

MONCEF ZAKI & ZAHID ELM'HAMED

**ÉLÉMENTS DE MESURES POUR UN ENSEIGNEMENT
DES TESTS STATISTIQUES**

Abstract. Elements of measure for teaching statistical tests – Statistical tests have traditionally been an important tool in the statistical analysis of data. However it is widely recognized, at least in the university medium, that the teaching of this concept is difficult for the teacher as well for the learner. Indeed, the literature of didactic research on statistical tests has revealed that there are various difficulties, often with respect to some misconceptions, which are encountered at every age and level of expertise. We start with the assumption according to which the teaching of statistical tests cannot succeed in the absence of situations allowing the apprehension of the meaning behind the procedures involved in this concept. This paper synthesizes some supporting measures that can be taken into account in the teaching of statistical tests, and presents the results of an exploratory study conducted on 3rd year students at a scientific university. These measures will, a priori, help students to deepen their understanding of statistical tests.

Key words. Statistical tests, Misconceptions, Frequentist approach, Bayesian approach, Fisher's significance tests, Neyman-Pearson's hypothesis tests, Teaching at university.

Résumé. Les tests statistiques sont à l'origine d'outils importants d'analyses statistiques de données. Cependant, il est largement reconnu, du moins dans le milieu universitaire, que leur enseignement reste un sujet difficile aussi bien pour l'enseignant que pour l'apprenant. En effet, la littérature de recherches en didactique sur les tests statistiques rapporte diverses difficultés, souvent en rapport avec des conceptions erronées, que l'on peut rencontrer chez tous les âges et à tous les niveaux d'expertise. Nous partons de l'hypothèse selon laquelle l'enseignement des tests statistiques ne peut être réussi en l'absence de situations permettant l'appréhension de la signification des procédures impliquées dans cette notion. Cet article synthétise quelques éléments de mesures qui pourraient être pris en considération lors de l'enseignement des tests statistiques, et dont le bien fondé a été examiné à l'aide d'une étude exploratoire auprès d'étudiants de troisième année universitaire scientifique. Ces mesures pourraient a priori aider les étudiants à mieux approfondir leur appréhension vis-à-vis des tests statistiques.

Mots clés. Tests statistiques, Conceptions erronées, Approche fréquentiste, Approche Bayésienne, Tests de signification de Fisher, Tests d'hypothèses de Neyman-Pearson, Enseignement à l'université.

1. Introduction

Le concept de test statistique est connu pour être à l'origine de l'élaboration d'outils fondamentaux de l'inférence statistique, notamment pour le traitement de

situations de décision à partir de bases de données issues de recherches expérimentales. Pourtant, il reste un objet mathématique souvent mal utilisé, d'une part par son application impropre, et d'autre part par une mauvaise interprétation des résultats qu'il peut fournir.

En fait, de nombreux travaux didactiques se sont attelés sur les difficultés dérivant de l'utilisation incorrecte de ce concept, et plusieurs études ont mis en avant l'existence d'erreurs d'interprétation des tests statistiques largement répandues chez les étudiants et les chercheurs en sciences expérimentales (Rosenthal, Gaito, 1963 ; Nelson, Rosenthal, Rosnow, 1986 ; Oakes, 1986 ; Falk, Greenbaum, 1995 ; Mittag, Thompson, 2000 ; Gordon, 2001 ; Poitevineau, Lecoutre, 2001 ; Zendera, 2004 ; Batanero, Diaz, 2006). Récemment, Haller et Krauss (2002) ont même démontré que des formateurs en méthodologie qui enseignent les outils statistiques de décision à des étudiants en psychologie, y compris des formateurs qui travaillent dans le domaine de la statistique, commettent les mêmes erreurs d'interprétation que leurs propres étudiants.

Les multiples problèmes dérivant de l'utilisation impropre des procédures de ce concept, nécessitent donc une réflexion particulière sur le besoin d'amélioration de leur enseignement en vue de leurs applications. Ce n'est qu'au cours des dix dernières années que certains chercheurs didacticiens se sont particulièrement penchés sur les difficultés d'enseignement et d'apprentissage des tests statistiques. Poitevineau (1998) par exemple, a présenté une étude très détaillée sur la méthodologie d'analyse expérimentale des données, dans laquelle il a mis l'accent sur les difficultés qui accompagnent l'utilisation et l'application de ces outils de décision. Ces difficultés trouvent souvent leur origine dans des pratiques d'enseignements qui font un amalgame entre deux approches totalement distinctes, les tests de signification introduits par Fisher et les tests d'hypothèses développés par Neyman-Pearson. D'ailleurs, de tels enseignements conduisent malheureusement les étudiants à ne pas assimiler les fondements des tests statistiques ; seules des procédures "automatisées" sont alors véhiculées par les étudiants, se résumant à des réponses dichotomiques de type : « résultat statistiquement significatif ou non significatif » ou encore « rejeter H_0 ou accepter H_0 ». De telles procédures restent bien entendu insuffisantes pour une application pertinente des tests statistiques.

L'amalgame entre les deux approches de Fisher et de Neyman-Pearson dans l'enseignement des tests statistiques a par ailleurs suscité plusieurs débats entre les chercheurs didacticiens conduisant à des controverses sur le concept même de test statistique (Cobb, Moore, 1997 ; Harlow, Mulaik, Steiger (Eds.) , 1997 ; Krantz, 1999 ; Batanero, 2000 ; Ben-Zvi, Garfield, 2004 ; Pfannkuch, Wild, 2004 ; Batanero, Diaz, 2006 ; Hubbard, Lindsay, 2008 ; ...) : ces controverses ont essentiellement porté sur quelques pratiques erronées dans l'utilisation de ce

concept par certains chercheurs et praticiens et des mauvaises interprétations qui en résultent. Des auteurs et associations pédagogiques (Kline, 2004 ; American Psychological Association, 1994) sont allés jusqu'à recommander la pure exclusion de ce concept de l'enseignement des statistiques inférentielles, ou du moins son enseignement accompagné d'autres méthodes statistiques tels que les intervalles de confiance, la taille de l'effet, la réplication ou la statistique bayésienne... Cependant, malgré tous ces débats, les tests statistiques continuent à être enseignés aux étudiants dans les universités et méritent à ce titre de faire encore l'objet de recherches didactiques.

Le travail que nous exposons dans cet article s'inscrit dans la continuité des réflexions et recherches menées sur la question d'enseignement des tests statistiques. Nous cherchons à y présenter, à la lumière de travaux antérieurs, l'essentiel d'une synthèse portant sur des mesures qui pourraient être prises en considération lors de l'enseignement des tests statistiques, pour mieux aider les étudiants à approfondir leur appréhension pour ce concept. Cette synthèse est aussi appuyée par les résultats d'une étude exploratoire que nous avons par ailleurs menée auprès d'étudiants de troisième année scientifique d'université. Ces mesures n'ont pas la prétention de vouloir remplacer un enseignement proprement dit des tests statistiques ; cependant, elles nous paraissent comme étant un élément important, sinon essentiel dans la mise en place d'un tel enseignement. Elles permettront en particulier aux étudiants de bien cerner les paradigmes de probabilités conditionnelles sur lesquels sont fondés les tests statistiques et de maîtriser la logique de mise en place des indicateurs de jugement et des conclusions qui en découlent, notamment entre les procédures de Fisher et de Neyman-Pearson. Cependant, nous n'irons pas jusqu'à étudier les questions de liaisons entre les conclusions issues d'un test statistique et les critères utilisés dans la mise en place d'un plan expérimental et encore moins les conséquences (effets) que peuvent avoir ces conclusions sur les hypothèses « scientifiques »¹ de départ : ces questions ont bien entendu beaucoup d'intérêt dans une étude faisant appel aux statistiques, mais elles ne sont pratiquement jamais prises en charge (à tort ou à raison...) dans un enseignement de statistique inférentielle.

Avant de développer ces mesures, nous souhaiterions rappeler au lecteur² les principales notions habituellement rencontrées dans les tests statistiques :

¹ Il s'agit de l'hypothèse étudiée par le chercheur au départ, qui à la suite d'une reformulation (modélisation), donne lieu à l'hypothèse alternative H_1 d'un test statistique.

² Cet article suppose que le lecteur est familiarisé avec les différentes notions habituellement impliquées dans une procédure de tests statistiques, à savoir : le niveau de signification, la p-value (ou niveau de signification atteint), les risques du 1^{er} type et du 2^{ème} type, la région critique, le seuil, l'hypothèse nulle, l'hypothèse statistique alternative,

l'hypothèse nulle H_0 , l'hypothèse alternative H_1 et les données observées D_0 (ou échantillon(s)); notions utilisées pour la construction d'outils d'évaluation permettant d'analyser la défaillance entre des données et des hypothèses statistiques.

2. Mesures préconisées pour l'enseignement des tests statistiques

2.1. Deux paradigmes basés sur la notion de probabilité conditionnelle

Deux paradigmes basés chacun sur une probabilité conditionnelle, sont formalisés pour évaluer la "défaillance" entre les données observées D_0 et les hypothèses statistiques H_i ($i = 0$ ou 1), il s'agit de l'approche fréquentiste relative aux "tests statistiques" et de l'approche subjectiviste relative aux "statistiques Bayésiennes" :

- le paradigme de tests statistiques évalue la probabilité $P(D^* | H_0)$, dite p-value³, d'obtenir les données D^* sachant que l'hypothèse statistique H_0 est vraie ; où D^* représente les données observées D_0 ou des données plus extrêmes⁴ que D_0 . L'expression du résultat des tests statistiques en fonction de la probabilité conditionnelle $P(D^* | H_0)$ induit deux éléments essentiels : d'une part, seules des argumentations basées sur la probabilité des données sont prises en considération dans la procédure du test statistique, d'autre part, l'hypothèse statistique H_0 est considérée comme étant un fait donné. Ce dernier élément implique en particulier que toute présentation des résultats du test statistique doit être formulée sous l'hypothèse d'acceptation de ce fait, c'est-à-dire "... sachant que H_0 est vraie" ;
- le paradigme de statistiques bayésiennes s'intéresse à la probabilité conditionnelle $P(H_i | D_0)$, $i=0$ ou 1 , à savoir la probabilité de l'hypothèse statistique H_i ($i=0$ ou 1) sachant les données observées D_0 . En fait,

ainsi que les notions de probabilité conditionnelle, d'échantillon et de population statistique.

³ Cette probabilité $P(D^* | H_0)$ est exactement la p-value que nous obtenons à partir des tests statistiques : la probabilité des données disponibles ou plus extrêmes (ou même moins probables), sachant que l'hypothèse nulle H_0 est vraie. Il est important aussi de noter que la notation de la p-value $P(D^* | H_0)$ n'a de sens que si, d'une part H_0 est interprétée comme étant un événement, et d'autre part H_0 représente une hypothèse simple. Dans le cas où H_0 représente une hypothèse composite, on calcule alors le maximum de la probabilité de rejet par rapport à l'ensemble des paramètres.

⁴ Des données plus extrêmes que D_0 sont des données qui s'écartent (plus que celles de D_0) de H_0 en faveur de H_1 . Le choix de ces données plus extrêmes pourrait dépendre des intentions de l'expérimentateur et du plan échantillonnal adopté.

l'approche bayésienne est fondée sur la *règle de Bayes* :

$$P(H|D_0) = P(H) \cdot \frac{P(D_0|H)}{P(D_0)}.$$

De ce fait, les statistiques bayésiennes représentent une méthodologie mesurant la défaillance entre les données observées D_0 et les hypothèses statistiques H_i ($i = 0$ ou 1), moyennant des argumentations à partir de probabilités (conditionnelles) des hypothèses⁵. Quant à la probabilité $P(H)$, elle mesure le degré de vérité de l'hypothèse H avant de collecter les données D_0 , elle est calculée sur la base d'informations extérieures qui sont indépendantes des données D_0 .

L'intérêt de préciser le fondement des statistiques bayésiennes à partir de la *règle de Bayes*, est de permettre aux étudiants de comprendre la signification de la probabilité $P(H_i | D_0)$, de ne pas la confondre avec l'approche des tests statistiques décrite ci-dessus (Carver, 1978 ; Oakes, 1986 ; Poitevineau, 1998), et de l'interpréter comme étant d'une certaine manière « l'inverse » du concept de *p-value*.

