FAUT-IL BRÛLER LES TESTS DE SIGNIFICATION STATISTIQUE ?

Ababacar MBENGUE
Laboratoire REPONSE, Université de Reims & Reims Management School
57bis rue Pierre Taittinger, 51096 Reims - France
Tel : +33 326 91 87 31 – Courriel : ababacar.mbengue@univ-reims.fr

Résumé :
La démarche d'inférence occupe une place centrale dans la recherche en management. Très souvent, le chercheur est amené à tirer des conclusions ou à procéder à des généralisations à partir de ses observations ou de ses résultats. Dans certains cas, la statistique peut lui permettre de le faire de manière rigoureuse. En effet, cette discipline accorde une grande place à la démarche d'inférence par laquelle le statisticien généralise une information collectée sur un échantillon à l'ensemble de la population dont est issu cet échantillon : c'est la *statistique inférentielle* dont le but est de tester des hypothèses formulées sur les caractéristiques d'une population grâce à des informations recueillies sur un échantillon issu de cette population. Les tests statistiques sont de ce fait au coeur de la statistique inférentielle.

Dès leur introduction, les tests de signification statistique ont fait l'objet de multiples controverses et critiques portant à la fois sur leur nature et sur leur rôle. Et si certains auteurs (Abelson, 1997 ; Hagen, 1997 ; Mulaik, Raju et Harshman, 1997 ; Wainer, 1999) ont dernièrement fourni une forte défense de ces tests, de nombreux autres (Cohen, 1994 ; Schmidt, 1996, Hunter, 1997 ; Krueger, 2001 ; Lecoutre, Poitevineau et Lecoutre, 2003 ; Armstrong, 2007a, 2007b ; Levine *et al.*, 2008) appellent à leur abolition pure et simple. Alors, faut-il oui ou non brûler les tests de signification statistique ? Afin de répondre à cette question, nous allons procéder en trois temps. Tout d'abord, nous allons rappeler la logique générale de ces tests statistiques, définir les notions fondamentales qui leur sont associées et préciser les étapes usuelles de l'usage desdits tests. Ce premier effort de clarification du contexte nous permettra ensuite d'aborder frontalement la question du sort à réserver aux tests statistiques en analysant les principales critiques qui leur sont adressées, les erreurs fréquemment commises dans leur usage ainsi que les raisons de leur popularité persistante en dépit des critiques. L'article propose finalement une discussion détaillée de plusieurs voies d'amélioration.
Cet article a voulu fortement attirer l'attention des chercheurs en management stratégique sur les dangers liés à l'usage irréfléchi des tests de signification statistique. Le principal danger pour le chercheur consisterait à s'abriter derrière l'image scientifique des tests statistiques, à céder à leur aura et au confort apparent lié à leur utilisation pour abdiquer sa responsabilité. Or, c'est le chercheur qui doit choisir s'il teste ou pas, ce qu'il teste et par quel moyen. Mais, plus encore, le chercheur doit garder à l'esprit que les tests de signification statistique ne sont qu'un instrument à l'intérieur d'un dispositif et d'une démarche de recherche : cette recherche commence avant l'éventuel test, se poursuit pendant le test et continue après le test. Quant au test lui-même, il n'est qu'un outil et, en tant que tel, il ne vaut que si on sait s'en servir et à bon escient. De ce point de vue, les questions récurrentes sur l'utilité des tests de signification statistique sont un bon stimulant et un garde-fou précieux pour l'exercice d'une saine activité de recherche.

Mots clés : méthodologie, inférence statistique, tests statistiques, signification statistique

INTRODUCTION

Dès leur introduction, les tests de signification statistique ont fait l'objet de multiples controverses et critiques portant à la fois sur leur nature et sur leur rôle. Et si certains auteurs (Abelson, 1997 ; Hagen, 1997 ; Mulaik, Raju et Harshman, 1997 ; Wainer, 1999) ont dernièrement fourni une forte défense de ces tests, de nombreux autres (Rozeboom, 1960 ; Cohen, 1994 ; Schmidt, 1996, Hunter, 1997 ; Gill, 1999 ; Krueger, 2001 ; Lecoutre, Poitevineau et Lecoutre, 2003 ; Armstrong, 2007a, 2007b ; Levine *et al.*, 2008) appellent à leur abolition pure et simple. Alors, faut-il oui ou non brûler les tests de signification statistique ? Afin de répondre à cette question, nous allons procéder en trois temps. Tout d'abord, nous allons rappeler la logique générale de ces tests statistiques, définir les notions fondamentales qui leur sont associées et préciser les étapes usuelles de leur usage. Ce premier effort de clarification du contexte nous permettra ensuite d'aborder frontalement la question du sort à réserver aux tests statistiques en analysant les principales critiques qui leur sont adressées, les erreurs fréquemment commises dans leur usage ainsi que les raisons de leur popularité persistante en dépit des critiques. L'article propose finalement une discussion détaillée de plusieurs voies d'amélioration.

1. LOGIQUE GÉNÉRALE DES TESTS STATISTIQUES

1.1. RECHERCHE, HYPOTHÈSE, INFÉRENCE, STATISTIQUE

1.1.1. Inférence et statistique

La démarche d'inférence occupe une place très importante dans la recherche en management. Très souvent, le chercheur est amené à tirer des conclusions ou à procéder à des généralisations à partir de ses observations ou de ses résultats. Dans certains cas, la statistique peut lui permettre de le faire de manière rigoureuse. En effet, cette discipline accorde une grande place à la démarche d'inférence. Cette dernière est au coeur du raisonnement par lequel le statisticien généralise une information collectée sur un échantillon à l'ensemble de la population dont est issu cet échantillon. Au demeurant, une branche entière de la statistique est dévolue à cette démarche : c'est la *statistique inférentielle*. Le but de la statistique inférentielle est de tester des hypothèses formulées sur les caractéristiques d'une population grâce à des informations recueillies sur un échantillon issu de cette population. Les tests statistiques sont de ce fait au coeur de la statistique inférentielle.

1.1.2. Hypothèse de recherche

Un corpus théorique préexistant, des résultats empiriques antérieurs mais aussi des impressions personnelles ou de simples conjectures peuvent constituer la source des hypothèses de recherche du chercheur. Une *hypothèse de recherche* n'est autre qu'une affirmation non prouvée à propos de l'état du monde. Par exemple, l'une des hypothèses de recherche de Robinson et Pearce (1983 : 201) était la suivante : « *Entre 1977 et 1979, les banques qui ont adopté des procédures formelles de planification auront des performances significativement supérieures à celles des banques qui ne l'ont pas fait* ». Pour passer d'une hypothèse de recherche à son test au moyen de la statistique, il faut préalablement la traduire en hypothèse statistique.

1.1.3. Hypothèse statistique

Une *hypothèse statistique* est un énoncé quantitatif concernant les caractéristiques d'une population (Baillargeon et Rainville, 1978). Plus exactement, elle est une affirmation portant sur la distribution d'une ou de plusieurs variables aléatoires (Dodge, 1993). Cette affirmation peut notamment concerner les paramètres d'une distribution donnée ou encore la loi de probabilité de la population étudiée.

On appelle *paramètre* d'une population un aspect quantitatif de cette population comme la moyenne, la variance, un pourcentage ou encore toute quantité particulière relative à cette population. Les paramètres d'une population sont généralement inconnus. Cependant, il est possible de les estimer de manière statistique à partir d'un échantillon issu de la population. Par convention, les paramètres des populations sont généralement représentés par des lettres grecques (μ, σ, π, etc.). On appelle *loi de probabilité* d'une population la forme générale de la distribution de fréquences de cette population. Plus explicite, sans doute, est l'expression anglo-saxonne équivalente : « probability distribution ». La loi de probabilité d'une population peut être définie plus techniquement comme un modèle représentant au mieux une distribution de fréquences d'une variable aléatoire (Dodge, 1993).

