58,0.64] » (Claudine Robert, 1995, page 221).

Les limites 0.58 et 0.64 sont des valeurs fixées par la réalisation du sondage, un évènement unique. La proportion π_1 se trouve, ou ne se trouve pas, dans la fourchette, et on ne peut pas attribuer de probabilités fréquentistes (autre que 0 ou 1) à ces énoncés. La conception

fréquentiste de traiter les données comme aléatoires, même quand elles sont connues leur paraît si étrange que la quasi-totalité des utilisateurs trouvent l'interprétation bayésienne, non seulement désirable, mais aussi correcte.

3.6 En conclusion de la section

L'interprétation bayésienne naïve ajoute encore au caractère hybride de l'amalgame de procédures que l'on trouve dans les publications expérimentales. En particulier, l'utilisation qui se répand des intervalles de confiance devrait conduire à une situation difficilement admissible sur le plan scientifique : « il ne serait pas scientifiquement sain de justifier une procédure par des arguments fréquentistes et de l'interpréter en termes bayésiens » (Rouanet, 2000, p. 54). Les défenseurs du « choix bayésien » peuvent donc avoir des raisons d'espérer que celui-ci deviendra tôt ou tard inévitable (Lecoutre, Lecoutre et Poitevineau, 2001). Mais les utilisateurs peuvent pour leur part penser que cet amalgame des justifications et des interprétations est légitimé par une longue pratique, laquelle est renforcée par les recommandations officielles, et qu'il leur donne « le meilleur » des différentes approches.

4. Prendre des décisions ou apprendre des données et de l'expérience?

4.1 Neyman-Pearson: Des décisions automatiques vues comme un comportement inductif

Même si Neyman et Pearson ont discuté le cas de « l'investigation scientifique », leur exemple principal pour illustrer leurs tests d'hypothèses concernait les processus de contrôle de qualité. Dans cette perspective, leur préoccupation essentielle était d'éviter de commettre des erreurs de décisions. Ils ont ainsi beaucoup insisté sur le fait qu'un test d'hypothèses n'a pas pour objet de porter un jugement sur la véracité (ou la fausseté) d'une hypothèse (Neyman et Pearson, 1933a), mais de décider de *se comporter comme si* l'hypothèse était vraie (ou fausse).

Leur approche est présentée comme seulement *déductive*. Ainsi quand il développa plus tard la notion d'intervalle de confiance, Neyman précisa : « tout le raisonnement qui est derrière cette méthode est clairement déductif » (Neyman, 1951, p. 85). Ceci l'amena à introduire l'expression « comportement inductif », « raisonnement inductif » étant pour lui « hors de propos » (Neyman, 1938).

4.2 Fisher: Apprendre à partir des données expérimentales

Fisher a reconnu l'utilité des tests « d'acceptation » pour la prise de décision dans certains domaines (« commerce », « technologie »). Mais, en ce qui concerne la recherche scientifique (sa principale préoccupation), il s'est toujours violemment opposé à une approche qui conduit à des décisions automatiques concernant « l'interprétation détaillée d'observations vérifiables » : « L'idée que cette responsabilité puisse être déléguée à un ordinateur géant programmé avec des Fonctions de Décision relève de l'imagination de milieux assez éloignés de la recherche scientifique » (Fisher, 1990c, p. 105).

En conséquence, il considérait que le but du test de signification n'était pas de prendre des décisions, mais « d'apprendre à partir des données expérimentales », et dans cette perspective, d'être utilisé comme « *une aide au jugement* » (Fisher, 1990a, p. 128). Il apparaît aussi que, pour Fisher, le rôle du test de signification est lié sa conception de la causalité (Lecoutre,

2004). « … la corrélation n'est pas la causalité. Le fait est que si deux facteurs, *A* et *B*, sont associés – clairement, positivement, avec comme je dis *une signification statistique* – il se peut que *A* soit une cause importante de *B*, il se peut que *B* soit une cause importante de *A*, il se peut que *A* qu'autre chose, disons *X*, soit une cause importante des deux. *Si, maintenant , A la cause supposée a été randomisée* – a été attribuée au hasard dans le dispositif ayant conduit à l'observation – alors on peut exclure d'un coup la possibilité que *B* cause *A*, ou que *X* cause *A*. Nous savons parfaitement ce qui cause *A* – le lancement du dé ou les chances des nombres obtenus par échantillonnage aléatoire, et rien d'autres » (Fisher, 1959, p. 14, italiques ajoutées). Ainsi, pour Fisher la randomisation est la « base physique » *suffisante* qui permet de dépasser la corrélation pour obtenir – *avec le test de signification* – une interprétation causale, au sens expérimental. Ceci peut expliquer son insistance sur le test de signification.

