**MONCEF ZAKI ET ZAHID EL M'HAMEDI**

# ASPECTS DE QUELQUES CRITIQUES NON FONDEES DE LA THEORIE DES TESTS STATISTIQUES

**Abstract. Aspects of some no founded critics of statistical tests theory**
Many critics have been issued since 1960 about significance testing theory of Fisher and hypothesis testing theory of Neyman-Pearson, notably in the domains of sociology and psychology. The controversy about the statistical tests has essentially bearing upon, on its use and on the interpretation of results that will generate. Some authors and professional associations proposed the gap of this concept in the teaching or at least be going with others inferential methods. We treat this subject in a purely didactic viewpoint, by analyzing the essential of these critics in the context of teaching these theories. One a priori analysis of these critics has allowed us to isolate many elements that demonstrate the no soundness. So, the exploratory study that we have led with students specialised in statistics, have involved some results that corroborate, if not confirm, the pertinence of the elements of our starting hypothesis «the majority of these critics must rather been imputed to one no mastery of statistical test concept by some users and to underlying interpretative abuses, and not to the considerations intrinsically linked to the concept of statistical test or to its teaching », that represent a plea of the statistical tests and of its teaching.

**Résumé.** Plusieurs critiques ont été émises à propos de la théorie des tests de signification de Fisher et celle des tests d'hypothèses de Neyman-Pearson, notamment dans les domaines de la sociologie et de la psychologie, et ce dès les années 1960. La controverse à propos des tests statistiques a porté essentiellement sur son utilisation et sur les interprétations qui en découlent. Certains auteurs et associations professionnelles en sont allés jusqu'à vouloir proposer la mise à l'écart de la théorie des tests statistiques de l'enseignement, ou du moins son accompagnement par d'autres méthodes inférentielles. Pour notre part, nous abordons ce sujet d'un point de vue purement didactique, en analysant l'essentiel de ces critiques dans le contexte d'enseignement de ces théories. Une analyse a priori de ces critiques nous a permis de dégager plusieurs éléments qui en démontrent le non bien fondé. Par ailleurs, des observations que nous avons conduites auprès d'étudiants se spécialisant en statistique, ont donné lieu à des résultats qui corroborent, sinon confirment, la pertinence des éléments de notre analyse a priori en faveur des tests statistiques. Ainsi, ce travail nous a permis de valider notre hypothèse de départ « la majorité de ces critiques devrait plutôt être imputée à une non maîtrise du concept de tests statistiques par certains utilisateurs et aux abus interprétatifs qui en découlent, et non pas à des considérations intrinsèquement liées au concept même des tests statistiques ou à leur enseignement », ce qui représente un plaidoyer de taille pour les tests statistiques et leur enseignement.

**Mots-clés :** Enseignement des tests statistiques, Test de signification, Test d'hypothèses, Approche Bayésienne, Controverse, Critiques non fondées

---

**ANNALES de DIDACTIQUE et de SCIENCES COGNITIVES**, volume 18, p. 139-172.

## 1. Introduction

Le concept de test statistique est la pierre angulaire de recherches empiriques dans des domaines aussi variés que l'éducation, la psychologie, la sociologie, la médecine, l'agronomie, l'écologie, le droit ou l'économétrie. Contrairement aux autres outils de l'inférence statistique, ce concept a fait l'objet de multiples controverses et a suscité beaucoup de débats [Jones, 1955 ; Rozeboom, 1960 ; Bakan, 1960 ; Morrisson et Henkel, 1970 ; Meehl, 1978; Oakes, 1986; Cohen, 1994; Krantz, 1999; Wilkinson, 1999; Batanero, 2000; Nickerson, 2000; Ben-Zvi et Garfieled, 2004; Pfannkuch et Wild, 2004; Batanero et Diaz, 2006; Hubbard et Lindsay, 2008;…]. Ces controverses ont essentiellement porté sur sa procédure d'utilisation par les chercheurs et praticiens, ainsi que sur l'interprétation des résultats qui en découlent [Denis, 2004]. Certains auteurs et associations professionnelles [Kline, 2004 ; American Psychological Association, 1994] en sont allés jusqu'au point de vouloir proposer la pure et simple exclusion des tests statistiques de l'enseignement, ou du moins leur accompagnement d'autres méthodes statistiques inférentielles tels les intervalles de confiance, la taille de l'effet, la réplication ou encore la statistique bayésienne. A ce propos, nous citerons Kline [2004] «*I argue that the criticism have sufficient merit to support the minimization or elimination of NHST in the behavioral sciences, …*», ou bien Hager [2000] «*As a consequence of this critique, it is recommended that statistical tests should be banned from science, or replaced by other methods, or supplemented either by other methods or judgment* …», ou encore Brandstätter et Kepler [1999] «*In order to avoid the problems posed by significance tests, various methods like graphic data analyses [Cohen, 1994; Tukey, 1977], meta-analyses [Schmidt, 1996], replications of studies, and confidence intervals [Cohen, 1994; Sedlmeier, 1996] have been proposed as alternatives to significance testing*». Cependant, et malgré les prétendus défauts décelés dans l'application de ce concept, ce dernier a été largement utilisé par le passé, continue de l'être aujourd'hui, et le restera certainement pour plusieurs années encore.

Les critiques avancées à l'encontre des tests statistiques ne peuvent être fondées que si celles-ci sont d'abord présentées dans un contexte d'utilisation bien déterminé. En effet, la méthodologie statistique dans laquelle va s'inscrire un test statistique est intimement liée à la discipline dans laquelle ce dernier sera utilisé. La formulation même des hypothèses va dépendre de la spécificité de la discipline et du modèle théorique qu'elle met en jeu, et ce n'est qu'à partir de cette formulation que l'on va convenir d'un choix pertinent d'une procédure de test statistique et de ses éléments de jugement, qui seront à la base des conclusions et interprétations. Enfin, il y va de soi qu'une bonne maîtrise des raisonnements et procédures sous-jacents aux tests statistiques joue un rôle très important pour un choix approprié et justifié de cet outil inférentiel.

Pour ce qui concerne cette étude, nous nous intéresserons aux tests statistiques dans leur contexte d'enseignement, domaine qui reste peu exploré dans les recherches en didactique des mathématiques, comme le souligne à juste titre Carranza [2011] « *... notre travail révèle les manques dans les recherches en didactique des mathématiques de travaux sur une notion aussi fondamentale que l'inférence statistique. De tels travaux devraient prendre plus en compte le rôle des interprétations tant pour l'architecture d'un test, que pour son interprétation* ». Ainsi, nous analyserons quelques critiques à l'encontre des tests statistiques relevées dans la littérature, que nous confronterons aux résultats d'observations auprès d'étudiants sur leurs approches conceptuelles relatives aux raisonnements et procédures sous-jacents aux tests statistiques, afin de questionner dans le contexte d'enseignement le bien-fondé de ces critiques.

Ces critiques sont généralement adressées aux utilisateurs de tests statistiques, et portent essentiellement sur les procédures utilisées et les interprétations qui en découlent. Notre hypothèse est que la majorité de ces critiques devrait être plutôt imputée à une non maîtrise du concept de tests statistiques par certains utilisateurs et aux abus interprétatifs qui en découlent, et non pas à des considérations intrinsèquement liées au concept même des tests statistiques ou à leur enseignement ; auquel cas, ces critiques s'avéreront non fondées.

Certes, certaines études ont révélé quelques difficultés conceptuelles relatives aux tests statistiques auxquels sont confrontés les étudiants [Zendrera, 2010], notamment autour de l'amalgame entre les *tests de signification* développés par Fisher[1] et les *tests d'hypothèses* (ou de décision) formalisés par les statisticiens Neyman[2] et Pearson[3] que certains auteurs [Gigerenzer, 1993] ont qualifié de « logique hybride de l'inférence scientifique » ; néanmoins, ces difficultés peuvent être surmontées en préconisant des éléments de mesures adéquats pour l'enseignement statistique, en l'occurrence à l'aide d'exemples pertinents qui appuient la différence de procédures chez Fisher et Neyman-Pearson [Zaki et El M'Hamedi, 2009 ; El M'Hamedi, 2010].

Dans cet article, nous présenterons une analyse *a priori* de l'essentiel des critiques s'opposant à l'utilisation du concept de test statistique[4], et que nous jugeons *non fondées*. Ensuite, nous présenterons les résultats d'observations menées auprès d'étudiants, qui vont appuyer notre analyse. Mais auparavant, nous donnerons un

[1]Sir Ronald Aylmer Fisher : (1890 - 1962), statisticien anglais.
[2]Jerzy Neyman : (1894 - 1981), statisticien et mathématicien polonais.
[3]Egon Pearson : (1895 - 1980), statisticien anglais, fils de Karl Pearson (1857 – 1936).
[4]Tout au long de cet article, l'expression ''test statistique'' représente à la fois le concept de test de signification de Fisher et celui de test d'hypothèses de Neyman-Pearson.

bref aperçu sur les procédures d'application des concepts de test de signification et de test d'hypothèses, moyennant une situation de test statistique usuellement utilisée dans l'enseignement. L'objectif en est, d'une part de mettre en évidence ces procédures, et d'autre part de rappeler quelques notions qui seront évoquées dans l'analyse des critiques.

## 2. La situation du test de conformité selon l'approche de Fisher et celle de Neyman-Pearson

Fisher [1925, 1935, 1956] et Neyman-Pearson [1928, 1933] sont à l'origine d'ouvrages et d'articles fondateurs des théories relatives aux tests de signification et aux tests d'hypothèses. Plus récemment, Carver [1953] et Poitevineau [1998] ont présenté les tests statistiques d'une manière plus simple et facile à comprendre.

Néanmoins, nous allons ci-après nous baser sur la situation de tests statistiques (dite de *conformité*) pour esquisser un bref aperçu sur les procédures sous-jacentes à ces deux types de tests, sachant que nous allons donner quelques éclaircissements dans la partie relative aux critiques non fondées que nous développerons plus loin. La situation est la suivante :

Soit X une variable aléatoire qui suit une loi normale de moyenne $\mu$ et d'écart type connu $\sigma$ :

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

Nous considérons les deux hypothèses statistiques suivantes $H_0$ et $H_1$ relatives à la moyenne $\mu$ de X, et auxquelles nous voulons appliquer ***un test statistique***[5] :

$$\begin{cases} H_0 : \text{“}\mu \leq \mu_0\text{”} \\ H_1 : \text{“}\mu > \mu_0\text{”} \end{cases}$$

Pour cela, nous avons recueilli à l'aide d'un échantillon $(X_1,...,X_{n_0})$ de taille $n_0$ et d'une épreuve (ou réalisation) $w_0$, les données $D_0$ : $(x_1^0,...,x_{n_0}^0)$ telles que

$\bar{x}_{ob} = \dfrac{x_1^0 + ... + x_{n0}^0}{n_0}$ (moyenne *ob*servée).

