ZAHID ELM'HAMEDI

EFFETS DU VOCABULAIRE ET DE L'AMBIGUÏTE LINGUISTIQUE SUR LA COMPREHENSION DES TESTS STATISTIQUES

Abstract. Effects of vocabulary and linguistic ambiguity on the understanding of statistical tests. It is widely recognized, at least in the university medium, that the teaching of statistical tests is difficult for the teacher as well as for the learner. Indeed, the literature of educational research on this tool has revealed that there are various difficulties, often associated to some misconceptions, which are encountered at every age and level of expertise. The research findings show that a student cannot generally understand and describe the fundamental ideas underlying the practice of statistical tests. Instead, students generally rely on mechanical calculus, which is based much more on memorization and computation, than on reflection and interpretation. The aim of this paper is to present some factors that bring difficulties in apprehension to this practice and that must be considered in its teaching. Very specifically, we focused in this paper on the effect that the *vocabulary and linguistic ambiguity factors*, used in statistical tests, could have on the interpretation by our students of the notions and expressions conveyed during the teaching of this practice.
Keywords: Statistical tests – Misconceptions – Frequentist approach – Bayesian approach – Linguistic ambiguity.

Résumé. Il est largement reconnu, du moins dans le milieu universitaire, que l'enseignement des tests statistiques exige de relever de grands défis, aussi bien pour l'enseignant que pour l'apprenant. En effet, une littérature abondante de travaux de recherches en didactique sur ce sujet indique diverses difficultés, souvent en rapport avec des conceptions erronées, que l'on peut rencontrer à tous les âges et à tous les niveaux d'expertise. Les conclusions de ces recherches s'accordent sur le fait qu'en général, un étudiant ne peut pas expliciter les fondements de la pratique de cet outil statistique, et que seul un calcul machinal est souvent restitué, résultant beaucoup plus de mémorisation que de réflexion et d'interprétation. L'objectif de cet article est de présenter un ensemble de facteurs qui sont susceptibles de contribuer aux difficultés d'appréhension de ce sujet et qui doivent être pris en considération lors de son enseignement. En particulier, nous nous sommes focalisés dans cet article sur l'effet que pourraient avoir *les facteurs liés au vocabulaire* et à *l'ambiguïté linguistique*, présents dans la pratique des tests statistiques, sur l'interprétation par nos étudiants des notions et des expressions véhiculées lors de l'enseignement de cette pratique.
Mots-clés : Tests statistiques – Conceptions erronées – Approche fréquentiste – Approche bayésienne – Ambiguïtés linguistiques.

1. Introduction

Il y a principalement deux cadres scientifiques dans lesquels la pratique des tests statistiques joue un rôle très important. Le premier, désigné comme la « *statistique de production* » ou « *statistique industrielle* », inclut le domaine du « *contrôle de qualité* » et, ce qui est pareil, vise l'aide à la prise de décision. Le deuxième est celui de la « *recherche scientifique* »; il a pour objectif une production de connaissance.

ANNALES de DIDACTIQUE et de SCIENCES COGNITIVES, volume 24, p. 133 - 181.

Etablir une distinction entre ces deux axes demeure une tâche délicate et prêtant à discussion. Grosso modo, ces deux axes se basent respectivement sur la théorie des « *tests d'hypothèses* (ou *de décision*) » de Neyman[1]-Pearson[2] et sur la théorie des « *tests de signification* » de Fisher[3]. Malheureusement, ce qui est généralement présenté aujourd'hui dans l'enseignement de la pratique des tests statistiques et ce dont traitent la plupart des ouvrages, même spécialisés en statistique, est un amalgame d'idées dérivant de ces deux théories considérées comme s'excluant mutuellement, sous la dénomination de « *logique hybride de l'inférence scientifique* » ou « *tests de signification de l'hypothèse nulle* » (Gigerenzer, 1993; Zaki, Elm'hamedi, 2009; Elm'hamedi, 2010). Cet amalgame est, en quelque sorte, un assemblage qui cache toujours son origine hybride et qui se présente souvent comme la logique monolithique et la recette unique de l'inférence scientifique toute entière (Gigerenzer, 2004), considérée comme un rituel.

Néanmoins, la pratique des tests statistiques[4] est devenue indispensable, et sur elle repose largement l'analyse statistique des données dans plusieurs champs disciplinaires comme la psychologie, la sociologie, l'économie, la médecine, l'écologie, le droit, etc. Malgré cela, il est largement reconnu, par comparaison avec les différents concepts de la boîte à outils statistiques, que cette pratique est source de conceptions erronées diverses, constituant un sujet difficile aussi bien pour l'enseignant que pour l'apprenant (Oakes, 1986; Falk, Greenbaum, 1995; Haller, Krauss, 2002). La littérature statistique montre de manière évidente l'existence de conceptions erronées chez les utilisateurs des tests statistiques, à tous les âges et à tous les niveaux d'expertise (Gigerenzer, 1993; Vallecillos, 1995; Poitevineau, Lecoutre, 2001; Haller, Krauss, 2002; Zendrera, 2004; Batanero, Diaz, 2006; Elm'hamedi, 2010; …). De plus, il existe de nombreux travaux qui se sont focalisés sur les difficultés d'application des tests statistiques en recherche expérimentale et particulièrement en sciences de l'éducation, découlant de leur utilisation incorrecte ou d'une interprétation fautive des résultats obtenus. Nous pouvons, à titre d'exemples, citer les travaux effectués par Morrison et Henkel (1970), Thompson(1996), Poitevineau(1998), Zendrera(2003), Zaki et Elm'hamedi (2009, 2013).

Évidemment, plusieurs facteurs sont générateurs d'une telle difficulté concernant les tests statistiques. Ainsi, nous sommes convaincus que l'obstacle de leur enseignement ne sera surmonté qu'à l'issue d'une étude approfondie des causes

[1] Jerzy Neyman: (1894 - 1981), statisticien et mathématicien polonais.
[2] Egon Pearson: (1895 - 1980), statisticien anglais, fils de Karl Pearson.
[3] Sir Ronald Aylmer Fisher: (1890 - 1962), statisticien anglais.
[4] Tout au long de cet article, l'expression « tests statistiques » se rapporte à la fois au test de signification de Fisher et au test d'hypothèses de Neyman-Pearson.

sous-jacentes. Il nous apparaît que les facteurs de difficultés sont pour la plupart d'entre eux raisonnablement repérés et bien compris, ou pour le moins reconnus. Mais quelques facteurs restent méconnus, voire ignorés. Cet article est focalisé sur l'exploration de l'effet chez des étudiants en dernière année de licences scientifiques, relevant d'universités du Maroc, de deux facteurs particuliers, souvent déconsidérés et négligés dans les études exploratoires publiées sur les éventuelles incompréhensions relativement à la statistique en général et de concepts associés aux tests statistiques en particulier. Depuis plusieurs années, ces deux facteurs ont suscité notre curiosité, ainsi qu'une préoccupation et un intérêt particuliers. Il s'agit spécifiquement du « *vocabulaire* » en usage dans la pratique des tests statistiques et de « *l'ambiguïté linguistique* » que les expressions habituellement utilisées lors de la mise en œuvre des procédures peuvent véhiculer.

Remarquons que le développement des concepts est intimement dépendant du vocabulaire qui est utilisé afin de les mobiliser dans les procédures sous-jacentes. Nous ne pouvons rien savoir des concepts sans les mots qui les désignent. Et ceux-ci peuvent être chargés de sens dans le langage courant. Pointons au passage le cas de l'étudiant qui lie toujours le mot « *population* » à un ensemble *d'êtres humains* seulement. De plus, les étudiants ont des difficultés notoires à appréhender les concepts statistiques, ne comprenant pas bien les concepts de base, que nous indiquerons plus avant dans cet article. D'une manière équivalente, ceci peut être exprimé de la façon suivante : « *Les étudiants ne maîtrisent pas bien le vocabulaire de base, et de ce fait éprouvent des troubles au contact de concepts plus complexes* ». Dans la plupart des cas, le vocabulaire statistique (de base) utilise des mots de tous les jours, mais souvent avec des significations spécialisées. Les étudiants devront alors accomplir deux opérations intellectuelles : connaître la signification spécialisée et se représenter la signification courante. Plus encore, beaucoup de ces concepts de base sont superficiellement simples, mais ils sont en réalité beaucoup plus complexes ou subtils. Par exemple, le terme « *population* », que nous avons déjà évoqué, peut signifier, dans le langage de tous les jours, soit l'ensemble des gens qui vivent dans un pays, soit leur effectif. En statistique, dans le cadre d'une étude donnée, la population désigne un ensemble (Mclean, 2002), celui des éléments considérés, nommés alors les individus de l'étude.

Ainsi, le caractère facilement trompeur du vocabulaire utilisé dans les tests statistiques constitue l'un des principaux facteurs induisant des difficultés de compréhension de ce concept. Ainsi, les deux mots « *Prouver* » et « *Vrai* », habituellement employés dans la pratique de tests statistiques, constituent à eux seuls des facteurs susceptibles de contribuer à une mauvaise appréhension de cette pratique (Mclean, 2002). On affirme familièrement : « *Vous pouvez prouver quelque chose à l'aide de la statistique inférentielle !* ». Mais la réalité est autre : « *Vous ne*

pouvez rien prouver[5] *à l'aide de cette discipline !* ». Il s'agit d'un problème sémantique. En effet, pour les mathématiciens (incluant beaucoup de statisticiens), prouver une assertion, signifie justifier sa véracité par une démonstration rigoureuse. Pour les autres chercheurs, notamment en sciences humaines, qui utilisent habituellement les tests statistiques dans leurs recherches, cela signifie plutôt : « *l'étude faite va à l'appui de l'assertion !* ». Notons au passage que même la démonstration mathématique qu'on croyait être une preuve universelle et indiscutable, ne l'est pas ainsi parmi les mathématiciens. Dans ce sens, Lamrabet (2001) a pu établir une liste non exhaustive d'expressions mettant en évidence le caractère relatif des démonstrations mathématiques. On peut citer comme exemple le caractère relatif « *à une époque donnée* », « *au courant de pensée mathématique adopté (formalisme, constructivisme, ...)* », « *aux moyens utilisés pour produire la démonstration (automatique, expérimental, ...)* », « *au degré d'exigence du lecteur* », ...

Pareillement, la notion de preuve, utilisée dans des domaines où les tests statistiques jouent un rôle très important (le domaine de droit, par exemple) n'est pas absolue. Une déclaration peut être considérée comme prouvée si l'argumentation avancée en sa faveur semble être « *assez solide* » et est une base satisfaisante pour une prise de décision. Ainsi, si on « *prouve* » qu'une personne est coupable d'un acte criminel, cela signifie que la « *preuve* » avancée contre cette personne est assez solide, allant au-delà du doute raisonnable, et qu'il apparaît donc justifié de prononcer une condamnation. « *Preuve* » dans ce cas signifie donc « *preuve au-delà du doute raisonnable* » et non pas « *au-delà du doute* ».

La tendance à concevoir un test statistique de façon mécanique est accentuée dans l'esprit de plusieurs étudiants et professionnels. Il est clair que des outils mathématiques, souvent complexes, sont utilisés dans la pratique des tests statistiques, mais la situation reste la même que dans le domaine de l'architecture, par exemple, où les calculs mathématiques sont seulement des aides à la réalisation du design d'un bâtiment. Le vocabulaire généralement utilisé, à savoir l'utilisation des mots « *prouver* » et « *vrai* » peut tromper. Le fait de parler de valeurs vraies d'un paramètre, de preuve, de règles de décision et de niveaux de signification 1%, 5%, est aussi trompeur. Un test statistique n'est pas directement concerné par la valeur d'un paramètre, mais par les modèles probabilistes qui traduisent au mieux la situation étudiée (Mclean, 2000). Accepter un modèle associé à une valeur spécifique d'un paramètre, signifie seulement qu'il est raisonnable de fonder nos prédictions sur un tel modèle ; le rejeter signifie qu'il est préférable d'utiliser un autre modèle, associé à une valeur différente de ce paramètre.

[5]Dans cette phrase, le mot « *prouver* » est utilisé dans le sens strict qu'il a en mathématiques.

D'un autre côté, l'ambiguïté linguistique constitue aussi un facteur majeur contribuant aux conceptions erronées relatives aux tests statistiques, en particulier dans l'utilisation des probabilités conditionnelles. Falk et Greenbaum (1995) ont signalé que ce facteur a été noté comme problématique pour comprendre de manière générale des relations à caractère probable (Bar-Hillel, Falk, 1982; Falk, 1992; Gillman, 1992; Margolis, 1987; Nickerson, 1996). De manière particulière, la distinction entre les deux probabilités conditionnelles[6] $P(\mathcal{D} \mid H)$ et $P(H \mid \mathcal{D})$ devient assez compliquée à cause du langage utilisé par les tests statistiques.

Considérons, par exemple, les deux expressions suivantes : « *La probabilité d'obtenir les données $\mathcal{D}$ par hasard* » et « *La probabilité que les données $\mathcal{D}$ soient obtenues par hasard* ». Il n'est pas surprenant que ces expressions soient considérées par beaucoup de gens comme traduisant une même signification. Pourtant, la première expression signifie : « *la probabilité qu'un processus considéré comme aléatoire, appelé H, peut produire $\mathcal{D}$* », alors que la seconde signifie : « *la probabilité qu'un événement connu $\mathcal{D}$, soit produit par un processus aléatoire H* ». La première expression fait alors référence à $P(\mathcal{D} \mid H)$, et la seconde à $P(H \mid \mathcal{D})$.

Il est aussi facile de trouver des expressions relatives à des relations à caractère probabiliste qui peuvent conduire à plus d'une seule interprétation. Par exemple, l'expression : « *la probabilité que le hasard produise les données $\mathcal{D}$* », peut être considérée comme étant équivalente à : « *sachant qu'un processus est aléatoire, la probabilité que les données $\mathcal{D}$ soient produites* ». Mais, il peut aussi signifier : « *la probabilité qu'un processus aléatoire puisse produire les données $\mathcal{D}$* ». Soit H désignant le processus aléatoire, la première expression peut être représentée par $P(\mathcal{D} \mid H)$, et la seconde peut être représentée par $P(H \cap \mathcal{D})$ (ou $P(H).P(\mathfrak{D} \mid H)$). Cette erreur est souvent commise en matière de tests statistiques. Par exemple, la définition de la puissance est : « *la probabilité de rejeter avec raison l'hypothèse nulle* ». Elle signifie donc la probabilité de rejeter l'hypothèse nulle, sachant qu'elle est fausse. Cependant, cette dernière interprétation n'est pas la seule, la puissance pourrait aussi être interprétée comme étant « *la probabilité de rejeter avec raison l'hypothèse nulle* », qui est, en fait, la probabilité conjointe des deux événements : « *l'hypothèse nulle est fausse* » *et* « *on rejette l'hypothèse nulle* ».

Pareillement, l'ambiguïté linguistique associée à l'expression « *erreur de 1ère espèce* » contribue aussi à la tendance à interpréter le niveau de signification, souvent

[6] $\mathfrak{D}$ désigne un événement déterminé en fonction des données construites ou non construites (exemples: les données construites ou les valeurs plus fortes dans le cas des tests de signification, et la région critique (RC) ou éventuellement la région d'acceptation de H_0 (RA) dans le cas des tests d'hypothèses), et H désigne l'événement « l'hypothèse nulle H_0 est vraie » ou bien l'événement « l'hypothèse alternative H_1 est vraie ».

noté α, qui est égal à la probabilité de la région critique RC sachant que l'hypothèse H_0 est vraie ($P(RC \mid H_0)$), comme étant (*erronément*) égal à la probabilité $P(H_0 \mid RC)$. Cette erreur vient du fait que le concept de niveau de signification est souvent présenté aux étudiants et aussi introduit dans la plupart des ouvrages, même spécialisés en statistique comme étant égal à la probabilité de l'erreur de 1ère espèce. Cette interprétation (ou cette présentation) se forme facilement dans les esprits des étudiants et semble être la racine de la confusion subséquente, étant donné que α est une probabilité conditionnelle bien définie, alors que l'expression « *erreur de 1ère espèce* » n'est pas conditionnellement formulée. De même, les étudiants savent que rejeter H_0 et commettre une erreur sont usuellement associés. Cette ambiguïté linguistique permet à la combinaison exacte des deux événements : « *rejet de H_0* » et « *commettre une erreur* » d'être ouverte à différentes interprétations : s'agit-il de leur conjonction ou d'un événement sachant le second ? Maintenant, quand H_0 est rejetée et que nous voulons vérifier la probabilité de l'erreur commise, nous nous demandons de quel type d'erreur il s'agit. Naturellement, le concept lâche « *erreur de 1ère espèce* » vient immédiatement à l'esprit. Ainsi, la distinction cruciale entre les deux directions opposées des probabilités conditionnelles devient floue via la médiation de l'expression mal définie « *erreur de 1ère espèce* ». L'emploi de cette expression de médiation devrait être accompagné de sa probabilité, sinon il est préférable d'invoquer le risque de première espèce. Comme des manuels scolaires, ouvrages et enseignants le signalent souvent, « *l'événement conditionnel* » $(A \mid B)$ n'est pas à proprement parler un concept légitime. Ce sont les probabilités de cet événement (c'est-à-dire, les probabilités conditionnelles) qui lui donnent une valeur convenable et sa légitimité, et permettent de le définir sans équivoque comme un rapport de probabilités d'événements. « *L'erreur de 1ère espèce* » est ainsi une expression malheureuse en tests statistiques (Falk, 1986).