2.2. Les tests statistiques, un amalgame de deux approches : Fisher et Neyman-Pearson

Les tests statistiques sont à leur tour conçus selon deux approches différentes, celle de Fisher et celle de Neyman-Pearson. L'enseignement des tests statistiques est souvent présenté à travers un amalgame de ces deux approches, que Gigerenzer (1993) qualifie de *logique hybride*, difficilement discernable chez les étudiants. En fait, les tests statistiques représentent deux sous paradigmes distincts :

- les tests de signification selon Fisher (1935 et 1956), et qui s'intéressent à la seule probabilité $P(D^* | H_0)$;
- les tests d'hypothèses selon Neyman-Pearson (1933), et qui s'intéressent aux deux probabilités $P(D^* | H_0)$ et $P(D^* | H_1)$;

En outre, Fisher et Neyman-Pearson procèdent selon deux approches différentes :

- Les tests de signification permettent de conclure au rejet (ou à l'échec du rejet) de l'hypothèse nulle H_0 avec un seuil (une probabilité) décidé a posteriori après le calcul de $P(D^* | H_0)$, sans pour autant permettre de conclure quant à l'acceptation ou au rejet de l'hypothèse alternative. Cependant, il faut préciser que l'hypothèse alternative H_1 , bien qu'elle représente implicitement chez Fisher l'hypothèse d'intérêt de la situation explorée, elle intervient directement dans la construction des

⁵ Pour une discussion détaillée de l'approche inférentielle bayésienne, voir par exemple, Edwards, Lindman & Savage (1963) ou Howson & Urbach (1989).

données extrêmes D^* , donc dans l'élaboration de la procédure de calcul des tests de signification ;

- quant aux tests d'hypothèses, ils permettent de construire, en fonction d'un seuil préalablement fixé, indépendamment des données observées D_0 , des règles de décisions permettant de choisir entre deux actions faisant référence, de manière explicite, respectivement à l'hypothèse nulle et à l'hypothèse alternative.

2.3. Interprétation des hypothèses statistiques en termes de modèles statistiques

Une difficulté majeure dans les tests statistiques est l'interprétation qu'on donne aux hypothèses statistiques dans les procédures de traitements mathématiques sous-jacentes à ces situations, particulièrement lorsqu'il s'agit de tests paramétriques. Que l'on soit dans le cadre des tests de signification de Fisher ou celui des tests d'hypothèses de Neyman-Pearson, l'hypothèse nulle H_0 et l'hypothèse alternative H_1 (uniquement chez Neyman-Pearson) renvoient respectivement chacune à un modèle statistique, qui en fait n'est autre qu'un espace probabilisé auquel on associe une famille de probabilités (celle-ci pouvant éventuellement être réduite à une seule probabilité, comme c'est le cas par exemple pour les tests d'hypothèses simples). Ainsi, tout traitement probabiliste sous l'hypothèse H_0 (respectivement sous H_1), en l'occurrence celui de $P(D^* | H_0)$ (respectivement $P(D^* | H_1)$), sous-entend que les calculs probabilistes mis en œuvre se font à partir du modèle probabiliste sous-jacent à cette hypothèse. En principe, cela ne devrait pas poser de problèmes aux étudiants, à condition de leur préciser que les hypothèses statistiques ne sont autres qu'une formulation simplifiée de modèles statistiques susceptibles de répondre le plus adéquatement possible au problème de modélisation des données observées. Mclean (2000) par exemple soutient une telle approche didactique, en suggérant de traduire les hypothèses statistiques mises en jeu dans certains problèmes de régression linéaire (tests sur une moyenne ou encore sur un rapport de variances), en termes de recherche de modèles statistiques qui soient les plus « proches » ou du moins les plus représentatifs des données recueillies.

Cependant, l'interprétation des hypothèses statistiques en termes de modèles statistiques n'est pas toujours immédiate chez les étudiants, bien que dans l'enseignement des tests statistiques toutes les précautions théoriques soient préalablement prises en compte. Ainsi par exemple, dans un cours de statistique mathématique introductif aux tests statistiques, il arrive assez souvent que les

étudiants ne fassent pas de rapport⁶ dans la formulation et l'interprétation des hypothèses H_0 et H_1 mises en jeu dans un test paramétrique, avec la propriété fondamentale de définition d'un modèle statistique, à savoir le fait que tout paramètre θ du modèle est *identifiable* – c'est-à-dire qu'à tout paramètre θ est associée une et une seule probabilité P_θ du modèle statistique.

Cette difficulté d'interprétation des hypothèses statistiques en termes de modèles statistiques peut engendrer d'autres interprétations erronées : en nous plaçant encore une fois dans le cadre des tests paramétriques, un test d'hypothèses simples de Neyman-Pearson ($H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$) peut donner lieu à de fausses interprétations des hypothèses statistiques mises en jeu. En effet, à l'issue de la réalisation de ce test, si les données observées conduisent par exemple à une action en faveur de l'hypothèse H_0 , les étudiants vont parfois conclure dans ce cas là que la vraie valeur du paramètre θ est θ_0 (Vallecillos, 1995 et 1996), en omettant de dire que le modèle P_{θ_0} est celui qui est en faveur des données observées. A notre avis, cette difficulté est non seulement liée à une question d'interprétation des hypothèses statistiques en termes de modèles statistiques, mais aussi à une mauvaise compréhension de la procédure même du test statistique. C'est ce que nous allons développer dans le prochain paragraphe.

2.4. Interprétation des procédures statistiques

Les tests statistiques sont fondés sur l'évaluation de la probabilité conditionnelle $P(D^* | H_0)$ chez Fisher, et sur les deux probabilités conditionnelles $P(D^* | H_0)$ et $P(D^* | H_1)$ chez Neyman-Pearson (cf. § II.2). Néanmoins, l'interprétation adéquate de ces probabilités conditionnelles dans la procédure même des tests statistiques peut poser quelques difficultés aux étudiants, voire même à certains chercheurs praticiens des statistiques - psychologues et autres -.

Le dernier exemple que nous venons de citer au paragraphe précédent, représente une variante de ces difficultés, qui en outre se trouve associée à une fausse interprétation de l'hypothèse nulle (voir l'exemple ci-dessus du test $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$) : certains étudiants qui ont une *approche déterministe* de la procédure du test statistique, vont aller déclarer dans le cas de cet exemple que le test vient de prouver (au sens d'une preuve mathématique) que le paramètre $\theta = \theta_0$ (Vallecillos, 1995 et 1996).

Une seconde variante de ces difficultés consiste à mettre en avant une *approche probabiliste erronée* due à une fausse interprétation de $P(D^* | H_0)$ et $P(D^* | H_1)$:

⁶ Expérience souvent vécue à la Faculté des Sciences Dhar El mehraz de Fès avec les étudiants de licence de Mathématiques (Option Probabilités - Statistiques) dans un cours de statistique mathématique introductif aux tests statistiques.

certaines étudiants vont renverser les conditions de ces probabilités, en passant de $\Pr(\text{Données} | \text{Hypothèse})$ à $\Pr(\text{Hypothèse} | \text{Données})$. La conséquence d'un tel renversement des hypothèses et des données dans les probabilités conditionnelles va favoriser de fausses interprétations des procédures des tests statistiques, comme par exemples :

« *La probabilité que l'hypothèse nulle soit vraie est égale à la p-value (ou au seuil du test)* » (Poitevineau, 1998),

ou encore,

« *La probabilité que les résultats soient dus au seul hasard est égale à la p-value (ou au seuil du test)* », ou de manière équivalente « *La probabilité que l'hypothèse alternative soit vraie est égale à 1 moins la p-value (ou à 1 moins le seuil du test)* » (Carver, 1978 ; Batanero, 2000).

L'approche probabiliste peut aussi prendre une autre forme d'interprétation erronée de la p-value $P(D^* | H_0)$ pour induire de fausses interprétations au niveau des conclusions d'un test statistique : l'expression « statistiquement significatif » renvoie à une conclusion en faveur de l'hypothèse d'intérêt⁷ (le rejet de H_0 chez Fisher ou l'acceptation de H_0 chez Neyman-Pearson), néanmoins, le résultat de la p-value ne devrait pas pour autant être interprété comme étant une mesure du degré de confiance (ou d'importance) que l'on va attribuer à l'hypothèse scientifique mise en jeu dans la situation étudiée (sachant que l'hypothèse statistique d'intérêt n'est autre qu'une formulation réduite de l'hypothèse scientifique). Pourtant, on va constater chez certains étudiants ou chercheurs-praticiens (Kerlinger, 1979 ; Pedhazur et Schmelkin, 1991) des conclusions erronées du type : plus le résultat d'un test est « statistiquement significatif », plus le résultat (hypothèse) scientifique étudié est « important ». En fait, ces conclusions incorrectes renvoient à deux conceptions erronées :

- certains étudiants ou chercheurs croient qu'une plus petite p-value reflète directement une influence (un effet) quantitative plus grande du facteur ou de la relation examinés et c'est en ce sens qu'ils considèrent qu'une plus petite p-value montre un résultat plus important (American Psychological Association (APA), 1994 ; Wilkinson & APA , 1999 ; Vacha-haase et al , 2000 , p. 599) ;
- certains confondent signification statistique et signification pratique et tentent de croire qu'une plus petite p-value montre un résultat qui a plus de signification ou d'importance pratique. Un autre aspect, aussi (ou même plus) important, de cette conception erronée est que certains

⁷ On rappelle que l'hypothèse d'intérêt est H_1 chez Fisher et H_0 chez Neyman-Pearson.

chercheurs sont tentés de croire que des résultats qui ne sont pas statistiquement significatifs n'ont pas d'intérêt ou d'importance pratique. Or de tels résultats peuvent très bien avoir de l'intérêt ou de l'importance pour l'étude menée, pour des raisons théoriques, sociales, ou en liaison avec d'autres études portant sur le même sujet (Nelder, 1999 ; Gliner & al, 2002 ; Lecoutre & al, 2003 ; Castro Sotos & al, 2007).

Les erreurs d'interprétation liées aux hypothèses ou aux conclusions dans la procédure d'un test statistique nécessitent à notre avis, du point de vue de l'enseignement, la prise en compte et la comparaison des deux paradigmes "tests statistiques" et "statistiques bayésiennes". Cette approche didactique, qui est aussi soutenue par Gigerenzer (1993), contribuerait d'une part à avoir une appréhension globale des approches développées pour les tests statistiques, et d'autre part à mieux percevoir et la nature et le statut des hypothèses et des conclusions dans chacune de ces approches.

On pourrait donc éventuellement penser à une présentation permettant aux étudiants d'accéder aux différents modèles d'organisation des tests d'hypothèses statistiques :

<i>Étape</i>	Fisher	Neyman-Pearson	Solution Bayésienne
<i>(1) Hypothèse d'intérêt</i>	H_1	H_0	H_1
<i>(2) Hypothèse(s) impliquée(s) dans la procédure de test statistique</i>	H_0	H_0 et H_1	H_i ($i=0$ ou 1)
<i>(3) Élément(s) de jugement</i>		Niveau de signification du test : $\alpha \in]0,1[$	Probabilités a priori : $P(H_i)$
<i>(4) Taille(s) Echantillon(s)</i>	(n_1, \dots, n_j)	(n_1, \dots, n_j)	(n_1, \dots, n_j)



Fisher	Neyman-Pearson	Solution Bayésienne
(5) <i>Données observées (Échantillon(s))</i> : D_0	(5) <i>Indicateur de jugement</i> : Calcul de règle de décision δ à partir du rapport de vraisemblances	(5) <i>Données observées (Echantillon(s))</i> : D_0
(6) <i>Indicateur de jugement</i> : Calcul de la p-valeur $P(D^* H_0)$ à partir d'estimateurs (dans le cas paramétrique) ou de statistiques subjectives	(6) <i>Données observées (Echantillon(s))</i> : D_0	(6) <i>Indicateur de jugement</i> : Probabilités a posteriori : $P(H_i D_0) = \frac{P(D_0 H_i)}{P(D_0)}$ $P(H_i) \cdot \frac{P(D_0 H_i)}{P(D_0)}$ (Règle de Bayes)



Étape	Fisher	Neyman-Pearson	Solution Bayésienne
(7) <i>Jugement</i>	Comparaison de $P(D^* H_0)$ à un seuil de signification α fixé a posteriori	$\delta(D_0)$	Comparaison de $P(H_1 D_0)$ à $P(H_0 D_0)$
(8) <i>Conclusions</i>	Si $P(D^* H_0)$ est petite (par rapport à α) : Rejeter H_0 Si $P(D^* H_0)$ est élevée (par rapport à α) : Echouer de rejeter H_0	Si $\delta(D_0)=1$: Rejeter H_0 Si $\delta(D_0)=0$: Accepter H_0 Si $\delta(D_0)=c$ ($\neq 0$ et 1) : Pile ou face	Si $P(H_1 D_0) < P(H_0 D_0)$: Rejeter H_1 Si $P(H_1 D_0) \geq P(H_0 D_0)$: Accepter H_1

Tableau 1 : Modèles d'organisation des tests d'hypothèses statistiques (Fisher, Neyman-Pearson et solution Bayésienne).