$\overline{X} = \dfrac{X_1 + .... + X_{n_0}}{n_0}$ désigne la moyenne d'échantillon $(X_1,...,X_{n_0})$ issu de la variable aléatoire X.

Tableau 1. Situation du test de conformité

[5] C'est-à-dire : appliquer un test de signification et un test d'hypothèses.

### 2.1. Tests de signification au sens de Fisher

Le rôle d'un test de signification est de *conclure,* avec une forte probabilité, si les données observées $D_0$ contredisent l'hypothèse nulle $H_0$. Les deux conclusions possibles sont "Oui" ou "Non" :

- Oui: $H_0$ est fausse (ou Rejeter $H_0$). On dit aussi que les données observées $D_0$ sont statistiquement significatives.

- Non : Echec dans le rejet de $H_0$ (ou les données $D_0$ sont statistiquement non significatives).

Pour ce faire, nous déterminons la probabilité $P(\overline{X} \geq \overline{x}_{ob} \mid H_0)$[6], appelée "p-value". Elle est égale à $P\left[ Y \geq \frac{\sqrt{n_0}\left(\overline{x_{ob}} - \mu_0\right)}{\sigma} \right]$, où $Y$ suit la loi N(0,1), car $\frac{\sqrt{n_0}(\overline{X} - \mu_0)}{\sigma}$ suit la loi N(0,1). Nous la comparons à un seuil de signification, noté α, choisi *a posteriori* (c'est-à-dire après avoir recueilli les données). Si cette probabilité est inférieure ou égale à α, nous concluons que $H_0$ est fausse. Si, au contraire, elle est supérieure à α, nous concluons que nous avons échoué dans le rejet de $H_0$.

### 2.2. Tests d'hypothèses au sens de Neyman-Pearson

Le rôle d'un test d'hypothèses est de *décider*, moyennant un risque d'erreur, entre $H_0$ et $H_1$. Les deux décisions auxquelles nous pouvons aboutir sont : "Rejeter $H_0$ et Accepter $H_1$" *ou bien* "Accepter $H_0$ et Rejeter $H_1$".

La procédure adoptée par Neyman-Pearson consiste à élaborer une règle de décision, en fonction de la taille $n_0$ de l'échantillon et d'une valeur[7] α appartenant à l'intervalle ]0,1[. Cette règle de décision donne lieu à la construction d'une région[8] de rejet de $H_0$ et d'acceptation de $H_1$. Cette région notée RC, est caractérisée par :

$$\begin{cases} P(RC|H_0) \text{ ne dépasse pas } \alpha \\ \text{et} \\ P(RC|H_1) \text{ est maximale} \end{cases}$$

[6]C'est la probabilité d'obtenir les données observées $D_0$ ou des données plus extrêmes que celles observées, sachant que $H_0$ est vraie.

[7]Appelée : "niveau de signification" du test.

[8]Appelée aussi : région critique du test.

Dans le cas de la situation que nous sommes en train de traiter :

RC=$\left\{(X_1,....,X_{n0})/\overline{X} \in \left]\mu_0 + \Phi_\alpha \frac{\sigma}{\sqrt{n_0}},+\infty\right[\right\}$, où $\Phi_\alpha$ est le quantile d'ordre (1-α) de la loi normale centrée réduite N(0,1).

Si $\bar{x}_{ob} \in \left]\mu_0 + \Phi_\alpha \frac{\sigma}{\sqrt{n_0}},+\infty\right[$ alors nous décidons de rejeter $H_0$ et d'accepter $H_1$ : le test est dit statistiquement significatif. Si $\bar{x}_{ob} \in \left]-\infty,\mu_0 + \Phi_\alpha \frac{\sigma}{\sqrt{n_0}}\right]$ alors nous décidons d'accepter $H_0$ et de rejeter $H_1$ : le test est dit statistiquement non significatif.

Le risque de 1ère espèce est la probabilité $P\left(\overline{X} \in \left]\mu_0 + \Phi_\alpha \frac{\sigma}{\sqrt{n_0}},+\infty\right[ \middle| H_0\right)$, et celui de 2ème espèce est la probabilité $P\left(\overline{X} \in \left]-\infty,\mu_0 + \Phi_\alpha \frac{\sigma}{\sqrt{n_0}}\right] \middle| H_1\right)$. La puissance du test est $P\left(\overline{X} \in \left]\mu_0 + \Phi_\alpha \frac{\sigma}{\sqrt{n_0}},+\infty\right[ \middle| H_1\right)$ : c'est le complément à 1 du risque de 2ème espèce.

### 3. Quelques aspects de critiques non fondées

La théorie des tests statistiques représente une partie de l'inférence inductive, qui utilise une logique mathématique très particulière, où les propositions inductives ne peuvent être traitées selon une logique formelle en vrai ou faux ; en fait, ces propositions ne peuvent être formulées que moyennant un degré de confiance, c'est-à-dire à l'aide d'une probabilité.

C'est cet aspect des tests statistiques qui a été à l'origine des premières critiques, avec bien entendu un ensemble de glissements au niveau des interprétations engendrés par des pratiques non maîtrisées. Cela a suscité une vive controverse dès le début des années 1960 dans des domaines comme la sociologie ou la psychologie, et qui ont fait l'objet d'études détaillées comme par exemples [Lecoutre B., Lecoutre M.-P. et Poitevineau, 2001], [Lecoutre et Poitevineau, 2000] ou [Morrison et Henkel, 1970].

Pour ce qui nous concerne, nous nous limiterons principalement aux critiques qui versent directement dans le contexte de l'enseignement des tests statistiques, contexte qui nous intéresse dans cette étude, et dont nous ferons une analyse *a priori* à travers laquelle nous mettrons en valeur des éléments qui remettent en question le bien-fondé de ces critiques.

### 3.1. Le raisonnement sous-jacent aux tests de signification est fallacieux

Pollard et Richardson [1987] ont critiqué la logique adoptée dans le raisonnement des tests de signification, à savoir le fait qu'une p-value est *faible implique* $H_0$ est *fausse*. Ils ont prétendu que cette logique était basée sur la suite de raisonnements $R_1$, $R_2$, $R_3$ et $R_4$ suivants :

***R1*** : Si A alors non B
Si B alors non A

***R2*** : Si A alors probablement non B
Si B alors probablement non A

***R3*** : Si A (l'hypothèse nulle est correcte), alors probablement non B (une différence (un effet) de taille égale à celle impliquée par les données observées $D_0$).
Si B alors probablement non A.

***R4*** : Si A (personne X est américaine), alors probablement non B (personne X est dans le Congres).
Si B alors probablement non A.

Cohen [1994], encore plus direct que Pollard et Richardson, a affirmé que le raisonnement sous-jacent aux tests de signification, qui apparait implicitement dans certains articles publiés, et explicitement dans certains ouvrages de statistique, est R3. Il a par ailleurs qualifié ce raisonnement de "*Illusion of attaining improbability*" ou encore "*Illusion of probabilistic proof by contradiction*".

Il est clair que le raisonnement $R_2$ n'a pas de sens défini en logique formelle, et que tout ce qui s'en suit est bien entendu fallacieux, en l'occurrence le raisonnement $R_3$. Or, le vrai raisonnement qui est sous-jacent aux tests de signification, et qui apparaît dans [Fisher, 1956, p. 42], est le raisonnement $R_5$ suivant :

***R5*** : Si A (l'hypothèse nulle est correcte) alors B (l'effet sous test) est improbable.
Si B alors soit (A et un événement improbable) soit non A.

Ce qui est équivalent à :

- Si $H_0$ est vraie alors l'effet produit par les données $D_0$ est négligeable (c'est-à-dire l'événement "$D \geq D_0$" est improbable).
- Si les données observées $D_0$ produisent un effet (c'est-à-dire si la p-value est faible) alors deux cas se présentent :
  (i) soit $H_0$ est vraie et (mais) l'événement "$(D \geq D_0)$" improbable (rare, inhabituel, ...) est réalisé,
  (ii) soit $H_0$ est fausse.

Ainsi, dans le cas d'une p–value faible, Fisher accepte (ii) et écarte (i), en utilisant intuitivement un raisonnement (faible) de type *aristotelicien*, considérant que la taille de l'effet observé (produit par les données $D_0$), ou sous test, est un signe faillible[9], et que cet effet ne doit être dû ni aux fluctuations d'échantillonnage, ni aux erreurs de mesures expérimentales.

### 3.2. L'hypothèse $H_0$ est presque toujours fausse

Les tests statistiques mettent souvent en jeu une hypothèse nulle simple du type: $H_0$: "$\theta=\theta_0$".Meehl [1967] a critiqué ouvertement les tests statistiques, en avançant qu'une telle hypothèse $H_0$ ne peut être vraie, même à la dixième décimale. Cohen [1977] à son tour a proclamé que dans un test d'hypothèses, $H_0$ est toujours fausse, en disant qu'il suffit pour cela d'augmenter la taille de l'échantillon. Pourquoi alors mettre en œuvre un test statistique ? Plus récemment, Loftus [1996] a déclaré que : "*rejeter une hypothèse nulle simple, c'est comme rejeter la proposition selon laquelle la lune est faite de fromage vert*".

Il s'agit là de critiques inconsistantes. En effet, le résultat apporté par un test statistique, à savoir "rejeter $H_0$", "accepter $H_0$" ou "échouer à rejeter $H_0$", est une

---

[9]Outre les signes faillibles, Aristote utilise aussi les signes infaillibles, les probabilités, les exemples, les analogies, ..., comme modes de raisonnements dans sa Rhétorique. Pour plus de détail, consulter Macdonald [2004].

D'ailleurs, la théorie basée sur ce raisonnement (les signes faillibles) est appelée : "Théorie *faible* de Fisher". La faiblesse de cette théorie pourrait aisément être expliquée à partir de la règle de Bayès : $P\left(H_0 \middle| (D \geq D_0)\right) = \frac{P(H_0)}{P((D \geq D_0)|H_0)P(H_0) + P((D \geq D_0)|H_1)P(H_1)} P((D \geq D_0)|H_0)$, du fait que la p-value $P((D \geq D_0) \mid H_0)$ est faible, ne garantit pas toujours que la probabilité $P(H_0 \mid (D \geq D_0))$ est faible (c'est-à-dire que $H_0$ est fausse). Il est intéressant de signaler que Fisher a aussi développé une autre théorie dite "Théorie *forte* de Fisher", qui se base sur la probabilité conditionnelle $P(H_0|D_0)$ et qui est utilisée dans le cas où l'on connait de manière *exacte* la probabilité *a priori* $P(H_0)$. L'existence chez Fisher des deux théories faible et forte, a permis aux statisticiens de les étiqueter respectivement de "*fréquentiste modéré*" et de "*quasi-bayésien*".

conclusion qui découle immédiatement de la distribution échantillonnale[10], et qui ne cherche aucunement à traiter la question de véracité ou de fausseté de $H_0$. En outre, $H_0$ : "$\theta=\theta_0$" n'est en fait qu'une écriture simplifiée du *modèle probabiliste* $H_0$ : "$P_\theta = P_{\theta 0}$", représentant une hypothèse ou une action scientifique, pour laquelle l'application d'un test statistique peut conduire à son acceptation, et ce non pas parce qu'il s'agit d'une hypothèse exacte, mais plutôt d'une hypothèse *robuste*, autrement dit capable de fournir de bonnes prédictions.