Cependant, pour Fisher, et contrairement à Neyman, l'inférence statistique met en jeu à la fois le raisonnement déductif et le raisonnement inductif : « L'examen statistique d'un corpus de données est ainsi logiquement similaire à l'alternance générale de la méthode inductive et de la méthode déductive dans les sciences » (Fisher, 1990a, p. 8). Il en résulte que l'estimation a un rôle encore plus fondamental que le test de signification : « Il faut reconnaître que, selon son approche de l'inférence statistique pour les données expérimentales, Fisher paraît avoir attribué aux tests un statut largement préparatoire à l'estimation » (Rosenkrantz, 1973, p. 304). Ce statut a malheureusement été ignoré dans la pratique.

4.3 Jeffreys: Une méthodologie générale pour apprendre des données et de l'expérience

Jeffreys est allé encore plus loin que Fisher et a affiché explicitement l'ambition de proposer une méthodologie générale pour *apprendre des données et de l'expérience*, applicable à la recherche dans tous les domaines de la science. Les probabilités bayésiennes sont successivement mises à jour quand de nouvelles données deviennent disponibles.

Jeffreys distinguait clairement, d'une part les *problèmes d'estimation*, et d'autre part les *tests de signification* qui « mettent en jeu une valeur tout particulièrement suggérée d'un *nouveau* paramètre » (Jeffreys, 1967, p. 246, italiques ajoutées). Il avait donc une conception des tests statistiques concernant ce qu'il est maintenant usuel d'appeler « sélection de modèles » : « la fonction des tests de signification est de fournir un moyen d'arriver, *dans les cas appropriés*, à une décision qu'au moins un nouveau paramètre est nécessaire pour fournir une représentation adéquate des données et des inférences valides pour des données futures » (Jeffreys, 1967, p. 245, italiques ajoutées). Mais, pour Jeffreys, la question posée dans un test de signification n'était pas pertinente dans les « expériences agricoles » (faisant référence à Fisher), qu'il considérait comme étant « très largement des problèmes de pure estimation. » Nous reviendrons plus loin sur ce point essentiel.

Notons encore que, contrairement à Fisher, Jeffreys rejetait le principe de causalité (ou de déterminisme, ou d'uniformité de la nature), sous toute forme telle que « des antécédents exactement semblables conduisent à des conséquences exactement semblables ».

Comme Fisher, Jeffreys était convaincu de l'insuffisance de la logique déductive pour la méthode scientifique : « Je rejette la tentative de réduire l'induction à la déduction » (Jeffreys, 1967, p. B). Même si le test est utilisé pour décider si un nouveau paramètre est nécessaire, ce n'est pas pour Jeffreys une procédure de prise de décision automatique. Mais la décision est basée sur une mesure de l'évidence contre l'hypothèse nulle qui conduit à *un jugement gradué*.

4.4 En conclusion de la section

L'analyse de l'usage effectif fait des tests statistiques dans la recherche expérimentale conduit à penser avec Rozeboom que « son erreur la plus fondamentale réside dans le fait de considérer à tort l'objectif d'une étude scientifique comme étant une décision, plutôt qu'une évaluation cognitive de propositions » (Rozeboom, 1960, p. 428). Il doit être reconnu que cet usage n'est pas en accord avec les points de vue de Fisher et Jeffreys. De fait, les tests, tels qu'ils sont couramment utilisés, n'apportent pas d'information: « Dans de nombreuses expériences il paraît évident que les différents traitements doivent avoir produit une certaine différence, aussi infime soit elle. Donc l'hypothèse qu'il n'y a pas de différence est irréaliste : le problème réel est d'obtenir des estimations de la grandeur des différences » (Cochran et Cox, 1957, p. 5).

5. Le rôle des probabilités bayésiennes

5.1 Jeffreys : bayésien déclaré

En tant que bayésien, Jeffreys considérait la probabilité comme le *degré de confiance* que nous pouvons raisonnablement avoir dans une proposition. Parce que notre degré de confiance change quand de nouvelles observations ou un nouvel élément de preuve devient disponible, il concevait la probabilité comme étant toujours conditionnelle : « il n'est pas davantage valide de parler de la probabilité d'une proposition sans indiquer les données qu'il le serait de parler de la valeur de x+y pour un x donné, sans tenir compte de la valeur de y » (Jeffreys, 1967, p. 15).