Nous procéderons tout au long de cet article à une étude exploratoire sur l'effet que pourraient impliquer *le vocabulaire* et *l'ambiguïté linguistique* sur la compréhension des concepts qui fondent les tests statistiques chez les étudiants susmentionnés. Mais auparavant, nous allons, dans un premier temps, partager avec le lecteur les autres facteurs, que nous estimons générateurs de difficultés d'appréhension. Signalons que nous avons pu déduire ces facteurs, qui se présentent souvent de manière assez subtile, d'une réflexion approfondie et de longue haleine, ainsi que d'une revue de littérature très diverse en rapport avec les tests statistiques. Cela a constitué pour nous un défi à surmonter un problème didactique qui n'a jamais cessé de nous préoccuper et avec lequel nous avons vécu depuis plusieurs années. Signalons qu'il existe des détracteurs qui critiquent l'usage des tests statistiques en recherche expérimentale, en proposant sa pure et simple exclusion de l'enseignement ou au moins son accompagnement par d'autres méthodes statistiques telles que les intervalles de confiance, la taille de l'effet, etc. Mais on doit prendre en considération, d'une part, que la pratique des tests statistiques a eu cours dans le

passé, continue d'être utilisée aujourd'hui et restera certainement en usage pour plusieurs années encore, et d'autre part, que cet outil est théoriquement solide, historiquement enraciné et qu'il est utilisé très généralement comme il devrait effectivement l'être (Mclean, 2002 ; Zaki, Elm'hamedi, 2013). Plusieurs alternatives aux tests statistiques ont été proposées. Cependant, chacune peut être évaluée pour sa capacité à remplacer ou compléter la procédure des tests statistiques d'aujourd'hui. Pareillement, il est à signaler qu'une théorie pourrait être considérée favorable jusqu'à ce qu'une autre théorie prouve elle-même être plus désirable scientifiquement et méthodologiquement. Cet article s'inscrit donc dans le cadre de l'idée réaliste suivante : « *sachant que les tests statistiques continueront d'être utilisés dans le futur, qu'est-ce qui peut être fait pour améliorer leur utilisation ?* ». Il est ainsi écrit avec l'idée que les tests statistiques constituent « *les meilleures* » alternatives disponibles dans la boîte à outils de l'inférence statistique.

2. Autres facteurs susceptibles d'avoir des effets sur la compréhension des tests statistiques

Les facteurs contribuant aux difficultés de compréhension des tests statistiques sont multi-dimensionnelles et assez nombreux. Ils reflètent la multitude de difficultés et de conceptions erronées détectées chez les apprenants et les utilisateurs à tout niveau d'expertise. Comme nous l'avons signalé dans l'introduction de ce travail, nous allons présenter dans cette section des facteurs autres que *le vocabulaire* et *l'ambiguïté linguistique* sur lesquels repose la présente étude. Nous avons répertorié ces autres facteurs en cinq classes essentielles, que nous présenterons dans ce qui suit, à savoir : « *la complexité de la nature de l'inférence statistique* », « *la complexité de la nature de l'outil des tests statistiques* », « *les pratiques usuelles d'enseignement de l'outil des tests statistiques* », « *l'usage aveugle des logiciels spécialisés en statistique* » et enfin « *les facteurs psychanalytiques* ».

2.1. La nature complexe de l'inférence statistique

Contrairement à la statistique descriptive, la statistique inférentielle sert, d'une part, à évaluer des hypothèses statistiques, et d'autre part, à estimer des paramètres de populations, et ce sur la base d'un ensemble de caractéristiques très spécifiques (Elm'hamedi, 2010). Elle est ainsi caractérisée par un ensemble de propriétés essentielles. Nous citerons, à titre d'exemples, les trois éléments suivants : « *la multitude d'écoles* » (ou de théories) pour traiter une situation-problème donnée, « *la non-unicité* » du résultat fourni par l'ensemble de ces écoles, et l'obligation d'utilisation d'une notion très complexe appelée « *jugement* » pour tirer une conclusion, quelle que soit la théorie utilisée.

2.1.1. La multitude d'écoles

Actuellement, nous avons *au moins* quatre écoles différentes de pensée de référence concernant la statistique inférentielle. Il y a « *l'approche fishérienne* », « *l'école de*

Neyman-Pearson », « *la théorie bayésienne* » et celle de « *vraisemblance* ». Une description assez détaillée et une évaluation entière de chacune d'elles sont présentées dans Oakes (1986). Il faut noter qu'aucune de ces approches statistiques n'est sans controverses et ni sans critiques. Toutefois, chaque école est un paradigme à part ayant son propre raisonnement, sa propre logique, sa propre force de preuve, ses concepts spécifiques, sous-jacents aux résultats auxquels il peut aboutir. Plus encore, dans une situation donnée d'inférence statistique, les résultats auxquels on pourrait aboutir pourraient éventuellement varier d'une école à l'autre, étant donné qu'elles se basent toutes sur un élément fondamental caractérisant la statistique inférentielle qui est « *le jugement argumenté* ».

2.1.2. La non-unicité du résultat fourni

Le caractère de non-unicité relatif au résultat fourni par application des outils de l'inférence statistique est un obstacle pour les étudiants, voire même les professionnels de la statistique. Ces derniers continuent à croire que la statistique inférentielle est semblable à la statistique descriptive ; de ce fait, ils considèrent que les résultats de cette discipline devraient *a priori* revêtir un caractère absolu. Cela est sans doute à l'origine de certaines conceptions erronées chez les différents utilisateurs. En fait, la statistique inférentielle utilise la notion de probabilités, et en particulier celle des probabilités conditionnelles : la probabilité des hypothèses H sachant des données $\mathfrak{D}$ ($P(H \mid \mathfrak{D})$) dans le cas de la statistique bayésienne et la probabilité des données sachant les hypothèses ($P(\mathfrak{D} \mid H)$) dans le cas des théories de Fisher et de Neyman-Pearson. Ces nuances introduites dans les paradigmes de base de ces différentes théories contribuent largement à la non-unicité du résultat fourni face à une situation donnée.

2.1.3. La complexité de la tâche de formulation de l'élément de jugement

Outre la non-unicité du résultat fourni par les différentes théories de la statistique inférentielle, *le jugement* du statisticien, qui s'appuie sur sa connaissance du domaine de recherche travaillé, reste prépondérant pour la détermination d'un résultat. Du point de vue de l'enseignement, les étudiants sont souvent en attente de règles de décision rigides et habillées de fer, telles par exemple : « *si la taille n de l'échantillon est supérieure ou égale à 30 alors la loi de la distribution dans l'échantillon sera normale, et si n est inférieur à 30 alors elle ne le sera pas* », ou encore « *si la p-valeur p est inférieure à 0,05 alors on rejettera H_0, sinon on échouera à rejeter H_0* », etc. Cependant, ils ont souvent des difficultés avec une approche qui dit : « *utilisez votre jugement pour conclure* ». Il faut dire que l'instauration de la valeur du jugement dans une situation d'inférence statistique, est assez complexe car elle dépend de la théorie adoptée : de Fisher, de Neyman-Pearson, de Bayes, … D'ailleurs, la détermination de la valeur seuil (α) prise pour le niveau de signification et de la probabilité *a priori* de H_0 ($P(H_0)$), utilisées respectivement comme éléments

de jugements dans le cas des tests statistiques et celui de la statistique bayésienne, est une tâche qui exige une réflexion approfondie sur plusieurs paramètres directement reliés aux contraintes du problème posé (Labovitz, 1970; Elm'hamedi, 2010; Zaki, Elm'hamedi, 2013).

2.2. La nature complexe des tests statistiques

La nature même des tests statistiques leur confère un caractère subtil, car leurs concepts de base sont à la fois abstraits et ouverts à plusieurs interprétations (Watts, 1991). Les fondements mathématiques et la terminologie véhiculés par les procédures sous-jacentes aux tests statistiques sont à l'origine de difficultés et de conceptions erronées chez les étudiants, voire même chez les utilisateurs professionnels (Oakes, 1986 ; Poitevineau, 1998 ; Haller, Krauss, 2002). Les notions de *vocabulaire* et d*'ambiguïté linguistique*, faisant l'objet du travail d'expérimentation rapporté dans cet article, relèvent toutes deux de cette catégorie importante et de la nature complexe de l'outil des tests statistiques.

2.2.1. Le caractère difficile du fondement mathématique

Le fondement mathématique, en particulier probabiliste, justifiant la construction des tests statistiques dépasse généralement les connaissances des programmes d'enseignement dans lesquels ces tests sont présentés. Ce n'est qu'au niveau d'un Master de mathématiques que l'on peut *a priori* établir en toute rigueur la théorie de cet outil (Zaki, Elm'hamedi, 2009, 2013). Par ailleurs, la notion de probabilités conditionnelles, considérée comme la pierre angulaire sur laquelle reposent les procédures sous-jacentes aux tests statistiques, est un exemple significatif de notion qui engendre beaucoup de difficultés chez les étudiants et les utilisateurs à tout niveau d'expertise. Par exemple, beaucoup d'étudiants utilisent les relations probabilistes erronées suivantes : $P(\mathcal{D}|H) = P(H|\mathcal{D})$, $P(\mathcal{D}|H) = P(H \cap \mathcal{D})$, … (Falk, 1986; Falk, Greenbaum, 1995; Nickerson, 2000).

De plus, il est probable que ce soit la complexité de ce fondement mathématique qui amène souvent les enseignants à ne pas présenter à leurs étudiants les tests d'hypothèses (surtout paramétriques) à partir d'un modèle statistique $(\Omega, \Gamma, P_\theta)_{\theta\in\Theta}$, en ne leur faisant pas connaitre, par exemple, que la région critique est non seulement un élément de la tribu Γ mais aussi que sa détermination nécessite, *sous certaines conditions*, la résolution d'un *problème d'optimisation* (probabiliste) induisant la minimisation simultanée de deux types de risques d'erreurs définis à partir d'une « *fonction coût* » et d'un « *espace d'actions* » fixés *a priori*.

2.2.2. La multitude de terminologies et de définitions

Une étude de la littérature de concepts particuliers impliqués dans les tests statistiques démontre l'existence de plusieurs inconsistances qui sont associées aux définitions et à la terminologie habituellement utilisées. Truran (1998), par exemple,

a observé un manque de consistance et de compréhensibilité dans une étude qu'il a menée sur les différentes définitions de l'hypothèse nulle H_0 qui existent dans les ouvrages et manuels de statistique. De leur côté, Freund et Perles (1993) ont aussi trouvé quatre définitions différentes du concept de p-valeur. Le niveau de signification a été nommé aussi alpha, et il est représenté par le symbole α ou (peu couramment) par le symbole $P_{critique}$ (Thompson, 1994). La p-valeur a été référée à « *la probabilité de signification* », « *le niveau de probabilité* » (Huberty, 1985), « *p* » (Carver, 1978), « $P_{calculée}$ » (Thompson, 1994), « *prob-value* », « *probabilité queue* », « *P* », « *P-value* » et « *niveau de signification descriptif* » (Freund, Perles, 1993). On comprend que de toutes ces différences peuvent résulter des troubles.

2.3. Les pratiques usuelles d'enseignement des tests statistiques

La difficulté de l'enseignement et aussi de l'apprentissage du sujet des tests statistiques est largement reconnue (Garfield, Ahlgren, 1988 ; Hawkins, 1991). Les travaux antérieurs remettent surtout en cause les pratiques usuelles d'enseignement de cet outil, car celles-ci restent essentiellement ancrées sur des formules et des techniques routinières (Ehrenberg, 1990), encourageant l'apprentissage par cœur et le calcul machinal, plutôt que la réflexion et l'interprétation des résultats (Hawkins, Jolliffe, Glickman, 1992). Malheureusement, ces pratiques d'enseignement présentent souvent les tests statistiques non seulement comme « *un amalgame entre les tests de signification et ceux d'hypothèses* », mais aussi « *de façon isolée, sans leur intégration dans un modèle plus général de l'inférence scientifique* », et enfin comme « *seul outil de la statistique inférentielle, qui sert à évaluer la discordance qui pourrait exister entre des données construites et des hypothèses scientifiques* ».

2.3.1. La présentation des tests statistiques comme amalgame entre les tests de signification et ceux d'hypothèses

Plusieurs auteurs ont déclaré que la base théorique des tests statistiques est largement mal comprise (e.g. Oakes, 1986 ; Gigerenzer, Murray, 1987 ; Pollard, Richardson, 1987 ; Lunt, Livingstone, 1989 ; Cohen, 1994 ; Dracup, 1995). Gigerenzer (1993) a argumenté que les conceptions de la plupart des étudiants de psychologie relatives à l'inférence statistique sont incohérentes parce qu'elles dérivent de deux théories mutuellement incompatibles, d'une part, celle de Fisher, et de l'autre, celle de Neyman-Pearson (Gigerenzer, Murray, 1987 ; Sedlmeier, Gigerenzer, 1989 ; Gigerenzer, 1993). Bien que la théorie de Neyman-Pearson se soit développée en dehors de celle de Fisher, il n'en reste pas moins qu'il y a eu beaucoup de débats entre ceux qui appuient la théorie de Fisher et ceux qui appuient celle de Neyman-Pearson, autour de la nature même de l'inférence statistique (Box, 1978; Kruskal, 1980). Dans le domaine de la psychologie, par exemple, où les tests statistiques jouent un rôle très important, Gigerenzer (1993) a étudié trente ouvrages de statistique destinés aux étudiants de cette discipline, sans relever aucune discussion

sur les deux approches de Fisher et de Neyman-Pearson. Ses conclusions étaient que ce qui est devenu institutionnel en psychologie et même en d'autres branches des sciences sociales n'était plus la statistique *fishérienne* ou *neyman-pearsonienne*, mais un *amalgame* incohérent d'idées incompatibles de Fisher et de Neyman-Pearson. Selon lui, cet amalgame est, en quelque sorte, la progéniture résultant d'un mariage forcé entre ces deux théories.

2.3.2. La présentation de façon isolée d'une procédure de tests statistiques

Il est important de rappeler qu'une procédure de tests statistiques n'est qu'une partie d'un processus plus global d'inférence scientifique. Une telle procédure ne doit donc pas être mise en œuvre sans être intégrée à une approche méthodologique de recherche plus générale et à un plan expérimental adéquat.

Malheureusement, une telle procédure est souvent présentée aux étudiants de la manière mécanique suivante : *a)* on se donne deux hypothèses statistiques H_0 et H_1, et des données D_0, *b)* on effectue un calcul probabiliste, et enfin *c)* on tire une conclusion de type : « *Rejeter* H_0 », « *Accepter* H_0 » (ou éventuellement : « *Echouer à rejeter* H_0 »). De cette façon de faire peuvent résulter plusieurs difficultés d'appréhension relatives aux tests statistiques. Notamment, peuvent être occultées les propriétés fondamentales suivantes :

- Les procédures sous-jacentes aux tests statistiques ne doivent pas être mises en œuvre comme une recette qui donne automatiquement et commodément l'une des réponses : « *Rejeter* H_0 », « *Accepter* H_0 » ou « *Echouer à rejeter* H_0 », à n'importe quel problème posé.
- Le rôle effectif d'un test statistique est d'une manière générale la mise en évidence des fluctuations d'échantillonnages et des erreurs de mesures expérimentales commises lors de la construction des données D_0.
  Cela ne sera pas garanti par cette façon de présenter les tests statistiques, puisque les données D_0 sont éventuellement imaginées par l'enseignant et non construites par l'étudiant.
- Les hypothèses statistiques H_0 et H_1 sont seulement deux modèles représentant chacun une hypothèse ou une action scientifique. En outre, les scientifiques ne sont intéressés ni par les tests statistiques, ni par les hypothèses testées par ces tests, mais plutôt par des hypothèses (d'étendue plus large) qui réfèrent à un phénomène scientifique, c'est-à-dire, à des hypothèses scientifiques. Ces dernières sont différentes des hypothèses testées par les tests statistiques. D'où l'importance de l'interprétation qui s'inscrit dans une connaissance du champ de recherche.
- Le résultat fourni par un test statistique, à savoir : « *Rejeter* H_0 », « *Accepter* H_0 » ou « *Echouer à rejeter* H_0 » n'est pas l'objectif ultime du chercheur. En pratique, il faut poursuivre la recherche en optant pour d'autres plans

expérimentaux bien construits, et en utilisant éventuellement des outils de l'inférence statistique autres que les tests statistiques, afin de mettre en évidence la signification pratique qui intéresse avant tout le chercheur.

2.3.3. La présentation des tests statistiques comme seul outil pour évaluer des hypothèses scientifiques (l'anti-pluralisme)

Les pratiques usuelles d'enseignement de l'inférence statistique présentent de manière générale les tests statistiques comme étant le seul outil existant pour évaluer des hypothèses scientifiques, inférer des connaissances nouvelles, ou prendre des décisions. La boîte à outils de l'inférence statistique contient plusieurs autres outils très pertinents qui peuvent aider à atteindre ces objectifs. Nous pouvons citer, entre autres exemples : *la statistique bayésienne*[7], *la réplication*, *la taille de l'effet*, *les intervalles de confiances*, … Hormis les tests statistiques et la statistique bayésienne, la plupart de ces outils peuvent être enseignés avec peu d'effort et sont faciles à appréhender. Il est vrai que des enseignants fournissent beaucoup d'efforts pour accompagner, dans leurs cours, les tests statistiques des intervalles de confiance, surtout dans le cas de statistiques paramétriques. Mais cet accompagnement risque de rester insuffisant, voire trompeur et pathogène, si ces enseignants :

- ne présentent pas à leurs étudiants la correspondance qui existe entre les tests d'hypothèses et les intervalles de confiances (i.e., comment peut-on construire à partir de la région critique d'un test d'hypothèses de taille α, un intervalle de confiance $(1-\alpha)$, et vice versa ?)[8],
- interprètent les intervalles de confiance comme étant des intervalles de *crédibilité* selon la statistique bayésienne (Lecoutre, 2006), en déclarant erronément que: « *[a,b] est un intervalle de confiance $(1-\alpha)$ du paramètre θ signifie que la probabilité $P(\theta \in [a,b])$ est égale à $(1-\alpha)$* ».