Quant à l'augmentation de la taille de l'échantillon, cela devrait plutôt renvoyer au registre des pratiques absurdes, sans pour autant constituer une critique à l'encontre des tests d'hypothèses : d'abord, parce qu'en pratique toute expérimentation a un coût, et par voie de conséquence une augmentation raisonnable de la taille d'un échantillon ne devient nécessaire que lorsque cela s'avère indispensable (par ex. approximation de la loi de la statistique mise en jeu ou encore estimation d'un paramètre du modèle sous-jacent). Par ailleurs, il est vrai que la puissance d'un test augmente en fonction de la taille de l'échantillon, mais encore une fois cela engendre un coût. C'est pourquoi, il est plus recommandé de commencer par calculer la taille de l'échantillon qui fournira la puissance recherchée, préalablement à la mise en place proprement dite du test. Maintenant, s'il arrive qu'un praticien doive recourir à un test très puissant, afin de déceler une signification statistique, *a priori* très faible, il est alors difficile d'attribuer un intérêt particulier à une telle signification.

En d'autres termes, un test statistique perd sa fonction et sa raison d'être si l'on augmente sans justification la taille de l'échantillon.

### 3.3. Le traitement asymétrique des hypothèses $H_0$ et $H_1$

Rozeboom [1960] a critiqué les tests statistiques du fait qu'ils sont asymétriques dans leur traitement des hypothèses statistiques $H_0$ et $H_1$. L'argument avancé est que dans un test statistique à deux hypothèses (i.e. $H_a$ : "un paramètre $\theta=0$" et $H_b$ : "$\theta=10$") l'acceptation d'une hypothèse reste arbitraire et surtout tributaire de ce qui est considéré comme $H_0$ et comme $H_1$.

La critique de Rozeboom n'est pas pertinente, aussi bien pour les tests de signification que pour les tests d'hypothèses. Dans l'approche de Fisher, on met plus de contrainte sur l'hypothèse $H_0$ en favorisant son rejet, et la question d'asymétrie ne se pose pas puisque la conclusion du test n'est exprimée qu'en fonction de $H_0$ : "rejeter $H_0$" ou "échec dans le rejet de $H_0$".

---

[10]La distribution d'échantillon dans le cas d'un test de conformité est celle de la statistique $Y = \frac{\sqrt{n_0}(\overline{X} - \mu_0)}{\sigma}$.

Dans l'approche de Neyman-Pearson, le test d'hypothèses joue un rôle de règle de décision entre deux hypothèses $H_0$ et $H_1$, qui ont respectivement un statut bien défini. Ces deux hypothèses représentent respectivement deux modèles, où l'hypothèse nulle $H_0$ renvoie au modèle de référence et donc à une action usuelle, que McLean [2000] traduit par « *aucune action* », alors que l'alternative $H_1$ renvoie à un nouveau modèle, et donc à une nouvelle *action*. En termes de décision, l'acceptation de $H_0$ ou de $H_1$ ne doit pas *a priori* être traitée sur le même pied d'égalité, celle de $H_1$ doit être traitée de manière plus rigoureuse et plus sévère que celle de $H_0$, car une telle acceptation conduit à une *nouvelle action* dont les conséquences peuvent être très coûteuses en cas de mauvaise décision.

Cela dit, les tests de signification et ceux d'hypothèses ne font *a priori* qu'une partie d'un processus global de l'inférence statistique. En cas de besoin, de telles procédures ne doivent donc pas être mises en œuvre isolément, il faut aussi les intégrer à un modèle de recherche plus général, en présence d'un plan expérimental adéquat.

### 3.4. L'arbitraire dans le choix du niveau de signification

La controverse sur les tests statistiques a aussi porté sur l'arbitraire du choix de niveau de signification $\alpha$, en reprochant le fait que certaines données peuvent être statistiquement significatives à un niveau donné et le contraire à un niveau différent.

Il serait intéressant tout d'abord de mettre un peu de lumière sur cette question d'un point de vue historique : Dans son ouvrage "*Design of experiments*", Fisher [1935] a suggéré pour l'étude de la signification statistique de résultats issus d'expériences, la valeur conventionnelle correspondant à 5%. Il faut aussi rappeler qu'à l'époque, la majorité des exemples donnés par Fisher, relevait d'un même et seul contexte, celui de l'agronomie. Ultérieurement, Fisher[1956, p. 42] déclare "*In fact, no scientific worker has a fixed level of significance at wich from year to year and in all circumstances, he rejects hypotheses*" ; en d'autres termes, le choix du niveau de signification est loin d'être arbitraire, mais dépend des contraintes de la problématique, ainsi que des circonstances et conditions expérimentales. Aujourd'hui, avec le développement de nouvelles méthodes inférentielles, comme par exemple le data mining [Lovell, 1983], le choix du niveau de signification sera dans ce cas bien plus grand que la valeur standard 5% (biais de Lovell), du fait que l'on est amené à utiliser une même base de données pour tester différentes hypothèses.

Par ailleurs, dans certains domaines de recherches expérimentales, quelques praticiens en sont allés jusqu'à proposer une sélection de critères qui permettent de procéder à un choix *adéquat* du niveau de signification. C'est le cas par exemple en sociologie, où entre autres, Labovitz [1970] va proposer une liste (bien entendu non

exhaustive) de critères qui relèvent notamment des « *conséquences pratiques* », de la « *plausibilité des alternatives* », du « *degré de contrôle du plan expérimental* », ou encore de la « *robustesse du test* ».

Il serait enfin intéressant de signaler aussi que le choix du niveau de signification $\alpha$ peut être à l'origine d'amalgame dans un contexte d'enseignement entre les procédures de Fisher et celles de Neyman-Pearson, que Zaki et El M'Hamedi [2009] ont qualifié de « *procédure hybride* ». Cependant, cette difficulté peut être surmontée auprès des étudiants, en présentant des situations pertinentes, comme par exemple celles de tests paramétriques relatifs aux modèles continus, où l'on va nuancer la différence entre les logiques utilisées par Fisher et Neyman-Pearson, et ce, au niveau de l'étape ultime concernant les conclusions fournies respectivement par un test de signification et un test d'hypothèses.

### 3.5. L'attention excessive en faveur du risque de 1ère espèce en dépit de celui de 2ème espèce

Les risques d'erreurs de 1ère espèce : P(Rejeter $H_0|H_0$ vraie) et de 2ème espèce : P(Accepter $H_0|$ $H_0$ fausse) sont inversement reliés. Si l'on augmente l'un, l'autre est diminué [Cohen, 1977]. On reproche aux tests d'hypothèses de porter davantage d'intérêt au risque de 1ère espèce en dépit du risque de 2ème espèce, à savoir le fait que l'on impose au 1er risque de ne pas dépasser un niveau de signification $\alpha$ fixé au départ, sans pour autant pouvoir maîtriser le 2ème risque.

Cette critique n'a pas de fondement, si l'on se réfère à la construction même d'un test d'hypothèses telle qu'elle a été proposée par Neyman-Pearson. En intégrant les deux hypothèses $H_0$ et $H_1$ dans sa procédure de test statistique, Neyman-Pearson va introduire deux types de risques, représentant respectivement sous ces deux hypothèses (mettons pour simplifier les choses, dans un cas de test paramétrique, en fonction de $\theta$) les coûts moyens $R(\theta,\delta)$ de la règle de décision $\delta$, faisant l'objet de la solution du problème : ce sont les risques de première et seconde espèces. Dans la recherche de $\delta$, 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 1ère espèce, *qui fait référence à son hypothèse d'intérêt $H_0$*, en fixant un seuil de signification $\alpha$, pour aller chercher parmi les règles de décisions répondant à ce seuil, celles dont le risque de seconde espèce est le plus faible (ou encore dont la puissance est maximale). Ainsi, nous constatons d'emblée dans la procédure de Neyman-Pearson le rôle dissymétrique des hypothèses $H_0$ et $H_1$, qui ne relève pas de leur statut respectif (cf. § 3.3), mais qui constitue plutôt une contrainte mathématique dans le traitement de la notion de risque au niveau de l'élaboration de la solution au problème de tests d'hypothèses.

### 3.6. Les tests statistiques ne sont pas informatifs sur la probabilité que $H_0$ soit vraie

Une des critiques qui revient le plus souvent à l'encontre de l'utilisation des tests statistiques, est que de telles procédures ne permettent pas aux chercheurs de répondre à la question "sachant les données observées $D_0$, qu'elle est la probabilité que $H_0$ soit vraie ? " [Cohen, 1994], question qui revêt une grande importance dans les investigations de ces derniers. Il est certain que le concept de test statistique ne traite pas de telles probabilités. Si maintenant les chercheurs formulent des attentes de cet ordre, en utilisant des tests statistiques ; de telles formulations sont bien entendu aberrantes, et ne peuvent aucunement donner lieu à des critiques à l'encontre des tests statistiques.

Maintenant, d'autres outils de l'inférence statistique peuvent *éventuellement* répondre à une telle question, en l'occurrence celui de la statistique *bayésienne* qui se fonde sur la fameuse règle de *Bayès* :

$$P(H_0|D_0) = P(H_0).\frac{P(D_0|H_0)}{P(D_0)}.$$

## 4. Etude exploratoire auprès des étudiants

### 4.1. Objectif de l'étude et expérimentation

Les éléments dégagés à la lumière de l'analyse *a priori* des critiques avancées à l'encontre de la théorie des tests statistiques, corroborent notre plaidoyer en faveur de cette théorie. Le contexte de notre étude étant celui de l'enseignement, il nous a alors semblé pertinent de confronter notre analyse *a priori* de ces critiques avec les conceptions d'étudiants, en explorant leurs approches à propos de questions faisant l'objet des critiques analysées. L'enjeu de cette exploration est d'abord de remettre en question ces critiques, du moins dans le contexte de l'enseignement, et d'autre part de confirmer notre hypothèse de départ (cf. § 1), selon laquelle il faudrait plutôt remettre en question la non maîtrise du concept de test statistique chez certains utilisateurs et praticiens. En validant cette hypothèse, la théorie des tests statistiques peut alors encore prétendre à sa pertinence et son intérêt, au moins du point de vue de son enseignement, mais alors à condition aussi que soient pris en compte quelques éléments de mesures adéquats pour que cet enseignement soit réussi [Zaki et El M'Hamedi, 2009].