5.2 Fisher : l'approche fiduciaire

Eviter l'usage des probabilités a priori. L'objectif clairement affiché de Fisher était d'éviter l'usage des probabilités *a priori* sur les hypothèses. Il a toujours reconnu cependant que l'argument bayésien devrait être utilisé « quand des connaissances *a priori* sous la forme d'énoncés de probabilité mathématique exacts sont disponibles » (Fisher, 1990b, p. 198). Ce qu'il contestait, c'est la pertinence de cette situation dans l'analyse des données expérimentales: « une question plus importante, cependant est de se demander si dans la recherche scientifique, et en particulier dans l'interprétation des expériences, il y a une raison convaincante pour introduire une expression correspondante représentant des probabilités *a priori* » (Fisher 1990c, p. 17).

Une définition du t de Student à l'opposé de la définition fréquentiste. Fisher considérait le seuil observé du test de signification (la *p-value*) comme caractérisant un « échantillon unique ». Il est révélateur de lire sa définition pour le test du « *t* de Student » usuel :

« Si *x* (par exemple la moyenne d'un échantillon) est une valeur distribuée normalement autour de zéro, et σ est sa vraie erreur type, alors la probabilité que x/σ excède toute valeur spécifiée peut être obtenue à partir de la table appropriée de la distribution Normale, mais si nous ne connaissons pas σ mais avons à sa place *s*, un estimateur de la valeur de σ la distribution requise sera celle de x/s et elle n'est pas Normale. La vraie valeur a été divisée par un facteur, s/σ ce qui introduit une erreur. […] la distribution de s/σ peut être calculée, et bien que σ soit inconnue, nous pouvons utiliser à sa place *la distribution fiduciaire de* σ *étant donné s* pour trouver la probabilité que *x* excède un multiple donné de *s* » (Fisher, 1990a, p. 118, italiques ajoutées).

Dans cette définition, la statistique *s* est clairement considérée comme une quantité fixe, alors que le paramètre σ est traité comme une variable aléatoire. Elle est à l'opposé de la définition fréquentiste traditionnelle, dans laquelle *x* et *s* sont des variables aléatoires alors que σ est une *quantité fixe*. Fisher utilisait l'argument fiduciaire (voir ci-après) pour dériver la distribution *a posteriori* de σ étant donné *s*. Le résultat est identique à celui de la conception fréquentiste – la distribution de *x*/*s* est une distribution de Student – mais la justification est complètement différente. C'est la distribution *a posteriori* (étant donné *s*) *prédictive* – et non la distribution d'échantillonnage – de la statistique de test *t* qui est considérée, conditionnellement à la valeur de la moyenne μ spécifiée par l'hypothèse nulle.

Puisqu'elle ne met en jeu que la distribution *a posteriori* du paramètre *parasite* σ, nous avons appelé la conception de Fisher un « test de signification semi-bayésien » (Lecoutre, 1985). Plus récemment, les bayésiens ont précisément introduit la notion de *p-value a posteriori prédictive*, vue comme « la moyenne *a posteriori* de la *p-value* classique [fréquentiste], moyennée sur la distribution *a posteriori* des paramètres (parasites) sous l'hypothèse nulle » (Meng, 1994, p. 1142). C'était précisément la conception de Fisher, même si cela n'a pas été reconnu par les bayésiens.

L'inférence fiduciaire. Plus tard, Fisher (1959) en est venu à écrire très explicitement qu'il utilisait la probabilité comme une mesure du degré d'incertitude: « L'objet d'un énoncé de probabilité si nous savons de quoi nous parlons, est singulier et unique; nous avons un certain degré d'incertitude sur sa valeur, et il se trouve que nous pouvons spécifier la nature et la mesure exactes de notre incertitude au moyen du concept de Probabilité Mathématique tel qu'il a été développé par les grands mathématiciens du 17 ème siècle Fermat, Pascal, Leibnitz, Bernoulli et leur successeurs immédiats » (Fisher, 1959, p. 22). Pour répondre à cette conception, il développa très tôt le concept de « *fiducial probability* » (cf Fisher, 1933). La traduction française *fiduciaire* est de Fisher lui-même (Fisher, 1948).