Aussi, une présentation diversifiée d'outils d'inférence statistique peut permettre aux étudiants non seulement de savoir que l'outil des tests statistiques n'est pas un générateur des probabilités conditionnelles $P(H|\mathfrak{D})$ propres à la statistique bayésienne, mais aussi de pouvoir « *appliquer une étendue variée de stratégies* » et de « *sélectionner et utiliser différents types de raisonnements et de preuves appropriés* ». Cependant, il est important de noter que *la pensée statistique* n'est pas fondée sur *la* méthode statistique que l'on doit appliquer. Bien entendu, toutes les

---

[7] L'enseignement de cette statistique est toujours réduit à la présentation de la fameuse règle de Bayes : $P(A|B) = P(A).\frac{P(B|A)}{P(B)}$, sans rendre compte que derrière cette règle il y a une théorie solide d'inférence statistique, avec un raisonnement propre, des concepts spécifiques, des résultats particuliers.

[8] Pour en savoir plus sur cette correspondance, consulter Dacunha-Castelle et Duflo (1990).

méthodes statistiques déjà mentionnées peuvent donner lieu à des recettes spécifiques appliquées de manière routinière. Or, une pensée statistique solide va au-delà de cette capacité, elle s'intéresse d'abord et avant tout à la justification de la procédure qui sera retenue. Effectuer des calculs statistiques suivant une procédure n'est pas ce que nous comprenons par pensée statistique. Pour ce faire, nous avons des logiciels. La pensée statistique est nécessaire au début - c'est-à-dire, le choix de la méthode inférentielle appropriée - et à la fin - c'est-à-dire, l'interprétation du résultat de tout processus inférentiel statistique.

Cependant, en enseignant des approches inférentielles variées, il est important de ne pas tomber dans le piège du dogmatisme et de ne pas se contenter de remplacer les tests statistiques par une autre méthode. Il n'y a pas de meilleure méthode inférentielle. Gigerenzer (1993) signale qu'il est notre devoir d'informer nos étudiants de l'existence de plusieurs chemins en inférence statistique, et de leur enseigner à utiliser le jugement qui conduit au choix de celui qui convient pour un problème particulier. De son côté, Iseler (1997) propose la métaphore qu'un couteau n'est pas approprié pour manger de la soupe, et qu'une cuillère ne permet pas facilement de couper ; il est heureux que nous ayons les deux, et nous ne devons pas jouer l'un contre l'autre.

Une analogie avec le domaine de l'art plastique peut clarifier la distinction entre connaître l'outil des tests statistiques et comprendre comment l'utiliser (Gall, 1995; Elm'hamedi, 2014). Penser à un pinceau parmi d'autres que le peintre utilise souvent. Au-delà des propriétés physiques (par exemple : la forme, la taille, …), qu'est ce que les peintres ont besoin de savoir à propos de leurs pinceaux ? La réponse est : « *probablement pas beaucoup* ». En effet, ce qui est plus important pour le peintre est de savoir les utilités des différents pinceaux. Par exemple, comment les utiliser pour obtenir des effets de surfaces désirés ou des peintures variées, ou bien savoir quand le pinceau n'est pas le meilleur outil pour exécuter une certaine tâche.

De même, nous pensons que les étudiants, d'une manière générale, doivent savoir qu'un test statistique est un outil mathématique utile dans certaines situations. De la même façon que le savoir-faire en art plastique va au-delà des aptitudes à tout simplement étaler la peinture sur des surfaces, les connaissances de procédures de tests statistiques par les étudiants doivent être situées dans un environnement. Elles nécessitent, en effet, plus que d'énoncer les propriétés mathématiques mobilisées ou d'utiliser une procédure comme une recette, et pour l'enseignant l'objectif ne se réduit pas à former des étudiants capables de traiter n'importe quelle série de nombres sans tenir compte de leur origine ou de leur nature. Comme pour les peintres qui utilisent les pinceaux, les besoins de base pour l'usage futur de tests statistiques sont que les étudiants sachent répondre aux questions suivantes.

- A quoi les tests statistiques sont-ils utiles ?

- Sous quelles conditions l'utilisation d'un test statistique donné est-elle justifiée, et pourquoi ?
- Que peut-il arriver si une procédure de tests statistiques est utilisée là où elle ne devrait pas l'être ?
- Quel est le rôle de l'outil tests statistiques en comparaison des autres outils de la boîte à outils de l'inférence statistique ? Les étudiants doivent savoir dans quelles situations les tests statistiques sont différents des autres outils ou similaires, et reconnaître que le test statistique peut ne pas nécessairement être le premier outil à utiliser, ni le seul outil dont on a besoin, ni le plus approprié.

L'analogie avec le domaine de l'art plastique pourrait ne pas être complète sans considérer qu'en plus des peintres qui créent l'art, il existe des amateurs d'art qui ont besoin de le comprendre. L'appréciation d'un tableau ne dépend pas de la connaissance générale des techniques de l'art plastique et n'oblige pas à savoir comment les peintres essayent de communiquer certaines impressions. De même existe, pour s'informer ou décider, le besoin d'information sur l'application de tests statistiques. Face à des tests statistiques, le lecteur n'a pas nécessairement besoin de savoir comment effectuer une recherche ou comment utiliser des techniques statistiques, mais il a besoin de savoir quoi penser de déclarations qui sont formulées à partir de l'application de tests statistiques.

Il est important de développer les capacités des apprenants à maîtriser les tests statistiques dans chacun des deux contextes, la production de tests et l'interprétation de tests. Malheureusement, bien que l'enseignement de la statistique soit souvent justifié en se référant à la nécessité de préparer les étudiants aux demandes d'une société d'information, les efforts d'enseignement pour satisfaire de telles demandes ne sont pas forcément à la hauteur souhaitable, d'une manière répandue. Bien des enseignants introduisent dans leurs cours des tâches qui ne demandent aux étudiants que d'effectuer des calculs. Ceci peut ne pas suffire à développer leur savoir-faire en statistique. Par exemple, il n'est pas sûr qu'ils soient en mesure de critiquer d'une façon pertinente des résultats de tests statistiques présentés dans des articles de revues, des informations radio-télévisées, des publicités, le monde du travail ou les publications de recherche

2.4. L'usage aveugle des logiciels spécialisés

L'usage aveugle des logiciels spécialisés n'est pas neutre dans les difficultés à appréhender les tests statistiques. En effet, des logiciels comme sphinx, R, Statistica, SPAD, SPSS, etc., destinés pour un large public, embraquent des procédures de tests statistiques de tout type et permettent de les appliquer à ce que l'utilisateur demande. En entrant un tableau de données, le logiciel donne automatiquement et commodément un résultat ; sans pour autant savoir si cela aura une valeur. Leur

usage donne accès à une réponse même si l'on ne connait pas les fondements des tests statistiques que l'on utilise.

2.5. Les facteurs psychanalytiques

Falk et Greenbaum (1995) déclarent que quelqu'un d'intelligence simplement moyenne pourrait facilement comprendre que les deux probabilités conditionnelles P(A|B) et P(B|A) sont différentes. Cependant, la confusion entre le niveau de signification faisant référence à P($\mathfrak{D}$|H) et la probabilité d'une hypothèse égale à P(H|$\mathfrak{D}$), reste une erreur fréquente chez les étudiants et les utilisateurs de statistiques. Ils ont ainsi conclu qu'il existe un mécanisme psychanalytique profond conduisant à commettre cette erreur et à croire, d'une part, qu'on élimine le facteur de chance, et d'autre part, qu'on minimise l'incertitude, lorsqu'on obtient un résultat statistiquement significatif par application d'une procédure de tests statistiques. Ils ont introduit, dans ce sens, la notion « *d'illusion de preuve par probable contradiction* », ou « *l'illusion d'atteindre l'improbabilité* », qui consiste en la croyance erronée que l'hypothèse nulle H_0 est improbable lorsqu'on aboutit à un résultat statistiquement significatif. Ce genre de preuve, se basant éventuellement sur une généralisation trompeuse du raisonnement déductif (Birnbaum, 1982; Lindley, 1993), est difficile à éradiquer dans les esprits, étant couramment présent dans plusieurs ouvrages de statistique (Falk, Greenbaum, 1995).

De plus, Acree (1978), Gigerenzer et *al*. (1989), Gigerenzer (1993) et Batanero (2000) ont tous considéré que, dans l'utilisation des tests statistiques, on a souvent recours à des *rituels*, marqués par une logique hybride, une procédure routinière ou un amalgame, engagés avant toute pensée statistique proprement dite. Ils ont expliqué cette conduite en établissant une analogie entre ce que Freud déclare des conflits inconscients et les comportements individuels dans la situation d'appliquer de tests statistiques. Le développement de cette analogie dépasserait le cadre de cet article, aussi nous nous en tiendrons là à son propos.

3. Etude exploratoire auprès des étudiants

Nous détaillons, dans ce qui suit, les étapes fondamentales de la mise en place de la partie expérimentale de cette étude.

3.1. Objectif de l'étude et expérimentation

Les conclusions dégagées à la lumière de l'analyse précédente concernant les facteurs contribuant aux difficultés qui concernent les tests statistiques, bien qu'elles soient fondées sur la revue de plusieurs travaux didactiques antérieurs, ne peuvent que gagner à être accompagnées d'observations effectives auprès des étudiants. Nous avons donc conduit à la fin du deuxième semestre de l'année scolaire 2017/2018 une expérimentation auprès de 195 étudiants de huit universités du Maroc, poursuivant leurs études en année de licence scientifique de l'une des spécialités: Mathématiques,

Physique, Chimie, Génie électrique, Informatique, Enseignement, et en Economie. Parmi ces sujets, âgés de 22 à 25 ans, 95 sont de sexe féminin et 100 de sexe masculin. Ils ont des baccalauréats de cinq types, à savoir: les Mathématiques, la Physique-Chimie, la Vie et Terre, les Sciences Techniques et l'Economie. A cela s'ajoute le fait que tous avaient déjà suivi au moins une fois un cours sur la probabilité globalement et sur la probabilité conditionnelle d'une manière particulière, étant donné que ce type de cours est indispensable pour passer l'examen du baccalauréat scientifique ou économique. De plus, 57% de ces sujets avaient déjà suivi un cours sur les tests statistiques pendant leurs séjours à l'université.

A ces sujets, nous avons administré un questionnaire de durée une heure, composé de cinq questions. Chaque question traite des significations susceptibles d'être données à un terme ou à des expressions usuellement utilisées dans les tests statistiques. Les quatre premières questions concernent le facteur *vocabulaire* et la cinquième le facteur *ambiguïté linguistique*. Chaque question est composée d'un ensemble d'items à choix multiple (double ou triple). Les items de chaque question tournent autour des différentes significations des termes ou des expressions en relation avec la question. Les modalités de chaque item sont quelques significations possibles (*justes ou fausses*) qu'on pourrait imaginer. Ces modalités sont éventuellement de trois types: juste, induisant implicitement ou explicitement l'erreur commise en tests statistiques, induisant un autre type d'erreur.

Nous avons recommandé aux sujets de lire attentivement chaque item avant de choisir la modalité qui leur semblait la plus convenable, puis de justifier brièvement, sans faire de calcul, chacune des réponses. Enfin, nous leur avons aussi indiqué qu'ils peuvent répondre de *manière indépendante* aux différents items de chaque question.

Il est important d'insister sur le fait que nous n'avons pas communiqué aux sujets l'objectif du questionnaire. D'ailleurs, ce dernier est construit de telle sorte que les termes dont nous demandons la signification et les expressions dont nous mettons en question leurs équivalents ne semblent pas relever du vocabulaire et du langage utilisés en tests statistiques.

3.2. Présentation et analyse *a priori* du questionnaire

Avant de procéder à l'analyse proprement dite des productions des étudiants, nous allons consacrer la prochaine section à l'analyse *a priori* du questionnaire, afin de mieux cerner l'interprétation (statistique) des réponses des étudiants. Pour chaque item des différentes questions, la modalité correcte (*selon la logique habituelle ou le contexte des tests statistiques*) est celle présentée cochée. Vu la nature de l'objectif fixé par la présente recherche, qui vise, en quelque sorte, l'exploration de l'effet impliqué par le *vocabulaire* et *l'ambiguïté linguistique* sur la signification qu'on pourrait éventuellement donner aux concepts utilisés en terme de procédures des tests statistiques, notre analyse se focalisera surtout sur les deux éléments essentiels

suivants: *a)* la justification *scientifique* proprement dite de la fausseté ou de la véracité de chaque modalité, en la liant éventuellement aux conceptions erronées usuelles relativement aux tests statistiques, et ce étant donné que ce présent travail s'adresse aux utilisateurs expérimentés et potentiels de cet outil statistique; *b)* la justification de la formulation par laquelle l'item d'une question donnée est construit et le terme (ou les termes) trompeur utilisé dans cette construction forçant un sujet qui n'a jamais suivi de cours sur les tests statistiques (ou même celui qui en a déjà suivi) à établir une correspondance erronée avec le terme ou l'expression objet de la question.

Question 1 : *Si l'on vous demande d'imaginer ce que vous comprenez par (ou ce que signifie pour vous) l'expression « Hypothèse Nulle », il survient à votre esprit :*

A. Une hypothèse dont sa véracité sera sans valeur et n'aura aucune importance dans la vie sociétale.	**O Oui**	**⊗ Non**
B. Une hypothèse dont sa véracité ne conduira à l'adoption d'aucune action nouvelle.	**O Oui**	**⊗ Non**
C. Une hypothèse prévue pour être falsifiée ou rejetée, afin de gagner un support susceptible de permettre un progrès scientifique.	**⊗ Oui**	**O Non**
D. Une hypothèse déjà connue d'être fausse avant même sa mise à l'épreuve.	**O Oui**	**⊗ Non**
E. Une hypothèse exprimée sous forme d'une égalité ou d'une différence nulle.	**O Oui**	**⊗ Non**

Items de la question 1 et réponses attendues

L'« *hypothèse nulle* » est spécialement réquisitionnée dans la théorie des tests de signification de Fisher. Dans cette théorie, une telle hypothèse est avancée pour être rejetée, afin d'acquérir un support d'acceptabilité de l'hypothèse scientifique (ou de recherche) qui intéresse le chercheur. Une telle hypothèse est souvent considérée par erreur comme une hypothèse dont la véracité ne conduit à l'adoption d'aucune action nouvelle, en la confondant ainsi avec l'hypothèse H_0 de la théorie de Neyman-Pearson, incompatible avec la théorie de Fisher. Notons au passage que l'hypothèse nulle est appelée ainsi, dans le sens de « *to be nullified* »; c'est-à-dire, à réfuter, et non nécessairement, comme on le trouve écrit dans certains manuels, d'une valeur de zéro pour le paramètre testé, même si c'est le cas le plus fréquent (Poitevineau, 1998). En général l'hypothèse nulle sera la négation de l'hypothèse à laquelle le chercheur s'intéresse réellement : « … *the hypothesis that the phenomenon to be demonstrated is in fact absent* » (Fisher, 1990/1935, p. 13).

De plus, avant sa mise à l'épreuve, l'hypothèse nulle devrait normalement être consistante avec l'état actuel de la connaissance scientifique. Malheureusement, plusieurs articles et chercheurs identifient dans leurs recherches l'hypothèse nulle

comme étant une hypothèse « *prête-nom* »[9], c'est-à-dire, une hypothèse qui est connue pour être fausse avant même que des données soient construites – et elle est souvent utilisée de cette façon. Par ailleurs, rejeter une hypothèse nulle connue pour être fausse avant qu'on construise les données ne fournit pas de nouvelle information scientifique : on a simplement démontré qu'une déclaration déjà connue pour fausse l'est effectivement. Mais si l'hypothèse nulle est consistante avec l'état actuel de la connaissance, alors la rejeter nous informera que nous avons trouvé quelque chose de nouveau, permettant à l'hypothèse de recherche (hypothèse nouvelle qui présente de l'intérêt) *d'avoir des chances* d'être acceptée (Granaas, 2002).

Pareillement, bien que le résultat d'un test de signification soit « *rejeter l'hypothèse nulle* » ou « *échouer à rejeter l'hypothèse nulle* », il est possible dans ce dernier cas d'approfondir la recherche pour étudier la possibilité de résistance d'une telle hypothèse, qui restera encore valide jusqu'à nouvel ordre. Par ailleurs, il y a trois critères pour accepter l'hypothèse nulle, à savoir : *(1)* l'hypothèse nulle devrait être possible, *(2)* les résultats devraient être consistants avec l'hypothèse nulle (ou une p-valeur faible), et *(3)* un bon effort devrait être fait pour chercher un effet (c'est-à-dire, pour rejeter l'hypothèse nulle). Selon Frick (1995), dire qu'un bon effort est fourni dans une expérimentation donnée est fonction du degré de satisfaction des trois critères suivants : *(a)* l'adoption d'un plan expérimental adéquat, *(b)* l'obtention d'intervalles de confiances très larges, pour le paramètre testé, et *(c)* l'existence (ou non) d'un effet (lié) dans une expérimentation similaire, utilisant le même plan expérimental que à celle pour laquelle il reste à trouver un effet.