Nous avons conduit au courant du deuxième semestre de l'année universitaire 2010/2011 une expérimentation auprès de 55 étudiants répartis de la façon suivante : 41 en $2^{ème}$ année «spécialité : Statistique» à l'Institut National de Statistique et d'Economie Appliquée de Rabat, et 14 en M1 du Master «Système d'Aide à la Décision et Management de Projets» à la Faculté des Sciences de Kenitra. Nous leur avons donc administré durant une demi-heure un questionnaire

en vrai - faux. Les étudiants interrogés ont tous suivi un enseignement classique de probabilités-statistique, d'un volume horaire qui dépasse les 100 heures, portant, entre autres, sur la théorie de base des probabilités (espaces probabilisés, variables aléatoires et lois de probabilités, lois des grands nombres, Théorème Centrale Limite,…), 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 élaboré sur la base d'une situation relevant d'un test paramétrique d'hypothèses statistiques simples, généralement présentée aux étudiants dans leur cours sur les tests statistiques. Il est composé de 4 questions (numérotées de A à D), faisant référence aux contenus des critiques analysées précédemment.

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

### 4.2. Présentation et analyse *a priori* du questionnaire

Les quatre questions (numérotées de A à D) qui composent le questionnaire sont toutes relatives à la situation de test paramétrique dont les éléments sont présentés dans le tableau 2 suivant :

- X est une variable aléatoire qui suit une loi de probabilité continue mais inconnue $P_\theta$, où $\theta \in IR$.
- $H_0$ : "$\theta = \theta_0$" *vs.* $H_1$ : "$\theta > \theta_0$" ($H_0$ : *hypothèse nulle* et $H_1$ : *hypothèse alternative*).
- $(X_1,...,X_{n_0})$ un échantillon de taille $n_0$, issu de la variable aléatoire X.
- $D_0$ : $(x_1^0,...,x_{n_0}^0)$ des données observées, recueillies à l'aide de $(X_1,...,X_{n_0})$ et d'une expérimentation $w_0$ $\left((x_1^0,...,x_{n_0}^0) = (X_1(w_0),...,X_{n_0}(w_0))\right)$.
- $\hat{\theta}$ : l'estimateur empirique de $\theta$, associé à l'échantillon $(X_1,...,X_{n_0})$.
- $\hat{\theta}_{obs}$ : la valeur de l'estimateur empirique $\hat{\theta}$ de $\theta$, associée aux données observées $D_0$.
- $(\hat{\theta}_{obs} > \theta_0)$ : condition réalisée.

Tableau 2. Eléments de la situation de test statistique retenue dans le questionnaire

**(A)** ***L'application d'un test statistique dans cette situation permet de conclure que $H_0$ est fausse si l'on estime que la probabilité $p = P(\hat{\theta} \geq \hat{\theta}_{obs} \mid H_0)$ est faible (c'est-à-dire si le calcul de p donne une valeur inférieure ou égale au niveau de signification α).***

***A votre avis, le (les) raisonnement(s) utilisé(s) dans les procédures des tests statistiques impliquant cette conclusion est (sont) :***

**(1)** [p est faible] *alors* [La probabilité de réalisation de $H_0$ sachant qu'on a observé les données $D_0$ est faible]. *Par conséquent*, $H_0$ est probablement fausse.

**□ Vrai □ Faux**

**(2)** [p est faible] *alors* [La probabilité que la différence observée $\left(\hat{\theta}_{obs} - \theta_0\right)$ soit due au seul hasard est faible]. *Par conséquent*, $H_0$ est probablement fausse.

**□ Vrai □ Faux**

**(3)** [p est faible] *implique* [Si $H_0$ est vraie alors probablement une différence telle celle impliquée par les données observées $D_0$, à savoir $\left(\hat{\theta}_{obs} - \theta_0\right)$, n'est pas réalisable] *ce qui est équivalent à* [La différence $\left(\hat{\theta}_{obs} - \theta_0\right)$, impliquée par les données observées $D_0$, est réalisable alors $H_0$ est probablement fausse].

**□ Vrai □ Faux**

**(4)** [p est faible] *alors* [La probabilité de réalisation de $H_0$, sachant qu'on a observé les données $D_0$ est égale à 0]. *Par conséquent*, $H_0$ est probablement fausse.

**□ Vrai □ Faux**

Notons d'abord que seul l'item A3 fait directement référence à la première critique portant sur le raisonnement fallacieux reproché aux tests statistiques. Il nous a cependant semblé intéressant d'intégrer trois autres raisonnements fallacieux A1, A2 et A4, identifiés chez les étudiants, et que leurs auteurs ont qualifiés respectivement de "*inverse probability error*" [Kline, 2004], "*odds-against-chance fantasy*" [Carver, 1953] et "*Le test statistique comme étant une procédure inductive qui permet de calculer la probabilité a posteriori de l'hypothèse nulle*" [Vallecilos, 1995]. Le recours à ces différents raisonnements fallacieux nous a semblé être intéressant, pour en étudier les liens d'une part, s'il en existe, et d'autre part mesurer la prépondérance des uns par rapport aux autres.

Les seules types de probabilités conditionnelles *autorisées* en théories des tests statistiques sont évidemment les probabilités des données D sachant la réalisation des hypothèses statistiques $H_i$ (i=0 ou 1), notées $P(D|H_i)$. Nous pouvons citer à titre d'exemple, la p-value, égale à la probabilité $P(\theta \geq \hat{\theta}_{obs}|H_0)$ dans le cas de la situation que nous sommes en train de traiter dans ce questionnaire. De telles probabilités ($P(D|H_i)$) sont bien entendu, *de manière générale*, différentes des probabilités inverses: $P(H_i|D)$, et il n'existe pas de relations établies pouvant lier les valeurs prises par ces deux probabilités conditionnelles [Gras et Totohasina, 1995]. Il est à rappeler que le calcul des probabilités $P(H_i|D)$ ne peut être déterminé que dans le cadre du paradigme alternatif aux tests statistiques : la *statistique bayésienne*. Néanmoins, ces types de probabilités ($P(H_i|D)$) sont utilisées à tort chez beaucoup d'étudiants, face à une situation de test statistique. Ces probabilités prennent des formes variées, en se manifestant soit de façon *directe* comme dans le cas de l'item A1 (la probabilité de réalisation de $H_0$ sachant qu'on a observé les données $D_0$), soit de façon *indirecte* comme c'est le cas de l'item A2 (la probabilité que la différence observée $(\hat{\theta}_{obs} - \theta_0)$ soit due au seul hasard)[11], ou encore de manière *déterministe* comme c'est le cas de l'item A4 (la probabilité de réalisation de $H_0$, sachant qu'on a observé les données $D_0$ est égale à 0).

Par ailleurs, l'item A3 impliquant un *raisonnement fondé sur la contraposée probabiliste*, est bien évidemment faux, comme nous l'avons déjà expliqué auparavant (cf. § 3.1). De ce fait, nous pouvons conclure que les quatre raisonnements des items A1, A2, A3 et A4 sont tous fallacieux.

**(B)*A votre avis, dans une situation de test statistique, le niveau de signification α est :***

**(1)** toujours égal à 5%, 1% ou 0.1%, selon si l'on veut tester si le résultat obtenu $(\hat{\theta}_{obs} - \theta_0)$ est peu significatif, significatif ou très significatif.

**□ Vrai □ Faux**

**(2)** choisi de manière arbitraire, dans la mesure où pour une même situation de tests statistiques, deux statisticiens sont libres de fixer leurs niveaux de signification à des valeurs trop éloignées, sans aucune contrainte à prendre en considération.

**□ Vrai □ Faux**

[11]L'événement « la différence observée $(\hat{\theta}_{obs} - \theta_0)$ est due au seul hasard » est en quelque sorte équivalent à l'événement : « $H_0$ est réalisée ».

Il est vrai que le chercheur choisit son niveau de signification. Néanmoins, ce choix reste tributaire des contraintes de la problématique mise en jeu, auxquelles s'ajoutent les conditions et les contraintes expérimentales (cf. § 3.4).

Par ailleurs, la p-value ($P(\hat{\theta} \geq \hat{\theta}_{obs} | H_0)$), à laquelle on compare le niveau de signification $\alpha$ pour pouvoir conclure, est un indicateur qui mesure sommairement la taille de l'effet impliquée par les données observées $D_0$ et la taille de l'échantillon dont proviennent ces données. Elle ne permet pas la mesure d'une seule de ces tailles. Par conséquent, des expressions directement liées au concept de taille de l'effet, telles « peu significatif », « significatif » ou « très significatif », n'ont aucun sens dans la théorie des tests statistiques.

En conclusion, les propositions des items B1 et B2 peuvent être considérées comme étant toutes les deux fausses.

***(C) Supposons qu'après application du test statistique nous ayons décidé de rejeter $H_0$:*** "***$\theta=\theta_0$***". ***Parmi les propositions suivantes, celle(s) qui est (sont) vraie (s) est (sont) :***

**(1)** Le rejet de $H_0$ : "$\theta=\theta_0$"est une information sans intérêt parce que si l'on augmente suffisamment la taille $n_0$ de l'échantillon dont dérivent les données observées $D_0$, le rejet de $H_0$ sera systématiquement réalisé.

**□ Vrai □ Faux**

**(2)** Le rejet de $H_0$ : "$\theta=\theta_0$" est une information très importante parce qu'un tel rejet signifie que $\theta$ est différent de $\theta_0$, et que $\theta$ est beaucoup plus grand que $\theta_0$.

**□ Vrai □ Faux**

Il est vrai que si l'on augmente la taille $n_0$ de l'échantillon, la région critique du test statistique s'élargit. Or, la puissance qui croit en fonction de la taille de l'échantillon, a quand même des conséquences coûteuses du point de vue expérimental. Maintenant, si une signification statistique est suffisamment petite, pour que seul un test très puissant peut identifier, il est alors difficilement acceptable de la considérer comme utile : dans ce cas-là, le test statistique perd sa fonction (cf. § 3.2). De ce fait, la proposition de l'item C1 peut être considérée comme fausse.

De la même façon, l'hypothèse nulle $H_0$ : "$\theta=\theta_0$" est en fait une écriture simplifiée du *modèle probabiliste* $H_0$ : "$P_\theta = P_{\theta 0}$", représentant une hypothèse ou une action scientifique, qui peut être acceptée par l'application d'un test statistique, non pas pour son exactitude mais plutôt pour sa *robustesse* et sa capacité de fournir de meilleures prédictions (cf. § 3.2). Par conséquent, la proposition de l'item C2 est fausse.