La présentation de ces modèles (ou de modèles similaires) est certainement nécessaire pour réussir un enseignement de statistiques inférentielles, mais elle n'est pas suffisante. En effet, dans les modèles précédents, nous voyons apparaître des paramètres tels que la taille de l'échantillon, les données D_0 , ou données extrêmes D^* et le niveau (ou seuil) de signification du test, qui vont jouer un rôle

très important dans le processus d'inférence statistique, se situant à des niveaux différents, avec différents statuts ; ils méritent à ce titre d'être traités avec attention dans un enseignement de tests statistiques, en particulier dans les procédures de Fisher et de Neyman-Pearson.

2.5. Niveau de signification α et amalgame entre les procédures de Fisher et de Neyman-Pearson

La procédure de recherche de l'indicateur de jugement en situation de tests statistiques diffère entre Fisher et Neyman-Pearson. Le premier construit son indicateur de jugement directement en fonction de données D^* , regroupant les données observées D_0 et celles qui sont plus extrêmes ; alors que le second établit une règle de décision (indicateur de jugement) en fonction uniquement de la taille n_0 de l'échantillon observé et du seuil de signification α qu'il se fixe a priori, indépendamment du corps même des données observées D_0 . Ce n'est qu'à l'étape du jugement effectif que Neyman-Pearson fait appel aux données observées D_0 , pour conclure ensuite.

Pourtant, dans les situations courantes, comme par exemple celles de tests paramétriques relatifs aux modèles continus, la distinction entre les deux procédures précédentes devient pratiquement inapparente, du moins au niveau de l'indicateur de jugement et du jugement lui-même : ce qui justifie en partie l'amalgame entre ces deux approches (cf. § II. 2), souvent rencontré dans un enseignement de tests statistiques.

En effet, en nous plaçant dans une situation de tests paramétriques dans le cas continu, il arrive très souvent que l'estimateur du paramètre testé, souvent utilisé par Fisher, coïncide avec la statistique issue du rapport de vraisemblances utilisé par Neyman-Pearson. C'est ainsi que, pour un seuil α fixé a priori, la recherche d'indicateur de jugement chez ce dernier conduit à une règle de décision δ , fonction de l'estimateur utilisé par Fisher, pour donner lieu à une région de rejet (voir valeur critique c de la figure 1).

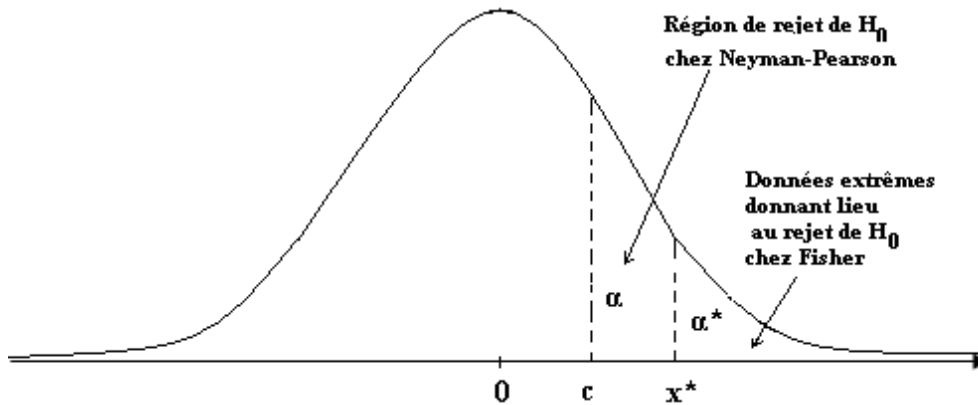


Figure 1 : Indicateurs de jugement chez Fisher et Neyman-Pearson.

Ainsi, les calculs de $\delta(D_0)$ ou $P(D^*|H_0) = \alpha^*$ relèvent d'indicateurs de jugement équivalents, le premier consiste à comparer une valeur x^* (issue des données D_0) à la valeur critique c définissant la région de rejet, et le second est basé sur la comparaison de α^* au seuil de signification α : ce qui est mathématiquement équivalent, car formellement cela se traduit en fait par une relation de type $P(D^*|H_0) = P^*([x^*, +\infty[) = \alpha^*$, et la fonction de répartition est bien entendu bijective dans le cas continu. Or, nous constatons que $P(D^*|H_0) = \alpha^*$ n'est autre que l'indicateur de jugement utilisé par Fisher, ce qui explique « *la logique hybride* » citée par certains auteurs (Gigerenzer, 1993), que nous préférons qualifier de *procédure hybride* dans la recherche d'indicateurs de jugement chez Fisher et Neyman-Pearson. Cependant, il faudrait souligner dans le cas présent que la différence entre les procédures de Neyman-Pearson et de Fisher, se situe plutôt au niveau de l'étape ultime concernant les conclusions : au vu des résultats obtenus à l'étape de jugement, le premier va soit accepter H_0 , soit la rejeter (autrement dit accepter H_1), alors que le second va soit rejeter H_0 , soit y échouer, sans pour autant engager de conclusions sur H_1 : cela correspond à deux logiques totalement différentes au niveau des conclusions de ces deux approches, logiques qui ne sont pas toujours nuancées dans l'enseignement, ce qui génère par conséquent un réel amalgame entre les procédures de Fisher et de Neyman-Pearson.

2.6. Exemple de test d'hypothèses appuyant la différence de procédures chez Fisher et Neyman-Pearson

Nous venons de soulever une question importante concernant la nuance qui existe dans les conclusions relatives aux procédures de Fisher et de Neyman-Pearson. Il serait alors intéressant de présenter un exemple de situations qui met en valeur la différence de procédures chez ces deux auteurs : les situations les plus riches dans

ce cas là sont celles de tests non paramétriques, faisant référence à des modèles statistiques discrets.

Considérons par exemple l'énoncé suivant :

« Une personne en état d'ébriété essaie d'ouvrir une porte à l'aide d'un porte clés composé de 5 clés. Elle réussit à ouvrir la porte au bout de 4 essais. Deux hypothèses sont mises en jeu dans cette situation :

H_0 : La personne est en état d'ébriété très avancé, auquel cas elle procède par essais avec remise.

H_1 : La personne n'est pas en état d'ébriété très avancé, auquel cas elle procède par essais sans remise. »

Pour traiter cette situation, nous considérons naturellement la variable aléatoire $X = \text{''Nombre d'essais effectués par la personne jusqu'à ouverture de la porte''}$. Sous les hypothèses H_0 et H_1 , la variable aléatoire X suit respectivement une loi géométrique de paramètre 1/5 et une loi uniforme sur $\{1, 2, 3, 4, 5\}$.

- La solution de Neyman-Pearson à ce problème, teste les hypothèses H_0 et H_1 , où H_0 représente l'hypothèse d'intérêt. Ainsi, le rapport de vraisemblances fournit une région de rejet de la forme :

$$RC = \left\{ k \geq 1 / \frac{P^{H_1}(X=k)}{P^{H_0}(X=k)} \geq c \right\},$$

où $c > 0$ représente une constante dépendant du seuil $\alpha \in]0,1[$ fixé au départ.

Sachant que sous H_0 , X suit une loi géométrique de paramètre 1/5, la région de rejet RC s'obtient alors par la relation :

$$P^{H_0} \left\{ X \in \{1, \dots, 5\} / X \geq \frac{\ln(c)}{\ln(5/4)} + 1 \right\} \leq \alpha$$

En se référant à la distribution de la loi géométrique de paramètre 1/5 pour les valeurs k allant de 1 à 5, nous obtenons pour $\alpha=0,05$, $RC = \emptyset$; pour $\alpha=0,1$, $RC = \{5\}$ et pour $\alpha=0,2$, $RC = \{4, 5\}$.

Ainsi pour cette situation, au seuil $\alpha=0,2$, la procédure de Neyman-Pearson va conclure au rejet de l'hypothèse H_0 .

- Pour Fisher, si nous considérons H_1 comme hypothèse d'intérêt, cette situation conduit alors au test de signification portant sur l'hypothèse H_0 .

Sachant que les données observées correspondent à l'ensemble $D_0 = \{X = 4\}$, l'ensemble D^* englobant les données observées et celles qui sont plus extrêmes que celles-ci, correspond aux données $D^* = \{1, 2, 3, 4\}$ qui sont en faveur de l'hypothèse H_1 . Par conséquent, l'indicateur de jugement chez Fisher va être $P^{H_0}(D^*) = 0,5904$. En comparant cet indicateur au seuil de signification $\alpha=0,2$ (en référence à la solution de Neyman-Pearson), le test de signification de Fisher va échouer pour cette situation dans le rejet de l'hypothèse H_0 . (Contrairement à la conclusion fournie par la solution de Neyman-Pearson)

La situation précédente constitue un exemple intéressant qui explique la différence de procédures entre Fisher et Neyman-Pearson ; il est donc important de présenter de telles situations aux étudiants lors d'un enseignement sur les tests d'hypothèses statistiques.

2.7. Niveau de signification α et rôle dissymétrique des hypothèses H_0 et H_1 dans la construction de tests chez Neyman-Pearson

En intégrant les deux hypothèses H_0 et H_1 dans la procédure de construction de tests, Neyman-Pearson va introduire deux types de risques, représentant respectivement sous ces deux hypothèses⁸ les coûts moyens $R(\theta, \delta)$ de la règle de décision δ , faisant l'objet de la solution du problème : ce sont les risques de première et de seconde espèces. Dans la recherche de δ , la solution du problème ne permet pas d'agir (minimiser) sur les deux risques simultanément. Ainsi, dans sa solution, Neyman-Pearson ne va contrôler que le risque de première espèce (qui fait référence à son hypothèse d'intérêt H_0), en fixant un seuil de signification α , pour aller chercher parmi les règles de décision répondant à ce seuil, celles dont le risque de seconde espèce est le plus faible (ou encore dont la puissance⁹ est maximale). Ainsi, nous constatons d'emblé dans la procédure de Neyman-Pearson un rôle dissymétrique des hypothèses H_0 et H_1 dans l'élaboration de la solution au problème de tests d'hypothèses.

Cette dissymétrie va effectivement contraindre le type de solution apporté au problème de test statistique, c'est le cas par exemple du test bilatéral, $H_0 : \theta \leq \theta_1$ ou $\theta \geq \theta_2$ vs $H_1 : \theta_1 < \theta < \theta_2$, où il existe pour les modèles à rapport de vraisemblances monotone un test (règle de décision) qui est uniformément le plus puissant (UPP) dans la classe des tests de seuil α ; alors qu'il n'en est rien pour le cas symétrique ($H_0 : \theta_1 \leq \theta \leq \theta_2$) : dans ce cas, la solution est limitée aux seuls

⁸ Ces deux risques sont naturellement fonction du paramètre θ respectivement sous les hypothèses H_0 et H_1 .

⁹ La puissance est le complémentaire à 1 du risque de seconde espèce (sous l'hypothèse H_1).

tests qui sont sans biais, c'est-à-dire dont le risque de première espèce est plus petit que la puissance (cf. par exemple C. Robert, 1992).

Par ailleurs, lorsqu'une règle de décision chez Neyman-Pearson conduit au rejet ou à l'acceptation d'une hypothèse H_0 , cette décision est conduite selon une certaine puissance du test. Ainsi, pour un seuil de signification α fixé, en faisant augmenter la taille n_0 de l'échantillon des données observées D_0 , il peut arriver que la puissance du test augmente, et que la région de rejet s'élargisse : par conséquent à un seuil de signification donné, l'augmentation de la taille n_0 de l'échantillon peut conduire à l'élaboration de règles de décision avec une plus forte puissance, mais pour lesquelles l'hypothèse H_0 a plus de chances d'être rejetée. A la limite, un échantillon de taille infinie va systématiquement conduire au rejet de l'hypothèse nulle H_0 ...

L'exemple simple de test unilatéral sur la moyenne, $H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$ d'un modèle gaussien $N(\mu, \sigma_0^2)$, décrit parfaitement la situation précédente. En effet, pour un seuil $\alpha \in]0,1[$ fixé, la région de rejet RC d'un échantillon de taille n_0 s'obtient par :

$$RC = \left\{ (X_1, \dots, X_{n_0}) / \bar{X} \in \left] \mu_0 + \Phi_\alpha \frac{\sigma_0}{\sqrt{n_0}}, +\infty \right[\right\}$$

où Φ_α désigne le quantile de la loi normale centrée - réduite $N(0,1)$ d'ordre $1 - \alpha$.

La puissance de ce test pour $\mu > \mu_0$ s'obtient par $P(N(0,1) \geq \Phi_\alpha + \frac{\sqrt{n_0}}{\sigma_0}(\mu_0 - \mu))$.