Il est vrai que l'hypothèse alternative H_1 est souvent celle que le *chercheur* (*ou le décideur*) souhaite prouver, mais il ne faut pas voir une absence d'information dans un résultat négatif. Nous savons qu'au rejet de l'hypothèse nulle H_0, impliquant la signification statistique, est attaché un grand prix, parce que ce rejet est vu comme conduisant à la théorie en développement, tandis que l'échec du rejet de H_0, ou même son acceptation, pourraient conduire à une impasse. La signification statistique devrait en revanche être considérée comme une mesure de satisfaction. Il est, par conséquent, plus utile d'attribuer à une recherche (ou une décision) une valeur qui est fonction de la qualité du plan expérimental adopté, dans la mesure où ce dernier réduit autant que possible les fluctuations d'échantillonnages et les erreurs de mesures expérimentales.

On peut remarquer que les différents items de cette question sont volontairement formulés à l'aide d'expressions traduisant une certaine nullité, qui peut être directement liée au terme « nulle », indiqué dans l'expression « hypothèse nulle ». On peut remarquer, par exemple, les expressions : « *sans valeur* », « *aucune importance* », « *ne conduira pas* », « *fausse* », « *égalité* », « *différence nulle* ». Cela

[9] « *Straw person* » en langue anglaise.

permettra d'étudier si ce terme constitue vraiment un obstacle pour donner une signification exacte à l'expression « hypothèse nulle ». Il est vrai que l'expression indiquée dans l'item E est celle la plus habituellement enseignée aux étudiants, mais cela n'empêche pas d'étudier la possibilité d'existence des autres expressions signalées dans les autres items.

Question 2 : *« Prouver la Véracité d'une Hypothèse » signifie pour vous :*

A. Présenter des arguments aussi solides que possible en faveur de la véracité de cette hypothèse, tout en fixant un seuil exigé et réfléchi d'évidence, et en exposant la force de preuve ou la puissance de votre raisonnement.	**⊗ Oui**	**O Non**
B. Présenter des arguments absolument convaincants comme dans le cas d'une démonstration purement mathématique envers la véracité de cette hypothèse.	**O Oui**	**⊗ Non**

Question 3 : *L'expression « Hypothèse Vraie » signifie pour vous :*

A. Une telle hypothèse est un modèle reflétant au mieux le monde réel, permettant à la plupart des gens rationnels de l'accepter comme modèle réalisable.	**⊗ Oui**	**O Non**
B. La véracité d'une telle hypothèse est seulement une affaire de jugement, dans la mesure où un individu peut considérer cette hypothèse vraie, alors que son voisin peut décider le contraire.	**⊗ Oui**	**O Non**
C. On a présenté des arguments absolument convaincants en faveur de la véracité de cette hypothèse, comme dans le cas d'une démonstration purement mathématique, et cette véracité est établie pour toujours.	**O Oui**	**⊗ Non**
D. Une telle hypothèse est un modèle dans lequel on tolère qu'il puisse exister quelques cas pour lesquels cette hypothèse ne tient pas, mais sans pour autant l'étiqueter comme étant non valide dans le monde réel.	**⊗ Oui**	**O Non**

Items des questions 2 et 3 et réponses attendues

Remarquons que l'analyse des deux expressions (prouver et vrai) objets de ces deux questions avait déjà commencé dans l'introduction de ce travail. Néanmoins, nous allons approfondir une telle analyse en présentant sur ce sujet quelques autres éléments de valeur. De plus, contrairement à l'expression « *hypothèse vraie* » qui est explicitement utilisée dans le langage usuel des tests statistiques, l'expression « *Prouver la Véracité d'une Hypothèse* » est implicitement emmagasinée dans l'esprit de l'utilisateur de cet outil qui sollicite les tests statistiques afin de lui permettre de prouver quelque chose. Signalons aussi que c'est volontairement que nous avons situé ces deux questions dans le même panier d'analyse. La raison en est que les deux termes sous-jacents sont étroitement liés, dans la mesure où *la véracité*

d'une hypothèse ne sera établie qu'en s'appuyant sur des éléments de *preuve* convaincants. De plus, les différents items de ces deux questions sont tous formulés avec des termes et des expressions induisant soit un caractère relatif considéré comme valide dans le contexte des tests statistiques et suivant la logique sous-jacente, soit un caractère absolu.

Il est à signaler qu'en termes de procédures de tests statistiques, il est préférable d'utiliser le terme « *modèle* » plutôt qu'hypothèse, si l'on veut souligner qu'en tests statistiques, on compare des modèles et non des hypothèses. En effet, des hypothèses (statistiques) sont des modélisations d'*hypothèses scientifiques* plus abstraites, reflétant le monde réel. Ainsi, accepter, par exemple, l'hypothèse nulle signifie que *le modèle nul* pourrait validement être utilisé. Un autre point que nous souhaitons soulever à propos des tests statistiques est que dans un projet de recherche, un test statistique est utilisé seulement pour *aider à décider* en s'aidant de l'analyse statistique. A cela s'ajoute le fait qu'il s'agit seulement d'un outil de la boîte des statistiques, parmi plusieurs qui peuvent être utilisés. Cela permet de conclure que son usage est limité par définition.

Il est vrai que les données construites supportent toujours le modèle alternatif. Si ce n'était pas le cas, aucun test statistique choisissant automatiquement le modèle nul ne devrait être utile ou nécessaire. Un test de signification, par exemple, fournit une mesure ayant force de preuve contre le modèle nul et en faveur de l'acceptabilité du modèle alternatif. L'indicateur de mesure généralement utilisé, est la *p-valeur* qui est égale à « *la probabilité d'obtenir le résultat de l'échantillon (ou même plus extrême) en supposant que le modèle nul est vrai* ». Cette approche en termes de modèles des tests statistiques est plus longuement discutée par McLean (2000).

Ainsi, l'usage de la terminologie telle que « *la valeur vraie de la moyenne* » conduit à des déclarations comme « *l'hypothèse nulle est presque toujours fausse* ». Bien que ceci puisse être vrai, c'est hors de propos. Un test statistique n'est pas directement concerné par la valeur vraie d'un paramètre, mais par les modèles escomptés. Accepter le modèle nul signifie seulement que l'usage de ce modèle, fondé sur la valeur spécifiée du paramètre, est avantageux (valable); le rejeter signifie qu'il est préférable d'utiliser un modèle fondé sur une valeur différente.

Cependant, juger de la qualité de la preuve pour conclure quant au rejet du modèle nul ou à l'échec de son rejet est une tâche subjective. Quelqu'un pourrait considérer une déclaration comme étant prouvée, alors que son proche voisin penserait le contraire. Si la preuve statistique est suffisamment forte, le lecteur pourra accepter le résultat comme « *prouvé* » dans un sens faible. Cet argument s'applique en particulier aux tests statistiques.

Par ailleurs, dans une situation de recherche, la décision est provisoire dans la mesure où il faut toujours avoir présent à l'esprit que l'acceptation ou le rejet d'une

hypothèse peuvent ne pas durer longtemps. Les recherches qui seront menées dans le futur pourront contredire les conclusions d'aujourd'hui.

Nous pouvons définir *la vérité* et *la preuve* pragmatiquement – une hypothèse ou une théorie est *vraie* si elle est suffisamment fondée sur de solides évidences observées, de sorte que *la plupart des gens raisonnables* l'acceptent comme un modèle. Cette vue pragmatique de la vérité décrit la science. Ainsi la théorie d'évolution peut être acceptée comme *vraie* parce que l'évidence est suffisamment solide et provient de larges sources variées. Pour la plupart des gens, cette acceptation est probablement implicite. Les conclusions de la justice sont de même nature : dire qu'une personne est *prouvée coupable* d'un crime revient simplement à dire que l'évidence factuelle est suffisamment solide pour que *la plupart des gens* (représentés par le jury) acceptent la culpabilité comme modèle de la réalité.

Question 4: *Si l'on vous demande ce que peut signifier pour vous l'expression « Résultat Significatif », vous penserez à un résultat qui:*

A. Est forcément remarquable et important dans la vie sociétale.	**O Oui**	**⊗ Non**
B. Peut ne pas être ni remarquable, ni important dans la vie sociétale.	**⊗ Oui**	**O Non**
C. N'est pas dû au hasard, dans la mesure où il existe une cause bien connue et bien définie qui l'a suscité.	**O Oui**	**⊗ Non**
D. Peut être remarquable et important dans la vie sociétale.	**⊗ Oui**	**O Non**

Items de la question 4 et réponses attendues

L'expression « *résultat significatif* » est très habituellement utilisée dans les procédures de tests statistiques. Elle est utilisée pour signifier qu'un test statistique a permis de rejeter l'hypothèse nulle. Evidemment, cette expression est utilisée à tort pour remplacer l'expression « *résultat statistiquement significatif* », dans la mesure où l'adverbe « *statistiquement* » y est introduit pour mettre en évidence la différence qui pourrait exister entre une « *signification statistique* » fournie par application d'un test statistique, et une « *signification pratique* »[10]. Dans ce sens, Kirk (1996) a défini une telle différence de la façon assez simple suivante : « *Statistical significance is concerned with whether a research result is due to chance or sampling variability; practical significance is concerned with whether the result is useful in the real world*» (Kirk, 1996, p. 746).

Par ailleurs, l'erreur majeure signalée dans l'interprétation des tests statistiques est la supposition naïve qu'un résultat statistiquement significatif est un résultat

[10]On l'appelle aussi: *signification sociale*, *signification théorique*, *signification scientifique*, *signification substantielle*, …

remarquable (Daniel, 1997). Depuis l'année 1931, Tyler avait déjà commencé à reconnaître l'étendue de la mauvaise interprétation de la signification statistique, en déclarant que les interprétations qui ont été généralement formulées à partir des études de l'époque nous contraignent à concevoir la signification statistique comme équivalente à la signification sociale, alors que ces deux termes sont essentiellement différents et ne devraient pas être confondus. Les différences qui sont statistiquement significatives ne sont pas toujours socialement importantes. La réciproque est aussi vraie : les différences qui ne s'avèrent pas être statistiquement significatives peuvent pourtant être socialement significatives.

De manière similaire, Schafer (1993) a noté que la plupart des chercheurs ne comprennent pas que (statistiquement) *significatif* et *important* sont deux choses différentes. Le terme *significatif* lui semble avoir été mal choisi. De plus, Meehl (1997) caractérise l'utilisation du terme *significatif* comme étant « *cancéreux* » et « *trompeur* » et recommande que les chercheurs interprètent leurs résultats en utilisant les intervalles de confiance de préférence aux p-valeurs impliquées par les tests de signification.

Certes, le résultat (existence ou non de la signification statistique) fourni par application d'un test statistique, à savoir: « *Rejeter* H_0 », « *Accepter* H_0 » ou «*Echouer à rejeter* H_0 », possède deux aspects favorisant le *non déterminisme*: D'une part, la région critique RC, dans le cas d'un test d'hypothèses, et la p-valeur, dans le cas d'un test de signification, sont déterminées par un calcul probabiliste, et d'autre part, le statisticien utilise son jugement (la valeur α, prise pour le niveau de signification) pour déterminer un tel résultat.

Par conséquent, l'omission, dans cette question, de l'adverbe *statistiquement* dans l'expression « *résultat statistiquement significatif* » et la formulation des items sous-jacents pourraient conduire non seulement un apprenant qui n'a jamais suivi de cours sur les tests statistiques, mais aussi un utilisateur expérimenté de l'outil statistique, à choisir des modalités erronées des différents items de cette question.

Le tableau 1 qui suit récapitule les éléments essentiels, caractéristiques des différents items relatifs aux questions 1, 2, 3 et 4 du questionnaire.

Question et « *Concept*[11] *impliqué* »	***Terme trompeur du concept***	***Signification*[12] *du concept***	***Items***	***Eléments d'ambiguïté du vocabulaire, impliqués par l'item***
Question1: « *Hypothèse nulle* »	*«Nulle»*	*Hypothèse prévue pour être falsifiée*	A	*sans valeur, aucune importance*
			B	*ne conduira pas, aucune action nouvelle*
			C	-
			D	*connue pour être fausse*
			E	*égalité, différence nulle*
Question 2: « *Prouver la véracité d'une hypothèse* »	*«Prouver»*	*Prouver de manière relative et non absolue*	A	*arguments aussi solides que possible, fixation d'un seuil exigé d'évidence*
			B	*arguments absolument convaincants, démonstration purement mathématique*
Question 3: « *Hypothèse vraie* »	*«Vraie»*	*Véracité à caractère relatif et non absolu*	A	*modèle reflétant au mieux le monde réel, la plupart des gens l'acceptent*
			B	*affaire de jugement de l'individu*
			C	*arguments absolument convaincants, démonstration purement mathématique*
			D	*modèle, on tolère*
Question 4: « *Résultat significatif*[13] »	*«Significatif»*	*Le résultat fourni par un test statistique pourrait être pratiquement*[14] *significatif, comme il se peut qu'il pourrait ne pas l'être*	A	*forcément remarquable et important*
			B	*peut ne pas être ni remarquable, ni important*
			C	*n'est pas dû au hasard, une cause bien définie l'a suscité*
			D	*peut être remarquable et important*

Tableau 1. Récapitulatif des éléments caractéristiques des différents items relatifs aux questions 1, 2, 3 et 4 du questionnaire

[11] Il s'agit du concept impliqué dans une procédure de tests statistiques.

[12] Il s'agit de la signification (*vraie*) du concept dans le vocabulaire usuel adopté par les procédures de tests statistiques.

[13] L'expression « *Résultat significatif* » est souvent utilisée à la place de l'expression « *Résultat statistiquement significatif* ».

[14] L'adverbe « *pratiquement* » peut être remplacé par : *socialement, théoriquement, scientifiquement, substantiellement, …*

Question 5 : *On désigne par D_0 l'ensemble des données issues d'une enquête policière, qu'on a pu construire de manière aussi adéquate que possible, et qui a priori vont dans le sens de la culpabilité du prévenu X. De telles données sont présentées au juge, qui a décidé de la culpabilité de ce prévenu.*

Présentation de la question 5

Les éléments de la question 5 s'inscrivent dans le cadre d'une situation de test d'hypothèses (ou de décision) de Neyman-Pearson. Ainsi, à la lumière des données construites D_0, on a contrasté à travers cette situation les deux hypothèses H_0 et H_1 qui sont respectivement les modélisations des hypothèses juridiques : « *Le prévenu X est innocent* » et « *Le prévenu X est coupable* ». La décision à laquelle on a abouti dans cette situation est : « rejeter H_0 ».

Chaque item de cette question est usuel dans le langage adopté par les tests d'hypothèses. Il met en exergue une erreur usuellement commise par les apprenants dans l'usage de tel outil statistique. Les expressions par lesquelles les items et leurs modalités sont construits impliquent toutes deux événements : $\mathcal{D}$ (relatif aux données) et H (relatif aux hypothèses statistiques). Modalités et items sont conçus de telle sorte qu'il n'est pas facile pour le sujet de connaitre qu'il s'agit de la probabilité d'un événement simple ou d'un événement conditionnel. Si par hasard ces sujets savent qu'il s'agit d'un événement conditionnel, il sera difficile pour eux de choisir, entre $\mathcal{D}$ et H, lequel est l'événement conditionné et lequel est l'événement conditionnant. Ainsi, les différentes difficultés auxquelles on s'est intéressé sont que la pratique de tests d'hypothèses génère de manière erronée les deux types de probabilités conditionnelles suivantes : *les probabilités des hypothèses H sachant les données $\mathcal{D}$ ($P(H \mid \mathcal{D})$)* et *les probabilités de données $\mathcal{D}'$ sachant d'autres données $\mathcal{D}$ ($P(\mathcal{D}' \mid \mathcal{D})$)*. Chacune des probabilités indiquées dans les items et dans les modalités correspondantes de la question 5 est relative à la conjonction ou au conditionnement des deux événements $\mathcal{D}$ et H. Ces derniers sont implicitement ou explicitement mentionnés dans les différentes formulations de tels items et modalités. Ces formulations, habituellement utilisés dans les tests d'hypothèses, sont conçues de façon que les sujets ne puissent pas mettre facilement en œuvre les deux événements indiqués, ainsi que le type de probabilité sous-jacente, à savoir conjointe ou conditionnelle. Notons au passage que le terme « *argumentation* », explicité dans la plupart de ces formulations, n'est pas utilisé dans le sens courant des mathématiques.

A. A votre avis, **« *La probabilité que la décision prise par le juge soit due au seul hasard* »** est la même que :

⊗ **a. « La probabilité que la décision prise par le juge soit produite par un processus de hasard ».**

O **b. « La probabilité d'obtenir par hasard la décision prise par le juge ».**

O **c. « La probabilité qu'un processus, considéré comme étant un processus de hasard, pourrait produire la décision prise par le juge ».**

Item 5A et réponses attendues

Cet item et les modalités qui le composent sont implicitement tous exprimés à l'aide de probabilités conditionnelles ne permettant pas aux sujets de saisir facilement lequel des deux événements conditionnels, $(H_0 \mid \mathcal{D})$ ou $(\mathcal{D} \mid H_0)$, est mis en avant. De telles probabilités (conditionnelles) sont très usuelles dans le langage adopté en théorie des tests statistiques.