Une hypothèse statistique se présente traditionnellement sous la double forme d'une première hypothèse appelée *hypothèse nulle* et d'une seconde hypothèse appelée *hypothèse alternative ou contraire*. L'hypothèse nulle désigne traditionnellement les situations d'absence de changement ou d'écart par rapport à un *statut quo* ou encore d'absence de différence entre des paramètres. C'est de là que provient la dénomination d'hypothèse *nulle* (Kanji, 1993 ; Dodge, 1993). Très souvent, l'objectif du chercheur est de réfuter cette hypothèse nulle au profit de

l'hypothèse alternative (Dodge, 1993 ; Sincich, 1996 ; Zikmund, 1994). L'hypothèse alternative est alors celle que le chercheur souhaite établir, celle à laquelle il croit (Sincich, 1996). Dans un tel cas, elle correspond à l'hypothèse de recherche du chercheur. Seule sa description formelle est différente : elle a souvent une formulation mathématique comme on le verra dans la suite du texte.

L'hypothèse nulle et l'hypothèse alternative ou contraire sont incompatibles et décrivent deux états complémentaires de la nature. L'hypothèse nulle est généralement notée H_0 et l'hypothèse alternative H_1 ou H_a. L'hypothèse alternative est celle qui sera acceptée si l'hypothèse nulle est rejetée. On notera que les tests statistiques sont conçus pour la réfutation et non la confirmation d'hypothèses. En d'autres termes, ces tests n'ont ni l'ambition ni le pouvoir de *prouver* des hypothèses : ils permettent de montrer qu'une hypothèse ne peut pas être acceptée parce qu'elle est associée à un niveau de probabilité trop faible (Kanji, 1993). De ce point de vue, il est important de formuler les hypothèses statistiques de telle manière que l'hypothèse alternative H_1 désigne l'hypothèse que l'on désire établir. Dès lors, plutôt que de tenter de prouver que l'hypothèse alternative est vraie, on essaye d'établir que l'hypothèse nulle est fausse et qu'il faut la rejeter.

1.2. TEST STATISTIQUE

L'évaluation de la validité d'une hypothèse statistique se fait au moyen d'un *test statistique* effectué sur des données issues d'un échantillon représentatif de la population étudiée. Ce test statistique est une procédure permettant d'aboutir, en fonction de certaines règles de décision, au rejet ou au non rejet d'une hypothèse de départ, en l'occurrence l'hypothèse nulle.

On distingue traditionnellement deux familles de tests statistiques : les *tests paramétriques* et les *tests non paramétriques*.

Un test paramétrique est un test statistique qui suppose une forme paramétrique particulière des distributions concernant les populations. C'est le cas, par exemple, lorsque les populations étudiées suivent une loi normale. Le test de Student est un exemple de test paramétrique. En effet, il vise à comparer les moyennes de deux populations qui suivent une loi normale.

Un test non paramétrique est un test statistique pour lequel il n'est pas nécessaire de spécifier la forme paramétrique de la distribution des populations. Des exemples de tests non paramétriques sont le test du signe, le test de Wilcoxon, le test de Mann-Whitney, le test de Kruskal-Wallis ou encore le test de Kolmogorov-Smirnov.

Dodge (1993) rappelle que les premiers tests statistiques ont eu lieu dans les sciences expérimentales et dans le domaine de la gestion. C'est ainsi que, par exemple, le test de Student a été conçu par William Sealy Gosset dit « Student » dans le cadre de son activité professionnelle aux brasseries Guinness. Mais ce sont Jerzy Neyman et Egon Shape Pearson qui ont développé la théorie mathématique des tests statistiques. Ces deux auteurs ont également mis en évidence l'importance de la prise en considération non seulement de l'hypothèse nulle mais aussi de l'hypothèse alternative (Dodge, 1993 ; Lehmann, 1991).

Dans le cas d'un test statistique portant sur la *loi de probabilité* suivie par la population, l'hypothèse nulle H_0 est celle selon laquelle la population étudiée suit une loi de probabilité donnée, par exemple la loi normale. L'hypothèse alternative H_1 est celle selon laquelle la population ne suit pas cette loi de probabilité donnée. Dans le cas d'un test statistique portant sur les *paramètres* d'une population, par exemple la moyenne ou la variance, l'hypothèse nulle H_0 est celle selon laquelle le paramètre étudié est égal à une valeur spécifiée alors que l'hypothèse alternative H_1 est celle selon laquelle le paramètre est différent de cette valeur.

La forme des tests statistiques dépend du nombre de populations concernées (une, deux ou davantage). Dans un test statistique portant sur une seule population, on cherche à savoir si la valeur d'un paramètre θ de la population est identique à une valeur présumée. L'hypothèse nulle qui est dans ce cas une supposition sur la valeur présumée de ce paramètre se présente alors généralement sous la forme suivante :

$H_0 : \theta = \theta_0$,

où θ est le paramètre de la population à estimer et θ_0 la valeur présumée de ce paramètre inconnu θ.

Quant à l'hypothèse alternative, elle pose l'existence d'une différence ou d'une inégalité. Par exemple, Robinson et Pearce (1983 : 201) font l'hypothèse d'une *supériorité* de performance des entreprises qui planifient formellement. Dans un tel cas, le test statistique qui sera effectué est un test dit *test unilatéral ou unidirectionnel à droite*. Si l'hypothèse était celle d'une *infériorité* de performance des entreprises planificatrices, il faudrait effectuer un *test unilatéral ou unidirectionnel à gauche*. Enfin, si l'hypothèse formulée par les deux auteurs devenait simplement celle d'une différence de performance sans plus grande précision, il faudrait effectuer un *test bilatéral ou bidirectionnel*. Il apparaît ainsi que l'hypothèse alternative peut prendre trois formes différentes :

- $H_1 : \theta > \theta_0$ (unilatéral ou unidirectionnel à droite)

- $H_1 : \theta < \theta_0$ (unilatéral ou unidirectionnel à gauche)

- $H_1 : \theta \neq \theta_0$ (bilatéral ou bidirectionnel)

1.3. RISQUES D'ERREUR

Les tests statistiques sont effectués dans le but de prendre une décision, en l'occurrence rejeter ou ne pas rejeter l'hypothèse nulle H_0. Mais parce que la décision est fondée sur une information partielle issue d'observations portant sur un *échantillon* de la population, elle comporte un risque d'erreur (Baillargeon et Rainville, 1978 ; Sincich, 1996 ; Zikmund, 1994). On distingue deux types d'erreurs dans les tests statistiques : l'*erreur de première espèce* notée α et l'*erreur de seconde espèce* notée β.

Les observations de l'échantillon peuvent conduire à rejeter l'hypothèse nulle H_0 alors que la population remplit effectivement les conditions de cette hypothèse. L*e risque (ou l'erreur) de première espèce, α,* mesure cette probabilité de rejeter l'hypothèse nulle H_0 alors qu'elle est vraie. Inversement, les observations de l'échantillon peuvent conduire à ne pas rejeter l'hypothèse nulle H_0 alors que la population remplit les conditions de l'hypothèse alternative H_1. L*e risque (ou l'erreur) de seconde espèce, β,* mesure cette probabilité de ne pas rejeter l'hypothèse nulle H_0 alors qu'elle est fausse.

Tableau 1 : Différents types d'erreurs dans un test statistique

		Situation dans la population	
		H_0 est vraie	H_0 est fausse
Décision	Ne pas rejeter H_0	Bonne décision	Erreur de 2e espèce (β)
	Rejeter H_0	Erreur de 1ère espèce (α)	Bonne décision

Puisque l'hypothèse nulle H_0 peut être vraie ou fausse et que le chercheur peut la rejeter ou ne pas la rejeter, seuls quatre cas mutuellement exclusifs sont possibles dans un test statistique, comme l'illustre le Tableau 1.

Il n'y a d'erreur que dans deux des quatre cas. Une erreur de première espèce ne peut survenir que dans les cas où l'hypothèse nulle est rejetée. De même, une erreur de seconde espèce ne peut avoir lieu que dans les cas où l'hypothèse nulle n'est pas rejetée. Par conséquent, soit le

chercheur ne commet pas d'erreur soit il en commet, mais d'un seul type. Il ne peut pas commettre à la fois les deux types d'erreur.

Le chercheur peut être tenté de choisir une valeur minimale de l'erreur de première espèce α. Malheureusement, une diminution de cette erreur de première espèce α s'accompagne d'une augmentation de l'erreur de seconde espèce β. D'une manière plus générale, la diminution de l'un des deux types d'erreur se traduit par l'augmentation de l'autre type d'erreur de même que l'augmentation de l'un des deux types d'erreur se traduit par la diminution de l'autre type d'erreur. Il ne suffit donc pas de diminuer α pour diminuer le risque global d'erreur dans la prise de décision. La seule manière de faire baisser simultanément α et β est d'augmenter la taille de l'échantillon étudié (Sincich, 1996). Sinon, il faut trouver un compromis entre α et β, par exemple en examinant la puissance du test (Dodge, 1993).