**(D)** ***Dans cette question, les données observées $D_0$ sont fixées. On considère alors les deux tests statistiques ci-après, servant à décider entre $H_0$ et $H_1$ :***

***- Le test statistique n°1 de $H_0$: "$\theta=\theta_0$" contre $H_1$ : "$\theta=\theta_1$" $(\theta_1 > \theta_0)$***

***et***

***- Le test statistique n°2 de $H_0$: "$\theta=\theta_1$" contre $H_1$: "$\theta=\theta_0$" $(\theta_1 > \theta_0)$***

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

**(1)** Si l'on fixe le même niveau de signification α pour ces deux tests statistiques, alors si l'on décide de rejeter "$\theta=\theta_0$" par le test statistique n°1 on doit décider d'accepter "$\theta=\theta_1$" par le test statistique n°2.

**□ Vrai □ Faux**

**(2)** Si le risque de 1ère espèce du test statistique n°2 est fixé à la valeur du risque de 2ème espèce du test statistique n°1, alors si l'on décide de rejeter "$\theta=\theta_0$" par le test statistique n°1 on doit décider d'accepter "$\theta=\theta_1$'' par le test statistique n°2.

**□ Vrai □ Faux**

Mettons que l'on se place par exemple dans le cas où la variable aléatoire X suit la loi normale $N(\mu,\sigma^2)$, $\theta = \mu$ et $P_\theta=[N(\mu,\sigma^2)]_\mu$, avec σ connu. Ainsi, les tests statistiques n°1 et n°2 testent respectivement $H_0$: "$\mu = \mu_0$" contre $H_1$: "$\mu = \mu_1$" et $H_0$: "$\mu = \mu_1$" contre $H_1$: "$\mu=\mu_0$", où $\mu_1>\mu_0$.

Si l'on fixe le même niveau de signification α pour ces deux tests statistiques, alors toutes les données D issues de l'échantillon $(X_1,...,X_{n_0})$ telles que $\overline{X} \in \left]\mu_0 + \Phi_\alpha \frac{\sigma}{\sqrt{n_0}}, \mu_1 + \Phi_\alpha \frac{\sigma}{\sqrt{n_0}}\right[$ (où $\overline{X} = \frac{X_1 + .... + X_{n_o}}{n_o}$), permettent[12] de rejeter "$\mu=\mu_0$" par le test statistique n°1 et en même temps de rejeter "$\mu= \mu_1$" par le test statistique n°2.

De même, si l'on choisit α comme étant le niveau de signification du test statistique n°1 et si le risque de 1ère espèce du test statistique n°2 est fixé à la valeur β du risque de 2ème espèce du test statistique n°1 ($\beta = P\left(\overline{X} \in \left]-\infty, \mu_0 + \Phi_\alpha \frac{\sigma}{\sqrt{n_0}}\right] \middle\| H_1\right)$, où $H_1$: "$\mu = \mu_1$"), alors en étant dans la condition

[12] Φα est le quantile d'ordre (1- α) de la loi normale centrée réduite N(0,1).

$(\mu_1 - \mu_0) > \frac{\sigma}{\sqrt{n_0}}(\Phi_\alpha - \Phi_\beta)$, avec un choix approprié de $\mu_0$ et $\mu_1$, toutes les données D issues de l'échantillon $(X_1,...,X_{n_0})$ telles que $\overline{X} \in \left]\mu_0 + \Phi_\alpha \frac{\sigma}{\sqrt{n_0}}, \mu_1 + \Phi_\beta \frac{\sigma}{\sqrt{n_0}}\right[$, permettent d'une part de rejeter "$\mu = \mu_0$" par le test statistique n°1, et d'autre part de rejeter "$\mu = \mu_1$"par le test statistique n°2 [Zaki et El M'Hamedi, 2009].

Par conséquent, les propositions des items D1 et D2 sont toutes les deux fausses.

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

| Libellé | Codage | Signification | RR[13] | RE[14] |
|---|---|---|---|---|
| Raisonnement. | A1 | • *Bayésien direct* | A1R | A1E |
| | A2 | • *Bayésien indirect* | A2R | A2E |
| | A3 | • *Contraposé probabiliste* | A3R | A3E |
| | A4 | • *Bayésien déterministe* | A4R | A4E |
| Niveau de signification. | B1 | • *Choix fixe* | B1R | B1E |
| | B2 | • *Choix arbitraire* | B2R | B2E |
| Rejet systématique de $H_0$. | C1 | • *Effet de la taille de l'échantillon* | C1R | C1E |
| | C2 | • *Pseudo-égalité.* | C2R | C2E |
| Asymétrie. | D1 | • *Au niveau des hypothèses statistiques* | D1R | D1E |
| | D2 | • *Au niveau des risques de 1ère et de 2ème espèces* | D2R | D2E |

Tableau 3. Codage des modalités de chaque item du questionnaire

[13] **R**éponse **R**éussie.
[14] **R**éponse **É**chouée (y compris les **N**on **R**éponses).

| Libellé de la critique | Items correspondants |
|---|---|
| Le raisonnement sous-jacent aux tests de signification est fallacieux | A1, A2, A3 et A4 |
| L'hypothèse $H_0$ est presque toujours fausse | C1 et C2 |
| Le traitement asymétrique des hypothèses $H_0$ et $H_1$ | D1 |
| L'arbitraire dans le choix du niveau de signification | B1 et B2 |
| L'attention excessive en faveur du risque de 1ère espèce aux dépens de celui de 2ème espèce | D2 |
| Les tests statistiques ne sont pas informatifs sur la probabilité que $H_0$ soit vraie | A1, A2 et A4 |

Tableau 4. Correspondances entre critiques non fondées et items du questionnaire

Pour le traitement des réponses des étudiants, l'analyse factorielle des correspondances multiples (AFCM) nous a semblé être un outil adéquat pouvant permettre de procéder à une analyse qui tient compte des croisements des réponses des étudiants, des types de liens qui existent entre ces réponses, et aussi de la prépondérance des réponses (en termes d'inertie relative dans le nuage des réponses), les unes par rapport aux autres.

Par ailleurs, pour ne pas biaiser l'analyse, vu le nombre important de modalités (20 au minimum) de l'ensemble du questionnaire, face au nombre « limité » d'étudiants interrogés (55), nous avons décidé de retenir un codage en bi-modalité « Réussite - Echec » pour tous les items du questionnaire.

Dans le tableau 3, nous avons résumé le codage des modalités des items du questionnaire, les libellés des items, ainsi que la signification de chaque modalité. Nous avons par ailleurs présenté dans le tableau 4 les correspondances qui existent entre les critiques que nous avons analysées et les items du questionnaire.

### 4.4. Analyse et interprétation des réponses des étudiants

#### *4.4.1. Valeurs propres et inertie totale*

L'analyse factorielle des correspondances multiples[15] appliquée au tableau disjonctif complet issu du codage des réponses des étudiants a conduit à des valeurs propres non nulles de moyenne 0,1 (l'inverse du nombre d'items qui est égal à 10). Par ailleurs, les valeurs propres supérieures à cette moyenne sont : $\lambda_1$=0,191,

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

$\lambda_2$=0,152, $\lambda_3$=0,138 et $\lambda_4$=0,101. L'inertie totale est égale à 1 (nombre de valeurs propres non nulles, divisé par le nombre d'items).

### *4.4.2. 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 statistique[16]; elle ne dépend pas des observations, elle dépend uniquement du nombre de variables (items) et du nombre de modalités impliquées. Par ailleurs, le calcul des taux d'inertie liés aux trois valeurs propres ci-dessus conduit à des pourcentages décroissant de 19,07% à pratiquement 4,36%, 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 d'items du questionnaire et $\lambda_k$ les valeurs propres obtenues par l'analyse du tableau disjonctif complet.

L'application de cette formule aux inerties initiales[17], donne pour les valeurs propres supérieures à 0,1 (égale à $\frac{1}{10}$) les résultats suivants : 0,010149, 0,003339, 0,001799 et 0,000003.

Le tableau 5 suivant illustre les pourcentages d'inerties corrigées et leurs cumuls, correspondant aux valeurs propres retenues et qui sont supérieures à la valeur moyenne $\frac{1}{10}$.

| **Valeurs propres** | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ |
|---|---|---|---|
| **% Inerties corrigées** | **66,38%** | **21,84%** | **11,77%** |
| **Cumul % Inerties corrigées** | **66,38%** | **88,22%** | **99,99%** |

Tableau 5. Pourcentages d'inerties corrigées et leurs cumuls

Par conséquent, le premier axe représente 66,38% de l'information totale, le deuxième axe 21,84% et le troisième axe 11,77%. Ces valeurs indiquent que

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

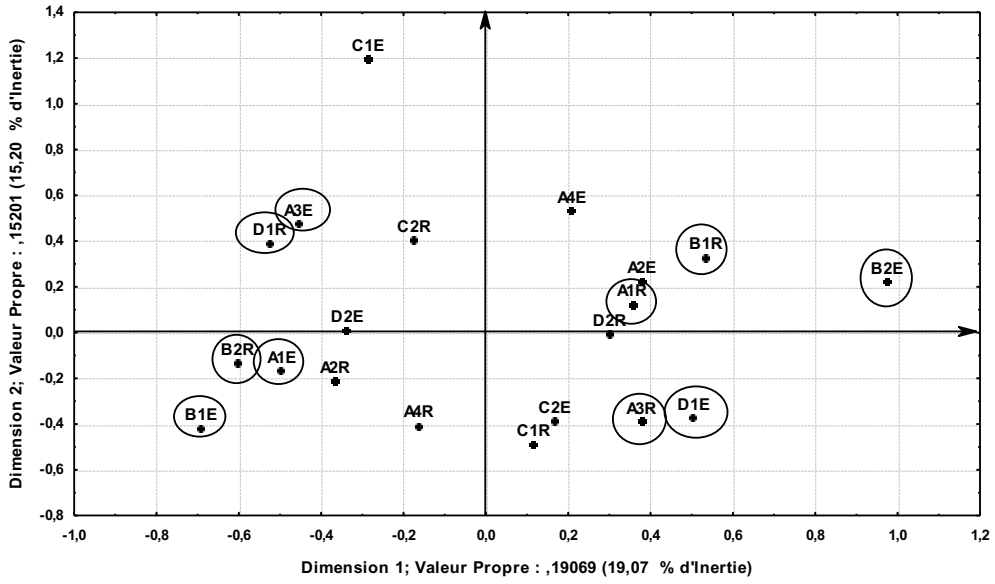
[17] Dans la suite de l'analyse, nous nous limiterons simplement aux 4 premières valeurs propres.

l'essentiel de l'information est porté par les trois premiers axes factoriels, qui représentent ensemble 99,99% de l'information totale.

Par ailleurs, ces trois valeurs propres sont différentes ; cela signifie que nous sommes pratiquement en présence de trois valeurs propres simples. Par conséquent, il va falloir interpréter séparément les axes : 1, 2 et 3.

### *4.4.3. Interprétation du 1er axe factoriel*

L'analyse factorielle de correspondances multiples conduit à une inertie[18] de 66,38% pour l'axe 1. Les modalités qui contribuent le plus à la construction de cet axe sont celles entourées dans le graphique 1 ci-dessous[19]. Elles ont une contribution relative comprise entre 3,8% et 19%. La qualité de représentation de ces modalités par rapport à cet axe est comprise entre 0,17 et 0,58.