Nous retiendrons seulement ici la *motivation* de l'inférence fiduciaire : obtenir une distribution *a posteriori* sur le paramètre en l'absence de connaissances *a priori*, ceci en accord avec la réticence de Fisher à spécifier une distribution *a priori*. Ainsi, « L'argument fiduciaire utilise les observations seulement pour changer le statut logique du paramètre, d'un statut où on ne connaît rien de lui, et où on ne peut donner aucun énoncé de probabilité, au statut d'une variable aléatoire ayant une distribution bien définie » (Fisher, 1990c, p. 54). L'interprétation est explicitement en termes de probabilités bayésiennes : « Le concept de probabilité en jeu est tout à fait identique avec la probabilité classique des premiers auteurs, tels que Bayes » (Fisher, 1990c, p. 54).

5.3 Neyman : priorité au fréquentisme

Neyman et Pearson, partageaient avec Fisher la préoccupation d'éviter l'utilisation des probabilités *a priori*. Leur objectif était de trouver des énoncés « qui ne seraient pas modifiés par un changement des probabilités *a priori* » (Neyman et Pearson, 1933b, p. 492), reconnaissant ainsi explicitement une conception dualiste de la probabilité. Après avoir cessé sa collaboration avec Pearson, Neyman exprima explicitement son opposition à Fisher, et défendit avec force une conception fréquentiste de la probabilité : « Pour moi elle est seulement la réponse à la question 'combien fréquemment ceci ou cela arrive' » (Neyman, 1952, p. 187).

Plus tard cependant, il insista aussi sur le fait que cela n'était pas une opposition systématique à l'utilisation de l'inférence bayésienne : « Peut-être à cause du manque de clarté de certain de mes papiers, certains auteurs semblent avoir l'impression que, pour une

certaine raison, je condamne l'utilisation de la formule de Bayes et que je suis opposé à toute considération des probabilités *a priori*. C'est un malentendu. Ce à quoi je suis opposé est le dogmatisme qui apparaît à l'occasion dans l'application de la formule de Bayes quand les probabilités *a priori* ne sont pas impliquées par le problème traité et sont la tentative d'un auteur d'imposer au consommateur des méthodes statistiques les probabilités *a priori* particulières inventées par lui-même pour cet usage particulier » (Neyman, 1957, p. 19).

Il faut souligner le fait que Neyman était clairement conscient des difficultés de l'interprétation fréquentiste de l'intervalle de confiance. Il se donna ainsi beaucoup de mal pour essayer « d'anticiper certains malentendus » en expliquant soigneusement son interprétation (Neyman, 1977, pp. 116-119). Il insista notamment sur le fait qu'il n'y a pas de probabilité fréquentiste attribuée à un intervalle de confiance unique calculé pour un échantillon particulier.

En conclusion de la section

Beaucoup de ceux qui analysent des données expérimentales ont un intérêt potentiel manifeste pour les méthodes bayésiennes. Mais, trop souvent encore, les présentations que font les tenants de ces méthodes les amènent à penser qu'il s'agit d'un idéal inatteignable, sans prise directe avec les réalités des pratiques expérimentales. Les tenants de l'approche bayésienne pour l'analyse des données expérimentales devraient notamment éviter tout dogmatisme excessif. L'opposition entre fréquentistes et bayésiens paraît de plus en plus radicalisée. Souvent les tenants de chacun deux camps rejettent les propositions faites par les autres en les jugeant selon les critères de leur propre approche, ce qui rend malheureusement difficile – sinon impossible – le dialogue. Il y aurait tout intérêt à prendre en compte les conceptions de Fisher et à revenir sur le point de vue de Jeffreys.

6. Jeffreys et le rôle de l'inférence statistique dans la recherche expérimentale

Les conceptions de Jeffreys sur le rôle de l'inférence statistique dans la recherche expérimentale ont été clairement exprimées, de manière détaillée, dans la troisième édition de sa *Théorie des Probabilités* (chapitre VII). Nous nous contenterons ici de résumer les points principaux pour notre propos.

6.1 Estimation et tests de signification

« Mais ce qui est appelé tests de signification dans les expériences agricoles me paraît être très largement des problèmes d'*estimation pure*. Quand un ensemble de variétés d'une plante sont testées pour leur productivité ou quand plusieurs traitements sont testés, il ne m'apparaît pas que *la question de présence ou d'absence de différence entre en quoi que ce soit en ligne de considération* » (Jeffreys, 1967, p. 389, italiques ajoutées). Clairement, ceci s'applique aux essais cliniques comme à la plupart des expériences en psychologie et dans bien d'autres domaines.