On appelle *puissance d'un test statistique* la probabilité (1-β) de rejeter l'hypothèse nulle H_0 alors qu'elle est fausse. La puissance d'un test est d'autant plus grande que l'erreur de deuxième espèce β est petite. On appelle *courbe d'efficacité* la courbe représentative des variations de β en fonction des valeurs de la statistique calculée pour lesquelles l'hypothèse alternative H_1 devrait être acceptée. Cette courbe indique la probabilité de ne pas rejeter l'hypothèse nulle H_0 -alors qu'elle est fausse- en fonction des valeurs du paramètre correspondant à l'hypothèse alternative H_1. On appelle *seuil de confiance d'un test statistique* la probabilité (1-α) d'accepter l'hypothèse nulle H_0 alors qu'elle est vraie.

Dans la pratique des tests statistiques, il est préférable de ne pas parler d'*acceptation* de l'hypothèse nulle mais de son *non rejet*. Cette nuance sémantique a son importance : si l'ambition était d'accepter H_0 la validité de la conclusion serait mesurée par l'erreur de seconde espèce β, c'est-à-dire la probabilité de ne pas rejeter l'hypothèse nulle H_0 alors qu'elle est fausse. Or, malheureusement, la valeur de β n'est pas constante. Elle dépend des valeurs spécifiques du paramètre et est très difficile à calculer dans la plupart des tests statistiques (Sincich, 1996). Du fait de cette difficulté de calculer β, la prise de décision sur la base de la puissance ou de la courbe d'efficacité des tests n'est pas chose facile. Il existe en fait une autre solution, plus pratique, qui consiste à choisir l'hypothèse nulle de sorte qu'une possible erreur de première espèce α soit beaucoup plus grave qu'une possible erreur de seconde espèce β. Par exemple, si l'on veut tester l'hypothèse de la culpabilité ou de l'innocence d'un accusé, il peut être préférable de choisir comme hypothèse nulle H_0 : « l'accusé est innocent » et comme hypothèse alternative H_1 : « l'accusé est coupable ».

Beaucoup de personnes conviendraient sans doute que, dans ce cas, une erreur de première espèce (condamner un innocent) est plus grave qu'une erreur de seconde espèce (acquitter un coupable). Dans un tel contexte, le chercheur peut se contenter de minimiser l'erreur de première espèce α.

L'erreur de première espèce est également appelée *seuil de signification* du test statistique. Il s'agit d'une grandeur que le chercheur peut fixer avant même la réalisation du test. Il est commun de trouver dans les recherches en management des seuils de signification fixés à 5% ou à 1%. Ces valeurs correspondent aux seuils de probabilités considérés comme étant trop petits pour qu'on ne rejette pas l'hypothèse nulle H_0. Autrement dit, toute probabilité d'occurrence des observations inférieure à ces seuils fixés d'avance signifie que les données suggèrent le rejet de l'hypothèse nulle H_0. Dans les recherches en management, les seuils de significations sont généralement mentionnés avec des signes, souvent des astérisques. On peut par exemple trouver le système de notation suivant (Horwitch et Thiétart, 1987) : $p < 0.10$* ; $p < 0.05$** ; $p < 0.01$*** ; $p < 0.001$****, ce qui signifie qu'une astérisque correspond à des résultats significatifs au seuil de 10%, deux astérisques à 5%, trois astérisques à 1% et quatre astérisques à 0.1% (*i.e.* un pour mille). L'absence de signe signifie que les résultats ne sont pas significatifs.

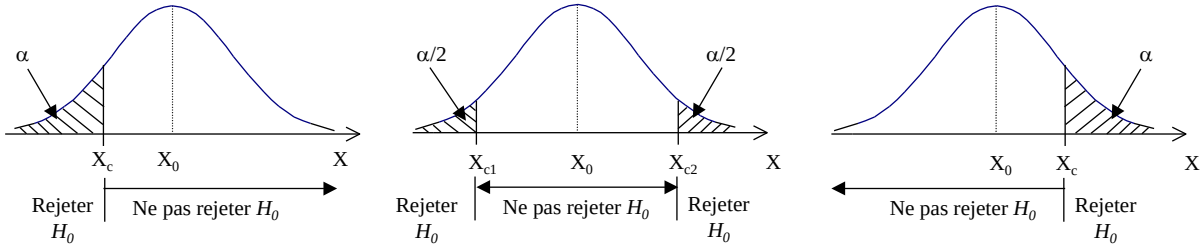
1.4. PRISE DE DÉCISION : STATISTIQUE UTILISABLE, RÉGION CRITIQUE ET SEUIL DE SIGNIFICATION OBSERVÉ

La décision de rejeter ou de ne pas rejeter l'hypothèse nulle H_0 est fondée sur le calcul d'une *statistique X*, c'est-à-dire d'une mesure calculée entièrement à partir des données issues d'un ou de plusieurs échantillons représentatifs d'une ou de plusieurs populations (Dodge, 1993 ; Kanji, 1993). Cette statistique X est une variable aléatoire. Elle doit être appropriée à l'hypothèse nulle H_0. Elle peut être relativement simple comme la moyenne ou la variance ou, au contraire, être une fonction complexe de certains de ces paramètres ou de plusieurs autres. Des exemples seront fournis dans la suite du texte. Une bonne statistique doit posséder trois propriétés (Kanji, 1993) : 1) elle doit se comporter différemment selon que c'est H_0 qui est vraie (et H_1 fausse) ou le contraire ; 2) sa loi de probabilité lorsque H_0 est vérifiée doit être connue et calculable ; 3) des tables procurant cette loi de probabilité doivent être disponibles.

La décision du rejet ou du non rejet de l'hypothèse nulle H_0 est prise au vu de la valeur de la statistique X. L'ensemble des valeurs de cette statistique qui conduisent au rejet de l'hypothèse nulle H_0 est appelé *région critique* ou encore *zone de rejet*. La région

complémentaire est appelée *zone d'acceptation* (en fait, de *non rejet*) On appelle *valeur critique* la valeur qui constitue la borne de la zone de rejet de l'hypothèse nulle H_0. Dans le cas d'un test unilatéral, il existe une seule valeur critique X_c. Dans le cas d'un test bilatéral, il en existe deux, X_{c1} et X_{c2}. La zone d'acceptation et la zone de rejet dépendent toutes les deux de l'erreur de première espèce α. En effet, α est la probabilité de rejeter H_0 alors que H_0 est vraie et 1-α et la probabilité de ne pas rejeter H_0 alors que H_0 est vraie. La Figure 1 illustre ce lien.

Figure 1 : Erreur de première espèce, zone d'acceptation et zone de rejet



Test unilatéral à gauche | Test bilatéral | Test unilatéral à droite

La règle de décision pour rejeter ou ne pas rejeter l'hypothèse nulle H_0 est la suivante : 1) dans le cas d'un test unilatéral à gauche, on rejette l'hypothèse nulle H_0 si la valeur de la statistique X est inférieure à une valeur critique X_c. Autrement dit, la zone de rejet sera constituée par des valeurs « trop petites » de X ; 2) dans le cas d'un test bilatéral, on rejette l'hypothèse nulle H_0 si la valeur de la statistique X est inférieure à une valeur critique X_{c1} ou supérieure à une valeur critique X_{c2}. Ici, la zone de rejet sera constituée par des valeurs soit « trop petites » soit « trop grandes » de X ; 3) enfin, dans le cas d'un test unilatéral à droite, on rejette l'hypothèse nulle H_0 lorsque la valeur de la statistique X est supérieure à une valeur critique X_c. La zone de rejet sera constituée par des valeurs « trop grandes » de X.