Par ailleurs, « *connaitre la probabilité que le résultat soit dû au seul hasard* », équivalente à la connaissance de la probabilité impliquée par cet item, est l'un des rôles d'un test statistique, parmi plusieurs habituellement déclarés aux étudiants lors d'un cours. De plus, le terme *résultat* indiqué dans un tel rôle prend une double signification accentuant l'amalgame des tests statistiques que nous avons indiqué dans l'introduction de ce travail, à savoir « *les données construites* D_0 » et « *la décision de rejet de* H_0 ».

Certes, ce n'est pas pour un tel rôle qu'un test statistique est conçu, du fait que « *la probabilité que le résultat soit dû au seul hasard* » n'est autre que « *la probabilité de réalisation de* H_0 *sachant les données* $\mathcal{D}$ » $(P(H_0 \mid \mathcal{D}))$, que les tests statistiques ne peuvent pas générer. La probabilité conditionnelle indiquée dans la modalité *a* de cet item est équivalente à la probabilité $P(H_0 \mid \mathcal{D})$, tandis que celles mentionnées dans les modalités *b* et *c* sont non seulement équivalentes mais aussi égales à la probabilité $P(\mathcal{D} \mid H_0)$, qu'il est possible de déterminer selon la logique des tests statistiques.

B. Selon vous, l'argumentation: **« *La probabilité que le juge décide correctement la culpabilité du prévenu X* »** est la même que l'argumentation:

⊗ **a. « La probabilité que le juge décide la culpabilité du prévenu X, sachant qu'il est coupable ».**

O **b. « La probabilité que le prévenu X soit coupable et le juge décide sa culpabilité ».**

O **c. « La probabilité que le prévenu X soit coupable, sachant que le juge décide sa culpabilité ».**

Item 5B et réponses attendues

« *La probabilité de rejeter correctement l'hypothèse H_0* » est la définition usuelle donnée par la théorie des tests d'hypothèses au concept de *puissance*. Elle est équivalente dans le cas de la présente situation à « *la probabilité que le juge décide correctement la culpabilité du prévenu X* ». Cette définition désigne vraiment « *la probabilité de rejeter l'hypothèse H_0 sachant qu'elle est fausse* » *($P(\mathcal{D} \mid H_1)$)*. Une telle probabilité est la même que celle indiquée dans la modalité *a* de cet item. Cependant, la façon d'exprimer cet item et les modalités correspondantes met en question une autre fois le problème du choix adéquat d'événements qui sont mis en œuvre dans les différentes probabilités signalées. En effet, la probabilité impliquée dans cette définition semble non conditionnelle, et elle pourrait conduire le sujet à croire à tort que c'est une probabilité conjointe (non conditionnelle) des deux événements: « *l'hypothèse nulle est fausse* » et « *on rejette l'hypothèse nulle* » *($P(\mathcal{D} \cap H_1)$)*, comme dans le cas de la modalité *b*, ou « *la probabilité de réalisation de H_1 sachant qu'on a décidé de l'accepter* » *($P(H_1 \mid \mathcal{D})$)*, comme dans le cas de la modalité *c*. Notons au passage que, d'une manière générale, la probabilité $P(A \mid B)$ est égale à $P(A \cap B)/P(B)$, et que l'égalité entre $P(A \mid B)$ et $P(A \cap B)$ supposerait que B soit un événement certain.

C. Les deux argumentations (1) et (2) suivantes sont-elles équivalentes ?

O Oui ⊗ Non

(1) *« En moyenne, 5 fois parmi toutes les 100 fois que ce juge rejettera l'innocence d'un prévenu, il se trompera ».*

(2) *« En moyenne, 5 fois parmi 100 qu'un prévenu est innocent, ce juge le déclare coupable ».*

Item 5C et réponse attendue

Cet item, déjà étudié par Birnbaum (1982), est une occasion de plus de mettre le sujet à l'épreuve, pour tester si l'ambiguïté linguistique peut le conduire à commettre l'erreur de considérer que les deux probabilités conditionnelles, *$P(\mathcal{D} \mid H_0)$* et *$P(H_0 \mid \mathcal{D})$*, sont équivalentes. Les deux argumentations (1) et (2) sont bien sûr différentes. En effet, la première est valide. Elle exprime l'interprétation donnée au concept de niveau de signification de valeur égale à 5%, et ce dans le cadre *fréquentiste radical* adopté par la théorie des tests d'hypothèses de Neyman-Pearson, sur laquelle nous nous basons dans cette question 5. Elle est équivalente au fait que la probabilité conditionnelle *$P(\mathcal{D} \mid H_0)$* soit égale à 0,05, où $\mathcal{D}$ est l'événement « *rejeter l'innocence d'un prévenu* » (c'est-à-dire, *rejeter H_0*), et H_0 est l'événement: « *un prévenu est innocent* » (c'est-à-dire, « *H_0 est vraie* »). En revanche, la deuxième argumentation est invalide dans le cadre de la théorie des tests d'hypothèses. Elle est, en quelque sorte, équivalente au fait que la probabilité conditionnelle *$P(H_0 \mid \mathcal{D})$* soit égale à 0,05.

D. En se basant sur les données D_0, le juge est convaincu de la culpabilité du prévenu X. Néanmoins, pour pouvoir mieux s'en assurer et afin de prendre une décision définitive, il a mis de côté les données D_0, et il a lancé en parallèle une nouvelle enquête indépendante de la première. Une telle enquête a donné lieu à d'autres données R.

A votre avis, ***« La probabilité que ces nouvelles données R permettent au juge de reprendre la même décision »***, est la même que:

O a. « Le complément à 1 de la probabilité de condamner le prévenu X sachant qu'il est innocent ».

⊗ b. « La probabilité de condamner le prévenu X, sachant qu'on avait déjà décidé qu'il est coupable ».

O c. « La probabilité que le prévenu X soit coupable, sachant qu'on avait déjà décidé qu'il est coupable ».

Item 5D et réponse attendue

Pareillement, cet item et les modalités qui en découlent sont tous exprimés de manière implicite ou explicite à l'aide de probabilités conditionnelles ne permettant pas aux sujets de choisir facilement lequel des trois événements conditionnels, $(H \mid \mathcal{D})$, $(\mathcal{D} \mid H)$ ou $(\mathcal{D}' \mid \mathcal{D})$, est mis en avant. De telles probabilités (conditionnelles) sont usuelles dans le langage des tests statistiques. Par ailleurs, « *connaître la probabilité que le résultat soit répliqué* », équivalente à la connaissance de la probabilité impliquée par cet item, est l'un des plusieurs rôles d'un test statistique habituellement déclarés aux étudiants lors d'un cours. Il est évident que ce n'est pas en vue d'un tel rôle qu'un test statistique est conçu, du fait que « *la probabilité que le résultat soit répliqué* » n'est autre que « *la probabilité de données $\mathcal{D}'$ sachant d'autres données $\mathcal{D}$* » $(P(\mathcal{D}' \mid \mathcal{D}))$, que les tests statistiques ne peuvent pas générer. En effet, toute présentation des résultats d'un test statistique doit être formulée sous la condition « *sachant que H_0 est vraie* », et non « *sachant des données* » (Haller, Krauss, 2002; Zaki, Elm'hamedi, 2009). La probabilité conditionnelle indiquée dans la modalité *b* de cet item est équivalente à la probabilité $P(\mathcal{D}' \mid \mathcal{D})$, tandis que celles mentionnées dans les modalités *a* et *c* sont respectivement équivalentes au complément à 1 de la probabilité $P(\mathcal{D} \mid H_0)$ et à la probabilité $P(H_1 \mid \mathcal{D})$.

E. A votre avis, l'assertion : **« *La probabilité que le juge décide incorrectement la culpabilité du prévenu X* »** est la même que l'assertion :

O a. « La probabilité que le prévenu X soit coupable et que le juge décide son innocence ».

O b. « La probabilité que le prévenu X soit innocent, sachant que le juge décide sa culpabilité ».

⊗ c. « La probabilité que le juge décide la culpabilité du prévenu X, sachant qu'il est innocent ».

Item 5E et réponse attendue

« *La probabilité de rejeter incorrectement l'hypothèse* H_0 » est la définition usuelle donnée dans la théorie des tests d'hypothèses pour le concept de *risque de 1ère espèce*. Elle est équivalente dans le cas de la présente situation à « *la probabilité que le juge décide incorrectement la culpabilité du prévenu X* ». Cette définition désigne vraiment *la probabilité de rejeter l'hypothèse* H_0 *sachant qu'elle est vraie* $(P(\mathcal{D}|H_0))$. Une telle probabilité est la même que celle indiquée dans la modalité *c* de cet item. Cependant, la façon d'exprimer cet item et les modalités correspondantes met en question une autre fois le problème du choix adéquat des événements mis en œuvre dans les différentes probabilités signalées. En effet, la probabilité impliquée dans cette définition semble non conditionnelle et pourrait conduire le sujet à croire à tort que c'est une probabilité conjointe (non conditionnelle) des deux événements : « *l'hypothèse nulle est vraie* » et « *on rejette l'hypothèse nulle* » $(P(\mathcal{D} \cap H_0))$, comme dans le cas de la modalité *a*, ou « *la probabilité de réalisation de* H_0 *sachant qu'on a décidé de la rejeter* » $(P(H_0|\mathcal{D}))$, comme dans le cas de la modalité *b*.

Nous présentons dans le tableau 2 qui suit le récapitulatif des éléments caractéristiques des différents items et modalités relatifs à la question 5 du questionnaire, en définissant ce que signifient les événements $\mathcal{D}$ (ou éventuellement $\mathcal{D}'$) et H, ainsi que les probabilités conditionnelles (ou conjointes) correspondantes. Nous rappelons que le langage (*ambigu*) par lequel les items et leurs modalités sont exprimés, pourrait éventuellement être la cause principale, d'une part, de la méconnaissance, par les étudiants, qu'il s'agit d'une probabilité conditionnelle, et d'autre part, de la confusion non seulement entre l'événement relatif aux données $\mathcal{D}$ et celui relatif aux hypothèses H, mais aussi entre l'événement conditionné et celui conditionnant dans la probabilité conditionnelle impliquée.

Item et *Concept impliqué*		***Evénement* (𝔇) ou (𝔇')**	***Evénement* (H)**	***Evénement conditionnel ou conjoint impliqué***
Item A: *Bayésien indirect*		*On a observé les données D_0* (D_0)	*La décision prise par le juge est due au seul hasard* (H_0)	**($H_0 \mid D_0$)**
Modalités	***a***	*On a observé les données D_0* (D_0)	*La décision prise par le juge est produite par un processus de hasard* (H_0)	**($H_0 \mid D_0$)**
	b	*Les données construites D_0 appartiennent à la région critique RC* (RC)	*Obtenir par hasard la décision prise par le juge* (H_0)	($RC \mid H_0$)
	c		*Un processus, considéré comme étant un processus de hasard, pourrait produire la décision prise par le juge* (H_0)	($RC \mid H_0$)
Item B: *Puissance*		*Le juge décide la culpabilité du prévenu X* (RC)	*Le prévenu X est coupable* (H_1)	**($RC \mid H_1$)**
Modalités	***a***	*Le juge décide la culpabilité du prévenu X* (RC)	*Le prévenu X est coupable* (H_1)	**($RC \mid H_1$)**
	b			($RC \cap H_1$)
	c			($H_1 \mid RC$)
Item C: *Niveau de signification*		*Le juge rejette l'innocence d'un prévenu* (RC)	*Le juge se trompera en rejetant l'innocence d'un prévenu* (H_0)	(1) ($RC \mid H_0$)
		Le juge accuse un prévenu d'être coupable (RC)	*Le prévenu est innocent* (H_0)	(2) ($H_0 \mid RC$)
Modalités	Oui			($RC \mid H_0$) = ($H_0 \mid RC$)

	Non			**$(RC \mid H_0) \neq (H_0 \mid RC)$**
Item D: *Réplicabilité*		*Les nouvelles données construites R permettent au juge de reprendre la même décision qu'il avait déjà prise en construisant les données D_0 (i.e., décider la culpabilité du prévenu X* (RC) *Le juge décide la culpabilité du prévenu X* (D_0)	-	**$(RC \mid D_0)$**
Modalités	a	*Décider que le prévenu X est coupable* (RC)	*Le prévenu X est innocent* (H_0)	$(RC \mid H_0)^c$
	b	*Décider que le prévenu X est coupable* (RC) *On avait déjà décidé que le prévenu X est coupable* (D_0)	-	**$(RC \mid D_0)$**
	c	*On avait déjà décidé que le prévenu X est coupable* (D_0)	*Le prévenu X est coupable* (H_1)	$(H_1 \mid D_0)$
Item E: *Risque de 1ère espèce*		*Le juge décide la culpabilité du prévenu X* (RC)	*Le prévenu X est innocent* (H_0)	**$(RC \mid H_0)$**
Modalités	a	*Le juge décide l'innocence du prévenu X* (RC)	*Le prévenu X est coupable* (H_1)	$(H_1 \cap RA)$
	b		*Le prévenu X est innocent* (H_0)	$(H_0 \mid RC)$
	c	*Le juge décide la culpabilité du prévenu X* (RC)		**$(RC \mid H_0)$**

Tableau 2. Récapitulatif des éléments caractéristiques des différents items et de leurs modalités relatifs à la question 5 du questionnaire

3.3. Codage et outil d'analyse retenus pour le traitement des réponses des étudiants

Pour le traitement des réponses des étudiants, l'analyse factorielle des correspondances multiples (AFCM), nous a semblé être un outil efficace, puisqu'il va permettre d'analyser et d'interpréter les réponses des étudiants en termes de liens qui existent entre les différentes modalités des différents items du questionnaire.

Dans les tableaux 3 et 4 suivants, nous résumons, les différents questions et items du questionnaire, ainsi que les significations et les codages des modalités associées. Ces deux tableaux sont, en quelque sorte, un complément de l'analyse *a priori* dont nous avons commencé à présenter l'essentiel dans la section 3.2. de cet article.

Question	***Items***	***Modalité correcte***	***Codes***[15]
Question1	A	Non	1Ar/1Ae
	B	Non	1Br/1Be
	C	Oui	1Cr/1Ce
	D	Non	1Dr/1De
	E	Non	1Er/1Ee
Question 2	A	Oui	2Ar/2Ae
	B	Non	2Br/2Be

Item	***Probabilités conditionnelles impliquées***	***Modalités***[16]	***Probabilités impliquées***[17]	***Codes***[18]
Item A	$P(H_0 \mid D_0)$	***a***	$P(H_0 \mid D_0)$	5Ar
		b	$P(RC \mid H_0)$	5Ab
		c	$P(RC \mid H_0)$	5Ac
Item B	$P(RC \mid H_1)$	***a***	$P(RC \mid H_1)$	5Br
		b	$P(RC \cap H_1)$	5Bb
		c	$P(H_1 \mid RC)$	5Bc
Item C	$P(H_0 \mid RC)$ *et* $P(RC \mid H_0)$	Oui	$P(H_0 \mid RC) = P(RC \mid H_0)$	5Ce

[15]Le codage adopté pour les modalités est de la forme nXy, où n est le *numéro de la question de la modalité,* X la *lettre désignant l'item de la modalité,* et y la *lettre désignant la modalité de l'item X appartenant à la question.* Les modalités de types nXr ou nXe (c'est-à-dire, y= r ou y= e) désignent respectivement la réussite ou l'échec à l'item X de la question numéro n. Ce dernier codage est utilisé pour les items en Oui/Non. De même, le codage nXr (c'est-à-dire, y= r) est aussi utilisé pour les modalités des items autres que ceux en Oui/Non et ce pour désigner qu'une telle modalité est celle qui est considérée correcte.

[16]Les lettres écrites en gras indiquent que les modalités correspondantes sont considérées comme correctes.

[17] Chacune des probabilités indiquées dans les items et dans les modalités correspondantes de la question 5 est relative soit à la conjonction ou au conditionnement des deux événements $\mathfrak{D}$ et H. Ces derniers sont implicitement ou explicitement mentionnés dans les différentes formulations de tels items et modalités. Ces formulations sont conçues de façon que les sujets ne puissent pas mettre facilement en évidence les deux événements indiqués, ainsi que le type de probabilité sous-jacentes, à savoir conjointe ou conditionnelle.

[18] C'est le même codage que celui signalé dans le tableau 3, pour les modalités des différents items des questions 1, 2, 3 et 4.

Question 3	A	Oui	3Ar/3Ae
	B	Oui	3Br/3Be
	C	Non	3Cr/3Ce
	D	Oui	3Dr/3De
Question 4	A	Non	4Ar/4Ae
	B	Oui	4Br/4Be
	C	Non	4Cr/4Ce
	D	Oui	4Dr/4De

Tableau 3. Codage des modalités des items des questions 1, 2, 3, 4

		Non	$P(H_0 \mid RC) \neq P(RC \mid H_0)$	5Cr
Item D	$P(RC \mid D_0)$	a	$1- P(RC \mid H_0)$	5Da
		b	$P(RC \mid D_0)$	5Dr
		c	$P(H_1 \mid D_0)$	5Dc
Item E	$P(RC \mid H_0)$	a	$P(H_1 \cap RA)$	5Ea
		b	$P(H_0 \mid RC)$	5Eb
		c	$P(RC \mid H_0)$	5Er

Tableau 4. Codage des modalités des items de la question 5

préta

Rappelons que nous avons choisi l'AFCM pour analyser les données de cette étude. Nous présentons, dans ce qui suit, les éléments nécessaires à l'application de cette méthode multi-dimensionnelle ainsi que les résultats d'analyse auxquels nous sommes arrivés à travers son application.