La distinction entre estimation et tests de signification est au cœur de la méthodologie de Jeffreys. Ses implications pour l'analyse des données expérimentales sont claires : pour comparer deux traitements (par exemple), nous ne devons pas utiliser un test de signification – au sens de Jeffreys – « si la question n'est pas de savoir si la différence est zéro » (plus généralement si un paramètre a une valeur spécifique). S'il n'y a pas de valeur numérique particulière du paramètre qui nous intéresse, ou, exprimé différemment, « s'il n'y a pas de

doute initialement sur la pertinence du paramètre », c'est un problème de *pure estimation*. En conséquence, la question suivante se pose, naturellement pour Jeffreys en termes bayésiens : « […] quelle est la distribution de probabilité de ce paramètre étant donné les observations? » (Jeffreys, 1967, p. 388). Même dans le cas où on dispose d'une théorie prédisant une valeur précise, Jeffreys mettait en avant la nécessité de définir clairement une hypothèse alternative, spécifiant une autre valeur précise ayant un intérêt théorique justifié, faute de quoi rejeter l'hypothèse nulle (et donc utiliser un test) est sans intérêt.

6.2 La réponse de Jeffreys aux questions posées par l'analyse des données expérimentales

La réponse de Jeffreys au problème de pure estimation – et par suite aux questions posées par l'analyse des données expérimentales – est de considérer la distribution *a posteriori* complète de la quantité à estimer à partir des données et de calculer la probabilité que cette quantité soit plus ou moins grande que certaines valeurs spécifiées. Dans le cas particulier familier de l'inférence sur la différence δ de deux moyennes sous le modèle normal, la solution de Jeffreys coïncide avec la solution fiduciaire de Fisher et la distribution *a posteriori* est une distribution du *t* de Student (généralisée). Il s'ensuit que la probabilité que δ soit de signe contraire à celui de la différence observée est exactement la *p-value* unilatérale. Il y a également une probabilité *a posteriori* $1-\alpha$ que δ soit contenu dans l'intervalle de confiance fréquentiste usuel. On trouve donc ici une *justification formelle* à l'interprétation bayésienne naïve.

La conclusion de Jeffreys est que « plusieurs des intégrales *P* [celles qui donnent les *p-values*] trouvent leur place dans la théorie présentée, *dans les problèmes d'estimation pure* » (Jeffreys, 1967, p. 387, italiques ajoutées).

6.3 Fisher et Jeffreys réconciliés ?

La conception de la probabilité de Fisher et ses travaux sur la théorie fiduciaire sont une contrepartie fondamentale à son insistance sur les tests de signification. Jeffreys l'a explicitement reconnu : « Mais il me semble que les cas qui concernent essentiellement Fisher sont des problèmes d'estimation, et pour ces cas l'approche fiduciaire et l'approche de probabilité inverse sont complètement équivalentes » (Jeffreys, 1940, p. 51). Même s'il peut paraître exagéré de parler de réconciliation entre Fisher et Jeffreys, il a manifestement existé entre eux un *gentleman agrément* : « L'accord général entre le Professeur R.A. Fisher et moi-même a été indiqué en de nombreuses occasions. Les différences apparentes ont été largement exagérées… » (Jeffreys, 1967, p. 393).

Même en ce qui concerne les tests de signification de l'hypothèse nulle, Jeffreys a écrit : « en dépit de la différence ce principe entre mes tests et ceux basés sur les intégrales *P* [les *p-values*],… il apparaît qu'il n'y a pas beaucoup de différence dans les recommandations pratiques. Les utilisateurs de ces tests parlent du point 5% sensiblement de la même façon que je parlerais du point $K=10^{-1/2}$ et du point 1% comme je parlerais du point $K=10^{-1}$. » Il reconnaissait que, pour des grands nombres d'observations, « il peut y a voir des décisions opposées », en ajoutant cependant « mais elles seront très rares » (Jeffreys, 1967, p. 435).

6.4 La conception de Student

Dans son célèbre article de 1908, William S. Gosset, chimiste à la brasserie Guinness, présentait sous le pseudonyme de 'Student' ce qui a été appelé « le test du *t* de Student » (Student, 1908). Il est remarquable de constater qu'il avait une conception similaire à celle de

Jeffreys. Clairement, l'objectif de la procédure était d'obtenir un jugement sur le signe (le mot hypothèse n'apparaît pas dans ce papier), et, même si la justification était différente de celle de Jeffreys, ce jugement était exprimé en termes de probabilités bayésiennes. La question posée par Student était de calculer « quelles sont les chances que la moyenne [il s'agissait en fait de la moyenne d'une différence] de la population dont cette expérience est un échantillon soit positive », et la réponse était « nous trouvons .8873 ou le rapport des chances que la moyenne soit positive est de .887 à .113 [où .113 est la *p-value* unilatérale]. » Student, comme Jeffreys, recherchait avant tout une inférence conditionnelle aux données (voir Zabell, 2008, p. 2). Cela est bien montré par les mots « échantillon unique » dans le titre d'un autre de ses articles (Student, 1917).