Graphique 1. Plan factoriel (1, 2)

En résumé, nous sommes en présence de l'opposition suivante : A1E, A3E, B1E, B2R et D1R vs A1R, A3R, B1R, B2E et D1E.

[18]Il s'agit de l'inertie corrigée.

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

Notons dans un premier temps que dans cette opposition, chaque modalité à forte contribution relative, qu'elle relève d'une réussite ou d'un échec, se trouve du côté opposé par rapport à l'axe 1 de sa modalité contraire correspondante, elle-même à forte contribution relative. Par conséquent, il suffira de donner une interprétation à l'un des groupes de modalités opposées, du fait que l'autre groupe aura tout simplement l'interprétation contraire de ce dernier.

Par ailleurs, le 1er axe oppose un mélange de modalités de réussites et d'échecs à leurs modalités correspondantes contraires. Dans cette opposition, nous n'avons pas d'un côté les réussites et de l'autre les échecs, comme c'est classiquement le cas lorsqu'on utilise un codage en bi-modalité réussite-échec. Ainsi, le 1er axe factoriel ne peut être interprété comme étant un axe de réussites-échecs, et qu'il va falloir aborder son interprétation selon une démarche qui répond à l'objectif même du questionnaire dispensé aux étudiants.

Le contenu du questionnaire relève d'une exploration auprès des étudiants sur leurs appréhensions conceptuelles relativement au raisonnement sous-jacent au test statistique, au niveau de signification d'un test et aux statuts de l'hypothèse nulle $H_0$ et de son alternative $H_1$, notamment à propos du rejet de $H_0$ et de l'asymétrie entre $H_0$ et $H_1$. Les items du questionnaire font donc référence à une exploration conceptuelle autour du concept de test statistique, et non pas aux performances des étudiants en situation de résolution de problèmes sur les tests statistiques. Il est donc normal que la plus grande part d'information (1er axe factoriel) contenue dans les réponses des étudiants ne se traduise pas en termes de performances des étudiants, autrement dit en termes de réussites-échecs.

Ainsi, nous aborderons l'interprétation du 1er axe factoriel, en portant une attention particulière aux modalités de réussites ayant les plus fortes contributions relatives par rapport à cet axe et qui relèvent des items A1, A2, A3, et A4. En effet, la bonne maîtrise du concept de test statistique commence d'abord par une bonne appréhension du paradigme sous-jacent de probabilité conditionnelle, à partir duquel est fondé le concept même de test statistique. Dans le cas présent (cf. Graphique 1 : Plan factoriel (1,2)), il s'agit de A1R et A3R. Ces deux modalités sont associées dans l'opposition du 1er axe factoriel aux modalités B1R, B2E et D1E, modalités à fortes contributions relatives par rapport à cet axe. Nous noterons au passage, que dans cette opposition de groupes de modalités à fortes contributions relatives, nous avons une opposition exclusive de modalités contraires, ce qui confère au questionnaire un caractère d'homogénéité.

Par ailleurs, la modalité A1R (Réussite au raisonnement Bayésien direct), constitue dans ce questionnaire une modalité discriminante pour l'appréhension du paradigme de probabilité conditionnelle sous-jacent à la construction même du concept de test statistique. Ainsi, dans l'analyse du 1er axe factoriel, pour avoir une idée complète sur les tendances de l'ensemble des items relevant du questionnaire

associées à la bonne appréhension du concept de test statistique (A1R), nous tiendrons compte aussi dans le groupe de modalités contenant A1R (dans l'opposition du 1er axe factoriel), de la nature (réussite ou échec) des modalités de l'ensemble du reste des items du questionnaire ((cf. Graphique 1 : Plan factoriel (1,2)). Pour ce groupe, nous obtenons donc le profil de modalités suivant :

A1R - A2E - A3R - A4E - B1R - B2E - C1R - C2E - D1E et D2R

Dans ce groupe de modalités, nous sommes en présence d'une certaine maîtrise du raisonnement sous-jacent au test statistique, puisqu'il n'y a pas de confusion avec un raisonnement bayésien direct (A1R), ou celui qui utilise la « contraposée probabiliste » (A3R), avec néanmoins une difficulté à identifier un raisonnement bayésien indirect ou déterministe (A2E et A4E), sans grand effet, puisque ces deux modalités n'ont pas une forte contribution relative au 1er axe factoriel.

Les deux modalités B1R et B2E (à fortes contributions relatives) semblent traduire une contradiction à propos de la maîtrise du choix du niveau de signification dans la procédure d'un test statistique : les étudiants reconnaissent bien que le choix du niveau de signification ne correspond pas systématiquement aux standards 5%, 1% ou 0,1%, mais ne vont pas jusqu'à maîtriser le fait que le choix de ce niveau est intimement lié aux contraintes de la problématique dont relève le test statistique. Cela n'est pas surprenant, car dans tout enseignement standard sur les tests statistiques, et surtout dans le traitement de situations-problèmes relevant de tests statistiques, généralement les contraintes du contexte (problématique) pour le choix du niveau de signification sont ignorées ; c'est souvent un aspect qui est passé sous silence dans l'enseignement. Par conséquent, nous pouvons conclure à une maîtrise partielle du traitement du niveau de signification.

La modalité D1E traduit un échec des étudiants à propos de l'asymétrie des hypothèses $H_0$ et $H_1$ dans le traitement d'un test d'hypothèses dans l'approche de Neyman-Pearson. Néanmoins, cette modalité est associée à C1R et D2R. Autrement dit, l'échec à propos de l'asymétrie de l'hypothèse nulle et de l'alternative mérite d'être nuancé. En effet, pour les étudiants, dans le traitement d'un test d'hypothèses, lorsqu'une hypothèse nulle $H_0$ est rejetée au seuil $\alpha$, la règle de décision conduit à accepter $H_1$ (sous-entendu au même seuil $\alpha$) : nous pensons que pour cet item, les étudiants en sont restés à cette étape dans leurs conclusions. Or, lorsqu'on intervertit le rôle de $H_0$ et de $H_1$, le test d'hypothèses change et la règle de décision ne conduit plus bien entendu aux mêmes conclusions, et ce même lorsqu'on garde le même niveau de signification $\alpha$ : en fait, les étudiants ont rarement l'occasion d'être confrontés à ce type de situations, ce qui explique leur échec à cet item. Par ailleurs, comme nous l'avons fait remarquer au-dessus, les modalités C1R et D2R, traduisent respectivement chez les étudiants, d'une part la maîtrise de l'effet de la taille de l'échantillon, et d'autre

part celle de l'asymétrie des risques de 1ère et de 2ème espèce, dans le traitement des tests d'hypothèses (Neyman- Pearson). Autrement dit, dans le cas présent, les étudiants font preuve d'une assez bonne maîtrise de la procédure relative au traitement des tests d'hypothèses selon l'approche de Neyman-Pearson. Quant à la modalité C2E (à faible contribution relative), elle traduit en fait une réponse chez ces étudiants non contradictoire avec C1R, à savoir que pour ces derniers le contenu de l'hypothèse nulle $H_0$ représente une information très importante. Sachant que l'hypothèse nulle $H_0$ représente chez Neyman-Pearson l'hypothèse d'intérêt, cela renforce auprès de ces étudiants l'idée d'une démarche selon l'approche de Neyman-Pearson dans leur traitement de test statistique.

En conclusion, les tendances expliquées par le premier axe factoriel renvoient au degré de maîtrise du paradigme de probabilité conditionnelle sous-jacent au concept de test statistique, associé au degré de la maîtrise procédurale des tests statistiques au sens de Neyman-Pearson, notamment à propos du choix du niveau de signification et l'asymétrie des risques de 1ère et 2ème espèces. Par conséquent, le premier axe factoriel peut être interprété comme étant un axe de *maîtrise du paradigme de test statistique selon l'approche de Neyman-Pearson.*

#### *4.4.4. Interprétation du 2ème axe factoriel*

L'analyse factorielle de correspondances multiples conduit à une inertie de 21,84% pour l'axe 2. En outre, les modalités qui contribuent le plus à la construction de cet axe sont celles entourées dans le graphique 2 ci-dessous. Elles ont une contribution relative comprise entre 5% et 27%. La qualité de représentation de ces modalités par rapport à cet axe est comprise entre 0,18 et 0,58.

Dans le deuxième plan factoriel, le 2ème axe factoriel présente l'opposition suivante :

A3E, A4E et C1E vs A3R, A4R et C1R.

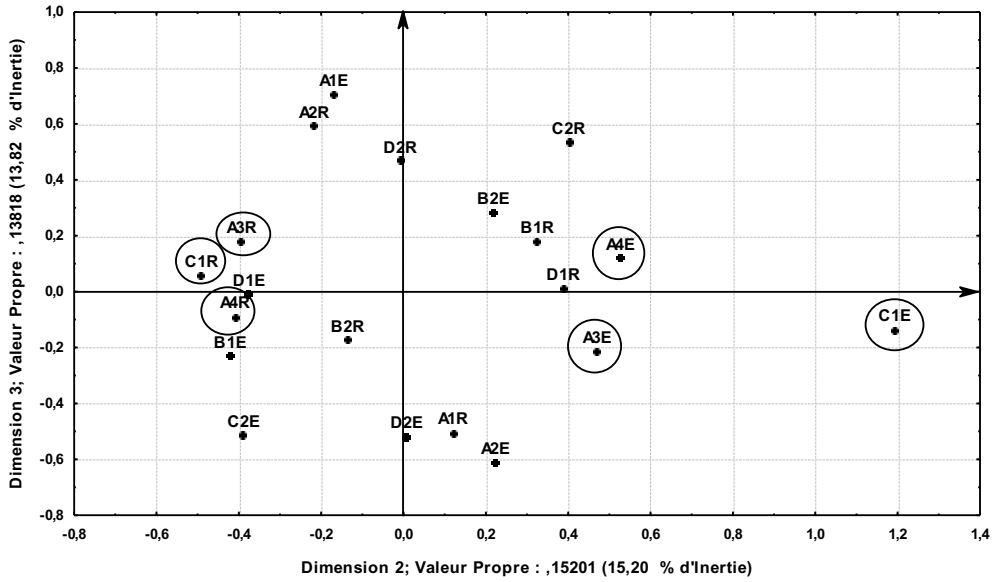
Dans cette opposition, nous avons d'un côté des modalités d'échecs et de l'autre leurs correspondantes de réussites, il suffira alors comme précédemment de donner une interprétation à l'un des groupes de modalités opposées, du fait que l'autre groupe aura tout simplement l'interprétation contraire de ce dernier. Nous noterons au passage, que dans cette opposition de groupes de modalités à forte contribution relative, nous avons une opposition exclusive de modalités contraires, ce qui renforce encore une fois le caractère d'homogénéité du questionnaire.