Par conséquent, lorsque la taille n_0 de l'échantillon augmente, la région de rejet RC s'élargit et la puissance du test augmente ($\mu_0 - \mu$ est négatif). A la limite, lorsque la taille n_0 de l'échantillon tend vers l'infini, l'hypothèse H_0 sera systématiquement rejetée, en vertu de la loi forte des grands nombres ($\bar{X} \xrightarrow{P_{\mu_0} - ps} \mu_0$), et la puissance du test devient maximale (elle tend vers 1).

Ainsi, nous constatons que le rôle dissymétrique des hypothèses H_0 et H_1 a un effet non négligeable sur les procédures d'élaboration de tests chez Neyman-Pearson, il en délimite la portée dans la recherche de solutions en statistiques inférentielles, au point de remettre en question la cohérence même de l'approche qui en est sous jacente.

3. Étude exploratoire auprès des étudiants

3.1. Objectif de l'étude et expérimentation

La portée des conclusions dégagées à la lumière de l'analyse précédente concernant quelques éléments de mesures préconisés pour l'enseignement des tests statistiques, bien qu'elle soit fondée sur la revue de plusieurs travaux didactiques antérieurs (Carver, 1978 ; Oakes, 1986 ; Gigerenzer, 1993 ; Vallecillos, 1995 et 1996 ; Poitevineau, 1998 ; Mclean, 2000 ; Zandrera, 2003, ...), ne pourrait que gagner en pertinence si elle est accompagnée d'observations effectives auprès des étudiants. Nous avons donc conduit au courant du deuxième semestre de l'année universitaire 2007/2008 une expérimentation auprès de 70 étudiants en 3^{ème} année de formation (38 étudiants de la Faculté des Sciences de Rabat-Agdal et 32 élèves ingénieurs de l'Institut Agronomique et Vétérinaire de Rabat), auxquels nous avons administré durant 1 heure un questionnaire en vrai - faux, avec une demande de justification écrite des réponses. Les étudiants interrogés¹⁰ ont tous suivi un enseignement classique de probabilités – statistiques, d'un volume horaire de 70 heures, portant sur la théorie de base des probabilités (espaces probabilisés, variables aléatoires et lois de probabilités, lois des grands nombres, ...), l'estimation ponctuelle et par intervalles de confiance, ainsi que les tests statistiques habituellement étudiés dans les cas paramétriques et non paramétriques.

Le questionnaire a été construit sur la base d'une situation relevant d'un test de *conformité* (cf. Tableau 2), généralement présentée aux étudiants dans leur cours sur les tests statistiques. Il est composé de 7 items- numérotés de A à G – recouvrant les différents aspects que nous avons analysés dans l'étude des éléments préconisés pour l'enseignement des tests statistiques, et dont le résumé est le suivant :

- rôle d'un test statistique (entre Fisher, Neyman-Pearson et Bayes) ;
- interprétation de l'hypothèse nulle et de l'alternative en termes de modèles probabilistes ;
- procédures statistiques et indicateurs de jugement chez Fisher et Neyman-Pearson ;
- procédures de test statistique et seuil de signification chez Fisher et Neyman-Pearson ;
- distinction et interprétation des conclusions des tests statistiques selon les approches de Fisher et Neyman-Pearson ;

¹⁰ Notons au passage, qu'il nous a été très difficile d'interroger des étudiants, du fait des programmes trop chargés et de la quasi indisponibilité des enseignants qui les encadrent.

- statut du seuil de signification et de la taille d'échantillon face à l'acceptation de l'hypothèse nulle ;
- dissymétrie de l'hypothèse nulle et de l'alternative en fonction de l'indicateur de jugement dans la construction d'un test chez Neyman-Pearson.

Avant de procéder à l'analyse proprement dite des productions des étudiants, nous allons consacrer le prochain paragraphe à l'analyse a priori du questionnaire, afin de mieux cerner l'interprétation (statistique) des réponses des étudiants.

Soit X une variable aléatoire qui suit une loi normale de moyenne μ et d'écart type connu σ :

$$X \sim N(\mu, \sigma^2)$$

On rappelle que la fonction densité f_μ de X est définie par :

$$f_\mu(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

On considère les deux hypothèses statistiques suivantes H_0 et H_1 relatives à la moyenne μ de X , et auxquelles on veut appliquer **un test statistique** :

$$\begin{cases} H_0 : \text{“}\mu = \mu_0\text{”} \\ H_1 : \text{“}\mu = \mu_1\text{”, avec } \mu_1 > \mu_0 \end{cases}$$

Pour cela, on a recueilli à l'aide d'un échantillon (X_1, \dots, X_{n_0}) de taille n_0 , les données $D_0 : (x_1^0, \dots, x_{n_0}^0)$ telles que $\bar{x}_{ob} = \frac{x_1^0 + \dots + x_{n_0}^0}{n_0}$ (moyenne observée).

Dans tout ce questionnaire, $\bar{X} = \frac{X_1 + \dots + X_{n_0}}{n_0}$ désigne la moyenne d'échantillon (X_1, \dots, X_{n_0}) issu de la variable aléatoire X .

Tableau 2 : Situation du test statistique retenu dans le questionnaire.

3.2. Analyse a priori du questionnaire

Les sept aspects (items numérotés de A à G) qui composent le questionnaire présenté aux étudiants, sont tous relatifs à la situation de test de *conformité* présenté ci-dessus :

(A) A votre avis, l'application d'un test statistique dans cette situation sert à :

- (1) Démontrer que μ est égal à μ_0 Vrai Faux
- (2) Démontrer que μ est égal à μ_1 Vrai Faux
- (3) Conclure, avec une forte probabilité, que les données observées D_0 rejettent l'hypothèse H_0 Vrai Faux
- (4) Déterminer la probabilité de réalisation de H_0 , sachant qu'on a observé les données D_0 Vrai Faux
- (5) Déterminer la probabilité pour que la réalisation de H_1 soit due au seul hasard Vrai Faux
- (6) Décider, moyennant un risque de se tromper, entre les hypothèses $H_0 : \mu = \mu_0$ et $H_1 : \mu = \mu_1$ Vrai Faux
-

Un test statistique ne peut jamais démontrer si une hypothèse est vraie ou fausse. Il fournit seulement des probabilités conditionnées par sa réalisation, et ne peut ainsi constituer une preuve mathématique proprement dite de sa véracité. Ces probabilités interviennent dans le calcul de la p-value ($P(\bar{X} \geq \bar{x}_{ob} | H_0)$) dans le cas des tests de signification (approche de Fisher) et dans la détermination de la région critique RC ($P(RC | H_0) = \alpha$) dans le cas des tests d'hypothèses (approche de Neyman-Pearson). De ce fait, les modalités (1) et (2) de cette question sont fausses.

En revanche, la modalité (3) est vraie. Elle indique le rôle joué par un test statistique selon l'approche de Fisher. Rappelons que ce rôle est de conclure, avec une forte probabilité, si les données observées D_0 rejettent ou non l'hypothèse nulle H_0 .

Par ailleurs, les probabilités conditionnelles impliquées dans une procédure de test statistique sont des probabilités de réalisation des données sachant que l'hypothèse nulle H_0 est vraie ($P(D_0 | H_0)$), et non le contraire ($P(H_0 | D_0)$). Ces deux types de probabilités sont bien entendu différents. Par conséquent, nous ne pouvons conclure à partir de l'application d'un test statistique que la réalisation d'une hypothèse est certaine (items (1) et (2)) ou qu'elle correspond à toute autre probabilité (item (4)). De ce fait, la modalité (4) est fausse¹¹.

¹¹ Néanmoins, certains outils statistiques permettent l'estimation des probabilités des hypothèses. Nous pouvons citer à titre d'exemple la statistique bayésienne.

La modalité (5) est fautive. En effet, dans une procédure de test statistique, H_0 est une hypothèse selon laquelle la réalisation de H_1 est due au seul hasard. Ensuite, les probabilités conditionnelles impliquées dans une telle procédure sont d'une part calculées en supposant a priori la réalisation certaine de H_0 ($p\text{-value} = P(\bar{X} \geq \bar{x}_{ob} | H_0)$ ou $P(RC | H_0) = \alpha$), et d'autre part utilisées pour décider a posteriori si nous acceptons ou rejetons l'idée de la réalisation certaine de H_0 .

Enfin, la modalité (6) est vraie. Elle indique le rôle joué par un test statistique selon l'approche de Neyman-Pearson. Cette dernière permettra de décider entre H_0 et H_1 , moyennant un risque de se tromper dans cette décision.

(B) Pour $i = 0$ ou 1 , l'hypothèse H_i " $\mu = \mu_i$ " signifie que :

(1) $\mu = \mu_i$ Vrai Faux

(2) La variable aléatoire X suit une loi $N(\mu_i, \sigma^2)$ Vrai Faux

H_i : " $\mu = \mu_i$ " est une écriture simplifiée de H_i : " $N(\mu, \sigma^2) = N(\mu_i, \sigma^2)$ ". Ceci est vrai puisque à chaque μ on associe une et une seule loi de probabilité $N(\mu, \sigma^2)$.

(C) Dans une procédure de test statistique, le calcul sous l'hypothèse H_0 de la probabilité

$$p = P \left[\frac{\sqrt{n_0}(\bar{X} - \mu_0)}{\sigma} \geq \frac{\sqrt{n_0}(\bar{x}_{ob} - \mu_0)}{\sigma} \right] = P \left[N(0, 1) \geq \frac{\sqrt{n_0}(\bar{x}_{ob} - \mu_0)}{\sigma} \right]$$

correspond à :

(1) La probabilité d'obtenir \bar{X} supérieure ou égale à \bar{x}_{ob} , sachant que μ est égal à μ_0 Vrai Faux

(2) La probabilité de réalisation de H_0 , sachant qu'on a observé les données D_0 Vrai Faux

(3) La probabilité $P \left[\frac{f_{\mu_1}(X_1) \dots f_{\mu_1}(X_{n_0})}{f_{\mu_0}(X_1) \dots f_{\mu_0}(X_{n_0})} \geq c \mid \mu = \mu_0 \right] = \alpha$, où c est une constante positive calculée en fonction d'un seuil α fixé préalablement, indépendamment des données observées D_0 Vrai Faux

La modalité (1) de cette question est vraie. En effet, la probabilité indiquée est la définition du concept de p-value ($P(\bar{X} \geq \bar{x}_{ob} | H_0)$), qui, après transformation, est

$$\text{égale à } P \left[N(0,1) \geq \frac{\sqrt{n_0}(\bar{x}_{ob} - \mu_0)}{\sigma} \right].$$

La modalité (2) est fautive (se reporter à l'explication donnée pour prouver la fausseté de la modalité (4) de l'item A).

$$\text{La probabilité } p = P \left[N(0,1) \geq \frac{\sqrt{n_0}(\bar{x}_{ob} - \mu_0)}{\sigma} \right] \text{ est l'indicateur de jugement selon}$$

l'approche de Fisher, que nous comparons à un seuil de signification α défini a posteriori, et à partir de laquelle on va tirer une conclusion, selon que cette probabilité est inférieure ou supérieure à α . Or, la probabilité

$$P \left[\frac{f_{\mu_1}(X_1) \dots f_{\mu_1}(X_{n_0})}{f_{\mu_0}(X_1) \dots f_{\mu_0}(X_{n_0})} \geq c \mid \mu = \mu_0 \right] \text{ indiquée dans la modalité (3) est fixée à } \alpha.$$

De ce fait, la modalité (3) est fautive.

(D) Une procédure de test statistique se base sur le concept de seuil de signification souvent noté α (ou ε). A votre avis, le seuil de signification α :

(1) Permet de construire, après avoir fixé la taille n_0 de l'échantillon, une règle de décision pour laquelle, au vu des données observées D_0 on décide de :

- rejeter l'hypothèse H_0 (et accepter l'hypothèse H_1)

Ou

- accepter l'hypothèse H_0 (et rejeter l'hypothèse H_1) **Vrai** **Faux**

(2) Est directement comparé à la probabilité $P \left[N(0,1) \geq \frac{\sqrt{n_0}(\bar{x}_{ob} - \mu_0)}{\sigma} \right]$ pour

conclure au rejet de H_0 , ou à l'échec dans le rejet de H_0 **Vrai** **Faux**

Les modalités (1) et (2) de cette question sont toutes les deux vraies. Elles permettent d'explicitier le rôle que peut jouer le concept de seuil de signification dans une procédure de test statistique selon l'approche de Neyman-Pearson (modalité (1)) et selon l'approche de Fisher (modalité (2)).