La plupart des logiciels d'analyse statistique fournissent une information très utile au chercheur : la probabilité associée à la valeur observée de la statistique X calculée. Cette probabilité est communément appelée *valeur p* (« p-value »). Plus exactement, il s'agit de la probabilité, calculée sous l'hypothèse nulle, d'obtenir un résultat aussi extrême (c'est-à-dire, selon les cas, soit plus petit ou égal, soit plus grand ou égal) que la valeur X obtenue par le chercheur à partir de son échantillon (Dodge, 1993). En termes plus concrets, la valeur p est le *seuil de signification observé*. L'hypothèse nulle H_0 sera rejetée si la valeur p est inférieure au seuil de signification fixé α (Sincich, 1996). Dans de plus en plus de publications, les chercheurs fournissent directement les valeurs p associées aux tests statistiques qu'ils ont

effectués (voir, par exemple, Horwitch et Thiétart, 1987). De ce fait, le lecteur peut comparer cette valeur p au seuil de signification α qui lui agrée et juger lui-même si l'hypothèse nulle H_0 aurait du être rejetée ou non. La valeur p a un intérêt supplémentaire : elle précise la localisation de la statistique X par rapport à la région critique (Kanji, 1993). Par exemple, une valeur p à peine inférieure au seuil de signification fixé α suggère qu'il existe dans les données des indications selon lesquelles l'hypothèse nulle H_0 ne devrait pas être rejetée, alors qu'une valeur p largement inférieure au seuil de signification α permet de conclure que les données fournissent de solides raisons de rejeter l'hypothèse nulle H_0. De même, une valeur p à peine supérieure au seuil de signification suggère l'existence dans les données d'indications selon lesquelles l'hypothèse nulle H_0 pourrait être rejetée, alors qu'une valeur p largement supérieure au seuil de signification α permet de conclure que les données fournissent de solides raisons de ne pas rejeter l'hypothèse nulle H_0.

1.5. ÉTAPES DE L'USAGE D'UN TEST STATISTIQUE

Dans les ouvrages de statistique (Baillargeon et Rainville, 1978 ; Ceresta, 1986 ; Dodge, 1993), la démarche présentée pour effectuer un test statistique à partir d'un échantillon est généralement la suivante :

1. Formuler les hypothèses (l'hypothèse nulle H_0 et l'hypothèse alternative H_1).
2. Choisir le seuil de signification α du test, c'est-à-dire le risque (généralement compris entre 1% et 10%) de rejeter l'hypothèse nulle H_0 alors qu'elle serait vraie.
3. Obtenir un échantillon d'observations aléatoires à partir de la population.
4. Pour les tests paramétriques, déterminer la loi de probabilité correspondant à la distribution d'échantillonnage (loi normale, loi de Poisson, etc.).
5. Déterminer une statistique X (c'est-à-dire un critère fonction des données) dont on connaît la loi de probabilité lorsque l'hypothèse nulle H_0 est vraie.
6. Calculer à partir du seuil de signification α les valeurs critiques (X_c ou X_{c1} et X_{c2}) et en déduire la région de rejet et la région d'acceptation de l'hypothèse nulle H_0.
7. Établir les règles de décision : 1) si la statistique observée sur l'échantillon appartient à la région d'acceptation, on ne rejettera pas l'hypothèse nulle H_0 ; 2) si la statistique observée sur l'échantillon appartient à la région de rejet, on rejettera l'hypothèse nulle H_0 au profit de l'hypothèse alternative H_1.

8. Calculer la statistique et déterminer si elle se situe dans la zone de rejet ou de non rejet de l'hypothèse nulle H_0.
9. Prendre la décision de ne pas rejeter ou de rejeter l'hypothèse nulle H_0 sur la base du test effectué sur l'échantillon étudié.

En fait, la tâche du chercheur sera beaucoup plus facile. En effet, la plupart des logiciels d'analyse statistique (SAS, SPSS, etc.) déterminent la statistique X appropriée au test choisi, procèdent à son calcul et indiquent la valeur p qui lui est associée. Certains logiciels comme Statgraphics vont même jusqu'à indiquer la décision à prendre (rejet ou non rejet de l'hypothèse nulle H_0) en fonction du seuil de signification fixé par le chercheur. En fait, pour le chercheur, la contrainte principale est de savoir choisir le test approprié. Mais ce choix n'est pas vraiment sans risque.

2. LE PROCÈS DES TESTS DE SIGNIFICATION STATISTIQUE

2.1. LES PRINCIPALES CRITIQUES

On reproche aux tests de signification statistique à la fois leurs défauts intrinsèques et les erreurs liées à leur usage par les chercheurs (Rozeboom, 1960 ; Sawyer et Peter, 1983 ; Cohen, 1994 ; Schmidt, 1996, Hunter, 1997 ; Poitevineau, 1998 ; Snyder et Thompson 1998; Gill, 1999 ; Nickerson, 1999, 2000 ; Krueger, 2001 ; Lecoutre, Poitevineau et Lecoutre, 2003 ; Morgan, 2003 ; Armstrong, 2007a, 2007b ; Gibbons, Crout et Healey, 2007 ; Levine *et al.*, 2008). Les erreurs commises par les chercheurs seront examinées ultérieurement. Concernant les défauts intrinsèques, les principales critiques sont les suivantes :

2.1.1 Une hypothèse vraiment… nulle

Plusieurs détracteurs des tests de signification statistique mettent en cause l'utilité même d'une hypothèse nulle (H_0) ponctuelle, c'est-à-dire une hypothèse attribuant au paramètre une valeur précise et non un intervalle. En effet, une telle hypothèse est pratiquement toujours fausse, ne serait-ce qu'à plusieurs décimales après la virgule. Ainsi, toute différence, même infinitésimale, deviendra statistiquement significative pour autant que la taille de l'échantillon soit suffisamment grande. Par exemple, pour un test de différence de moyenne entre deux groupes, toute différence non rigoureusement nulle pourra être « rendue » significative pour autant que la taille des groupes soit suffisamment grande. Une autre façon de voir les choses est de partir des populations et non plus des échantillons. Il est très probable qu'à l'échelle des populations, l'hypothèse nulle soit fausse (il y aura une différence, même infinitésimale) Dès lors, tout test

conduira à un résultat significatif et l'information apportée par le test est donc quasi-nulle. Cela conduit plusieurs détracteurs des tests à la conclusion de l'inutilité de recueillir des données et de procéder à de tels tests dont les résultats sont connus d'avance.

2.1.2. Une démarche anti-scientifique

L'accent mis par les tests de signification statistique sur l'hypothèse nulle (hasard, absence de différence, *statu quo*, absence d'effet, etc.) aurait conduit à affaiblir la démarche scientifique. En effet, celle-ci consiste essentiellement à confronter les données recueillies à l'hypothèse de recherche (H_1), puis, quand elles semblent incompatibles avec H_1, à envisager d'autres hypothèses (dont celle du hasard, éventuellement). Au contraire, dans le cas des tests de signification statistique, l'hypothèse du hasard (H_0) est généralement mise en avant; elle est la première testée, quel que soit son (in)intérêt scientifique. Par conséquent, l'hypothèse de recherche ne sera même pas examinée si le test est non significatif, alors même qu'elle pourrait présenter une bonne compatibilité avec les données. Si le choix d'assimiler l'hypothèse nulle aux situations de hasard, d'absence de différence et d'effet, etc. présente de multiples avantages techniques comme nous l'avons évoqué dans la première section de ce texte, il n'en demeure pas moins que cela éloigne de la préoccupation principale du chercheur qui est sa question de recherche (souvent H_1).

2.1.3. Hypothèse nulle, hypothèse alternative et hypothèse de recherche

Lorsque l'hypothèse de recherche (celle à laquelle le chercheur s'intéresse réellement) conduit à une prédiction précise du paramètre, il est possible de l'identifier à l'hypothèse nulle et non à l'hypothèse alternative. C'est notamment le cas lorsqu'il s'agit de validation de modèles. Une illustration peut ainsi être trouvée dans le test des modèles de causalité. Mais, quel que soit le choix adopté par le chercheur pour l'hypothèse de recherche (H_0 ou H_1), les difficultés demeurent : si on identifie l'hypothèse de recherche à l'hypothèse alternative (cas classique), alors, comme déjà évoqué, il suffit de choisir un échantillon suffisamment grand pour être sûr d'obtenir un résultat favorable (significatif) ; si, par contre, on identifie l'hypothèse de recherche à l'hypothèse nulle, on se trouve confronté au dilemme suivant (Poitevineau, 1998) :

- monter un design de recherche très sensible (par exemple en choisissant un grand échantillon) ; c'est se ramener au cas précédent de rejet presque certain de l'hypothèse, alors même que celle-ci peut constituer une très bonne approximation de la réalité, voire la meilleure disponible ;

- monter un design de recherche très peu sensible (par exemple en choisissant un petit échantillon) ; c'est permettre une corroboration facile et artificielle de l'hypothèse de recherche.