3.4.1. Valeurs propres et inertie totale

L'AFCM[19] appliquée au tableau disjonctif complet issu du codage des réponses des étudiants, a conduit à des valeurs propres non nulles de moyenne 0,05 (l'inverse du nombre d'items qui est égal à 20). Par ailleurs, les valeurs propres supérieures à cette moyenne sont présentées dans le tableau 5 suivant:

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}
0,111	0,100	0,092	0,086	0,075	0,072	0,066	0,063	0,058	0,053	0,052

Tableau 5. Valeurs propres

L'inertie totale est égale à 1,20 (nombre de valeurs propres non nulles, divisé par le nombre d'items).

3.4.2. Nombre d'axes retenus dans l'analyse

Nous rappelons tout d'abord que l'inertie totale en AFCM appliquée à un tableau disjonctif complet n'a pas de signification statistique[20]; elle ne dépend pas des observations, elle dépend uniquement du nombre de variables et du nombre de modalités impliquées. Par ailleurs, le calcul des taux d'inertie liés aux onze valeurs propres ci-dessus conduit à des pourcentages décroissant de 9,25% à pratiquement 4,33%, ce qui donne une idée assez pessimiste des parts d'informations obtenues par l'analyse factorielle. Nous avons donc eu recours à la formule proposée par Benzécri

[19] Cette analyse a été conduite à l'aide du logiciel Statistica (Version 6).

[20] Ce n'est pas le cas pour le calcul d'inertie en analyse de correspondances simples d'un tableau de Burt.

(1979), qui dans une telle situation, va permettre une meilleure appréciation des taux d'inertie :

Inertie corrigée : $\left(\frac{s}{s-1}\right)^2\left(\lambda_k-\frac{1}{s}\right)^2$ pour $\lambda_k > \frac{1}{s}$, où s est le nombre d'items du questionnaire et λ_k les valeurs propres obtenues par l'analyse du tableau disjonctif complet.

L'application de cette formule aux inerties initiales[21], donne pour les valeurs propres supérieures à 0,05, les résultats présentés dans le tableau 6 qui suit.

Valeurs propres	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
Inerties corrigées	0,004118	0,002796	0,001934	0,001421	0,000682	0,000556

Valeurs propres	λ_7	λ_8	λ_9	λ_{10}	λ_{11}
Inerties corrigées	0,000281	0,000186	0,000072	0,000011	0,000004

Tableau 6. Inerties corrigées associées aux valeurs propres

Le tableau 7 suivant fournit les pourcentages d'inerties corrigées et leurs cumuls, correspondant aux valeurs propres supérieures à la valeur moyenne $\frac{1}{20}$:

Valeurs propres	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}
% Inerties corrigées	34	23	16	12	6	5	2	2	1	≈ 0	≈ 0
Cumul % Inerties corrigées	34	57	73	85	91	96	98	≈ 100	≈ 100	≈ 100	≈ 100

Tableau 7. Pourcentages d'inerties corrigées et leurs cumuls

Par conséquent, le premier axe représente 34% de l'information totale, le deuxième axe 23%, le troisième axe 16% et le quatrième axe 12% (les axes suivants chutent à une part d'information représentant moins de la moitié du quatrième axe). Ces valeurs indiquent que l'essentiel de l'information est lié aux quatre premiers axes factoriels qui représentent ensemble 85% de l'information totale.

Par ailleurs, les deux premières valeurs propre λ_1=0,111 et λ_2=0,100 sont très proches ; cela signifie que nous sommes pratiquement en présence d'une valeur propre double. En revanche, λ_3=0,092 est significativement supérieure à λ_4=0,086. Par conséquent, nous allons étudier, d'une part, le plan (1,2), et d'autre part, les axes 2 et 3, plutôt que de chercher à donner séparément une interprétation à chacun des 3 premiers axes. Avant d'entamer une telle étude, nous allons consacrer la prochaine section à la présentation de quelques éléments d'analyse globale que nous avons pu déduire en mettant en œuvre les différents plans et axes factoriels que nous avons pu retenir.

[21] Dans la suite de l'analyse, nous nous limiterons simplement aux 11 premières valeurs propres.

3.4.3. Eléments d'analyse globale

En premier lieu en nous basant sur des éléments élémentaires de la statistique descriptive, nous pouvons signaler que *le vocabulaire et le langage* adoptés par les procédures de tests statistiques sont nettement ouverts à plusieurs interprétations (fallacieuses) par nos sujets. Ceci est déduit des pourcentages élevés (compris entre 27% et 85%) des réponses erronées aux différents items du questionnaire. Nous rappelons au passage que ces réponses erronées sont repérées à la fois chez les sujets qui avaient déjà suivi un cours sur les tests statistiques et chez ceux qui n'en avaient pas suivi. Ces pourcentages sont faibles tant pour les items relatifs à *l'ambigüité linguistique* que pour les items concernant *le vocabulaire,* excepté l'item 5C relatif à la distinction, *dans un cadre fréquentiste*, entre la probabilité $P(\mathcal{D}|H)$ qui désigne le concept de niveau de signification, et la probabilité inverse $P(H|\mathcal{D})$, pour lequel ce pourcentage a pu atteindre les 56%.

De plus, les items du questionnaire font référence à une exploration conceptuelle à propos des tests statistiques, et non pas aux performances des étudiants en situation de résolution de problèmes sur ce sujet. Il est donc normal que la plus grande part de l'information (1er plan factoriel) contenue dans les réponses des étudiants ne se traduise pas en termes de performances des étudiants, autrement dit en termes de réussites-échecs. Notons aussi au passage, que dans les oppositions de groupes de modalités à fortes contributions relatives au plan et aux axes factoriels retenus pour l'analyse, nous n'avons pas une opposition nette de modalités contraires. Ce qui confère au questionnaire une hétérogénéité de degré plus ou moins moyen.

Cette hétérogénéité peut être expliquée, d'un côté, par la nature de notre questionnaire qui est formé non seulement d'items en bi-modalités mais aussi en tri-modalités, et d'un autre côté, par l'hétérogénéité des sujets (le public cible) questionnés, dans la mesure où il existe parmi eux ceux qui avaient déjà suivi un cours sur les tests statistiques et ceux qui n'en avaient jamais suivi. Une telle hétérogénéité peut aussi être due à la façon selon laquelle les différents items sont construits, qui ne permet pas aux sujets de facilement reconnaître qu'ils sont face à un questionnaire (homogène) qui tourne autour d'un objectif unique et bien défini, à savoir dans notre cas, l'exploration de l'effet que pourraient avoir le vocabulaire et l'ambiguïté linguistique sur la compréhension de la procédure des tests statistiques.

Sans doute, une telle hétérogénéité est l'une des causes principales du caractère non discriminant[22] des items 1D, 1E et 5E, dont les modalités n'ont que de très faibles contributions relatives aux quatre axes factoriels retenus. Du caractère non discriminant de 1D et 1E, on ne peut pas extraire beaucoup d'information, sinon qu'il

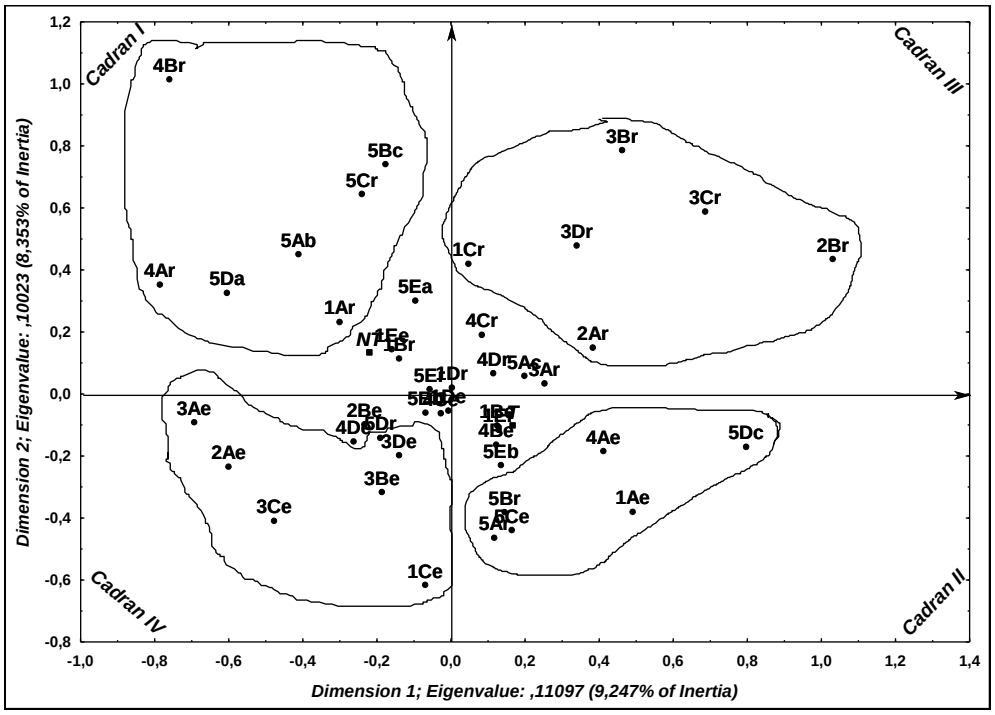
[22] Cela revient à dire que leur présence ou leur absence n'ont pas d'influence sur les réponses des sujets au questionnaire tout entier. Par conséquent l'analyse du questionnaire ne changera pas si on ne les considère pas.

n'y a pas de fortes corrélations entre les réponses individuelles aux différents items du questionnaire et à chacun de ces deux items. Au contraire, celui de l'item 5E conduit à conclure que le langage habituellement utilisé a vraiment un effet négatif sur l'interprétation des concepts et des expressions utilisés relativement aux tests statistiques. En effet, les expressions utilisées pour formuler les items 5B et 5E ne diffèrent que par le fait que le terme *correctement* a été remplacé par le terme *incorrectement*. Contrairement à l'item 5E, l'item 5B est tout à fait discriminant (voir modalités à fortes contributions relatives dans les graphiques 1, 2 et 3). Par ailleurs, les probabilités conditionnelles impliquées par les items 5B et 5E sont respectivement $P(RC \mid H_1)$ et $P(RC \mid H_0)$. Cela signifie que l'échange des deux termes *correctement* et *incorrectement* fait passer d'une probabilité conditionnelle à H_1 à une probabilité conditionnelle à H_0. Mais ce passage n'est pas effectué par les étudiants qui n'ont pas choisi en même temps la modalité (correcte ou fausse) pour l'item 5B et la modalité correspondante (correcte ou fausse) pour l'item 5E. En effet, dans les graphiques 1, 2 et 3, en portant l'attention sur les proximités des modalités de l'item 5B de celles de l'item 5E (sans tenir compte de l'ordre de grandeur de leurs contributions relatives à la construction des axes factoriels retenus), on peut remarquer que les modalités 5Br, 5Ba et 5Bc sont très éloignées respectivement des modalités 5Er, 5Ea et 5Eb. Ceci explique une fois de plus l'effet que pourrait avoir l'ambiguïté linguistique sur la compréhension des tests statistiques.

De même, les items 5C et 5E impliquent tous les deux le même concept, à savoir le niveau de signification (ou le risque de 1ère espèce), identifié par la probabilité conditionnelle $P(RC \mid H_0)$. Ceci devrait être concrétisé par des proximités, dans les graphiques 1, 2 et 3, entre les deux modalités 5Ce et 5Eb et entre les deux modalités 5Cr et 5Er. Mais c'est le constat que les modalités 5Cr et 5Ce sont respectivement proches des modalités 5Ea et 5Eb qui retient l'attention. Une explication acceptable tient au langage dans lesquels les deux items et leurs modalités sont exprimés: *fréquentiste* pour l'item 5C et *probabiliste* pour l'item 5E. De plus, cette proximité remarquée entre 5Cr et 5Ea nous rend conscients de l'existence d'une difficulté conceptuelle très importante dans notre public cible, en raison de l'existence de sujets qui comprennent bien que les probabilités $P(RC \mid H_0)$ et $P(H_0 \mid RC)$ soient différentes, mais qui en même temps confondent $P(RC \mid H_0)$ avec $P(H_1 \cap RA)$. Ceci explique à nouveau l'effet que pourrait avoir l'ambiguïté linguistique sur l'interprétation des expressions usuelles relativement aux tests statistiques.

3.4.4. Interprétation du plan factoriel (1, 2)

L'AFCM conduit à des inerties de 34% pour l'axe 1 et de 23% pour l'axe 2. Le graphique 1 suivant représente la projection de l'ensemble des modalités par rapport au plan (1,2):



Graphique 1. *Plan factoriel (1, 2)*

Les modalités correspondant aux plus fortes contributions relatives (comprises entre 3,3% et 9,6%) dans la construction des axes 1 et 2, et ayant une bonne représentation, sont celles qui sont entourées dans le graphique 1. Ces modalités, distribuées en quatre regroupements (voir cadrans de I à IV, graphique 1), donnent lieu, dans le plan factoriel (1,2), à deux oppositions : *Cadran I vs Cadran II* et *Cadran III vs Cadran IV*. Ainsi, avant d'attribuer des significations à ces deux oppositions, qui aideront à mettre en place une interprétation convenable de ce 1er plan factoriel qui contient une part très importante d'information des réponses des étudiants, nous allons étudier quelques proximités remarquées, dans le graphique 1, entre les différentes réussites et les différents échecs aux items de notre questionnaire.

Les différents items des questions 2, 3 et 4 sont soit de caractère *relatif*, soit de caractère *absolu*. Bien qu'ils soient formulés de façons différentes, chacun d'entre eux est évidemment associé à l'un de ces deux caractères. Ces items reflètent, en quelques sortes, les interprétations variées (*vraies* ou *fausses*) usuellement données par les étudiants, et même par les professionnels, aux différents concepts habituellement utilisés relativement aux tests statistiques. Ces concepts sont mobilisés par les questions sous-jacentes, à savoir : « *Prouver la véracité d'une hypothèse* » pour la question 2, « *Hypothèse vraie* » pour la question 3 et « *Résultat significatif* » pour la question 4. On en déduit alors que non seulement les réussites

aux différents items composant ces trois questions devraient être proches sur le graphique 1, mais aussi que les échecs correspondants devraient être proches. Un tel constat est satisfait pour la plupart des items, sauf pour les modalités (4Ar, 4Br) et (4Ae, 4Be), qui forment chacune un groupe, et qui se trouvent respectivement éloignées des autres réussites et des autres échecs aux items des trois questions susmentionnées. Ceci est sans doute dû au vocabulaire trompeur, utilisé en tests statistiques, à savoir l'expression « *le résultat est significatif* » qui figure dans la question 4 du questionnaire. En effet, une telle expression est habituellement utilisée pour indiquer que le résultat fourni par le test statistique est *le rejet de l'hypothèse* H_0, ou *l'acceptation de l'hypothèse* H_1 dans le cas des tests d'hypothèses. Cette dernière expression est, en quelque sorte, équivalente à l'expression « *l'hypothèse* H_1 *est vraie* », figurant dans la question 3. Cependant, contrairement aux items de cette dernière question, les réussites (ou les échecs) aux items de la question 4 ne sont pas tous voisins des réussites (ou des échecs) aux différents items des questions 2, 3 et 4.

De même, les proximités observées dans le graphique 1, entre 5Ab, 5Bc, 5Cr et 5Da, d'un côté, et entre 5Ar, 5Br, 5Ce et 5Dc, de l'autre, montrent de manière nette l'existence de sujets qui distinguent entre les deux probabilités $P(\mathcal{D}|H)$ et $P(H|\mathcal{D})$ (5Ar, 5Br et 5Cr) et qui en même temps commettent l'erreur induisant non seulement à considérer que ces deux probabilités sont égales (5Ab, 5Bc et 5Ce) mais aussi que la probabilité $P(\mathcal{D}|H)$ est égale au complémentaire de la probabilité $P(\mathcal{D}'|\mathcal{D})$ (5Da et 5Dc). Ceci est sans doute dû au langage par lequel les items et les modalités de la question 5 sont exprimés, dans la mesure où les autres facteurs susceptibles d'induire de telles difficultés, surtout ceux à caractère psychanalytique (cf. § 2.5), pourraient être exclus ici, à cause de l'hétérogénéité de nos sujets, d'une part, et de leur ignorance de l'ignorance de l'orientation du questionnaire vers les tests statistiques, d'autre part. Ceci explique une autre fois l'effet possible de *l'ambiguïté linguistique* sur l'interprétation des concepts utilisés dans les tests statistiques. Le positionnement des deux groupes de modalités (5Ab, 5Bc, 5Cr et 5Da) et (5Ar, 5Br, 5Ce et 5Dc) renforce une telle conjecture, dans la mesure où ils impliquent tous deux les deux difficultés déjà indiquées, alors même qu'ils se placent dans des cadrans *opposés* (cadran I et cadran II) du graphique1.

Après cette étude succincte des proximités observées entre les différentes modalités dans ce 1er plan factoriel, nous consacrons le paragraphe suivant à l'extraction d'informations supplémentaires à partir des réponses des sujets, pour enfin attribuer à ce plan une signification qui lui convient. Rappelons qu'il y a quatre groupes de modalités qui ont de fortes contributions relatives à la construction de ce 1er plan factoriel et qui s'opposent deux à deux[23]: d'une part, 1Ar, 4Ar, 4Br, 5Ab, 5Bc, 5Cr

[23] Voir Cadran I, Cadran II, Cadran III et Cadran IV du graphique 1.

et 5Da contre 1Ae, 4Ae, 5Ar, 5Br, 5Cr et 5Dc, et d'autre part, 1Cr, 2Ar, 2Br, 3Br, 3Cr et 3Dr contre 1Ce, 2Ae, 3Ae, 3Be, 3Ce et 3De.

