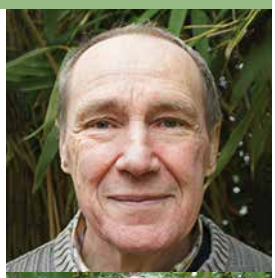
Tester les nouveaux médicaments :
les statisticiens et la règlementation[1]



Bruno Lecoutre[2] et Gérard Derzko[3]

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, qui réunit les autorités réglementaires et l'industrie pharmaceutique de l'Europe, du Japon et des Etats-Unis, a rédigé en 1999 des lignes directrices (« *guidelines* ») qui fournissent les principes statistiques à suivre dans un essai clinique. Ces lignes directrices reposent sur plusieurs points de vue théoriques pas toujours concordants, et de ce fait peuvent entraîner des difficultés conceptuelles pour les utilisateurs, et des difficultés théoriques pour les statisticiens.

En guise d'introduction

Cet article évoque l'utilisation de méthodes d'inférence statistique diverses. Nous en rappelons les principes fondamentaux pour ceux qui ne maîtrisent pas ce sujet.

Dans le *test de signification* de Fisher, une seule hypothèse, appelée « l'hypothèse nulle » – souvent, mais pas nécessairement, l'hypothèse que le paramètre a une valeur nulle – est considérée. On utilise une statistique de test appropriée dont la distribution d'échantillonnage, quand l'hypothèse est vraie, est exactement connue. La probabilité que la statistique de test « excède par chance » la valeur observée pour l'échantillon, *si l'hypothèse nulle est vraie* donne le seuil (ou niveau) de signification, aujourd'hui appelée la « *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* »).

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

1. Nous remercions les rédacteurs en chef de la revue pour leur aide à la rédaction d'un texte accessible à un lecteur non spécialiste. Nous assumons l'entière responsabilité des difficultés de lecture qui pourraient subsister.
2. 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
3. Président du groupe Biopharmacie-Santé de la Société française de statistique, 64, rue du jardin des violettes, 34070 Montpellier, gerard.derzko@numericable.fr

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. Le rôle du test est de minimiser « sur le long terme » la proportion de décisions erronées. Des erreurs de deux types sont prises en compte : 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 fixé et est donc égale à 1-β. Plus généralement la puissance est une fonction du paramètre. Le 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 »).

C'est Neyman qui, dans le cadre de l'estimation statistique, a introduit la conception *fréquentiste* de l'intervalle de confiance (opposée à la conception bayésienne), en liaison avec la notion de région d'acceptation. De fait, la plupart des intervalles de confiance utilisés en pratique peuvent être obtenus par « *inversion* » d'un test d'hypothèses (c'est le cas de ceux considérés ici) : 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 » pour lesquels la probabilité d'échantillonnage fluctue autour de 95%.

En conclusion de cette introduction, nous soulignerons le fait que 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.

Les essais de supériorité

La plus grande partie des lignes directrices de 1999 (réf. [3]) concerne 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.

Dans un pareil contexte, le processus d'approbation des médicaments a, par tradition, un caractère décisionnel : acceptation/rejet. En conséquence l'utilisation traditionnelle des tests d'hypothèses (Neyman-Pearson) est toujours très prégnante dans l'industrie pharmaceutique. On choisit comme hypothèse à tester, hypothèse « privilégiée », l'absence de différence entre l'effet du nouveau médicament et l'effet du médicament de référence. On choisit comme hypothèse alternative l'existence d'une certaine différence. Cette différence traduit une hypothèse « de travail » : elle doit être justifiée, 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 l'erreur de type I – rejeter à tort l'hypothèse privilégiée (absence de différence) si elle est vraie – est généralement fixée de manière conventionnelle à α=5% (éventuellement moins). Avant de commencer l'expérimentation, on détermine le nombre d'essais nécessaire pour que la probabilité de rejeter l'hypothèse privilégiée lorsque l'hypothèse alternative est vraie soit suffisamment grande : au moins 0,8 ou

0,9 (soit respectivement β=20% et β=10%). C'est ce qu'on appelle la « condition de puissance » du test.

Mais 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 la supériorité alors que la différence vraie avec le médicament de référence est en fait triviale.