2.1.4. Probabilité de l'hypothèse ou probabilité des données ?

Les tests de signification statistique fournissent une probabilité des données observées conditionnellement à la véracité de l'hypothèse nulle ou Pr(Données|H_0). Pourtant, ce qui intéresse vraiment le chercheur, c'est plutôt la probabilité des hypothèses (H_0 et/ou H_1) conditionnellement aux données observées Pr(H_0|Données) ou Pr(H_1|Données). Cette position peu naturelle et peu intuitive conduit du reste à des erreurs d'interprétation constantes comme nous le verrons plus en détail par la suite. De nombreux utilisateurs interprètent ainsi les résultats observés Pr(Données|H_0) comme la probabilité conditionnelle de l'hypothèse nulle, à savoir Pr(H_0|Données). Pourtant les valeurs de ces deux probabilités peuvent être très différentes l'une de l'autre (par exemple, 0,005 et 0,82).

2.1.5. α est arbitraire, valeur p est ambiguë

Le seuil de signification α retenu (généralement 5%, mais aussi parfois 10% ou 1%, etc.) est totalement arbitraire et, pourtant, il conduit directement à décider du rejet ou du non rejet des hypothèses. Nous avons évoqué la valeur p (ou « seuil de signification observé ») et montré comment elle permettait de pallier le caractère brutal d'un seuil de signification ponctuel fixé α. En effet, la valeur p est caractéristique des données observées et indicatrice du degré de réfutation de l'hypothèse nulle : si p est jugé suffisamment faible, on rejette l'hypothèse nulle, on considère qu'on a réussi à en montrer la fausseté et le résultat est déclaré significatif (Poitevineau, 1998 ; Nickerson, 2000). Seulement, la valeur p elle-même n'est pas exempte de critiques. En particulier, elle dépend de la taille de l'échantillon : plus l'échantillon est grand, plus les valeurs p seront faibles, toutes choses égales par ailleurs. Il devient donc difficile de distinguer, dans une valeur p donnée, ce qui provient de la grandeur de l'effet testé (effect size) de ce qui provient de l'effet taille de l'échantillon (sample size). Dans tous les cas, qu'il s'agisse du seuil de signification α ou de la valeur p, la position du curseur qui va déterminer la frontière entre ce qui est « significatif » et ce qui ne l'est pas restera une position (valeur) arbitraire.

2.1.6. La grandeur de l'effet

La grandeur de l'effet (effect size) mesure l'amplitude ou la force de la relation entre deux ou plusieurs variables dans la population. Les tests ont souvent négligé la grandeur de l'effet.

Comme le souligne Poitevineau (1998), un résultat significatif n'est qu'une indication de l'existence de l'effet supposé; un résultat non significatif un constat d'ignorance. Aller plus loin sur la seule base du test serait assimiler à tort significativité statistique et grandeur de l'effet. En fait, le chercheur est surtout intéressé par la grandeur et la précision des effets. Par exemple, est-ce qu'une différence de moyenne est triviale, faible, moyenne ou forte ? Cette différence fait-elle sens, naturellement ? Est-elle suffisamment importante pour être incluse dans un modèle plus large ? Toutes ces questions de recherche importantes ne sont pas traitées par les tests de signification statistique car elles en dépassent le cadre.

2.2. DES ERREURS FRÉQUENTES À L'USAGE

Au-delà de leurs défauts intrinsèques, on reproche aux tests de signification statistique d'être la source d'erreurs fréquentes commises par les chercheurs même si on pourrait estimer que ces derniers sont le plus à blâmer. Voici les principales erreurs rencontrées (Sawyer et Peter, 1983 ; Cohen, 1994 ; Schmidt, 1996, Hunter, 1997 ; Poitevineau, 1998, 2004 ; Snyder et Thompson 1998; Gill, 1999 ; Nickerson, 1999, 2000 ; Krueger, 2001 ; Lecoutre, Poitevineau et Lecoutre, 2003 ; Morgan, 2003 ; Armstrong, 2007a, 2007b ; Gibbons, Crout et Healey, 2007 ; Levine *et al.*, 2008) :

2.2.1 Le renversement des conditions

Une première erreur, la plus typique, et que nous avons déjà évoquée, est de considérer la valeur p (seuil observé) ou α comme une probabilité concernant l'hypothèse ($Pr(H_0|Données)$) et non plus conditionnelle à celle-ci $Pr(Données|H_0)$. Ce qui donne les affirmations erronées du style :

- « la probabilité que l'hypothèse nulle soit vraie est p (ou α) » ;
- « la probabilité que les résultats soient dus au seul hasard est p (ou α) » ;
- « la probabilité que l'hypothèse alternative soit vraie est 1-p (ou 1-α) ».

Ces erreurs s'expliquent par l'écart entre ce que les chercheurs attendent et ce que les tests fournissent. Les chercheurs ont recours aux tests de signification statistique pour décider si les résultats obtenus confirment ou infirment leur hypothèse. Or les tests indiquent la probabilité p d'obtenir les résultats observés (sachant que l'hypothèse nulle est vraie) et non la probabilité de l'hypothèse nulle (au regard des données). Nous avons déjà signalé que les valeurs de ces deux probabilités pouvaient différer très sensiblement même s'il reste vrai que plus p est petit, plus les preuves contre l'hypothèse nulle sont grandes.

On trouve des erreurs similaires concernant les intervalles de confiance (ou « fourchettes »). Souvent, les chercheurs affirment à propos d'un intervalle de confiance à 95% [X ; Y] : « le paramètre π a une probabilité 0.95 de se trouver dans la fourchette (ou l'intervalle) [X ; Y] ». Cette interprétation naturelle est néanmoins fausse. Les paramètres sont généralement inconnus mais ils ont une valeur fixe, non aléatoire. L'événement « X< π <Y » est vrai ou faux (car π est fixé), et on ne peut pas lui attribuer de probabilité (sinon 1 ou 0). L'interprétation correcte de l'intervalle de confiance 95% est la suivante : « 95% des intervalles calculés sur l'ensemble des échantillons possibles (tous ceux qu'il est possible de tirer dans la population) contiennent la vraie valeur π ». Chaque intervalle particulier a une probabilité 0 ou 1 de contenir la vraie valeur. Ici, ce n'est pas le paramètre qui est aléatoire mais les bornes de l'intervalle de confiance (ou « fourchette ») qui varient d'un échantillon à l'autre.

2.2.2. La probabilité de reproduction du résultat

Une deuxième erreur est de considérer 1-p (ou 1-α) comme la probabilité de reproduire le résultat observé. Les théories traditionnelles des tests ne fournissent aucune indication de la probabilité de reproduire le résultat observé. Cette reproductibilité peut être envisagée de deux manières :

- soit elle ne concerne que la significativité du résultat, c'est le cas le plus courant; ce qui donne un énoncé du type : « la probabilité qu'une réplique soit significative est 1-p (ou 1-α) » ;
- soit la reproductibilité concerne la valeur même de l'effet : « il y a 95% de chances d'observer un même résultat dans les travaux ultérieurs ».

Ces énoncés sont bien entendus erronés, même s'il est exact que plus le seuil observé p est faible et plus la reproductibilité est assurée.

2.2.3. Significativité statistique et significativité substantielle

Une troisième erreur est de confondre la significativité statistique avec la significativité substantielle. C'est considérer que plus un résultat est statistiquement significatif, plus il est scientifiquement intéressant, et/ou que plus l'effet correspondant dans la population est grand. Les paragraphes précédemment consacrés à la grandeur de l'effet (effect size) et à l'effet de la taille de l'échantillon (sample size) ont montré la différence entre les deux significativités. C'est le chercheur qui a réfléchi à son hypothèse de recherche et les tests sont uniquement des outils à sa disposition. C'est au chercheur de réfléchir à ce qui peut faire sens, et c'est à lui de décider des hypothèses statistiques à tester. Par exemple, c'est à lui de savoir si, du point de vue

de la signification, il ne vaut pas mieux choisir pour hypothèse nulle qu'entre les moyennes de deux groupes la différence est égale à une constante non nulle plutôt qu'une différence strictement nulle.