On remarquera que l'usage de la probabilité de réplication p_{rep} proposé par Killeen (voir plus haut) a exactement le même objectif d'obtenir un jugement sur le signe de la différence, mais exprimé de manière prédictive.

6.5 Le test bayésien de Jaynes

Jaynes (1976), un autre physicien, a soutenu une perspective bayésienne proche de celle adoptée par Jeffreys. Pour comparer les moyennes *b* et *a* (avec ses notations) de deux distributions normales, il utilisait la distribution *a posteriori* de Jeffreys mentionnée plus haut. Il paraît avoir été une évidence pour lui que « si la question est de savoir si *b>a* », le test bayésien consiste à « calculer la probabilité que b>a, conditionnellement aux données disponibles » (Jaynes, 1976, p. 182). Par ailleurs, il a fermement argumenté contre l'utilisation d'un seuil de signification préassigné et il considérait que la différence entre son test bayésient et le test de signification de Fisher utilisant la *p-value* était dans ce cas « seulement un désaccord verbal sur le fait de savoir si nous devrions utiliser le mot 'probabilité' ou le mot 'signification' » (Jaynes, 1976, p. 185). Jaynes défendait aussi l'idée que « le meilleur intervalle de confiance pour tout paramètre de position ou d'échelle » était l'intervalle bayésien de probabilité *a posteriori*.

6.6 En conclusion de la section

De nos jours, l'approche de Jeffreys a été intégrée dans un cadre théorique décisionnel bayésien, qui perpétue la dichotomie « rejet/acceptation » des tests d'hypothèses, sans considération pour le rôle prééminent que Jeffreys assignait à l'estimation dans l'analyse des données expérimentales. Ainsi, dans la revue très détaillée du livre de Jeffreys faite par Robert, Chopin et Rousseau (2009), les auteurs ont seulement retenu de la section 7.2 du chapitre VII considérée ici (qu'ils présentent comme la plus connue du chapitre) la critique de Jeffreys concernant *l'interprétation fréquentiste* de la *p-value*. Ils ont mis en avant son insistance sur la nécessité d'hypothèses alternatives, en relation avec leur propre conviction que tester des hypothèses est *le problème central*, mais ils ont omis de mentionner la conception de Jeffreys que ce problème n'est pertinent que quand la théorie prédit une valeur précise. On ne trouve pas davantage mention du point de vue de Jeffreys concernant l'analyse des données expérimentales dans les commentaires des éminents statisticiens qui ont discuté cette revue.

Suivant Jeffreys (et bien d'autres), l'analyse des données expérimentales doit être regardée comme un problème de « pure estimation », et les tests d'hypothèses précises devraient avoir un rôle très limité. Si l'on est d'accord avec cette perspective, il n'y a pas de sens à rechercher une interprétation de la *p-value* fréquentiste comme probabilité de l'hypothèse nulle.

Au contraire, face à l'amalgame des procédures fréquentistes, on peut défendre l'idée que la conception de l'approche bayésienne de Jeffreys fournit une solution cohérente et unifiée. La distribution *a posteriori*, apporte à elle seule les réponses aux principales questions qui se posent.

7. Conclusion générale

Pourquoi les méthodes bayésiennes sont-elles si peu utilisées dans les publications expérimentales? Cette question renvoie avant tout à une autre question : pourquoi les tests de signification de l'hypothèse nulle d'absence d'effet ont-ils résistés à toutes les critiques et sont-ils toujours l'approche dominante ? Cette dernière n'ayant jamais reçu de réponse satisfaisante, il serait prétentieux de prétendre apporter une réponse définitive à la première question. Nous avons seulement considéré un certain nombre d'éléments qui peuvent contribuer à la compréhension d'un état de fait. Pour reprendre la métaphore de Mark Twain à propos du tabac, « une habitude est une habitude, et on ne s'en débarrasse pas en la jetant par la fenêtre, mais en lui faisant descendre les marches une à une. » Pour les tenants de l'approche bayésienne, la situation est encore plus difficile, car il leur faut aussi introduire une nouvelle habitude, qu'on n'imposera certainement pas par décret : il faut donc aussi « remonter les marches une à une. »