Le texte des lignes directrices de 1999 prend en compte ces objections et reconnaît explicitement l'insuffisance d'une décision en tout ou rien. Plusieurs suggestions complémentaires sont donc faites. Il est ainsi recommandé de rapporter la valeur exacte du seuil observé des tests, la « *p-value* ». Mais ce seuil, outre le fait qu'il n'est pas compatible avec l'approche de Neyman-Pearson (voir introduction), dépend des nombres d'observations et n'est donc qu'un indicateur indirect de la différence vraie. La procédure, plus satisfaisante, recommandée est de rapporter également une estimation de la différence (la « taille de l'effet » du changement de médicament) accompagnée d'un intervalle de confiance. L'usage est de considérer un intervalle bilatéral, généralement de confiance 95%, soit le même α que le test bilatéral.

Les essais d'équivalence

Il y a une trentaine d'années, le développement de nouveaux médicaments a posé aux statisticiens de l'industrie pharmaceutique une question aussi étrangement simple et pragmatique dans sa formulation qu'instructive par ses solutions : quelle procédure statistique doit-on utiliser pour être assuré, avec un niveau de garantie imposé, que deux médicaments ont les mêmes effets sur un patient ? Il faut réaliser qu'une substance active nouvelle peut présenter de nombreuses variantes au cours de son développement comme candidat médicament ou comme médicament reconnu : on peut en effet en modifier à plusieurs époques la formulation, la forme d'administration, les posologies, etc. Le premier souci de la communauté médicale et des agences de santé est toujours d'apporter la garantie de *l'équivalence thérapeutique* des formes successives. Toute nouvelle variante doit donc être comparée à une forme de référence au moyen d'une ou de plusieurs *études pharmacocinétiques*, qui mesurent par un certain nombre de paramètres les évolutions temporelles des concentrations de substances dans le sang. De plus, lorsque le brevet protégeant un médicament vient à expiration, tout fabriquant peut mettre sur le marché une préparation similaire, de sa fabrication, dite « générique », à condition d'en avoir démontré la bio-équivalence avec le médicament en fin de brevet. En plus des études pharmacocinétiques, les promoteurs de nouveaux médicaments ont réalisé par la suite des essais *cliniques* spécifiques d'équivalence, ou plus fréquemment encore de *non infériorité*. Quand un nouveau produit est supérieur en termes de tolérance ou de facilité d'administration au produit de référence, ces études ont pour objectif de démontrer qu'en termes d'efficacité clinique il lui est équivalent, ou de préférence qu'il n'est pas moins efficace (non infériorité).

Au début des années 1980, la pratique courante était d'utiliser pour établir une « démonstration d'équivalence » pour les paramètres pharmacocinétiques le même test d'absence de différence que pour les études de supériorité. Cette pratique se basait sur le raisonnement indirect prenant en compte a *posteriori* ce qu'aurait été la puissance du test statistique si les médicaments n'étaient pas équivalents. Les statisticiens ont été nombreux à faire observer que ce raisonnement n'est pas satisfaisant et ne garantit pas nécessairement l'équivalence des traitements dans le cas où le test ne permet pas de rejeter l'hypothèse d'absence de différence. Il faut pour démontrer l'équivalence une procédure spécifique appropriée.

Aussi, les « lignes directrices » de 1999 traitent différemment des essais de supériorité les essais d'équivalence et de *non-infériorité*. L'approche des tests d'hypothèses de Neyman-Pearson n'a pas été retenue pour l'analyse de ces essais. Dans les deux situations, il est recommandé de baser « normalement » l'analyse statistique sur un intervalle de confiance. Pour les essais d'équivalence il est nécessaire de spécifier à l'avance une « marge de petitesse », c'est-à-dire une quantité *cliniquement négligeable* qui doit être justifiée scientifiquement (et non une référence conventionnelle). Cette marge définit une « région d'équivalence » qui doit tenir compte de la grandeur relative des différences, en comparaison notamment avec les résultats des essais de supériorité. La procédure décisionnelle recommandée est d'utiliser un intervalle de confiance bilatéral. De nombreux auteurs ont montré que, pour le niveau de confiance traditionnel 95%, l'intervalle approprié est l'intervalle de confiance 90%, et non 95%. Ainsi la procédure consiste à calculer l'intervalle bilatéral usuel de confiance 90% et à conclure à l'équivalence s'il est entièrement contenu dans la région d'équivalence. Le fait, admis par les autorités réglementaires, de fonder la décision sur l'usage des intervalles 90% peut se justifier de la façon suivante : si, dans le cas où les deux limites de cet intervalle sont de même signe (ce qui, au seuil unilatéral 5%, rejette l'hypothèse d'une différence nulle), on étend l'intervalle jusqu'à la valeur 0 – *ce qui ne change pas la décision* – la procédure résultante est un intervalle de confiance 95%[4].