2.2.4. L'acceptation de l'hypothèse nulle

Une quatrième erreur est de conclure à la véracité de l'hypothèse nulle en cas de résultat non significatif. Par exemple, un chercheur examine l'hypothèse de recherche selon laquelle « la mise en place d'un plan de formation augmente la performance du personnel ». Un test de différence de moyennes le conduit à observer des résultats non significatifs : il n'observe pas de différence significative entre les groupes ayant suivi ou pas un plan de formation. Il doit en conclure qu'il ne peut pas rejeter l'hypothèse nulle et ne doit surtout pas affirmer qu'il « accepte » cette hypothèse nulle. Encore une fois, il peut simplement dire qu'il ne peut pas refuser l'hypothèse nulle. Par contre, il ne doit pas affirmer : « la mise en place d'un plan de formation n'augmente pas la performance du personnel ». Ceci équivaut à une accepter l'hypothèse nulle et signifie qu'on conclut par inférence (à l'ensemble de la population) à l'absence d'effet, sans se contenter de jugements descriptifs incontestables. De même, la phrase « la valeur 0 étant comprise dans l'intervalle de confiance on ne peut pas refuser l'hypothèse nulle selon laquelle les deux séries de valeurs ont la même moyenne » est correcte car il s'agit d'une description concernant les échantillons. Cependant, on ne doit pas affirmer : « l'ensemencement n'a pas eu d'effet sur la prise des pêcheurs » car il s'agit là d'une inférence.

2.2.5. L'omission d'ajustement en situation de tests simultanés

Une cinquième erreur d'usage très fréquente, y compris dans les meilleures publications, concerne l'absence de correction du seuil de significativité en fonction du nombre de tests simultanés. C'est très souvent le cas à l'occasion de tests de différences de moyennes entre de multiples groupes. Pourtant, la plupart des logiciels statistiques (par exemple, le logiciel SPSS) indiquent dans leur documentation ou dans l'aide en ligne du logiciel si le test effectue les corrections ou pas.

2.3. UNE POPULARITÉ PERSISTANTE

Les erreurs précédentes sonnent comme autant de critiques supplémentaires des tests : une méthode donnant lieu à tant d'erreurs dans son application, même chez des usagers avertis, n'aurait-elle rien à se reprocher ? Pourtant, la popularité des tests de signification statistique reste grande auprès des chercheurs. En dépit des nombreuses critiques dont leur usage fait constamment l'objet, ces tests sont conventionnellement acceptés comme une preuve de la

validité des conclusions et sont une norme incontournable pour la publication des résultats de recherche. Tout se passe comme si on était en présence d'une pratique critiquable aux plans théorique et méthodologique mais sociologiquement adaptée, d'un outil mal utilisé car son mode d'emploi se révèle particulièrement trompeur mais bénéficiant néanmoins d'une aura jusque-là intacte. Poitevineau (1998 ; 2004) identifie plusieurs raisons à ce paradoxe apparent :

- L'ambiguïté de la terminologie : les tests de signification statistique sont souvent appelés des « tests de signification », ce qui renvoie à « significatif », à quelque chose qui donne du sens, qui a de l'importance, etc. Ce faisant, la confusion entre significativités statistique et substantielle est induite.
- L'objectivité : les chercheurs souhaitent disposer de méthodes objectives et formalisées leur permettant de savoir si un jeu de données présente des variations aléatoires ou systématiques. Et ils estiment important de ne pas devoir s'en remettre à leurs seules intuition et subjectivité pour déterminer la part d'aléatoire et de systématique dans les données. Dès lors, les tests confèrent aux conclusions des chercheurs cette impression d'objectivité qui est chez eux un souci crucial.
- La scientificité : dans des disciplines comme le management qui souffrent plus ou moins d'un complexe de non scientificité, du moins par rapport à des sciences plus « dures », l'appareillage mathématique et le formalisme des tests fournissent à bon compte une apparence de scientificité. En outre, la rigueur des mathématiques et l'aura dont elles jouissent sont sensées diffuser sur l'ensemble de la recherche, assurant *de facto* sa validité.
- Le renfort de Karl Popper : les tests de signification statistique offrent une grande ressemblance avec l'idée de Popper selon laquelle la démarcation entre énoncés scientifiques et non scientifiques est réalisée sur la base du caractère réfutable, ou non, de ces énoncés. Une hypothèse scientifique est une hypothèse qui peut être empiriquement « testée ». La théorie des tests a ainsi pu bénéficier du succès des idées de Popper.
- Un confort assuré et une économie : les tests assurent un confort certain à leurs utilisateurs. En semblant fournir une procédure automatique, les tests dispensent d'une réflexion supplémentaire. Les tests déchargent le chercheur de la tâche d'interprétation. Dans ces conditions, le test, avec sa possibilité de déclarer « significatif » un effet, est vu comme une solution, déchargeant le chercheur de la tâche d'interprétation, comme si la significativité statistique se suffisait à elle-même.

C’est donc comme si le succès persistant des tests de signification statistique était dû à un formidable malentendu : une apparence d’objectivité et de scientificité ainsi qu’une illusion d’adéquation aux besoins des chercheurs permise par l’ignorance que ces derniers ont de la nature et des conditions d’utilisation desdits tests. Pourtant, les critiques -qui ne sont pas nouvelles- sont si constantes que Lecoutre et Poitevineau (2000 : 683) ont voulu prédire : « il y a de bonnes raisons de penser que le rôle des tests de signification usuels dans la recherche en psychologie sera considérablement réduit dans un proche avenir. Les résultats des analyses statistiques traditionnelles devraient être systématiquement complétés ("au delà des seuls seuils observés p") pour inclure systématiquement la présentation d'indicateurs de la grandeur des effets et leurs estimations par intervalles. Ces procédures pourraient rapidement devenir de nouvelles normes de publication ». Cette prédiction devient progressivement réalité. L’élément déclencheur est non pas un réveil brusque ou une prise de conscience subite des utilisateurs mais plutôt le relais des critiques pris ces dernières années par des institutions comme l’American Psychological Association ou des comités éditoriaux de revues scientifiques. Dans cette perspective, quelles sont les principales voies d’amélioration possibles ?

3. QUELQUES VOIES D’AMÉLIORATION

Plusieurs voies d’amélioration sont envisageables. Une façon pratique de les aborder est de commencer par les recommandations de la *Task Force* chargée par le bureau des affaires scientifiques de l'*American Psychological Association* d'étudier le rôle du test de signification dans la recherche en psychologie. On pourra ensuite explorer des voies complémentaires.

3.1. Les recommandations de l’American Psychological Association

- Tests d’hypothèses : il est difficile d’imaginer une seule situation dan laquelle une décision binaire d’acceptation/refus serait préférable au fait de reporter les valeurs p ou, mieux encore, un intervalle de confiance. N’utilisez jamais l’expression malheureuse : « accepter l’hypothèse nulle ». Toujours fournir une mesure de la grandeur de l’effet quand on reporte une valeur p.
- Intervalles : des intervalles devraient être fournis pour toute grandeur d’effet concernant les résultats principaux. Fournir de tels intervalles pour les corrélations et les indices d’association ou de variation à chaque fois que possible.
- Grandeur des effets : toujours présenter les grandeurs d’effets pour les résultats bruts. Si les unités de mesure ont un sens pratique (par exemple, nombre de cigarettes fumées

par jour), préférer une mesure non standardisée (coefficient de régression ou différence de moyenne) à une mesure standardisée.

- Puissance et taille de l'échantillon : fournir l'information sur la taille de l'échantillon et le processus qui a conduit au choix d'une telle taille. Expliciter les postulats concernant la grandeur des effets, l'échantillonnage et la mesure des variables de même que les procédures analytiques utilisées pour le calcul de la puissance. Dans la mesure où le calcul de la puissance fait davantage sens lorsqu'il est effectué avant la collecte et l'examen des données, il est important de montrer comment des estimations de la grandeur des effets ont été déduites des recherches et théories antérieures pour écarter la soupçon qu'elles ont pu être extraites des données de l'étude en cours ou, pis encore, qu'elles ont été construites pour justifier un échantillon donné.