Est-il possible de *prescrire* à ceux qui publient les résultats de données expérimentales des méthodes bayésiennes « de routine » pour les situations courantes? Ces méthodes doivent répondre à leur besoin d'objectivité, mais aussi pouvoir être faciles à comprendre, et donc relativement simples, pour pouvoir être acceptées par la communauté scientifique. La question reste ouverte, mais pour notre part nous partageons entièrement la position d'Efron : « Une théorie bayésienne objective largement acceptée, ce que l'inférence fiduciaire entendait être, serait d'une importance théorique et pratique immense. Pour connaître le succès, une théorie bayésienne objective devrait avoir de bonnes propriétés fréquentistes dans les situations familières, par exemple des probabilités de couverture raisonnables pour ce qui remplacerait les intervalles de confiance » (Efron, 1998, p. 106). Plus encore, nous pensons qu'une telle théorie n'est pas un point de vue spéculatif, mais au contraire un projet, non seulement désirable, mais aussi parfaitement faisable. Les conceptions et les travaux de Fisher et de Jeffreys en fournissent les éléments fondateurs. Nous avons fait par ailleurs des propositions concrètes : voir notamment Lecoutre, 1996, 2008 ; Lecoutre et Poitevineau, 2014.

8. Références

American Psychological Association, *Publication Manual of the American Psychological Association* (6ème édition). Washington, DC., American Psychological Association, 2010.

J.O. Berger, « Could Fisher, Jeffreys and Neyman have agreed on testing? », *Statistical Science*, 18, 2003, p. 1-32.

W.G. Cochran & G.M. Cox, *Experimental Designs* (2 ème édition), New York, John Wiley & Sons, 1957.

B. Efron, « R.A. Fisher in the 21st century (avec discussion) », *Statistical Science*, 13, 1998, p. 95-122.

R.A. Fisher, « The concepts of inverse probability and fiducial probability referring to unknown parameters », *Proceedings of the Royal Society of London. Series A*, 139, p. 343-348.

R.A. Fisher, « Conclusions fiduciaires ». *Annales de l'institut Henri Poincaré*, 10, 1948, p. 191-213

R.A. Fisher, « Mathematical probability in the natural sciences », *Technometrics*, 1, 1959, p. 21-29.

R.A. Fisher, *Statistical Methods for Research Workers* (réimpression de la 14ème édition, 1970), *in* J.H. Bennett (ed.) *Statistical Methods, Experimental Design, and Scientific Inference*, Oxford, Oxford University Press, 1990a.

R.A. Fisher, *The Design of Experiments* (réimpression de la 8ème édition, 1966), *in* J.H. Bennett (ed.) *Statistical Methods, Experimental Design, and Scientific Inference*, Oxford, Oxford University Press, 1990b.

R.A. Fisher, *Statistical Methods and Scientific Inference* (réimpression de la 3ème édition, 1973), *in* J.H. Bennett (ed.) *Statistical Methods, Experimental Design, and Scientific Inference*, Oxford, Oxford University Press, 1990c.

Food and Drug Administration: Guidance for Industry and FDA Staff, *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials*, U.S. Department of Health and Human Services Food and Drug Administration Center for Devices and Radiological Health, 2010.

G. Gigerenzer, « The superego, the ego, and the id in statistical reasoning », *in* G. Keren & C. Lewis (eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, Hillsdale, NJ., Erlbaum, 1993, p. 311-339.

ICH E9 Expert Working Group, « Statistical principles for clinical trials: ICH harmonized tripartite guideline », *Statistics in Medicine*. 18, 1999, p. 1905-1942.

E.T. Jaynes, « Confidence intervals vs Bayesian intervals (with discussion) », *in* W. L. Harper & C.A. Hooker (eds.) *Statistical Inference and Statistical Theories of Science Volume 2*, Dordrecht, D. Reidel, 1976, p. 175-257.

H. Jeffreys, « Note on the Behrens-Fisher formula », *Annals of Eugenics*, 10, 1940, p. 48-51.

H. Jeffreys, *Theory of Probability* (3ème édition corrigée, 1ère édition 1939), Oxford, Clarendon, 1967 (réimprimé en 1998).

P.R. Killeen, « An alternative to null-hypothesis significance tests », *Psychological Science*, 16, 2005, p. 345-353.

B. Lecoutre, « Reconsideration of the F test of the analysis of variance: The semi-Bayesian significance tests », *Communications in Statistics A-Theory and Methods*, 14, 1985, p. 2437-2446.

B. Lecoutre, *Traitement statistique des données expérimentales: Des pratiques traditionnelles aux pratiques bayésiennes*, Paris, DECISIA Editions, 1996.