Une règlementation hybride qui devrait évoluer

Les lignes directrices de 1999 apparaissent ainsi comme un amalgame de différentes procédures plus ou moins compatibles (voir Lecoutre et Poitevineau, 2014) :

- 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 dite « de Fisher » avec l'usage du seuil observé, exclu par l'approche décisionnelle ;
- l'utilisation d'intervalles de confiance, soit comme procédure complémentaire dans les essais de supériorité, soit comme procédure décisionnelle principale dans les essais d'équivalence.

Ce caractère hybride des règles imposées pour l'analyse statistique fait apparaître deux catégories de difficultés, conceptuelles pour les utilisateurs et théoriques pour les statisticiens.

Difficultés pour les utilisateurs

Cet amalgame de procédures peut conduire à des situations conflictuelles :

- Dans les essais de supériorité, le test peut aboutir à la décision de « supériorité », alors que l'intervalle de confiance ne permet pas de conclure que la différence est pertinente au sens clinique. Cela renvoie à la distinction entre « significativité statistique » et « significativité clinique ».
- Dans les essais d'équivalence, la conclusion peut être « l'équivalence » des deux traitements à comparer, alors que l'un d'eux est supérieur à l'autre au sens du test utilisé dans les essais de supériorité.

Pour éviter ces incohérences apparentes, il faudrait remettre en cause les règles édictées pour l'établissement de la supériorité. 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

4. Cet intervalle peut être obtenu par inversion de la procédure de test consistant à utiliser simultanément deux tests unilatéraux pour tester l'hypothèse (composite) que la différence est en dehors des de la région d'équivalence, d'un côté ou de l'autre, contre l'hypothèse alternative (composite) que la différence est à l'intérieur de la région. Celle-ci est connue sous le nom de procédure des « deux tests unilatéraux » (« *Two One-Sided Tests* »).

l'hypothèse privilégiée est rejetée, permet seulement de penser que la différence est non nulle. Si l'hypothèse alternative de travail correspond à l'effet minimal pertinent au sens clinique, c'est elle – et non l'hypothèse d'absence d'effet – qui devrait être testée, soit directement par un test, soit par l'intermédiaire d'un intervalle de confiance. Cela serait cohérent avec l'approche utilisée pour les essais d'équivalence et de non infériorité.

De l'importance de bien choisir les marges

Il faudrait donc nécessairement spécifier à l'avance une marge scientifiquement acceptable qui déterminerait une « région de supériorité clinique ». Comment choisir cette région ? Jusqu'à présent cette question a été plus ou moins éludée, puisqu'on se contente de tester l'hypothèse d'absence de différence, mais cette situation n'est pas vraiment satisfaisante. Le lecteur comprendra aisément, sans autre considération technique, que le choix sera d'autant plus exigeant et coûteux pour l'expérimentateur, autrement dit qu'il faudra d'autant plus d'unités expérimentales dans l'essai, que la marge requise pour pouvoir parler de supériorité sera grande. Mais une marge trop petite, et donc entraînant un coût moindre, ne garantit pas la supériorité clinique. En pratique, une conséquence importante serait la nécessité d'utiliser des effectifs plus élevés dans les études de supériorité, ce qui entraînerait l'augmentation du coût de ces études qui sont de loin les plus nombreuses.

Pour les essais d'équivalence, et également de ceux de non infériorité, la situation est inverse : la procédure est d'autant plus coûteuse que la marge d'équivalence sera choisie petite, alors qu'une région d'équivalence trop grande ne garantit pas l'équivalence aux yeux des autorités réglementaires et de l'usager. L'obligation faite par les lignes directrices de 1999 de spécifier la région d'équivalence a donc soulevé une difficulté conceptuelle importante.
Pour les essais d'équivalence ou de non-infériorité clinique, le choix de la marge est discuté et convenu avec les agences publiques responsables des autorisations de mise sur le marché des médicaments. Des aménagements peuvent être obtenus dans chaque cas particulier, lorsque le contexte l'exige.

Dans la situation spécifique des études pharmacocinétiques, la difficulté a cependant été rapidement résolue de façon conventionnelle : une région d'équivalence de ±20% autour de l'effet de la référence est le standard par défaut.

En résumé

La figure 1 résume les procédures pour démontrer, d'une part la supériorité clinique d'un nouveau médicament, et d'autre part l'équivalence de deux médicaments, avec un intervalle de confiance pour la différence. Selon la réglementation, on utilise respectivement des intervalles de confiance 95% et 90%. Dans le cas de la supériorité, on fixe une marge minimale qui détermine la région de supériorité clinique. La supériorité est démontrée si la limite inférieure de l'intervalle de confiance 95% est supérieure à cette marge. Notons que dans le deuxième exemple de la figure 1, où on ne peut pas démontrer la supériorité clinique, le test usuel de supériorité permet de rejeter l'hypothèse d'une différence nulle (« significativité statistique »).