- *Cadran I vs Cadran II:*

Dans ces deux cadrans, nous sommes en présence de l'opposition : 1Ar, 4Ar, 4Br, 5Ab, 5Bc, 5Cr et 5Da *vs* 1Ae, 4Ae, 5Ar, 5Br, 5Cr et 5Dc. Ainsi, comme nous l'avons déjà indiqué dans les paragraphe précédents, les deux groupes de modalités (5Ab, 5Bc, 5Cr et 5Da) et (5Ar, 5Br, 5Cr et 5Dc), qui figurent respectivement dans le cadran I et le cadran II, impliquent tous les deux l'existence de sujets qui distinguent entre les deux probabilités $P(\mathcal{D}|H)$ et $P(H|\mathcal{D})$ dans des situations, mais qui commettant non seulement l'erreur induisant que ces deux probabilités soient égales dans d'autres situations similaires, mais aussi l'erreur consistant à avancer que la probabilité $P(\mathcal{D}|H)$ soit égale au complémentaire de la probabilité $P(\mathcal{D}'|\mathcal{D})$. De plus, la modalité 4Br est une conséquence logique de la modalité 4Ar, dans la mesure où il est rare de trouver un sujet caractérisé par 4Ar et 4Be à la fois, du fait que l'item 4B est, en quelques sortes, une négation de l'item 4A. Nous traduisons ceci par le fait que ces deux cadrans opposent la réussite à l'échec au fait que l'expression r*ésultat significatif* désigne vraiment un résultat à caractère *relatif* et non *absolu*. Enfin, nous remarquons aussi que les deux modalités 1Ar et son opposée 1Ae se trouvent dans ces mêmes cadrans. Pareillement, cela signifie que ces deux cadrans, I et II, opposent aussi la réussite à l'échec au fait que le terme *nulle* indiqué dans l'expression *hypothèse nulle* ne signifie en aucun cas « *sans valeur* ».

- *Cadran III vs Cadran IV:*

De même, dans ces deux cadrans, nous sommes en présence de l'opposition (réussites-échecs) de 1Cr, 2Ar, 2Br, 3Br, 3Cr et 3Dr *vs* 1Ce, 2Ae, 3Ae, 3Be, 3Ce et 3De. Contrairement aux deux cadrans, I et II, dans lesquels un mélange de modalités relatives au *vocabulaire* et à *l'ambiguïté linguistique* impliqués par les tests statistiques, est mis en jeu, les deux cadrans III et IV sont concernés par le vocabulaire *seul*.

La formulation d'un item donné de la question 3 explicite soit le caractère *absolu* (*erroné*) de la véracité d'une hypothèse comme dans le cas de l'item 3C, soit le caractère *relatif* (*correct*) de cette véracité comme dans le cas des autres items : 3A, 3B, 3D et 3E. Ce caractère relatif est exprimé dans ces items par différents aspects qui tous l'impliquent. Ces aspects sont soit *intrinsèques* à l'hypothèse elle-même comme dans le cas de l'item 3D, soit liés à *la personne* qui décide de cette véracité tout en utilisant son *jugement argumenté*, son *raisonnement rationnel*, …, comme dans le cas des items 3A et 3B. On en déduit alors que la *réussite* (*respectivement l'échec*) à n'importe quel item de cette question 3 implique le caractère *relatif* (*respectivement absolu*) de la notion de *véracité d'une hypothèse*. De même, alors que la modalité 2Be ne se trouve pas parmi le groupe de modalités entourées dans le

cadran IV du graphique 1, sa contribution relative à la construction du 1er plan factoriel n'est pas faible comparée aux contributions relatives des autres modalités, comme 2Ae. Nous pouvons donc dire que ces deux cadrans opposent les deux groupes de modalités (2Ar et 2Br) et (2Ae et 2Be) et nous interprétons ceci comme le fait que le *raisonnement adopté*[24] pour montrer la véracité d'une hypothèse est à caractère *relatif* et non *absolu*[25].

Enfin, remarquons aussi que chacune des modalités 1Cr et son opposée 1Ce est nettement située dans un cadran, avec une forte composante sur l'axe 2. Cela signifie que les deux cadrans III et IV opposent aussi la réussite à l'échec à la définition *correcte*, donnée par *Fisher*, au concept *d'hypothèse nulle*, à savoir : « *C'est une hypothèse prévue pour être falsifiée ou rejetée, afin de gagner un support susceptible de permettre un progrès scientifique* ».

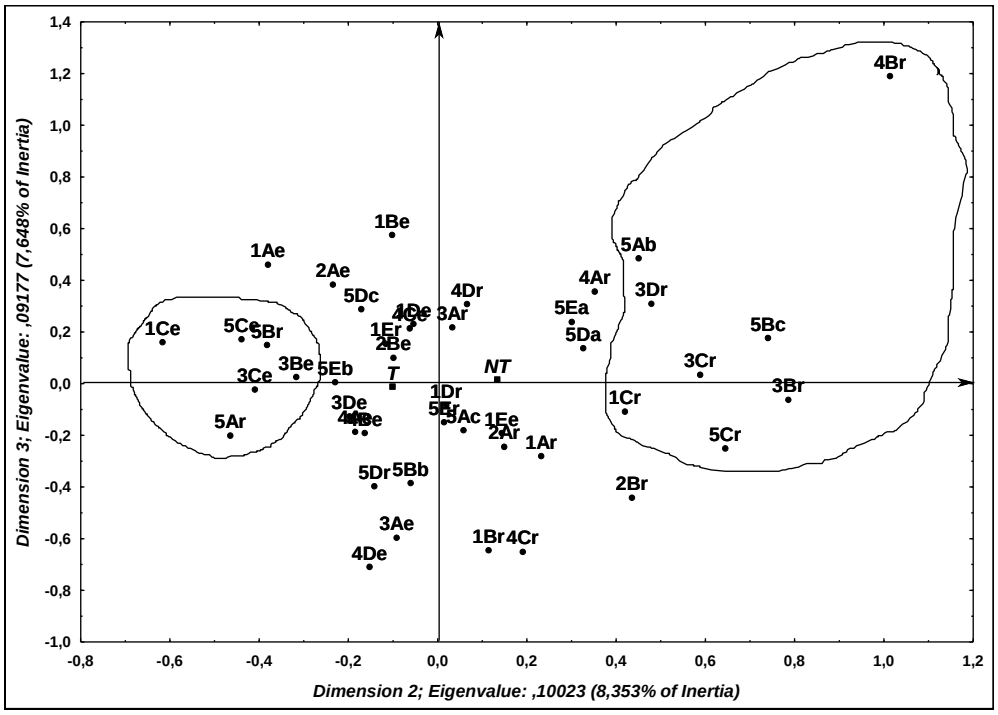
En conclusion, le plan (1,2) est *un plan d'ambiguïté linguistique et de réussite-échec en vocabulaire utilisé*, qui se traduit par :

- Une distinction mal assimilée entre les trois probabilités $P(\mathcal{D}|H)$, $P(H|\mathcal{D})$ et $P(\mathcal{D}'|\mathcal{D})$, associée à une réussite-échec au fait que l'expression « *résultat significatif* » désigne un résultat à caractère *relatif* et non *absolu*, et au fait que le qualificatif *nulle* dans l'expression « *hypothèse nulle* » est tout à fait différent du qualificatif « *sans valeur* » *(Cadran I vs Cadran II)*;
- Une réussite ou un échec au fait que la notion de « *véracité d'une hypothèse* » et celle de *raisonnement* adopté pour montrer une telle véracité sont toutes les deux à caractère *relatif* et non *absolu*, est associée à une réussite ou un échec à la définition correcte (*donnée par Fisher*) au concept d*'hypothèse nulle (Cadran III vs Cadran IV).*

3.4.5. Interprétation du 2ème axe factoriel

L'AFCM conduit à une inertie de 23% pour l'axe 2. Les modalités qui contribuent le plus à la construction de cet axe sont celles entourées dans le graphique 2 ci-après[26]. Elles ont une contribution comprise entre 3,3% et 8,9%. La qualité de représentation de ces modalités par rapport à cet axe est comprise entre 0,095 et 0,290.

[24] Selon la logique des tests statistiques.
[25] Ibid.
[26] Notons au passage que ces modalités occupent des positions correspondant aux modalités ayant les plus grandes coordonnées en valeur absolue par rapport au 2ème axe factoriel. Les autres modalités occupent des positions quasiment intermédiaires, avec en outre de faibles contributions relatives.



Graphique 2. *Plan factoriel (2, 3)*

Dans le deuxième plan factoriel, le 2ème axe factoriel présente l'opposition suivante: 1Ce, 3Be, 3Ce, 5Ar, 5Br et 5Ce *vs* 1Cr, 3Br, 3Cr, 3Dr, 4Br, 5Ab, 5Bc et 5Cr.

Cet axe oppose donc une absence de relativité conférée à la véracité d'une hypothèse (3Be, 3Ce), à la présence d'une telle relativité sous ses différents aspects, à savoir: celui *intrinsèque* à l'hypothèse (3Br), celui lié à *la personne* qui utilise son jugement argumenté pour pouvoir décider de la véracité (*ou non*) (3Dr), et enfin celui relatif au pouvoir d'être vigilant à conclure sur cette véracité, bien qu'elle soit exprimée par des termes absolus habituellement utilisés, comme par exemple: « *résultat significatif* » (4Br). Notons au passage que la présence de la modalité 3Cr du côté des modalités 3Br, 3Dr et 4Br n'est qu'un support qui consolide notre interprétation concernant la relativité de l'hypothèse.

De plus, l'axe 2 oppose les deux groupes de modalités (5Ar, 5Br et 5Ce) et (5Ab, 5Bc et 5Cr). Chaque modalité d'un groupe donné trouve son opposé dans l'autre groupe. Le mélange de modalités, de réussites et d'échecs, composant chacun des deux groupes nous a permis de conclure à l'effet que peut avoir *l'ambiguïté linguistique* sur l'interprétation des concepts utilisés en tests statistiques. En effet,

pour le groupe de modalités[27]: (5Ar, 5Br et 5Ce), il est vrai que la réussite à l'item 5A permet d'emblée de supposer que le langage, impliquant une approche *bayésienne indirecte*, habituellement utilisé pour définir le rôle pour lequel un test statistique est conçu, à savoir: déterminer « *la probabilité que la décision prise par le juge soit due au seul hasard* », n'a pas d'effet sur la confusion, par les sujets, des probabilités: $P(H \mid \mathcal{D})$ et $P(\mathcal{D} \mid H)$ mobilisées par les différentes modalités de cet item. Il est vrai aussi que la réussite à l'item 5B permet de supposer que le langage habituellement utilisé pour définir la puissance statistique : « *la probabilité que le juge décide correctement la culpabilité du prévenu X* », n'a pas d'effet sur la confusion des probabilités $P(H \mid \mathcal{D})$, $P(\mathcal{D} \mid H)$ et $P(\mathcal{D} \cap H)$ mobilisées dans les différentes modalités de cet item. Cependant, l'échec à l'item 5C montre que cette supposition n'est pas valide dans tous les contextes. En effet, en exprimant, dans *un cadre fréquentiste*, les deux probabilités $P(H \mid \mathcal{D})$ et $P(\mathcal{D} \mid H)$, tout en utilisant la probabilité $P(\mathcal{D} \mid H)$ pour déterminer un *niveau de signification*, les sujets qui paraissaient avoir distingué la nuance qui pourrait exister entre ces probabilités conditionnelles, en ayant auparavant réussi à répondre aux items 5A et 5B, commettent encore une telle confusion. Ceci fait à nouveau apparaître l'effet du langage utilisé sur l'interprétation des concepts et expressions mobilisés par un test statistique.

Enfin, ce 2ème axe confirme encore l'opposition de la réussite (1Cr) à l'échec (1Ce) sur la définition *correcte* (donnée par *Fisher*) de ce que signifie le concept « *d'hypothèse nulle* ».

En conclusion, les tendances expliquées par le 2ème axe factoriel renvoient à une *présence-absence* de relativité du vocabulaire utilisé, associé à *réussite-échec* sur la définition *correcte* (donnée par *Fisher*) du concept *d'hypothèse nulle*, accompagnée d'une ambiguïté linguistique impliquant la confusion entre les probabilités conditionnelles $P(\mathcal{D} \mid H)$, $P(H \mid \mathcal{D})$ et $P(\mathcal{D} \cap H)$. Par conséquent, le 2ème axe factoriel peut être interprété comme *un axe de confirmation de l'ambiguïté linguistique et de la réussite-échec en vocabulaire utilisé, déjà trouvés dans le plan (1,2).*

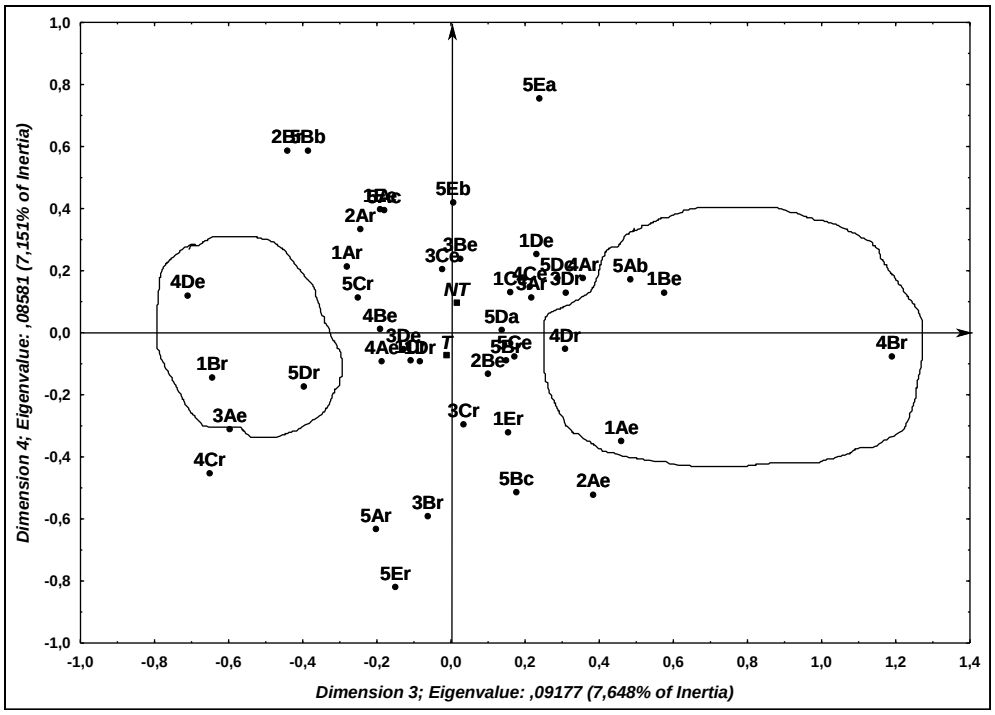
3.4.6. Interprétation du 3ème axe factoriel

L'AFCM conduit à une inertie de 16% pour l'axe 3. Les modalités qui contribuent le plus à la construction de cet axe sont celles entourées dans le graphique 3 ci-après[28]. Elles ont une contribution comprise entre 3,0% et 10,7%. La qualité de

[27] Le même raisonnement est valable pour l'autre groupe de modalités (5Ab, 5Bc et 5Cr).

[28] Comme précédemment, notons au passage que ces modalités occupent des positions correspondant aux modalités ayant les plus grandes coordonnées en valeur absolue sur le 3ème axe factoriel. Les autres modalités occupent des positions quasiment intermédiaires, avec en outre de faibles contributions relatives.

représentation de ces modalités par rapport à un tel axe est comprise entre 0,084 et 0,370.



Graphique 3. *Plan factoriel (3, 4)*

Dans le troisième plan factoriel, le 3ème axe factoriel présente l'opposition de 1Br, 3Ae, 4De et 5Dr *vs* 1Ae, 1Be, 4Br, 4Dr et 5Ab.

Cet axe oppose donc l'absence à la présence d'une certaine relativité quant à la notion de *résultat significatif* (4De *vs* 4Dr). D'un côté, une telle présence est consolidée par la réussite à l'item 4B, qui exprime autrement une telle relativité. De l'autre côté, une telle absence est supportée par l'échec à l'item 3A, revenant à avancer que « *la véracité d'une hypothèse* », qui est, en quelque sorte, une autre façon d'exprimer un *résultat significatif*, a aussi un caractère relatif.

Notons aussi que l'axe 3 oppose la réussite (1Br) à l'échec (1Be) à la définition correcte du concept *d'hypothèse nulle*, qui diffère de celle donnée à « *l'hypothèse testée* » dans la théorie de Neyman-Pearson, à savoir : « *c'est une hypothèse dont la véracité ne conduira à l'adoption d'aucune action nouvelle* ». Un tel échec est accompagné par un autre échec (1Ae) consistant à prétendre que le qualificatif de *nulle* dans l'expression « *hypothèse nulle* » signifie « *sans valeur* ».