(E) Après application d'un test statistique, la (les) conclusion(s) qu'on peut tirer est (sont) :

- | | | |
|---|--------------------------------------|--------------------------------------|
| (1) Rejeter H_0 OU Accepter H_0 | <input type="checkbox"/> Vrai | <input type="checkbox"/> Faux |
| (2) Rejeter H_0 OU Echouer de rejeter H_0 | <input type="checkbox"/> Vrai | <input type="checkbox"/> Faux |
| (3) Rejeter H_1 OU Echouer de rejeter H_1 | <input type="checkbox"/> Vrai | <input type="checkbox"/> Faux |
| (4) Accepter H_1 OU Echouer de rejeter H_0 | <input type="checkbox"/> Vrai | <input type="checkbox"/> Faux |
-

Les modalités vraies de cette question sont bien sûr (1) et (2) :

- (1) si le test statistique en question est vu selon l'approche de Neyman-Pearson ;
- (2) si le test statistique en question est vu selon l'approche de Fisher.

Cependant, en nuancant les réponses, si nous sommes face à un test statistique selon l'approche de Fisher, la modalité (1) devient fausse, du fait que la conclusion "Accepter H_0 " est différente "d'échouer de rejeter H_0 ". De même que, si nous sommes face un test statistique selon l'approche de Neyman-Pearson, la modalité (2) devient fausse, du fait que la conclusion "Echouer de rejeter H_0 " n'implique pas "Accepter H_0 ". En fait, il y a deux raisons pour cela :

- la première est qu'il peut y avoir d'autres raisons que celles explicitées dans l'hypothèse H_0 pour que les données soient conformes aux prédictions de H_0 , en particulier, dans le cas où il y aurait un manque de spécificité de H_0 : des hypothèses H_i , autres que H_0 , pourraient être elles qui soient vraies (et non H_0) ;
- la seconde est que le test peut manquer de puissance. Il peut arriver en effet que l'hypothèse H_1 soit vraie, alors que le dispositif expérimental ou le plan échantillonnal ne permettent pas de le détecter.

Bien évidemment, la modalité (3) est fausse selon les deux approches.

La modalité (4) est à son tour fausse selon les deux approches, pour les raisons suivantes :

- Selon Fisher, la conclusion "Rejeter H_0 " ne permet pas d'affirmer que le mécanisme impliqué dans l'hypothèse H_1 est démontré. Les observations peuvent en effet résulter d'un mécanisme d'action différent de celui qui est invoqué dans H_1 . Rejeter H_0 implique seulement que H_1 est acceptable.
- Selon Neyman-Pearson, la décision "Rejeter H_1 " est évidemment différente "d'Echouer de rejeter H_0 ".

(F) Supposons qu'après application du test statistique nous avons décidé d'accepter H_0 . A votre avis, cela veut dire que :

(1) L'acceptation de H_0 est vraie quelle que soit la taille n_0 de l'échantillon dont dérivent les données observées D_0 , et quelle que soit la valeur du seuil α

Vrai Faux

(2) L'acceptation de H_0 dépend uniquement de la taille n_0 de l'échantillon dont dérivent les données observées D_0

Vrai Faux

(3) L'acceptation de H_0 dépend uniquement de la valeur du seuil α

Vrai Faux

(4) L'acceptation de H_0 n'est que provisoire : en fixant le seuil α et en augmentant suffisamment la taille n_0 de l'échantillon dont dérivent les données observées, on finit toujours par rejeter l'hypothèse H_0

Vrai Faux

Nous sommes dans cette question face à un test d'hypothèses, à travers lequel nous devons décider entre H_0 et H_1 . La région de rejet de l'hypothèse H_0 de ce test est

$$RC = \left\{ (X_1, \dots, X_{n_0}) / \bar{X} \in \left[\mu_0 + \Phi_\alpha \frac{\sigma}{\sqrt{n_0}}, +\infty \right) \right\}, \text{ où } \Phi_\alpha \text{ représente le quantile d'ordre}$$

$(1-\alpha)$ de la loi normale centrée réduite $N(0,1)$.

Par ailleurs, la borne $\mu_0 + \Phi_\alpha \frac{\sigma}{\sqrt{n_0}}$ est une fonction de (α, n_0) . Elle diminue, et par

conséquent la région de rejet s'élargit, si nous augmentons la taille n_0 de l'échantillon ou/et si nous augmentons la valeur α du niveau du test. Par conséquent, les modalités (1), (2) et (3) de cet item sont toutes les trois fausses, et seule la modalité (4) est vraie.

(G) Dans cette question, les données observées D_0 et la valeur du seuil de signification α sont fixées. On considère alors les deux tests statistiques suivants :

- Le test statistique n°1 de H_0 : " $\mu = \mu_0$ " contre H_1 : " $\mu = \mu_1$ "

et

- Le test statistique n°2 de H_0 : " $\mu = \mu_1$ " contre H_1 : " $\mu = \mu_0$ "

Parmi les propositions suivantes, celle(s) qui est (sont) vraie (s) est (sont):

(1) Si nous décidons de rejeter " $\mu = \mu_0$ " par le test statistique n°1 alors nous devons décider d'accepter " $\mu = \mu_1$ " par le test statistique n°2

Vrai Faux

(2) Si nous décidons de rejeter “ $\mu = \mu_1$ ” par le test statistique n°1 alors nous devons décider d’accepter “ $\mu = \mu_0$ ” par le test statistique n°2

Vrai Faux

(3) Les hypothèses H_0 et H_1 ne jouent pas un rôle symétrique dans une procédure de test statistique, par conséquent, le résultat fourni par le test statistique n°1 n’est pas le même que celui fourni par le test statistique n°2 Vrai Faux

Dans cette question, nous sommes devant deux sortes de tests d’hypothèses, dans lesquels nous voulons décider entre H_0 et H_1 .

La région de rejet du test d’hypothèses n°1 est :

$$RC_1 = \left\{ (X_1, \dots, X_{n_0}) / \bar{X} \in \left[\mu_0 + \Phi_\alpha \frac{\sigma}{\sqrt{n_0}}, +\infty \right[\right\}, \text{ celle du test d’hypothèses n°2 est :}$$

$$RC_2 = \left\{ (X_1, \dots, X_{n_0}) / \bar{X} \in \left[-\infty, \mu_1 + \Phi_\alpha \frac{\sigma}{\sqrt{n_0}} \right[\right\}, \text{ avec } \mu_1 > \mu_0.$$

Les réponses correctes aux différentes modalités de cet item sont présentées dans le tableau suivant :




	$-\infty$	$\mu_0 + \Phi_\alpha \frac{\sigma}{\sqrt{n_0}}$	$\mu_1 + \Phi_\alpha \frac{\sigma}{\sqrt{n_0}}$	$+\infty$
Décision du test d’hypothèses n°1	Accepter : “ $\mu = \mu_0$ ”	Rejeter : “ $\mu = \mu_0$ ”	Rejeter : “ $\mu = \mu_0$ ”	Rejeter : “ $\mu = \mu_0$ ”
Décision du test d’hypothèses n°2	Rejeter : “ $\mu = \mu_1$ ”	Rejeter : “ $\mu = \mu_1$ ”	Rejeter : “ $\mu = \mu_1$ ”	Accepter : “ $\mu = \mu_1$ ”
	 La modalité (2) est vraie ¹²	 La modalité (1) est fausse  La modalité (3) est fausse		

Tableau 3 : Réponses aux modalités de l’item G.

¹² C’est seulement dans ces deux cas particuliers de tests d’hypothèses -que nous sommes en train de traiter- que la modalité (2) pourrait être considérée comme étant vraie. Mais, d’une manière générale, ceci n’est pas toujours le cas.

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

Les aspects étudiés dans ce questionnaire présentent, de part la nature même du contenu des tests statistiques, des liens conceptuels inhérents aux modèles d'organisation des tests d'hypothèses, notamment pour ce qui concerne les approches de Fisher et de Neyman-Pearson¹³ : à ce propos, nous avons déjà présenté dans le tableau 1 (cf. §II.4 *Interprétation des procédures statistiques*), des exemples de modèles d'organisation qui illustrent bien les relations conceptuelles qui existent entre ces différents aspects.

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

Par ailleurs, pour ne pas biaiser l'analyse, vu le nombre important de modalités (24) de l'ensemble des items (7) du questionnaire, face au nombre « limité » d'étudiants interrogés (70), nous avons décidé de retenir un codage en bimodalité « Réussite - Echec » pour toute modalité de chaque item du questionnaire.

Dans le tableau 4 suivant, nous résumons le codage des modalités des items du questionnaire, les libellés des items, ainsi que la signification associée à chaque modalité :

Libellé	Codage	Signification	RR ¹⁴	RE ¹⁵
Conceptions relatives au rôle d'un test statistique	A1	• <i>Conceptions bayésiennes :</i>		
	A2	- <i>Conception déterministe :</i>		
	A4	• <i>En termes de vérité absolue de H_0.</i>	A1R	A1E
	A5	• <i>En termes de vérité absolue de H_1.</i>	A2R	A2E
	A3	- <i>Conception bayésienne directe.</i>	A4R	A4E
	A6	- <i>Conception bayésienne indirecte.</i>	A5R	A5E
		• <i>Conception Fishérienne.</i>	A3R	A3E
		• <i>Conception Neyman-Pearsonienne.</i>	A6R	A6E

¹³ Ce sont ces deux approches qui nous intéressent le plus, vue que l'approche bayésienne est quasi absente de l'enseignement des tests statistiques au niveau de l'enseignement initial universitaire.

¹⁴ Réponse Réussie.

¹⁵ Réponse Echouée (y compris les Non Réponses).

Conception concernant la signification des hypothèses statistiques H_i	B1 B2	<ul style="list-style-type: none"> • Conception paramétrique : $\theta = \theta_i$. • Conception en termes de modèle probabiliste : $P_0 = P_{0i}$. 	B1R B2R	B1E B2E
Conception sur la p -value exprimée en fonction d'une statistique test	C1 C2 C3	<ul style="list-style-type: none"> • $P(D \geq D_0 H_0)$. • Probabilité inverse : $P(H_0 D_0)$. • Probabilité de la région critique selon Neyman-Pearson. 	C1R C2R C3R	C1E C2E C3E
Localisation du seuil de signification dans une procédure de tests statistiques	D1 D2	<ul style="list-style-type: none"> • Avant collecte des données (approche de Neyman-Pearson). • Après collecte des données (approche de Fisher). 	D1R D2R	D1E D2E
Résultat d'un test statistique	E1 E2 E3 E4	<ul style="list-style-type: none"> • Résultats en conformité avec l'approche de Neyman-Pearson. • Résultats en conformité avec l'approche de Fisher. • Résultats en symétrie totale avec celles de l'approche de Fisher. • Résultats en symétrie partielle avec celles de l'approche de Fisher. 	E1R E2R E3R E4R	E1E E2E E3E E4E
Aspect provisoire du résultat d'un test d'hypothèses : statut du seuil de signification α et de la taille n_0 de l'échantillon face à l'acceptation de H_0	F1 F2 F3 F4	<ul style="list-style-type: none"> • Résultat indépendant de la taille n_0 de l'échantillon et de la valeur seuil α. • Résultat dépendant uniquement de la taille n_0 de l'échantillon. • Résultat dépendant uniquement de la valeur seuil α. • Résultat provisoire, dépendant du seuil α, de la taille n_0 de l'échantillon et des données observées D_0. 	F1R F2R F3R F4R	F1E F2E F3E F4E
Aspect asymétrique du résultat d'un test d'hypothèses	G1 G2 G3	<ul style="list-style-type: none"> • Asymétrie à droite, du résultat : Rejet de H_0 par le test n°1 implique une acceptation de H_0 par le test n° 2. • Asymétrie à gauche, du résultat : Rejet de H_1 par le test n°1 implique une acceptation de H_1 par le test n° 2. • Asymétrie totale du résultat: Résultat du test n°1 est en général différent de celui du test n°2. 	G1R G2R G3R	G1E G2E G3E

Tableau 4 : Codage des modalités de chaque item du questionnaire.

3.4. Analyse et interprétation des réponses des étudiants

3.4.1. Valeurs propres et inertie totale

L'analyse factorielle des correspondances multiples¹⁶ appliquée au tableau disjonctif complet issu du codage des réponses des étudiants, a conduit à des valeurs propres non nulles de moyenne 0,042 (l'inverse du nombre de modalités qui est égal à 24). Par ailleurs, les valeurs propres supérieures à cette moyenne sont : $\lambda_1=0,122$, $\lambda_2=0,095$, $\lambda_3=0,089$, $\lambda_4=0,083$, $\lambda_5=0,069$, $\lambda_6=0,067$, $\lambda_7=0,055$, $\lambda_8=0,051$ et $\lambda_9=0,047$. L'inertie totale est égale à 1 (nombre de valeurs propres non nulles, divisé par le nombre de modalités).

3.4.2. Nombre d'axes retenus dans l'analyse

Nous rappelons tout d'abord que l'inertie totale en analyse factorielle de correspondances multiples appliquée à un tableau disjonctif complet n'a pas de signification statistique¹⁷; 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 neuf valeurs propres ci-dessus conduit à des pourcentages décroissant de 12% à pratiquement 5%, 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 (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 $\lambda_k > \frac{1}{s}$, s étant le nombre de modalités du questionnaire et λ_k les valeurs propres obtenues par l'analyse du tableau disjonctif complet.