B. Lecoutre, « Expérimentation, inférence statistique et analyse causale. *Intellectica*, 38, 2004, p. 193-245.

B. Lecoutre, « Bayesian methods for experimental data analysis », *in* C.R. Rao, J. Miller & D.C. Rao (eds.), *Handbook of statistics: Epidemiology and Medical Statistics* (Vol 27), Amsterdam, Elsevier, 2008, p. 775-812.

B. Lecoutre & P. Killeen, « Replication is not coincidence: Reply to Iverson, Lee, and Wagenmakers, (2009) », *Psychonomic Bulletin & Review*, 17, 2010, p. 263-269.

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

B. Lecoutre, M.-P. Lecoutre & J. Poitevineau, « Killeen's probability of replication and predictive probabilities: How to compute, use and interpret them », *Psychological Methods*, 15, 2010, p. 158-171.

B. Lecoutre & J. Poitevineau, *The Significance Test Controversy Revisited: The Fiducial Bayesian Alternative*. SpringerBriefs in Statistics, 2014.

X.-L. Meng, « Posterior predictive p-values », *Annals of Statistics* 22, 1994, p. 1142-1160.
J. Neyman, « L'estimation statistique traitée comme un problème classique de probabilité », *Actualités Scientifiques et Industrielles*, 739, 1938, p. 25-57.
J. Neyman, « Foundations of the general theory of estimation », *Actualités Scientifiques et Industrielles*, 1146, 1951, p. 83-95.
J. Neyman, *Lectures and Conferences on Mathematical Statistics and Probability* (2ème édition), Washington, Graduate School U.S. Department of Agriculture, 1952.
J. Neyman, « "Inductive behavior" as a basic concept of philosophy of science », *Revue de l'Institut International de Statistique*, 25, 1957, p. 7-22.
J. Neyman, « Frequentist probability and frequentist statistics », *Synthese* 36, 1977, p. 97-131.
J. Neyman & E.S. Pearson, « On the problem of the most efficient tests of statistical hypotheses », *Philosophical Transactions of the Royal Society of London*, *Series A*, 231, 1933, p. 289-337.
J. Neyman & E.S. Pearson, « The testing of statistical hypotheses in relation to probabilities a priori », *Proceedings of the Cambridge Philosophical Society*, 29, 1933, p. 492-510.
M.R. Novick & P.H. Jackson, *Statistical Methods for Educational and Psychological Research*. NewYork, McGraw-Hill, 1974.
L.D. Phillips, *Bayesian Statistics for Social Scientists*. London: Nelson, 1973.
Cl. Robert, *L'Empereur et la girafe,* Paris, Diderot Editeur, 1995.
C.P. Robert, N. Chopin & J. Rousseau, « Harold Jeffreys's theory of probability revisited (avec commentaires) », *Statistical Science*, 24, 2009, p. 141-194.
R.D. Rosenkrantz, « The significance test controversy », *Synthese*, 26, 1973, p. 304-321.
H. Rouanet, « 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* (2ème edition), Berne, Peter Lang, 2000, p. 29-64.
H. Rouanet, D. Lépine & J. Pelnard-Considère, « Bayes-fiducial procedures as practical substitutes for misplaced significance testing: An application to educational data », *in* D.N.M. De Gruijter & L.J.T. Van Der Kamp (eds.), *Advances in Psychological and Educational Measurement*, New York, Wiley, 1976, p. 33-50.
W.W. Rozeboom, « The fallacy of the null hypothesis significance test », *Psychological Bulletin*, 57, 1960, p. 416-428.
Student, « The probable error of a mean », *Biometrika*, 6, 1908, p. 1-25.
Student, « Tables for estimating the probability that the mean of a unique sample of observations lies between - ∞ and any given distance of the mean of the population from which the sample is drawn », *Biometrika*, 11, 1917, p. 414-417.
L. Wilkinson and Task Force on Statistical Inference, APA Board of Scientific Affairs, « Statistical Methods in Psychology Journals: Guidelines and Explanations », *American. Psychologist* 54, 1999, p. 594-604.
R.L. Winkler, « Statistical analysis: theory versus practice », *in* C.-A.S. Staël Von Holstein (ed.), *The Concept of Probability in Psychological Experiments*, Dordrecht, D. Reidel, 1974, pp. 127-140.
S.L. Zabell, « On Student's 1908 paper "The probable error of a mean." », *Journal of the American Statistical Association*, 103, 2008, p. 1-7.