Dans le cas de l'équivalence, on fixe une région d'équivalence (ou « petitesse »), et l'équivalence est démontrée si l'intervalle de confiance 90% est entièrement contenu dans cette région.

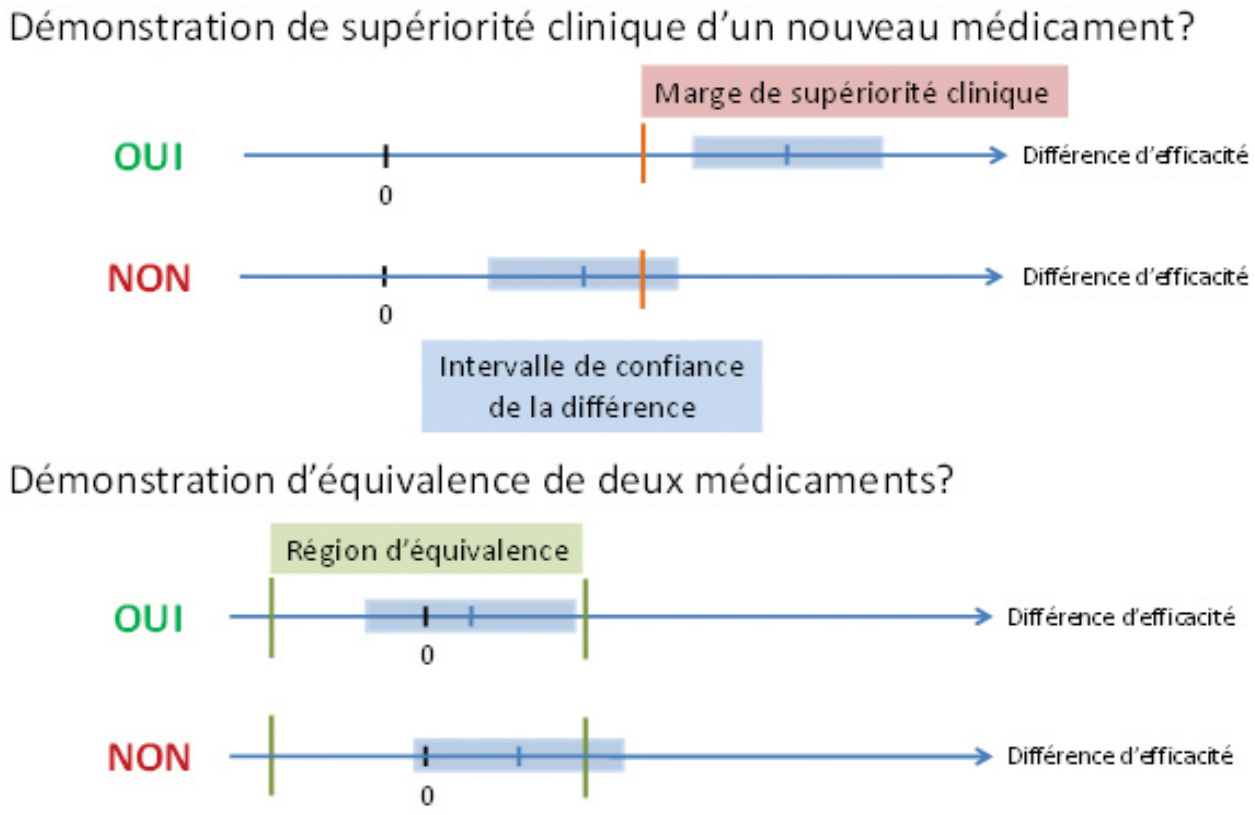


Figure 1 - Procédures pour démontrer la supériorité clinique d'un nouveau médicament et l'équivalence de deux médicaments avec un intervalle de confiance pour la différence.

Théorie contre pratique

Dans les essais d'équivalence, les directives réglementaires prennent bien en compte le fait que, quand l'hypothèse d'absence de différence ne peut pas être rejetée, le test classique de cette hypothèse ne permet pas à lui seul de conclure à une réelle équivalence. D'où la procédure spécifique recommandée. Mais celle-ci ne satisfait pas le souci permanent des statisticiens d'utiliser des outils d'inférence *optimaux*. En effet elle est *conservatrice* : la probabilité de conclure à tort que les médicaments ne sont pas équivalents alors qu'ils le sont est supérieure à 5%. Par ailleurs, la principale critique des statisticiens concernant cette procédure est son manque relatif de puissance.