L'application de cette formule aux inerties initiales¹⁸, donne pour les valeurs propres supérieures à 0,042 (égale à $\frac{1}{24}$) les résultats suivants :

0,006994, 0,003104, 0,002433, 0,001837, 0,000809, 0,000683, 0,000191,
0,000095 et 0,000033

Le tableau 5 suivant illustre les pourcentages d'inerties corrigées et leurs cumuls, correspondant aux valeurs propres supérieures à la valeur moyenne $\frac{1}{24}$:

¹⁶ Cette analyse a été conduite à l'aide du logiciel Statistica (Version 6).

¹⁷ Ce n'est pas le cas pour le calcul d'inertie en analyse de correspondances simples d'un tableau de contingence.

¹⁸ Dans la suite de l'analyse, nous nous limiterons simplement aux 9 premières valeurs propres.

Valeurs propres	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9
% Inerties corrigées	43%	19%	15%	11%	5%	4%	1%	0,5%	0,2%
Cumul % Inerties corrigées	43%	62%	77%	89%	94%	98%	99%	99,5%	≈ 100%

Tableau 5 : Pourcentages d'inerties corrigées et leurs cumuls.

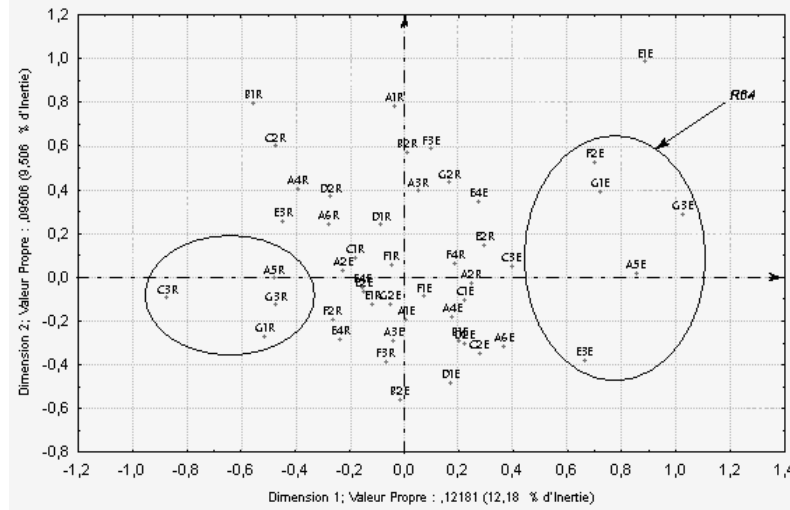
Par conséquent, le premier axe représente 43% de l'information totale, le deuxième axe 19%, le troisième axe 15% et le quatrième axe 11% (les axes suivants chutent à une part d'information représentant moins de la moitié du quatrième axe). Ces valeurs indiquent que l'essentiel de l'information est lié aux quatre premiers axes factoriels qui représentent ensemble 89% de l'information totale.

Par ailleurs, la première valeur propre $\lambda_1=0,122$ est significativement supérieure à $\lambda_2=0,095$. Par contre, $\lambda_2=0,095$, $\lambda_3=0,089$ et $\lambda_4=0,083$ sont très proches ; cela signifie que nous sommes pratiquement en présence d'une valeur propre multiple d'ordre 3. Par conséquent, il va falloir étudier d'une part l'axe 1, et d'autre part le plan formé par les axes 2 et 3 et celui formé par les axes 3 et 4, plutôt que de chercher à donner séparément une interprétation à chacun de ces 3 derniers axes.

3.4.3. Interprétation du premier axe factoriel : (Axe de réussite/Echec)

L'analyse factorielle des correspondances multiples conduit à une inertie¹⁹ de 43% pour l'axe 1. En outre, presque toutes les réussites ont une coordonnée négative sur cet axe, par opposition aux échecs, qui y ont tous une coordonnée positive (voir graphique 1). Par conséquent, l'axe 1 s'interprète comme étant l'axe de *réussite-échec*.

¹⁹Il s'agit de l'inertie corrigée.



Graphique 1 : Plan factoriel (1,2).

Les modalités qui contribuent le plus à l'axe 1 sont celles entourées dans le graphique 1 ci-dessus²⁰. Elles ont une contribution comprise entre 4,6% et 11,3%. La qualité de représentation de ces modalités par rapport à cet axe est comprise entre 0,18 et 0,48. Nous pouvons donc considérer que ce sont les items A5, C3, E3, F2, G1 et G3, correspondant à ces modalités, qui sont les plus représentatifs de l'ensemble du questionnaire.

Premièrement, nous constatons, comme le montre le tableau 6 ci-dessous, que la réussite aux items C3, G1, G3 et A5, et l'échec aux items F2, G1, A5 et G3, sont directement liés au degré d'appréhension du concept de probabilités conditionnelles $P(A|B)$ usuellement impliqué dans une procédure de tests statistiques, où A et B désignent les événements D (Données observées ou données plus extrêmes) ou H (Hypothèses statistiques H_i , avec $i=0$ ou 1).

Item	Éléments d'appréhension du concept de probabilités conditionnelles
C3	<ul style="list-style-type: none"> • La <i>p-value</i>, $P(D H)$, exprimée, selon l'approche de Fisher, à partir d'une statistique test, est une fonction des données observées, et non égale à une valeur fixe.

²⁰ Notons au passage que ces modalités occupent des positions correspondant aux modalités ayant les plus grandes coordonnées en valeur absolue par rapport au 1^{er} axe factoriel. Sinon, le reste des modalités occupent quasiment des positions intermédiaires, avec en outre de faibles contributions relatives.

G1 et G3	<ul style="list-style-type: none"> • $[P(RC H_0) \text{ faible et } P(RC H_1) \text{ élevée}]$ n'est pas toujours équivalente à $[P(RC^c H_1) \text{ faible et } P(RC^c H_0) \text{ élevée}]$; avec RC : Région Critique, RC^c : Complémentaire de RC.
A5	<ul style="list-style-type: none"> • La probabilité conditionnelle qui pourrait être déterminée dans le cadre du paradigme de tests statistiques est $P(D H)$ et non $P(H D)$.
F2	<ul style="list-style-type: none"> • La région critique (RC), déterminée par le biais de la résolution de l'équation : $P(RC H_0) \approx \alpha$, ne dépend pas seulement de la taille n_0 de l'échantillon, mais aussi d'autres facteurs.

Tableau 6 : Éléments d'appréhension du concept de probabilités conditionnelles.

De ce fait, ce facteur s'interprète comme étant « un gradient » d'appréhension du concept de probabilités conditionnelles, impliqué dans une procédure de tests statistiques. En effet, d'une part, cet axe oppose la réussite et l'échec dans la distinction entre le paradigme de tests statistiques et celui de statistiques bayésiennes (cf. item A5 dans le tableau 6) : la probabilité conditionnelle susceptible d'être déterminée dans le cadre des procédures de tests statistiques est $P(D|H)$ et non $P(H|D)$, qui elle, est relative à l'approche bayésienne.

D'autre part, l'opposition des modalités C3R « la valeur de la p-value n'est pas égale à une valeur fixe » et E3E « le résultat fourni par application d'un test statistique est en symétrie totale avec celui fourni par l'application d'un test de signification », permet de conclure que cet axe distingue la réussite de l'échec dans l'appréhension du paradigme de tests de signification (approche de Fisher).

De la même façon, l'opposition de la réussite aux items G1 et G3 « le résultat fourni par l'application d'un test d'hypothèses est caractérisé par un aspect asymétrique », et l'échec à ces items ainsi qu'à l'item F2 « dans un test d'hypothèses, l'acceptation de H_0 dépend uniquement de la taille n_0 de l'échantillon », permet de conclure que cet axe distingue aussi la réussite de l'échec dans l'appréhension du paradigme de tests d'hypothèses (approche de Neyman-Pearson)

Enfin, la réussite ou l'échec²¹ à l'item C3 « le concept de p-value est différent du niveau de signification – seuil – permettant la construction d'une région de rejet », traduit le fait que cet axe oppose aussi la réussite et l'échec dans la distinction entre le paradigme de tests de signification et celui de tests d'hypothèses.

²¹ Noter sur le graphique 1 que la modalité C3E est à proximité du groupe de modalités d'échec ayant fortement contribué à la construction du premier facteur.

Pour confirmer ces interprétations, nous avons calculé les coordonnées par rapport au premier axe de deux individus supplémentaires : le premier individu (de coordonnée **-1,664**) est caractérisé par les modalités C3R, G1R, G3R et A5R. C'est un étudiant ayant une bonne appréhension des aspects relatifs au concept de probabilités conditionnelles, impliqué dans une procédure de tests statistiques, à la distinction entre le paradigme de tests statistiques et celui de statistiques bayésiennes, à l'approche de tests de signification et celle de tests d'hypothèses et à la distinction entre ces deux paradigmes. Le deuxième individu (de coordonnée **1,809**) est caractérisé par les modalités G1E, G3E, A5E, C2E et F2E ; c'est un étudiant ayant une mauvaise appréhension de tous ces aspects.

Par ailleurs, nous avons consulté les copies de tous les étudiants ayant répondu incorrectement aux items A5, E3, G1 et G3, l'étudiant **R64** représente l'individu ayant la plus forte contribution à l'axe 1, avec une coordonnée positive sur cet axe (voir graphique 1) : sur sa copie, nous avons pu relever en particulier les trois erreurs illustrées dans le tableau 7 suivant :

Type d'erreur	Item	Réponse de l'individu R64
<i>Erreur d'ambiguïté linguistique</i>	A5	<ul style="list-style-type: none"> • La probabilité indiquée dans cette question est : "la probabilité que le résultat observé soit dû au seul hasard". Cette probabilité est en fait équivalente à $P(N(0,1) \geq \sqrt{n_o}(\bar{x}_{ob} - \mu_o) / \sigma)$.
<i>Erreur d'interprétation</i>	E3	<ul style="list-style-type: none"> • Le résultat d'un test statistique est : Accepter H_1 ou Echouer de rejeter H_0. Or l'acceptation de H_1 est équivalente à l'échec dans le rejet de H_1, et l'échec dans le rejet de H_0 est équivalent à l'acceptation de H_0 (c'est-à-dire le rejet de H_1).
<i>Erreur de probabilités conditionnelles</i>	G1	<ul style="list-style-type: none"> • Puisque le rejet de H_0 implique l'acceptation de H_1 alors l'item G1 est vrai.
	G3	<ul style="list-style-type: none"> • Puisque G1 et G2 sont tous les deux vrais, alors G3 est faux. Donc, les hypothèses H_0 et H_1 jouent un rôle symétrique dans une procédure de test statistique.

Tableau 7 : Erreurs significatives relevées dans l'interprétation de l'axe 1.

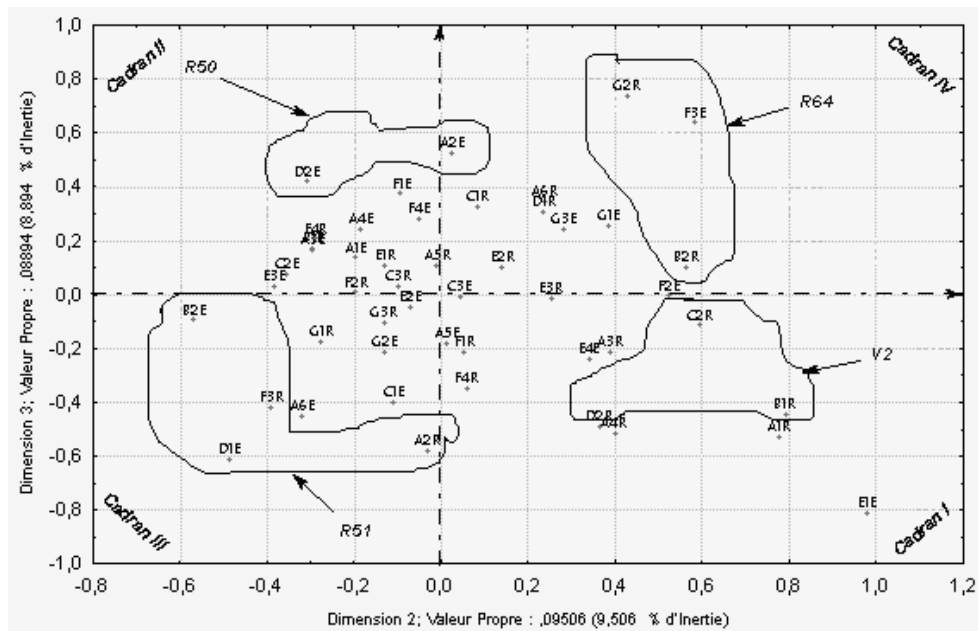
En conclusion, l'axe 1 s'interprète comme étant *un axe de réussite/échec* dans l'appréhension :

- du concept de probabilités conditionnelles impliqué dans une procédure de tests statistiques ;

- de la distinction entre le paradigme de tests statistiques et celui de statistiques bayésiennes ;
- du paradigme de tests de signification (approche de Fisher) ;
- du paradigme de tests d'hypothèses (approche de Neyman-Pearson) ;
- de la distinction entre le paradigme de tests de signification et celui de tests d'hypothèses.