Enfin, cet axe oppose aussi la réussite (5Dr) à la distinction entre les deux probabilités: $P(\mathcal{D}'|\mathcal{D})$ et $P(\mathcal{D}|H)$ lorsque $P(\mathcal{D}'|\mathcal{D})$ est exprimée de la façon habituellement utilisée, mais mal interprétée, en tests statistiques, à savoir: « *la probabilité de réplication de la décision* », équivalente à **«** *la probabilité que ces nouvelles données R permettent au juge de reprendre la même décision* **»** dans l'item 5D de notre questionnaire, et l'échec (5Ae) induisant la confusion entre les deux probabilités: $P(H|\mathcal{D})$ et $P(\mathcal{D}|H)$, et ce lorsque $P(H|\mathcal{D})$ est exprimée, comme nous avons déjà indiqué en interprétant le 2ème axe factoriel, sous une façon impliquant *l'approche bayésienne indirecte*, habituellement utilisée pour définir le rôle pour lequel un test statistique est instauré, à savoir: déterminer « *la probabilité que la décision prise par le juge soit due au seul hasard* », objet de l'item 5A de notre questionnaire.

En conclusion, les tendances expliquées par le troisième axe factoriel renvoient à *la présence-absence* du caractère relatif du vocabulaire utilisé, à la *réussite-échec* en termes de langage utilisé, associée à une *réussite-échec* quant à la définition *correcte* (donnée par *Fisher*), qui est différente de celle donnée à *l'hypothèse testée* dans la théorie de Neyman-Pearson. Par conséquent, le 3ème axe factoriel peut être interprété comme *un axe de réussite-échec dans l'utilisation du vocabulaire et du langage.*

4. Discussion, conclusions et perspectives didactiques

Nous avons passé en revue, dans cet article, un ensemble de facteurs susceptibles d'être principalement en cause pour les difficultés que rencontrent les étudiants, et même les professionnels, dans la compréhension des tests statistiques. Ce texte s'est focalisé sur les facteurs de *vocabulaire* et *d'ambiguïté linguistique*, qui sont souvent déconsidérés dans les cours dispensés aux étudiants par la plupart des enseignants de statistique (et probabilités). Pareillement, ces enseignants sous-estiment leur ampleur et leurs effets néfastes sur la compréhension des concepts statistiques, lesquels passent habituellement pour être davantage liés à la vie de tous les jours que les autres concepts de mathématiques. Par ailleurs, toute l'importance donnée dans ce travail à ces deux facteurs ne devrait pas faire oublier les difficultés qui peuvent être causées par les autres facteurs indiqués dans la partie théorique introductive de cette étude. Ces derniers peuvent entraîner des conséquences néfastes, à ne pas sous-estimer, et méritent aussi des études approfondies.

Certes, la nature de l'instrument de mesure (questionnaire) sur lequel nous nous sommes appuyés ne permet pas à nos sujets, surtout ceux qui avaient déjà suivi un cours sur les tests statistiques, de reconnaitre facilement qu'ils sont face à des situations impliquant le vocabulaire et le langage utilisés en procédures de tests statistiques. Ceci est un garant plus ou moins fort de l'obtention de réponses qui ne soient pas influencées par d'autres facteurs, surtout de type psychanalytique (§ 2.5.). Ainsi, les résultats d'analyse factorielle multi-dimensionnelle des réponses des

étudiants interrogés ont donné lieu à un plan et des axes factoriels avec des interprétations en termes de *maîtrise du vocabulaire et du langage utilisés en tests statistiques*. Plan et axes se déclinent selon l'importance de leur part d'inertie absolue (i.e., importance de leur part d'information contenue dans l'ensemble des réponses) suivant les interprétations respectives qui suivent.

- *Une ambiguïté linguistique et une réussite-échec en vocabulaire utilisé (Plan (1,2))*, se traduisant par :
  - *Une distinction non assimilée* entre les trois probabilités $P(\mathcal{D} \mid H)$, $P(H \mid \mathcal{D})$ et $P(\mathcal{D}' \mid \mathcal{D})$, associée à une *réussite-échec* au fait que l'expression « *résultat significatif* » désigne un résultat à caractère *relatif* et non *absolu*, et au fait que le qualificatif *nulle* indiqué dans l'expression « *hypothèse nulle* » est différent du qualificatif « *sans valeur* » *(Cadran I vs Cadran II)*;
  - *Une réussite-échec* au fait que les notions de *véracité d'une hypothèse* et de *raisonnement* adopté pour montrer une telle véracité sont toutes les deux à caractère *relatif* et non *absolu*, accompagnée d'une *réussite-échec* à la définition correcte (*donnée par Fisher*) du concept *d'hypothèse nulle (Cadran III vs Cadran IV).*
- *Une confirmation de l'ambiguïté linguistique et de la réussite-échec en vocabulaire utilisé, déjà trouvées dans le plan (1,2)*, se traduisant par une *présence-absence* de relativité du vocabulaire utilisé, associé à *réussite-échec* à la définition *correcte* (donnée par *Fisher*) du concept *d'hypothèse nulle*, accompagnée d'une ambiguïté linguistique impliquant la confusion entre les probabilités conditionnelles $P(\mathcal{D} \mid H)$, $P(H \mid \mathcal{D})$ et $P(\mathcal{D} \cap H)$ (*Axe 2*).
- *Une réussite-échec en termes de vocabulaire et de langage utilisés*, se traduisant par *présence-absence* du caractère relatif du vocabulaire utilisé, une *réussite-échec* en termes de langage utilisé, associée à une *réussite-échec* quant à la définition *correcte* (donnée par *Fisher*), différente de celle donnée à *l'hypothèse testée* dans la théorie de Neyman-Pearson (*Axe 3*).

Finalement, notons au passage que ces facteurs de vocabulaire et d'ambiguïté linguistique pourraient *a priori* être pris en considération par l'enseignement, contrairement à ce qui se fait habituellement et plus généralement dans toute approche didactique des tests statistiques. Cela permettra certainement d'explorer des perspectives de recherches d'ingénieries didactiques sur les tests statistiques.

Nous avons placé en individus supplémentaires sur les trois premiers plans factoriels les centres de gravité respectifs du groupe d'étudiants ayant déjà suivi un cours sur les tests statistiques, le centre de gravité (*T*), et celui des étudiants qui n'en ont pas suivi, le centre de gravité (*NT*). Il est apparu qu'aucun de ces deux centres de gravité (*T*) et (*NT*) ne s'est situé du côté d'une bonne maitrise du vocabulaire et du langage

habituellement utilisés en tests statistiques. Est-ce l'effet d'un enseignement qui présente les matières (surtout scientifiques) comme une batterie de connaissances statiques et non dynamiques ? Cela mérite en tout cas une étude didactique approfondie en termes d'approches d'enseignement des sciences dans notre système éducatif. Celle-ci pourrait permettre, lorsque des anomalies sont repérées, d'élaborer de nouveaux curricula prenant en considération le caractère relatif de la science et l'ambiguïté linguistique que pourraient causer les expressions et concepts véhiculés. Dans de tels curricula, on devrait aussi insister sur le fait d'introduire en cours les concepts scientifiques en les présentant aux étudiants dans les différents registres sémiotiques possibles (*verbal*, *graphique, tabulaire, ...*), en insistant surtout sur le registre verbal qui est assez constructif mais souvent négligé, et en apprenant aux étudiants à passer d'un registre sémiotique à un autre de manière fluide en toute facilité (Duval, 1993).

BIBLIOGRAPHIE

ACREE, M. C. (1978) *Theories of statistical inference in psychological research: A historicocritical study*. Ann Arbor, MI: University Microfilms International. (University Microfilms No. H790 H7000).

BAR-HILLEL, M. A. & FALK, R. (1982) Some teasers concerning conditional probabilities, *Cognition*, **11**, 109-122.

BATANERO, C. (2000) Controversies around the role of statistical tests in experimental research, *Mathematical Thinking and Learning*, **2(1-2)**, 75-98.

BATANERO, C. & DIAZ, C. (2006) Methodological and Didactical Controversies around Statistical Inference, *Proceedings of 38th Conference of the French Statistical Conference*, Paris: SFDE. CDROM.

BENZECRI, J. P. (1979) Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, *Cahiers de l'analyse des données*, **4**, 377-378.

BIRNBAUM, I. (1982) Interpreting statistical significance, *Teaching Statistics*, **4**, 24-27.

BOX, J. F. (1978) *R. A. Fisher: The life of a scientist*, New York: Wiley.

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

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

DACUNHA-CASTELLE, D. & DUFLO, M. (1990) *Probabilités et Statistiques, Tome 1, Problèmes à temps fixe* (2ème édition), Masson.

DANIEL, L. G. (1997) Kerlinger's research myths: An overview with implications for educational researchers, *Journal of Experimental Education*, **65**, 101-112.

DRACUP, C. (1995) Hypothesis testing –what it really is, *The psychologist*, **8**, 359-362.

DUVAL, R. (1993) Registre de représentation sémiotique et fonctionnement cognitif de la pensée, *Annales de Didactique et de Sciences Cognitives,* IREM de Strasbourg, **5**, 37-65.

EHRENBERG, A. S. C. (1990) A hope for the future of statistics: MSOD, *The American Statistician*, **44(3)**, 195-196.

ELM'HAMEDI, Z. (2010) *Contribution à une ingénierie didactique pour l'enseignement et l'apprentissage des tests statistiques à l'université*. Doctorat. Université Sidi Mohammed Ben Abdellah- Faculté des Sciences Dhar El Mehraz, Fès, Maroc.

ELM'HAMEDI, Z. (2014) Effets d'un apprentissage empirique sur la compréhension du concept de moyenne arithmétique, *Annales de Didactique et de Sciences Cognitives*, IREM de Strasbourg, **19**, 129-169.

FALK, R. & GREENBAUM, C. W. (1995) Significance test die hard, *Theory & Psychology*, **5(1)**, 75-98.

FALK, R. (1986) Misconceptions of statistical significance, *Journal of Structural Learning*, **9**, 83-96.

FALK, R. (1992) A closer look at the probabilities of the notorious three prisoners, *Cognition*, **43**, 197-223.

FISHER, R. A. (1990) - The Design of Experiments. London: (Réimp. 8ème édition de 1966) In J. H. BENNETT (Ed.), Statistical Methods, Experimental Design, and Scientific Inference. Oxford: Oxford University Press. (1rst edition: 1935, London: Oliver and Boyd.)

FREUND, J. E. & PERLES, B. M. (1993) Observations on the definition of p-values. *Teaching Statistics*, **15(1)**, 8-9.

FRICK, R. W. (1995) Accepting the null hypothesis, *Memory & Cognition*, **23(1)**, 132-138.

GALL, I. (1995) Statistical Tools and Statistical Literacy: The Case of The Average, *Teaching Statistics*, **17(3)**, 97-99.

GARFIELD, J. & AHLGREN, A. (1988) Difficulties in learning basic concepts in probability and statistics: Implications for research, *Journal for Research in Mathematics Education*, **19(1)**, 44-63.

GIGERENZER, G. & MURRAY, D. J. (1987) *Cognition as intuitive statistics,* Hillsdale, NJ: Erlbaum.

GIGERENZER, G. (1993) *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.

GIGERENZER, G. (2004) Mindless statistics, *The Journal of Socio-Economics*, **33**, 587-606.

GIGERENZER, G., SWIJTINK, Z., PORTER, T., DASTON, L., BEATTY, J. & KRÜGER, L. (1989) *The empire of chance: How probability changed science and everyday life*, Cambridge, UK: Cambridge University Press.

GILLMAN, L. (1992) The car and the goats, *American Mathematical Monthly*, **99**, 3-7.

GRANAAS, M. (2002) Hypothesis testing in psychology: throwing the baby out with the bathwater? *ICOTS 6.*

HALLER, H. & KRAUSS, S. (2002) Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, **7(1)**, 1-20.

HAWKINS, A. (1991) Success and failure in statistical education –a UK perspective. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (Vol. **1**, pp. 24-32), Voorburg, The Netherlands: International Statistical Institute.

HAWKINS, A., JOLLIFFE, F. & GLICKMAN, L. (1992) *Teaching statistical concepts*, London: Longman.

HUBERTY, C. J. (1985) *On statistical testing*, Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

ISELER, A. (1997) Signifikanztests: Ritual, guter Brauch und gute Gründe. Diskussion von Sedlmeier 1996. *Methods of Psychological Research Online.*

KIRK, R. E. (1996) Practical Significance: A concept whose time has come, *Educational and Psychological Measurement*, **56**, 746-759.

KRUSKAL, W. (1980) The significance of Fisher: a review of R. A. Fisher: the life of a scientist, *Journal of the American Statistical Association*, **75**, 1019-1030.

LABOVITZ, S. (1970) Criteria for selecting a significance level: A note on the sacredness of .05. In D. E. Morrison et R. E. Henkel (Eds.), *The significance test controversy* (pp. 166-171). Chicago: Aldine.

LAMRABET, D. (2001) La démonstration mathématique : structure et situations, *Actes du Congrès international : Universalité et Localité en Sciences*, Faculté des Lettres et Sciences Humaines. Rabat, mars 2001.

LECOUTRE, B. (2006) Training students and researchers in Bayesian methods for experimental data analysis, *Journal of Data Science*, **4**, 207-232.

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

LUNT, P. K. & LIVINGSTONE, S. M. (1989) Psychology and statistics: testing the opposite of the idea you first thought of, *The psychologist*, **2**, 528-531.

MARGOLIS, H. (1987) *Patterns, thinking, and cognition: A theory of judgment.* Chicago: University of Chicago Press.

MCLEAN, A. L. (2000) *On the nature and role of hypothesis tests*, Department of Econometrics and Business Statistics Working Paper 4/2001.

MCLEAN, A. L. (2002) Statistacy: Vocabulary and hypothesis testing, *ICOTS 8*.

MEEHL, P. E. (1997) The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. I., Harlow, S. A., Mulaik et J. H., Steiger (Eds.), *What if there were no significance tests?* (pp. 393-426). Mahwah, NJ: Erlbaum.

MORRISON, D. E. & HENKEL, R. E. (Eds). (1970) *The Significance Tests Controversy*, A reader, Chicago: Aldine.

NICKERSON, R. S. (1996) Ambiguities and unstated assumptions in probabilistic reasoning, *Psychological Bulletin*, **120**, 410-433.

NICKERSON, R. S. (2000) Null hypothesis significance testing: A review of an old and continuing controversy, *Psychological Methods*, **5(2)**, 241-301.

OAKES, M. (1986) *Statistical Inference:* A commentary for the social and behavioural sciences, Chichester, England: Wiley.

POITEVINEAU, J. & LECOUTRE, B. (2001) The interpretation of significance levels by psychological researchers: The .05-cliff effect may be overstated, *Psychonomic Bulletin and Review*, **8**, 847-850.

POITEVINEAU, J. (1998) *Méthodologie de l'analyse des données expérimentales : Etude de la pratique des tests statistiques chez les chercheurs en psychologie, approches normative, perspective et descriptive*, Ph. D, Université de Rouen.

POLLARD, P. & RICHARDSON, J. T. E. (1987) On the probability of making Type I errors, *Psychological Bulletin*, **102**, 159-163.

SCHAFER, J. P. (1993) Interpreting statistical significance and nonsignificance, *Journal of Experimental Education*, **61(4)**, 383-387.

SEDLMEIER, P. & GIGERENZER, G. (1989) Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, **105**, 309-316.

THOMPSON, B. (1994) The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results, *Journal of Personality*, **62**, 157-176.

THOMPSON, B. (1996) AERA editorial policies regarding statistical significance testing: Three suggested reforms, *Educational Researcher*, **25(2)**, 26-30.

TRURAN, J. M. (1998) The development of the idea of the null hypothesis in research and teaching. In L. Periera-Mendoza, L. S., Kea, T. W., Kee & W., Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching Statistics* (pp. 1067-1073), Voorburg, the Netherlands: ISI Permanent Office.

TYLER, R. W. (1931) What is statistical significance? *Educational Research Bulletin*, **10**, 115-118.

VALLECILLOS, A. (1995) Comprensión de la lógica del contraste de hipótesis en estudiantes universitarios, (University students' understanding of the logic of hypothesis testing), *Recherches en Didactique des Mathématiques*, **15(3)**, 53-81.

WATTS, D. G. (1991) Why is introductory statistics difficult to learn? And what can we do to make it easier? *The American Statistician*, **45(4)**, 290-291.

ZAKI, M. & ELM'HAMEDI, Z. (2009) Eléments de mesures pour un enseignement des tests statistiques, *Annales de Didactique et de Sciences Cognitives*, IREM de Strasbourg, **14**,153-194.

ZAKI, M. & ELM'HAMEDI, Z. (2013) Aspects de quelques critiques non fondées de la théorie des tests statistiques, *Annales de Didactique et de Sciences Cognitives*, IREM de Strasbourg, **18**, 139-171.

ZENDRERA, N. (2003) Difficultés d'apprentissage liées aux tests statistiques. Le cas des tests auprès des étudiants en sciences humaines, *Actes des XXXV^{ème} JdS (2)*, 895-899, SFdS et Université Lyon, 2-6 juin 2003.

ZENDRERA, N. (2004) Difficultés et obstacles rencontrés dans l'apprentissage des tests paramétriques : objets d'une recherche en didactique de la statistique, *Actes des XXXVI^{ème} JdS*, SFdS, Université Montpellier II et Ecole supérieure agronomique. Montpellier, 24-28 mai 2004.

ZAHID ELM'HAMEDI
Centre Régional des Métiers de l'Education et de la Formation, Région Casablanca-Settat, Section Provinciale de Settat. Hay El Farah II, Avenue Taha Hussein, B.P: 3066, Ville de Settat, Maroc.
E-mail: mhamdizahid@yahoo.fr