Ici aussi, nous nous mettrons du côté de la bonne appréhension du paradigme de probabilité conditionnelle sous-jacent au concept de test statistique, à savoir la modalité A1R (Réussite au raisonnement bayésien direct), et ce pour les mêmes raisons que nous avons déjà soulevées dans l'interprétation du 1er axe factoriel. Par ailleurs, nous chercherons aussi à donner une interprétation au 2ème axe factoriel, qui tiendrait compte dans le groupe de modalités contenant A1R (dans l'opposition

du 2ème axe factoriel), non seulement des modalités à forte contribution relative, mais aussi de la nature (réussite ou échec) du reste des modalités des items du questionnaire (cf. Graphique 2 : Plan factoriel (2,3)). Pour ce groupe, nous obtenons donc le profil de modalités suivant :

A1R - A2E - A3E - A4E - B1R - B2E - C1E - C2R - D1R - D2E

Du côté de la modalité A1R, nous avons les modalités A3E et A4E qui sont à fortes contributions relatives, ce qui signifie que malgré une certaine maîtrise du raisonnement bayésien direct, les étudiants font des raisonnements fallacieux de types « contraposée probabiliste » et « bayésien direct déterministe ». Autrement dit, ces deux types de raisonnements fallacieux, auxquels s'ajoute implicitement un raisonnement bayésien indirect (A2E), vont certainement avoir un effet sur leurs conceptions procédurales à propos des tests statistiques.




Graphique 2. Plan factoriel (2, 3)

Les modalités B1R et B2E sont du même côté que A1R, comme c'était le cas dans l'opposition du 1er axe factoriel ; en revanche, les items C1, C2, D1 et D2 ont tous du côté de A1R dans l'opposition du 2ème axe, exactement les modalités contraires à celles relevées du côté de A1R dans l'opposition du 1er axe, avec C1E comme étant la seule modalité à forte contribution relative. Le fait que les étudiants ne reconnaissent pas (C1E) le principe de « l'effet de la taille de l'échantillon » sur le rejet de $H_0$ dans le cas d'un test d'hypothèses n'est pas en faveur d'une bonne maîtrise procédurale d'un test statistique au sens de Neyman-Pearson. Dans le cas présent (B1R, B2E), les étudiants maîtrisent partiellement le principe de choix du niveau de signification, avec la même appréhension que celle que nous avons

relevée dans l'interprétation du 1er axe factoriel ; en revanche, la modalité D2E qui traduit la non reconnaissance de l'asymétrie des risques de 1ère et 2ème espèce dans le cas des tests d'hypothèses, corrobore la non maîtrise des étudiants d'une procédure de test statistique au sens de Neyman-Pearson. Par conséquent, la réussite D1R à propos de l'asymétrie de $H_0$ et $H_1$ (dans un test d'hypothèses) ne peut que traduire une approche correcte, non pas du point de vue des tests au sens de Neyman-Pearson, mais plutôt au sens de Fisher : en effet, le rejet de $H_0$ dans le cas de tests de signification ne permet pas *a priori* de conclure quant à l'acceptation de $H_1$. En conséquence, l'approche adoptée dans le cas présent par les étudiants dans le traitement de test statistique se trouve plutôt orientée vers celle des tests de signification, traduisant ainsi une assez bonne maîtrise de la procédure de traitement de test statistique au sens de Fisher.

Enfin, pour ce qui concerne la modalité C2R (à faible contribution relative), elle traduit tout simplement le fait que les étudiants considèrent l'hypothèse nulle $H_0$ comme étant sans intérêt (contrairement à ce que nous avions obtenu pour le 1er axe factoriel), autrement dit qu'ils considèrent plutôt $H_1$ comme étant l'hypothèse d'intérêt, ce qui est effectivement le cas dans la procédure de Fisher, et cela renforce justement auprès de ces étudiants l'idée d'une démarche selon l'approche de Fisher dans leur traitement de test statistique

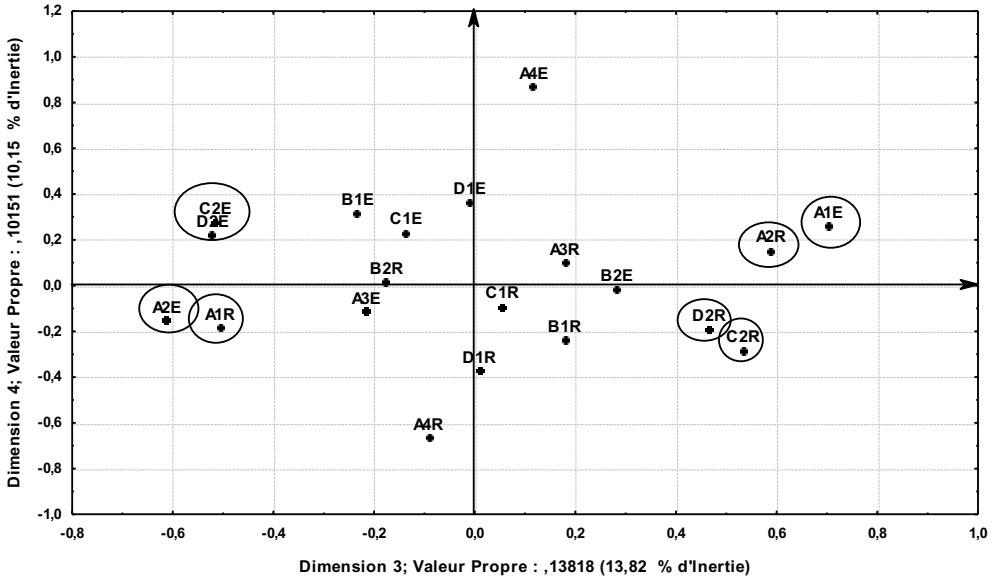
En conclusion, les tendances expliquées par le deuxième axe factoriel renvoient au degré de maîtrise du paradigme de probabilité conditionnelle sous-jacent au concept de test statistique, associé au degré de maîtrise procédurale des tests statistiques au sens de Fisher, notamment à propos du choix de niveau de signification. Par conséquent, le 2ème axe factoriel peut être interprété comme étant un axe de *maîtrise du paradigme de test statistique selon l'approche de Fisher.*

### *4.4.5. Interprétation du 3ème axe factoriel*

L'analyse factorielle de correspondances multiples conduit à une inertie de 11,77% pour l'axe 3. Les modalités qui contribuent le plus à la construction de cet axe sont celles entourées dans le graphique 3 ci-après. Elles ont une contribution relative comprise entre 8,3% et 14,9%. La qualité de représentation de ces modalités par rapport à cet axe est comprise entre 0,24 et 0,36.

Dans le troisième plan factoriel, le 3ème axe factoriel présente l'opposition suivante :

A1R, A2E, C2E et D2E vs A1E, A2R, C2R et D2R.




Graphique 3. Plan factoriel (3, 4)

Comme pour les deux plans factoriels précédents, nous avons d'un côté des modalités d'échec et de réussite et de l'autre les modalités correspondantes contraires. Dans ce cas aussi, nous nous limiterons à une interprétation de l'un de ces groupes de modalités, du fait que le groupe opposé aura l'interprétation contraire.

Enfin, nous constatons que même pour le 3ème plan factoriel, dans l'opposition des modalités à fortes contributions relatives, nous sommes en présence d'une opposition exclusive de modalités contraires, ce qui confirme maintenant de manière très nette (presque 100% de l'inertie (corrigée) totale, en comptant le cumul des inerties absolues des 3 premiers axes factoriels) le caractère d'homogénéité des items du questionnaire.

Du côté de la modalité A1R concernant le raisonnement bayésien direct, nous avons la modalité A2E (raisonnement bayésien indirect) qui est à forte contribution relative ; autrement dit, les étudiants ne présentent ici qu'une appréhension partielle du paradigme de probabilité conditionnelle sous-jacent au concept de test statistique. Quant aux modalités A3E « contraposée probabiliste » et A4R « raisonnement déterministe bayésien », au vu de leurs faibles contributions relatives, elles vont sensiblement renforcer l'aspect de non maîtrise totale du raisonnement sous-jacent au test statistique. L'échec au raisonnement bayésien indirect A2E (à forte contribution relative) est quasiment accompagné ici (hormis B2R qui du reste est à faible contribution relative) d'un échec au reste des items du

questionnaire. Par ailleurs, les modalités qui sont à fortes contributions relatives sont C2E et D2E, et elles traduisent respectivement, d'une part une non maîtrise procédurale aussi bien des tests de signification que ceux d'hypothèses (Fisher et Neyman-Pearson) à propos de l'interprétation du rejet de l'hypothèse nulle $H_0$ (approche déterministe), et d'autre part une non maîtrise procédurale des tests d'hypothèses (au sens de Neyman-Pearson) à propos de l'asymétrie des risques de 1ère et de 2ème espèces.

En conclusion, les tendances expliquées par le troisième axe factoriel renvoient au degré de maîtrise du paradigme de probabilité conditionnelle sous-jacent au concept de test statistique, associé au degré de la maîtrise procédurale des tests statistiques, selon simultanément les deux approches : Fisher et Neyman-Pearson. Par conséquent, le 3ème axe factoriel peut être interprété comme étant un axe de *maîtrise du paradigme de test statistique selon simultanément les deux approches : Fisher et Neyman-Pearson.*

## 5. Discussions, conclusions, et perspectives didactiques

L'analyse *a priori* des critiques que nous avons passées en revue à l'encontre de la théorie des tests statistiques, nous a permis de dégager des éléments qui remettent en question le bien-fondé de ces critiques. Si nous avions à résumer la nature de ces éléments de manière non exhaustive, nous dirions que ces éléments renvoient soit à des confusions au niveau du fondement théorique des tests statistiques, soit à des abus d'interprétation dans leur mise en pratique, voire éventuellement les deux à la fois :

a) Le raisonnement fallacieux dans le traitement inductif sous-jacent au test de signification de Fisher (cf. 3.1), la non interprétation de l'hypothèse nulle (de même pour l'alternative) en termes de modèle(s) probabiliste(s) (cf. 3.2), la non maîtrise de l'asymétrie entre $H_0$ et $H_1$ dans les deux approches : Fisher et Neyman-Pearson (cf. 3.3), l'amalgame procédural entre les deux approches de Fisher et Neyman-Pearson induit par le niveau de signification (cf. 3.4), la confusion au niveau du traitement des risques de 1ère et 2ème espèces dans la procédure de Neyman-Pearson (cf. 3.5) et le calcul de $P\left(H_0 \middle| D_0\right)$ moyennant des procédures de tests statistiques, renvoient tous *a priori* à des confusions à caractère théorique.

b) Le rejet de $H_0$ par augmentation de la taille de l'échantillon (cf. 3.2), l'asymétrie dans le traitement de $H_0$ et $H_1$ dans l'approche de Fisher (cf. 3.3), l'arbitraire dans le choix du niveau de signification dans un test statistique (cf. 3.4), renvoient *a priori* à des abus d'interprétation d'un point de vue pratique.

c) Le rejet de $H_0$ par augmentation de la taille de l'échantillon (cf. 3.2) et l'asymétrie dans le traitement de $H_0$ et $H_1$ dans l'approche de Fisher (cf. 3.3),

nous semblent *a priori* relever à la fois d'une confusion à caractère théorique et d'un abus d'interprétation d'un point de vue pratique.