3.4.4. Interprétation du plan (2,3) : (Plan de maîtrise des tests statistiques)

L'analyse factorielle des correspondances multiples conduit à des inerties²² de 19% pour l'axe 2 et de 15% pour l'axe 3. Le graphique 2 suivant représente la projection de l'ensemble des modalités par rapport au plan (2,3) :



Graphique 2 : Plan factoriel (2,3).

Les modalités correspondant aux plus fortes contributions relatives (comprises entre 3,4% et 7,4%) dans la construction des axes 2 et 3, et ayant une bonne représentation, sont celles qui sont entourées dans le graphique 2.

En cherchant parmi ces modalités les groupements qui présentent les plus fortes oppositions, nous constatons qu'il y a quatre groupes de modalités qui s'opposent

²²Il s'agit d'inerties corrigées.

deux à deux²³ : d'une part B1R, C2R et D2R contre A2E et D2E, et d'autre part A2R, B2E, D1E et F3R contre B2R, F3E et G2R :

-Cadran I vs Cadran II

L'opposition de la réussite aux items B1, C2 et D2 face à l'échec aux items A2 et D2 traduit l'opposition d'une bonne maîtrise de l'approche de Fisher (C2R et D2R) favorisant le non déterminisme (B1R) d'un test statistique face à une mauvaise maîtrise de cette approche (D2E) favorisant le déterminisme (A2E) d'un test statistique.

Pour confirmer cette interprétation, nous avons consulté les réponses de l'individu **R50** (voir graphique 2) qui a la plus forte contribution par rapport à l'axe 1, parmi les individus ayant une forte contribution par rapport à l'axe 2 et/ou 3 et se trouvant dans le cadran II, chez qui nous avons relevé les erreurs illustrées dans le tableau 8 ci-dessous :

Type d'erreur	Item	Réponse de l'individu R50
<i>Erreur de déterminisme</i>	A2	<i>Dans une procédure de tests statistiques, le rejet de H_0 est équivalent à la probabilité $P(H_0 D_0)$ est égale à 0, c'est-à-dire que $P(H_1 D_0)$ est égale à 1. Ce qui signifie que "$\mu=\mu_1$".</i>
<i>Erreur procédurale</i>	D2	<i>Comparer $P(N(0,1) \geq \sqrt{n_o}(\bar{x}_{ob} - \mu_o) / \sigma)$ à α n'est pas suffisant pour conclure.</i>

Tableau 8 : Erreurs significatives relevées dans l'interprétation de l'opposition des cadrans I et II du plan (2,3).

Nous pouvons donc conclure que les cadrans I et II du plan (2,3) opposent *des erreurs procédurales favorisant le déterminisme dans l'approche des tests statistiques* contre *une bonne maîtrise de l'approche de Fisher favorisant une procédure non déterministe*.

- Cadran III vs Cadran IV

Ces deux cadrans opposent l'échec aux items B2 et D1 et la réussite aux items A2 et F3, contre la réussite aux items B2 et G2 et l'échec à l'item F3 : il s'agit ici d'une opposition entre une maîtrise partielle (A2R, F3R, B2E et D1E) des tests d'hypothèses (approche de Neyman-Pearson), contre une maîtrise de ces tests (B2R et G2R) accompagnée d'un « automatisme procédural » (F3E).

²³ Voir Cadran I, Cadran II, Cadran III et Cadran IV du graphique 2.

En effet, les réponses de l'individu **R51** (voir graphique 2) qui a la plus forte contribution par rapport à l'axe 1, parmi les individus ayant une forte contribution par rapport à l'axe 2 et/ou 3 et se trouvant dans le cadran III, vont dans le sens de cette interprétation (voir tableau 9) :

Type d'erreur	Item	Réponse de l'individu R51
<i>Erreur implicite</i>	B2	H_i ne signifie pas que $X \sim N(\mu_i, \sigma^2)$ parce que, par hypothèse, X suit déjà $N(\mu, \sigma^2)$ et nous avons $\mu \neq \mu_i$.
<i>Erreur d'amalgame</i>	D1	La valeur de α est directement comparée à $P(N(0,1) \geq \sqrt{n_o}(\bar{x}_{ob} - \mu_o)/\sigma)$ pour conclure au rejet ou à l'échec dans le rejet de H_0 .

Tableau 9 : Erreurs significatives relevées dans l'interprétation du cadran III du plan (2,3).

L'individu **R64** (voir graphique 2) qui a la plus forte contribution par rapport à l'axe 1, parmi les individus ayant une forte contribution par rapport à l'axe 2 et/ou 3 et se trouvant dans le cadran IV, a échoué à l'item F3. Sa réponse à cet item a permis d'identifier une *erreur d'automatisme* : "L'acceptation de H_0 dépend uniquement de α puisque à cette seule valeur (α) qu'on compare la probabilité $P(N(0,1) \geq \sqrt{n_o}(\bar{x}_{ob} - \mu_o)/\sigma)$ pour conclure, et non à d'autres valeurs".

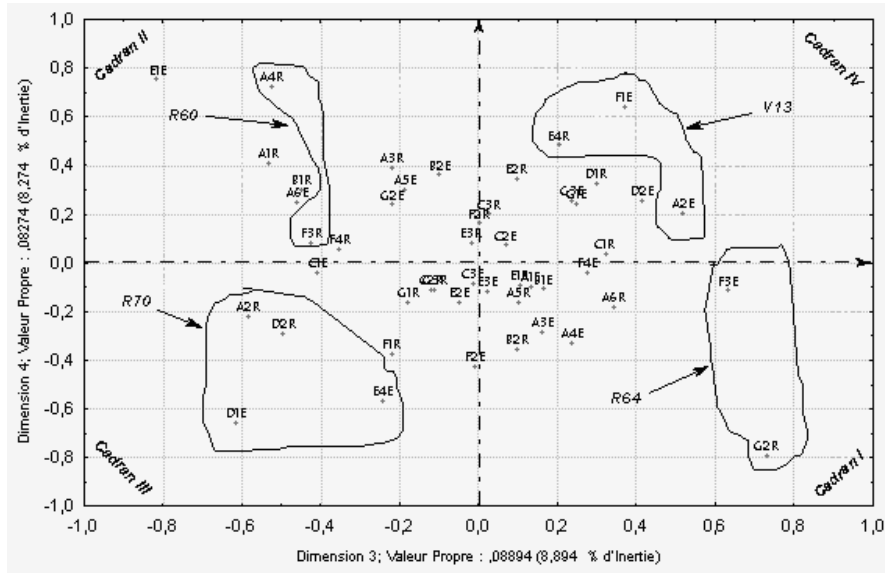
Ainsi, nous pouvons conclure que les cadrans III et IV du plan (2,3) opposent *des erreurs d'amalgame et d'implicite à une erreur d'automatisme procédural*.

En conclusion, le plan (2,3) est *un plan de maîtrise des tests statistiques* (degré d'appréhension des tests de signification et des tests d'hypothèses) qui se traduit par :

- *l'opposition d'une bonne maîtrise de l'approche de Fisher favorisant une procédure non déterministe contre une mauvaise maîtrise de cette approche favorisant un déterminisme procédural (CADRAN I vs CADRAN II) ;*
- *l'opposition d'une maîtrise partielle de l'approche de Neyman-Pearson contre une maîtrise de cette approche accompagnée d'automatisme procédural (CADRAN III vs CADRAN IV).*

3.4.5. Interprétation du plan (3,4) : (Plan de distinction entre les trois paradigmes)

L'analyse factorielle conduit à une inertie²⁴ de 15% pour l'axe 3 et de 11% pour l'axe 4. Le graphique 3 suivant représente la projection des modalités par rapport au plan (3,4) :



Graphique 3 : Plan factoriel (3,4).

Les modalités correspondant aux plus fortes contributions relatives (comprises entre 5% et 8,2%) dans la construction des axes 3 et 4, et ayant une bonne représentation, sont celles qui sont entourées dans le graphique 3.

Ces modalités, distribuées en quatre regroupements (voir cadrans de I à IV, graphique 3), donnent lieu dans le plan factoriel (3,4) à deux oppositions :

- Cadran I vs Cadran II

L'opposition de l'échec à l'item F3 et de la réussite à l'item G2, contre la réussite aux items A4 et F3 traduit l'opposition d'une maîtrise partielle de l'approche de tests d'hypothèses (G2R et F3E) contre une maîtrise de cette approche (F3R), avec la distinction entre le paradigme de tests statistiques et celui de statistiques bayésiennes.

En allant consulter la réponse (incorrecte) à l'item F3 de l'individu **R64** (voir graphique 3) qui a la plus forte contribution par rapport à l'axe 1, parmi les

²⁴Il s'agit de l'inertie corrigée.

individus ayant une contribution très élevée par rapport à l'axe 3 et/ou 4 et se trouvant dans le cadran I, nous avons repéré une *erreur d'automatisme procédural*, exprimée de la façon suivante :

“L'acceptation de H_0 dépend uniquement de α puisque à cette seule valeur (α) qu'on compare la probabilité $P(N(0,1) \geq \sqrt{n_o}(\bar{x}_{ob} - \mu_o) / \sigma)$ pour conclure, et non à d'autres valeurs”

Nous pouvons donc conclure que les cadrans I et II du plan (3,4) opposent *une erreur d'automatisme procédural contre une maîtrise de l'approche de tests d'hypothèses accompagnée d'une distinction entre le paradigme de tests statistiques et celui de statistiques bayésiennes.*

- Cadran III vs Cadran IV

L'opposition de l'échec aux items D1 et E4 et de la réussite aux items A2 et D2, contre la réussite à l'item E4 et l'échec aux items A2 et F1, traduit l'opposition d'une maîtrise partielle de l'approche de tests de signification (D2R et E4E), accompagnée d'une procédure non déterministe du paradigme de tests statistiques (A2R), contre une non maîtrise de l'approche de tests d'hypothèses (D1E), et d'une maîtrise partielle de l'approche de tests de signification (E2R) accompagnée de procédure déterministe (A2E et F1E).

Les réponses (incorrectes) aux items D1 et E4 de l'individu **R70** (voir graphique 3) qui a la plus forte contribution par rapport à l'axe 1, parmi les individus ayant une contribution très élevée par rapport à l'axe 3 et/ou 4 et se trouvant dans le cadran III, a permis de repérer une *erreur d'amalgame* illustrée dans le tableau 10 suivant :

Item	Réponse de l'individu R70
D1	<i>α ne participe pas à la construction d'une règle de décision. Mais, nous devons la comparer à la probabilité $P(N(0,1) \geq \sqrt{n_o}(\bar{x}_{ob} - \mu_o) / \sigma)$ pour conclure.</i>
E4	<i>Le résultat d'un test statistique est soit rejeter H_0 soit accepter H_0. Or le rejet de H_0 est équivalent à l'acceptation de H_1, et l'acceptation de H_0 est équivalente à l'échec dans le rejet de H_0.</i>

Tableau 10 : Erreurs significatives relevées dans l'interprétation du cadran III du plan (3,4).

Par ailleurs, en consultant les réponses (incorrectes) aux items A2 et F1 de l'individu **V13** (voir graphique 3) qui a la plus forte contribution par rapport à l'axe 1, parmi les individus ayant une contribution très élevée par rapport à l'axe 3

et/ou 4 et se trouvant dans le cadran IV, nous avons repéré une *erreur procédurale de déterminisme* illustrée dans le tableau 11 suivant :

<i>Item</i>	<i>Réponse de l'individu R70</i>
A2	<i>Un test statistique sert à démontrer que "$\mu=\mu_0$" ou que "$\mu=\mu_1$".</i>
F1	<i>Le résultat d'un test statistique est de démontrer que "$\mu=\mu_0$" ou que "$\mu=\mu_1$" indépendamment de la taille de l'échantillon n_0 et du seuil α.</i>

Tableau 11 : Erreurs significatives relevées dans l'interprétation du cadran IV du plan (3,4).

Nous pouvons donc conclure que les cadrans III et IV du plan (3,4) opposent une *erreur d'amalgame* contre une *erreur de déterminisme procédural*.