3.2. Les méthodes statistiques complémentaires

Poitevineau (1998 ; 2004) présente de manière détaillée, entre autres, deux séries de méthodes statistiques prometteuses (Sawyer et Peter, 1983 ; Gill, 1999 ; Nickerson, 1999, 2000 ; Levine *et al.*, 2008) : les méthodes de vraisemblance et les méthodes bayésiennes.

- Les méthodes de vraisemblance : dans le cas simple de deux hypothèses ponctuelles H_0 et H_1, la méthode du rapport de vraisemblance consiste à calculer le rapport des densités de probabilité de la statistique observée (x) sous H_0 et sous H_1, à savoir $f(x|H_0) / f(x|H_1)$. Ce rapport exprime, sur la base des résultats observés, les « chances » d'une hypothèse relativement à l'autre. On peut éventuellement retenir H_0 ou H_1 selon que ce rapport est supérieur ou inférieur à une constante choisie arbitrairement (un, par exemple, si l'on ne privilégie aucune hypothèse). La méthode du rapport de vraisemblance présente l'avantage de ne faire intervenir ni probabilités *a priori*, ni éléments non observés, Si le rapport de vraisemblance permet de juger de la force probante des données entre deux hypothèses ponctuelles, il est malheureusement très rare, dans la pratique, que le chercheur soit confronté à un tel cas.
- Les méthodes bayésiennes : utilisée en tant que méthode d'inférence statistique, la méthode bayésienne consiste à calculer, au moyen du théorème de Bayes, la distribution *a posteriori* pour le paramètre auquel on s'intéresse, à partir :
    - des données observées ;
    - d'un modèle d'échantillonnage associé ;
    - des probabilités *a priori* sur le paramètre.

L'approche bayésienne a été, et est encore, beaucoup critiquée comme une méthode trop subjective car elle nécessité de spécifier des probabilités *a priori*. Toutefois, le poids de la distribution *a priori* dans la distribution *a posteriori* diminue d'autant que la masse des données s'accroît. Ainsi, deux chercheurs partant de distributions *a priori* différentes s'accorderont sur leurs conclusions si les données sont suffisantes. Il est d'ailleurs recommandé de varier les distributions a priori (position optimiste, neutre, pessimiste) et d'analyser la sensibilité des résultats. En fait, les méthodes bayésiennes semblent disposer de nombreux atouts pour s'imposer comme véritables « challengers » des tests traditionnels. Il est à noter que l'analyse bayésienne est de plus en plus utilisée en sciences de gestion, notamment en finance et, dans une moindre mesure, en marketing. Il est également remarquable que de plus en plus de grands éditeurs de logiciels statistiques incorporent des modules d'analyse bayésienne dans leurs programmes (c'est, par exemple, le cas de l'éditeur de logiciels statistiques SPSS qui a incorporé un module bayésien dans son programme AMOS).

3.3. Reprendre la posture de chercheur

Au-delà de la compréhension des recommandations des institutions comme l'American Psychological Association ou de la considération de méthodes nouvelles d'inférence statistique (comme les méthodes bayésiennes), une troisième voie d'amélioration –et celle-là nous paraît, de loin, la plus importante !– concerne l'attitude même du chercheur. Nous avons vu qu'il manquait parfois de prise de distance et d'esprit critique vis-à-vis de son environnement, en particulier des outils à sa disposition. Or, la meilleure recherche, celle susceptible de produire les résultats les plus intéressants, nécessite sans doute d'aller au-delà du mimétisme basique et de l'usage irréfléchi des dispositifs les plus usités. Certes, nous avons conscience que le test est de nos jours encore un critère important dans la sélection des articles soumis à publication, dans le sens où un résultat non significatif a très peu de chances d'être publié, et dans ce cas, il faut très souvent que la puissance soit forte. Par exemple, Cohen (1998) recommande un seuil d'acceptation de 20. Le faible nombre de résultats non significatifs publiés peut d'ailleurs aussi bien résulter d'une sélection opérée par les chercheurs eux-mêmes que d'une politique éditoriale délibérée : les résultats non significatifs ne sont acceptés que si la puissance est forte. Or, comme cela est rarement le cas, il s'ensuit un très faible taux de publication des ces résultats non significatifs. À ce propos, il est également question de réplication concernant les compléments à apporter aux tests de signification statistique. Mais réfléchissons à la situation suivante : si plusieurs chercheurs testent, indépendamment les uns des autres, une même

hypothèse nulle H_0 vraie, environ 5% d'entre eux trouveront un résultat significatif (au seuil de 5%) et seront pratiquement les seuls à même de publier, laissant ainsi croire à la réalité du phénomène étudié (rejet de H_0). On se retrouverait alors uniquement avec des « faux résultats » dans la littérature. Dans ce cas, les réplications des chercheurs peu audacieux ne feront qu'aggraver la situation : seuls des résultats significatifs seront publiés et diffusés. Quelle assurance avons-nous de ne pas être dans un tel cas de figure lorsque nous procédons à une revue de la littérature ? Pratiquement aucune. C'est là, encore une fois, une illustration du devoir de vigilance, d'audace et d'esprit critique de la part du chercheur. Et nous dirons que ces qualités sont d'autant plus nécessaires que les outils deviennent plus nombreux et sophistiqués. Cette exigence individuelle gagnerait à être accompagnée d'une action collective visant à réduire un des travers de notre environnement professionnel : obtenir que les résultats « non significatifs » puissent sortir de l'ombre.

CONCLUSION

Cet article a voulu fortement attirer l'attention des chercheurs en management stratégique sur les dangers liés à l'usage irréfléchi des tests de signification statistique. Il prend appui sur un travail que nous consacrons depuis une décennie aux tests statistiques (Mbengue, 1999, 2007) et sur une série de publications qui ont nourri et continuent d'alimenter la controverse sur l'utilité des tests de signification statistique. Ces publications ont concerné pratiquement tous les champs disciplinaires : la psychologie, d'abord (Rozeboom, 1960 ; Schmidt, 1996 ; Hagen, 1997 ; Hunter, 1997 ; Poitevineau, 1998, 2004 ; Snyder et Thompson, 1998 ; Wainer, 1999 ; Nickerson, 1999, 2000 ; Lecoutre et Poitevineau, 2000 ; Krueger, 2001 ; Lecoutre, Poitevineau et Lecoutre, 2003) mais également le marketing (Sawyer et Peter, 1983), les sciences politiques (Gill, 1999), la prospective (Armstrong, 2007a, 2007b), l'écologie (Gibbons, Crout et Healey, 2007), la communication (Levine, Weber, Hullett, Hee Sun Park et Lindsey, 2008), etc.

On remarque aisément que la controverse a connu un essor à partir de la deuxième moitié des années 1990 et qu'elle s'est progressivement étendue de la psychologie aux autres champs disciplinaires... à l'exception notable du management stratégique Par conséquent, la motivation principale du présent article était de porter cette controverse à la connaissance de la communauté francophone des chercheurs en management stratégique. En effet, notre expérience directe nous avait montré que peu de membres de notre communauté connaissaient la controverse autour des tests de signification statistique et encore moins sa teneur. À propos de cette controverse, il convient de souligner qu'il est devenu difficile de parler de débats. C'est ainsi que, s'agissant des tests de signification statistique, notre article présente leur

procès (leurs limites intrinsèques et les problèmes liés à leur usage) sans aucune opinion divergente sur ces aspects. Cela tient au fait qu'il n'y a plus véritablement débat sur les dangers (de mauvais usage) des tests de signification statistique. Le seul débat porte sur l'identité du coupable : les tests ou les chercheurs qui les utilisent à mauvais escient ? D'un côté, ceux qui clament qu'il faut bannir les tests car ils sont dangereux et contreproductifs ; de l'autre, ceux qui estiment qu'il ne faut pas jeter le bébé avec l'eau du bain ! C'est à cela qu'est actuellement réduit l'essentiel du « débat » sur les tests de signification statistique !