Ainsi, quelle que soit la nature des critiques, il s'agit d'abord et avant tout d'une question de *maîtrise du concept de test statistique.* En effet, les résultats d'analyse factorielle multidimensionnelle des réponses des étudiants interrogés ont donné lieu à des axes factoriels avec des interprétations en termes de *maîtrise du concept de test statistique*, axes qui se déclinent selon l'importance de leur part d'inertie absolue (i.e. importance de leur part d'information contenue dans l'ensemble des réponses) suivant les interprétations respectives ci-après :

- *Maîtrise du paradigme de test statistique selon l'approche de Neyman-Pearson (Axe 1)*
- *Maîtrise du paradigme de test statistique selon l'approche de Fisher (Axe 2)*
- *Maîtrise du paradigme de test statistique selon simultanément les deux approches : Fisher et Neyman-Pearson (Axe 3)*

Si maintenant, nous regardons les modalités (de réussite ou d'échec) qui ont le plus contribué à la construction de ces axes factoriels (modalités à fortes contributions relatives), nous constatons que celles-ci renvoient à des éléments de la théorie des tests statistiques, qui sont de natures soit « fondement théorique », soit « pratique », soit les deux à la fois (cf. Tableau 6).

Ces modalités, lorsqu'elles contribuent à un axe factoriel, elles sont exprimées à la fois en termes de réussite et d'échec, autrement dit, l'interprétation des axes ne peut être exprimée qu'en termes de *maîtrise*, dont la spécification est entièrement déterminée par le groupe de modalités ayant contribué à la construction de chaque axe. Ainsi, les interprétations des axes factoriels qui renvoient à *la maîtrise du paradigme de test statistique selon l'approche de Neyman-Pearson, ou celle de Fisher, ou encore des deux à la fois*, constituent en fin de compte une validation de notre hypothèse de départ, à savoir que « *la majorité des critiques devrait être plutôt imputée à une non maîtrise du concept de tests statistiques par certains utilisateurs et aux abus interprétatifs qui en découlent, et non pas à des considérations intrinsèquement liées au concept même des tests statistiques ou à leur enseignement* ».

Maintenant, si nous nous référons à la nature des modalités qui ont donné lieu aux interprétations des axes factoriels, nous constatons qu'elle met plus en avant l'aspect « fondement théorique » que l'aspect « pratique ». Cela traduit une fois de plus, le fait que la maîtrise du concept de test statistique est davantage exprimée par rapport aux éléments de fondements théoriques de ce concept, et par voie de conséquence à un « *bon* » enseignement de ce dernier.

| Axe | Libellé et signification des modalités | Nature |
|---|---|---|
| 1 | A1 : « fondement théorique » | Raisonnement *bayésien* direct |
| | A3 : « fondement théorique » | Contraposition probabiliste |
| | B1 : « fondement théorique » | Choix de niveau de signification |
| | B2 : « pratique » | Choix de niveau de signification |
| | D1 : « fondement théorique » | Asymétrie dans le traitement de $H_0$ et $H_1$ dans une procédure de Neyman |
| 2 | A3 : « fondement théorique » | Contraposition probabiliste |
| | A4 : « fondement théorique » | Raisonnement *bayésien* direct déterministe |
| | C1 : « fondement théorique » et « pratique » | Effet de la taille de l'échantillon sur le rejet de $H_0$ |
| 3 | A1 : « fondement théorique » | Raisonnement *bayésien* direct |
| | A2 : « fondement théorique » | Raisonnement *bayésien* indirect |
| | C2 : « fondement théorique » | Rejet de $H_0$ dans la procédure de Fisher (approche déterministe) |
| | D2 : « fondement théorique » | Asymétrie des risques de 1ère et 2ème espèces |

Tableau 6. Nature(s) des modalités ayant contribué à la construction des axes factoriels

Néanmoins, il ne faudrait pas pour autant sous-estimer les autres aspects « pratiques » d'utilisation des tests statistiques, notamment en ce qui concerne le choix du niveau de signification ou l'effet de la taille de l'échantillon sur le rejet de l'hypothèse nulle, comme le souligne notre analyse factorielle. Ces aspects devraient *a priori* aussi être pris en charge par l'enseignement, contrairement à ce qui se fait habituellement, et plus généralement dans toute approche didactique des tests statistiques ; cela permettra certainement d'explorer des perspectives très intéressantes d'investigations en termes d'ingénierie didactique sur les tests statistiques.

A ce propos, en plaçant en individus supplémentaires sur les trois premiers plans factoriels les centres de gravité respectifs du groupe d'élèves ingénieurs et celui des étudiants en Master : le premier centre de gravité s'est toujours positionné du côté de la 'bonne maîtrise' pour les 3 premiers axes factoriels, contrairement au deuxième centre de gravité qui s'est systématiquement positionné du côté de la 'mauvaise maîtrise'. Est-ce un effet d'enseignement ? Cela mérite en tout cas une

étude didactique approfondie en termes d'approche d'enseignement des tests statistiques, autrement dit, l'exploration d'une nouvelle piste de recherche en didactique de la statistique.

## BIBLIOGRAPHIE

AMERICAN PSYCHOLOGICAL ASSOCIATION. (1994) *Publication manual of the American Psychological Association* (4th ed.). Washington, DC.

BAKAN, D. (1960) The test of significance in psychological research, *Psychological Bulletin* 66, 423-437.

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.

BEN-ZVI, D. & GARFIELD, G. (2004) Statistical Literacy, Reasoning, and Thinking: Goals, Definitions, and Challenges in D. Ben-Zvi et G. Garfield (eds), *The Challenge of Developing Statistical Literacy. Reasoning and Thinking*, 3-16. Dordrecht.

BRANDSTÄTTER E. & KEPLER, J.(1999) Confidence Intervals as an Alternative to Significance Testing, *Methods of Psychological Research Online* 4(2).

CARRANZA, P. (2011) Dualité dans l'enseignement de la probabilité. Apport pour l'enseignement de la statistique, *Recherches en Didactique des Mathématiques* 31-2, 229-259.

CARVER, R. (1953) The case against statistical significance testing, *Harvard Educational Review* 48, 378-399.

COHEN, J. (1977). *Statistical Power Analysis for the Behavioral Sciences* (2ème édition). Hillsdale, NJ: Erlbaum.

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

DENIS, D. J. (2004) The Modern Hypothesis Testing Hybrid: R. A. Fisher's Fading Influence. With Discussion by Michel Armatte, Bernard Bru, Michael Friendly, Jeff Gill, Ernest Kwan, Bruno Lecoutre, Marie-Paul Lecoutre, Jacques Poitevineau and Stephen Stigler. *Journal de la Société Française de Statistique* 145(4). 5-68.

ELM'HAMEDI, Z. (2010) *Contribution à une ingénierie didactique pour l'enseignement et l'apprentissage des tests statistiques à l'université*. Thèse de

Doctorat. Université Sidi Mohammed Ben Abdellah- Faculté des Sciences Dhar El Mehraz, Fès, Maroc.

FISHER, R. (1925). *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd.

FISHER, R. (1935) *The Design of Experiments*, Edinburgh: Oliver & Boyd.

FISHER, R. (1956). *Statistical Methods and Scientific Inference*, Edinburgh: Oliver & Boyd.

GIGERENZER, G. (1993) The superego, the ego, and the id in statistical reasoning, in G. Keren & C. Lewis (Eds.) *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, (p. 267-288). Hillsdale, NJ: Erlbaum.

GRAS, R. & TOTOHASINA, A. (1995). Chronologie et causalité, conceptions sources d'obstacles épistémologiques à la notion de probabilité conditionnelle. *Recherches en Didactique des Mathématiques* 15/1, 49-95.

HAGER, W. (2000) About some misconceptions and the discontent with statistical tests in psychology, *Methods of Psychological Research Online* 5(1).

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.

JONES, L. V. (1955). Statistical theory and research design. *Annual Review of Psychology* 6. 405-430.

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.

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

LECOUTRE, B., LECOUTRE M.-P. & POITEVINEAU, J. (2001). Uses, abuses and misuses of significance tests in the scientific community: won't the Bayesian choice be unavoidable? *International Statistical Review* 69. 399-418.

LECOUTRE, B. & POITEVINEAU, J. (2000). Aller au-delà des tests de signification traditionnels : vers de nouvelles normes de publication [Beyond traditional significance tests: Prime time for new publication norms]. *L'année Psychologique* 100. 683-713.

LOFTUS, G. (1996). Psychology will be much better science when we change the way we analyze data. *Current Directions in Psychological Science* 5. 161-170.

LOVELL, M.C. (1983). Data mining. *Review of Economics and Statistics* 65. 1-12.

MACDONALD, R. R. (2004). Statistical inference and Aristotle's rhetoric. *British Journal of Mathematical and Statistical Psychology* 57. 193-203.

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

MEEHL, P. E. (1967). Theory testing in psychology and physics: A methodological paradox, *Philosophy of Science* 34. 103-115.

MEEHL, P. E. (1978) Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology* 46. 806-834.

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

NEYMAN, J. & PEARSON, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika* 20A. 175-240.

NEYMAN, J & PEARSON, E. S. (1933) On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, Series A 231. 289-337.

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

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

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* (p. 17-46). Dordrecht.

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.

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

ROZEBOOM, W. W. (1960). The fallacy of null hypothesis significance testing, *Psychological Bulletin* 57. 416-428.

SCHMIDT, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology. Implications for training of researchers. *Psychological Methods* 1. 115-129.

SEDLMEIER, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. *Methods of Psychological Research Online* 1. 41-63.

TUKEY ,J. W., (1977) *Exploratory data analysis*. Reading, MA: Addison-Wesley.

WILKINSON, L., TASK FORCE ON STATISTICAL INFERENCE (1999). Statistical methods in psychological journal, *American Psychologist* 54. 594-604.

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

ZENDRERA, N. (2010) *Enseignement et apprentissage des tests d'hypothèses paramétriques : difficultés rencontrées par les étudiants en sciences humaines. Une contribution à l'éducation statistique.* Thèse de Doctorat de l'Université Sherbrooke (Québec, Canada).

Moncef ZAKI
Moncef Zaki <zaki.moncef@yahoo.fr>

Zahid El M'HAMEDI

Université Sidi Mohammed Ben Abdellah
Faculté des Sciences Dhar El Mehraz
Fès, Maroc