En résumé, le plan (3,4) est un *plan de distinction entre les trois paradigmes* : tests de signification, tests d'hypothèses et statistiques bayésiennes. Cela se traduit par :

- *l'opposition d'une maîtrise partielle de l'approche de tests d'hypothèses contre une maîtrise de cette approche favorisant la distinction entre le paradigme de tests statistiques et celui de statistiques bayésiennes (CADRAN I vs CADRAN II) ;*
- *l'opposition d'une maîtrise partielle de l'approche de tests de signification et d'une non maîtrise de celle de tests d'hypothèses, accompagnées de procédure non déterministe, contre une maîtrise de l'approche de tests de signification, accompagnée de procédure déterministe (CADRAN III vs CADRAN IV).*

4. Conclusions et perspectives de recherche

Les tests statistiques occupent une place importante dans l'inférence statistique, cependant leur enseignement est souvent réduit à une simple présentation de recettes, que les étudiants s'approprient et restituent de manière mécanique, sans que cela représente en soi un réel objectif. En fait, l'intérêt de l'enseignement des tests statistiques, est avant tout de développer chez les étudiants *une façon de penser* pour répondre à des questions d'inférence statistique, tout en leur fournissant les outils mathématiques qui permettent d'atteindre cet objectif.

Les tests statistiques cherchent à évaluer la défaillance qui existe entre hypothèse(s) H et données D_0 observées. Les deux paradigmes développés pour répondre à cette question font référence aux probabilités conditionnelles $P(D_0|H)$ et $P(H|D_0)$, pour donner lieu respectivement aux deux approches suivantes :

- l'approche fréquentiste, qui est à la base des tests statistiques et dont le développement a donné lieu à deux types de procédures, les tests de signification de Fisher et les tests d'hypothèses de Neyman-Pearson ;
- l'approche subjectiviste, qui a donné lieu aux statistiques bayésiennes qui ont aussi apporté des solutions aux tests statistiques.

La présentation de ces deux approches pour l'introduction des tests statistiques peut s'avérer bénéfique auprès des étudiants, elle peut jouer le rôle de support de base d'interprétations, en particulier pour ce qui est de :

- une bonne interprétation des procédures mises en jeu, elle permet d'éviter par exemples des confusions dues à l'inversion des conditions des probabilités conditionnelles mises en jeu dans cette approche : *approche probabiliste erronée* (cf. tableaux 6 et 7 dans l'interprétation du premier axe factoriel), ou encore des conclusions erronées à propos de la p-value (cf. tableau 6 dans l'interprétation du premier axe factoriel et tableau 10 dans l'interprétation du plan factoriel (3,4)), qui peuvent aussi se traduire en termes d'importance (ou non) de l'hypothèse (scientifique) étudiée ;
- une interprétation des hypothèses statistiques en termes de modèles statistiques, permettant d'éviter entre autres chez les étudiants de privilégier une *approche déterministe* dans leur interprétation de la procédure d'un test statistique (cf. tableau 11 dans l'interprétation du plan factoriel (3,4)).

L'enseignement des tests statistiques peut conduire à un amalgame entre les deux approches de Fisher et de Neyman-Pearson (cf. tableau 6 dans l'interprétation du plan factoriel (2,3)). Cet amalgame peut ne pas être apparent dans beaucoup de situations classiquement traitées auprès des étudiants, comme c'est le cas pour certains tests paramétriques relevant de modèles continus. En effet, dans ce type de situations, les deux approches utilisent une *procédure hybride* dans la recherche d'indicateurs de jugement. Néanmoins, les deux approches renvoient à deux logiques fondamentalement différentes au niveau de leurs conclusions, qu'il est important de souligner auprès des étudiants à travers des exemples appropriés.

Les tests d'hypothèses statistiques de Neyman-Pearson, largement enseignés à l'université, sont construits à partir de procédures spécifiques mettant en avant deux risques, un de première espèce et un de seconde espèce (puissance), ne faisant pas jouer aux deux hypothèses mises en jeu dans le test un rôle symétrique. Il est alors important de souligner auprès des étudiants cette question de dissymétrie (cf. tableau 7 dans l'interprétation du premier axe factoriel), sous ses divers aspects, à travers divers exemples, et d'en expliquer l'effet et les limites dans l'élaboration des tests chez Neyman-Pearson.

Nous avons présenté ici les diverses approches élaborées pour traiter des problèmes de tests statistiques, en soulevant quelques difficultés et fausses interprétations que l'on pourrait préconiser chez les étudiants en situation d'enseignement, d'ailleurs, à ce propos, les résultats d'analyse de productions des étudiants que nous avons interrogés le confirment largement. En fait, ce travail constitue une étape fondamentale de notre recherche, il représente une analyse a priori importante en vue d'une élaboration de séquences d'enseignement sur les tests d'hypothèses statistiques, que nous prévoyons expérimenter auprès d'étudiants de licences de filières scientifiques à l'université²⁵. En outre, nous pensons aussi intégrer lors de ces séquences d'enseignement un logiciel²⁶ (tableur numérique) permettant aux étudiants de traiter diverses situations de tests statistiques, tout en faisant varier les paramètres des situations étudiées, base de données, tailles des échantillons, seuil α de signification, etc.

L'objectif de notre recherche est d'étudier l'effet d'un enseignement de tests statistiques pour lequel on aura introduit deux actions de natures différentes :

- la première est qualitative, il s'agit de tenir compte des éléments mis en avant par l'analyse a priori précédente, ainsi que des résultats de l'analyse des productions des étudiants interrogés.
- la seconde est quantitative, elle concerne l'introduction d'un environnement informatique permettant des traitements de situations de tests statistiques, avec la possibilité de variation des paramètres qui y interviennent.

A l'issue de ce travail, nous pensons être en mesure de fournir des propositions concrètes pour l'élaboration d'ingénieries didactiques sur les tests statistiques.

²⁵ Le Maroc a adopté depuis 2003 pour l'université le système LMD, avec une organisation semestrielle des enseignements.

²⁶ En nous inspirant de Excel 2000 (J.-P. Georgin et M. Gouet, 2000), nous allons développer notre propre tableur sur les tests statistiques.

Bibliographie

- AMERICAN PSYCHOLOGICAL ASSOCIATION (1994), *Publication manual of the American Psychological Association* (4th ed.), Washington, DC: Author.
- 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.
- BEN-ZVI D. & GARFIELD G. (2004), "Statistical Literacy, Reasoning, and Thinking: Goals, Definitions, and Challenges" in D. Ben-Zvi & G. Garfield (eds), *The Challenge of Developing Statistical Literacy. Reasoning and Thinking*, Dordrecht: 3–16.
- CARVER R.P. (1978), The case against significance testing. *Harvard Educational Review*, **48**, 378–399.
- CASTRO SOTOS, A.E., VANHOOF, S., VAN DEN NOORTGATE, W. & ONGHENA, P. (2007), "Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education", *Educational Research Review*, **2**, 98–113.
- COBB, G. & MOORE, D. (1997), "Mathematics, Statistics, and Teaching", *The American Mathematical Monthly*, **104(9)**, 801–823.
- EDWARDS, W., LINDMAN, H. & SAVAGE, L.J. (1963), Bayesian statistical inference for psychological research, *Psychological Review*, **70**, 193–242.
- FALK, R. & GREENBAUM, C.W. (1995), Significance tests die hard. The amazing persistence of probabilistic misconception, *Theory and Psychology*, **5**, 75–98.
- FISHER, R. A. (1935), *The design of experiments*, Oliver and Boyd, Edinburgh.
- FISHER, R. A. (1956), *Statistical methods and scientific inference*, Oliver and Boyd, Edinburgh.
- GEORGIN, J.-P. & GOUET, M. (2000), *Statistiques avec Excel 2000*, Eyrolles, Paris.
- 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 behavioural sciences: Methodological issues*. Hillsdale, NJ: Erlbaum.

- GLINER, J. A., LEECH, N.L. & MORGAN, G.A. (2002), Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education*, **71**(1), 83–92.
- GORDON, H.R.D. (2001), American Vocational Education Research Association members' perceptions of statistical significance tests and other statistical controversies, *Journal of Vocational Education Research*, **26**(2).
- HALLER, H. & KRAUSS, S. (2002), Misconceptions of significance: A problem students share with their teachers? *Methods of Psychological Research*, **7**(1).
- HARLOW, L.L., MULAİK, S.A. & STEIGER, J.H. (Eds.) (1997), *What if there were no significance tests?* London: Lawrence Erlbaum Associates, Publishers.
- HOWSON, C. & URBACH, P. (1989), *Scientific Reasoning, The Bayesian Approach* Open Court Publishing Company.
- HUBBARD, R. & LINDSAY, R.M. (2008), "Why P Values are not a useful measure of evidence in statistical significance testing", *Theory & Psychology*, **18**(1), 69–88.
- KERLINGER, F.N. (1979), *Behavioral research: a conceptual approach*, Holt, Rinehart and Winston, New York.
- KLINE, R.B. (2004), *Beyond significance testing: Reforming data analysis methods in behavioral research*, Washington, DC: American Psychological Association.
- KRANTZ, D.H. (1999), The null hypothesis testing controversy in psychology, *Journal of the American Statistical Association*, **44**, 1372–1381.
- LECOUTRE, M.P., POITEVINEAU, J. & LECOUTRE, B. (2003), Even statisticians are not immune to misinterpretations of null hypothesis significance tests, *International Journal of Psychology*, **38**(1), 37–45.
- MCLEAN, A.L. (2000), On the nature and role of hypothesis tests. *Department of Econometrics and Business Statistics Working Paper 4/2001*. Available at: <http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/2001/wp4-01.pdf>.
- MITTAG, K.C. & THOMPSON, B. (2000), A national survey of AERA members' perceptions of statistical significance tests and other statistical issues, *Educational Researcher*, **29**, 14–20.
- NELDER, J.A. (1999), " Statistics for the Millennium: From statistics to statistical science" (with comments), *The Statistician*, **48**, 257–269.
- NELSON, N., ROSENTHAL, R. & ROSNOW, R.L. (1986), Interpretation of significance levels and effect sizes by psychological researchers, *American Psychologist*, **41**, 1299–1301.

- NEYMAN, J. & PEARSON, E.S. (1933), On the problem of the most efficient tests of statistical hypotheses, *Phil. Trans.* **A 231**, 289–337.
- OAKES, M. (1986), *Statistical inference: a commentary for the social and behavioural sciences*, Willey, New York.
- PEDHAZUR, E.J. & SCHMELKIN, L.P. (1991), *Measurement, design and analysis: an integrated approach*, Hillsdale, NJ, Erlbaum.
- PFANNKUCH, M. & Wild, C. (2004), "Towards an Understanding of Statistical Thinking" in D. Ben-Zvi & G. Garfield (eds), *The Challenge of Developing Statistical Literacy. Reasoning and Thinking*, Dordrecht: 17–46.
- 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*. Thèse de Doctorat. Université de Rouen.
- POITEVINEAU, J. & LECOUTRE, B. (2001), The interpretation of significance levels by psychological researchers: The .05-cliff effect may be overstated, *Psychonomic Bulletin and Review*, **8**, 847–850.
- ROBERT, C. (1992), *L'analyse statistique bayésienne*, Economica, Paris.
- ROSENTHAL, R. & GAITO, J. (1963), The interpretation of significance levels by psychological researchers, *Journal of Psychology*, **55**, 33–38.
- VACHA-HAASE, T., NILSSON, J. E., REETZ, D.R., LANCE, T.S. & THOMPSON, B. (2000), Reporting practices and APA editorial policies regarding statistical significance and effect size, *Theory & Psychology*, **10**, 413–425.
- 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.
- VALLECILLOS, A. (1996), Students' conceptions of the logic of hypothesis testing, *Hiroshima Journal of Mathematics Education*, **4**, 43–61.
- WILKINSON, L. & TASK FORCE ON STATISTICAL INFERENCE (1999), Statistical methods in psychological journal, *American Psychologist*, **54**, 594–604.
- ZENDRERA, N. (2003), Difficultés d'apprentissage liées aux tests statistiques. Le cas des tests paramétriques auprès des étudiants en sciences humaines, *Actes des XXXI^{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.

MONCEF ZAKI

Laboratoire Interdisciplinaire de Recherche en
Didactique des Sciences et Techniques (LIRDIST)
Université Sidi Mohammed Ben Abdellah
Département de Mathématiques
B.P. 1796 Fès-Atlas 30 000, Maroc
zaki.moncef@yahoo.fr

ZAHID ELM'HAMEDI

Etudiant de Doctorat à l'UFR de
Didactique des Mathématiques et de la Physique
Université Sidi Mohammed Ben Abdellah de Fès, Maroc
Département de Mathématiques
B.P. 1796 Fès-Atlas 30 000, Maroc