Nous avons, dans la deuxième partie de cet article, accordé une place très importante aux travaux de Jacques Poitevineau (1998, 2004) et de toute l'équipe rouennaise dirigée par Bruno Lecoutre (http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris.html) car ces chercheurs ont produit un travail très complet de critique des tests de signification statistique classiques (dits tests « fréquentistes ») et développé un plaidoyer argumenté pour une perspective alternative bayésienne. Pourtant, il faut garder à l'esprit que l'ensemble des griefs portés contre les tests de signification statistique constitue à présent de la sagesse conventionnelle qui a été construite de manière cumulative (Rozeboom, 1960 ; Sawyer et Peter, 1983 ; Cohen, 1994 ; Schmidt, 1996, Hunter, 1997 ; Poitevineau, 1998 ; Snyder et Thompson 1998; Gill, 1999 ; Nickerson, 1999, 2000 ; Krueger, 2001 ; Lecoutre, Poitevineau et Lecoutre, 2003 ; Morgan, 2003 ; Armstrong, 2007a, 2007b ; Gibbons, Crout et Healey, 2007 ; Levine *et al.*, 2008). En d'autres termes, il existe un accord général sur les dangers (de l'usage) des tests de signification statistique.

Un premier danger pour le chercheur serait d'ignorer leur mode d'emploi, c'est-à-dire leurs conditions d'utilisation. Ce danger devient particulièrement menaçant compte tenu de la disponibilité croissante des logiciels statistiques. Un autre danger pour le chercheur consisterait à s'abriter derrière l'image scientifique des tests statistiques, à céder à leur aura et au confort apparent lié à leur utilisation pour abdiquer sa responsabilité. Or, c'est le chercheur qui doit choisir s'il teste ou pas, ce qu'il teste et par quel moyen. Mais, plus encore, le chercheur doit garder à l'esprit que les tests de signification statistique ne sont qu'un instrument à l'intérieur d'un dispositif et d'une démarche de recherche : cette recherche commence avant l'éventuel test, se poursuit pendant le test et continue après le test. Quant au test lui-même, il n'est qu'un outil et, en tant que tel, il ne vaut que si on sait s'en servir et à bon escient. De ce point de vue, les questions récurrentes sur l'utilité des tests de signification statistique sont un bon stimulant et un garde-fou précieux pour l'exercice d'une saine activité de recherche.

BIBLIOGRAPHIE

Abelson R.P. (1997), "A retrospective on the significance test ban of 1999 (if there were no significance tests, they would be invented)", *in* Harlow, L. L., Mulaik, S. A. et Steiger, J. H. (eds), *What if there were no significant tests?*, Mahwah, NJ, Lawrence Erlbaum, pp. 117-141.

Armstrong J.S. (2007a), "Significance tests harm progress in forecasting", *International Journal of Forecasting*, Volume 23, Issue 2, April-June, pp321-327.

Armstrong J.S. (2007b), "Statistical significance tests are unnecessary even when properly done and properly interpreted: Reply to commentaries", *International Journal of Forecasting*, Volume 23, Issue 2, April-June, pp335-336.

Baillargeon G. et Rainville J. (1978), *Statistique appliquée*, Tome 2, 6e édition, Trois-Rivières, Les Éditions SMG.

Boursin J.L. et Duru G. (1995), *Statistique*, Paris, Vuibert.

Ceresta (1986), « Aide-mémoire pratique des techniques statistiques », *Revue de statistique appliquée*, volume XXXIV, numéro spécial.

Chow G.C. (1960), "Tests for equality between sets of coefficients in two linear regressions", Econometrica, 28,3, p.591-605.

Cohen J. (1994), "The earth is round ($p < .05$)", *American Psychologist*, *49*, pp. 997-1003.

Cohen J. (1998), *Statistical power analysis for the behavioral sciences*, Hillsadle, New Jersey, Lawrence Erlbaum Associates Publishers.

Dodge Y. (1993), *Statistique : Dictionnaire encyclopédique*, Paris, Dunod.

Gibbons J.M., Crout N.M et Healey J.R. (2007), "What role should null-hypothesis significance tests have in statistical education and hypothesis falsification?", *Trends in Ecology & Evolution*, Volume 22, Issue 9, September, pp445-446.

Gill J. (1999), "The Insignificance of Null Hypothesis Significance Testing", *Political Research Quarterly*, Vol. 52, No. 3 , pp. 647-674.

Hagen R.L. (1997), "In praise of the null hypothesis statistical test", *American Psychologist*, 52, pp. 15-24.

Harlow L.L., Mulaik S.A. et Steiger J.H. (eds) (1997), *What if there were no significance tests?*, Mahwah, NJ, Erlbaum.

Horwitch M. et Thietart R.A. (1987), “The effect of business interdependencies on product R&D-intensive business performance”, *Management Science*, 33,2, pp.178-197.

Hunter J. (1997), “Needed: A ban on the significance test”, *Psychological Science*, 8, pp.3-7.

Kanji G.K. (1993), *100 Structural tests*, Thousand Oaks : Sage publications.

Kaufmann P. (1994), *Statistique*, Paris, Dunod

Krueger J. (2001), “Null Hypothesis Significance Testing”, *American Psychologist*, Vol. 56 Issue 1, p16-26

Lecoutre B., Poitevineau J. (2000), « Aller au delà des tests de signification traditionnels : vers de nouvelles normes de publication », *L'Année Psychologique*, 100, pp. 683-713.

Lecoutre M-P., Poitevineau J., Lecoutre B. (2003), "Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests", *International Journal of Psychology*, Vol. 38, Issue 1, p37-45

Lehmann E.L. (1991), *Testing statistical hypotheses*, Pacific Grove, California, Wadsworth & Brooks.

Levine T., Weber R., Hullett C., Hee Sun Park et Lindsey L. (2008), “A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research”, *Human Communication Research*, Vol. 34 Issue 2, p171-187.

Mbengue A. (1999), « Tests de comparaison », in "*Méthodes de recherche en management*", R.A. Thiétart (Ed.), Dunod, (3e édition, 2007).

Morgan P.L. (2003), “Null Hypothesis Significance Testing: Philosophical and Practical Considerations of a Statistical Controversy”, Exceptionality, Vol. 11 Issue 4, p209-221.

Mulaik S.A., Raju N.S.et Harshman R.A. (1997), “There is a time and place for significance testing”, *in* Harlow L.L., Mulaik S.A. et Steiger, J.H. (eds), *What if there were no significant tests?*, Mahwah, NJ, Lawrence Erlbaum, pp. 65-116.

Nickerson R.S. (1999), “Statistical Significance Testing: Useful Tool or Bone-Headedly Misguided Procedure? Review of What If There Were No Significance Tests? by Lisa L. Harlow, Stanley A. Mulaik, and James H. Steiger (Eds.)”, *Journal of Mathematical Psychology*

Nickerson R.S. (2000), “Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy”, *Psychological Methods*, Vol. 5, No. 2, pp. 241-301.

Poitevineau J. (1998), *Méthodologie de l'analyse des données expérimentales : étude de la pratique des tests statistiques chez les chercheurs en psychologie, approches normative, prescriptive et descriptive*, Thèse de Doctorat, Université de Rouen.

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

Robinson R.B. et Pearce J.A. (1983), "The impact of formalized strategic planning on financial performance in small organizations", *Strategic Management Journal*, 4, p.197-207.

Rozeboom W.W. (1960), "The Fallacy of The Null-Hypothesis Significance Test", *Psychological Bulletin*, Vol. 57, No. 5, pp.416-428.

Sawyer A.G. et Peter J. P. (1983), "The Significance of Statistical Significance Tests in Marketing Research", *Journal of Marketing Research*, Vol. 20 Issue 2, p122-133

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

Sincich T. (1996), *Business statistics by example*, 5e édition, Upper Saddle River, New Jersey, Prentice-Hall.

Snyder P.A. et Thompson B. (1998), "Use of Tests of Statistical Significance and Other Analytic Choices in a School Psychology Journal: Review of Practices and Suggested Alternatives", *School Psychology Quarterly*, Vol. 13, No. 4, pp. 335-348.

Toyoda T. (1974), "Use of the Chow test under heteroscedasticity", Econometrica, 42, 3, p. 601.

Wainer H. (1999), « One cheer for null hypothesis significance testing », *Psychological Methods*, 6, pp. 212-213.

Zikmund W.G. (1994), *Business research methods*, 4e édition, Orlando, Florida, The Dryden Press.