Cela a conduit à une multitude de travaux pour définir un test uniformément plus puissant pour démontrer l'équivalence. Les difficultés à résoudre viennent du fait que l'hypothèse à rejeter est composite, constituée d'une infinité de points, et non plus d'un seul point comme dans le cas de l'hypothèse conventionnelle d'absence de différence dans les essais de supériorité.

Mais ces travaux ont reçu en retour de nombreuses critiques : les tests optimaux définis par les statisticiens théoriciens ont plusieurs propriétés indésirables (voir par exemple Lecoutre et Derzko, 2001, qui introduisent une discussion critique d'un point de vue bayésien). En particulier, la région critique peut inclure des valeurs de la différence observées (pour lesquelles on conclut donc à l'équivalence) qui sont en dehors de la région d'équivalence. De plus, pour une différence observée donnée, la région critique varie de façon non monotone en fonction de l'effectif. Les défenseurs de ces tests optimaux minimisent les conséquences de ces propriétés indésirables. Leur argument est qu'en fixant une taille minimale de l'échantillon le risque d'inférences

inacceptables est en pratique très limité. Cependant on ne peut malheureusement éviter qu'une expérimentation mal planifiée puisse aboutir à une conclusion fausse, quoiqu'en apparence bien établie, et que la procédure puisse être toujours suspecte. De fait, les tests optimaux pour l'équivalence ont toujours été considérés comme inacceptables par les statisticiens praticiens, et ceci explique pourquoi ils n'ont pas été retenus dans les « lignes directrices » de 1999.

Conclusion

Plus clairement que les études de supériorité, les études d'équivalence et de non-infériorité mettent en évidence l'insuffisance de procédures statistiques optimales exclusivement décisionnelles dans l'acceptation de la mise sur le marché d'un nouveau médicament : la définition, préalable à l'étude, d'une région d'équivalence, qui n'est pas du ressort du statisticien, fait apparaître l'intérêt de l'estimation de la taille et de la variabilité de l'effet (intervalle de confiance) dans l'inférence statistique.

Par ailleurs, la question de la démonstration de l'équivalence de deux médicaments illustre les difficultés de communication, voire les incompréhensions, qui peuvent exister entre les statisticiens « praticiens » et les statisticiens « académiques ». Les premiers considèrent que les méthodes optimales proposées par les seconds sont contraires à l'intuition et que c'est le bon sens qui justifie le choix des procédures. Ceci n'a pas manqué d'alimenter des débats théoriques. Par exemple, Berger et Hsu (1996), ont affirmé que les arguments des praticiens découlaient de contre-intuitions et ils ont continué à défendre la supériorité des tests optimaux : « Nous croyons que les notions de taille, puissance, et de sans-biais sont plus fondamentales que « l'intuition » » (traduit de (Berger et Hsu, 1996, page 292)). Au contraire, Perlman et Wu (1999) ont sérieusement mis en question ces tests optimaux et ont défendu la position qu'ils sont « scientifiquement inappropriés ». Leur réponse à Berger et Hsu est que « A notre avis, une telle déclaration entraîne de sérieux risques sur la crédibilité de la science statistique dans la communauté scientifique. En effet, si nous enseignons à nos étudiants d'ignorer l'intuition dans la recherche scientifique, alors il faut de manière urgente une réévaluation fondamentale de la mission des statistiques mathématiques » (traduit de Perlman et Wu, 1999, page 366).

Ces débats illustrent bien la nécessité du dialogue entre les statisticiens et les organismes, tels l'industrie pharmaceutique, qui utilisent la statistique à l'appui d'investigations scientifiques. On pourra ici méditer sur le fait que ce sont les autorités réglementaires qui ont tranché et légitimé les procédures retenues par les praticiens, en dépit des arguments formels avancés par les théoriciens.

Références

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

ICH E9 Expert Working Group (1999). Statistical principles for clinical trials: ICH harmonised tripartite guideline. *Statistics in Medicine*, **18**, 1905–1942. ICH : International Conference on Harmonization of Technical Requirements for Registration of Pharmaceutical for Human Use.

Lecoutre, B. et G. Derzko (2001) Asserting the smallness of effects in ANOVA. *Methods of Psychological Research*, **6**, 1–32.

Lecoutre, B. et J. Poitevineau (2014). *The Significance Test Controversy Revisited: The Fiducial Bayesian Alternative*. (SpringerBriefs in Statistics). A paraître.

Perlman, M. D. et L. Wu (1999). The emperor's new tests. *Statistical Science*, **14**, 355–369.