HAL
archives-ouvertes.fr

Quelques obstacles rencontrés dans l'apprentissage de l'analyse statistique implicative

Jean-Claude Oriol, Jean-Claude Régnier

▶ **To cite this version:**

Jean-Claude Oriol, Jean-Claude Régnier. Quelques obstacles rencontrés dans l'apprentissage de l'analyse statistique implicative. 41èmes Journées de Statistique, SFdS, Bordeaux, 2009, Bordeaux, France, France. 2009. <inria-00386688>

HAL Id: inria-00386688
https://hal.inria.fr/inria-00386688

Submitted on 22 May 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

QUELQUES OBSTACLES RENCONTRÉS DANS L'APPRENTISSAGE DE L'ANALYSE STATISTIQUE IMPLICATIVE

Jean-Claude Oriol* & Jean-Claude Régnier**

* Université de Lyon IUT Lumière-CERRAL Jean-claude.oriol@univ-lyon2.fr
** Université de Lyon UMR 5191 ICAR Jean-claude.regnier@univ-lyon2.fr

L'analyse statistique implicative est une méthode d'analyse développée par Régis Gras qui se propose de donner un sens statistique à des expressions du type a => b dans le cas où l'implication est non stricte, c'est-à-dire que lorsque la variable a est vraie, b a peu de chances d'être fausse. Notre activité s'est développée dans le cadre de projets d'études statistiques d'une durée de 70 heures menés avec des étudiants de deuxième année du DUT STID (Statistique et Traitement Informatique des Données), durant trois années successives. Ces étudiants ont en principe une culture statistique assez riche tant sur l'axe des contenus (statistique descriptive, statistique inférentielle, approche de la classification, ADD, ACP, AFC, etc.) que sur celui des logiciels (SAS, SPSS, SPAD, Mapinfo, Excel, etc…) mais l'analyse statistique implicative leur est complètement étrangère.
Notre propos ici n'est pas de développer directement ce type d'analyse mais de montrer les obstacles rencontrés par nos étudiants lors de l'apprentissage de cette analyse et les objets d'enseignement mis en place afin de les surmonter.

1. Le contexte des projets d'études statistiques

Les projets d'études statistiques regroupent cinq ou six étudiants qui travaillent sur un sujet proposé par un commanditaire partenaire du département. Ces projets durent environ 70 heures entre janvier et juin. Pour chaque groupe un enseignant ou un professionnel (différent du commanditaire) joue le rôle d'animateur du groupe. Le groupe peut éventuellement faire appel à un "expert" afin d'avoir une aide ponctuelle sur un point délicat (interprétation de résultats par exemple). A la mi parcours l'ensemble des étudiants participe à un "séminaire" au cours duquel chaque groupe expose devant l'ensemble des étudiants et des animateurs l'état d'avancement des travaux. Le travail se termine par une soutenance devant le commanditaire, l'animateur et un ou plusieurs enseignants du département. Cette soutenance est complétée par la remise d'un dossier au commanditaire et la remise d'un résumé de 4 à 8 pages à fin de publication, validé par le commanditaire afin d'assurer la confidentialité des données. La soutenance, le rapport et le résumé sont pris en compte afin d'attribuer une note de projet comptée dans l'attribution du semestre 4. Chaque année un groupe d'étudiant travaille sur un projet en utilisant l'analyse statistique implicative (ASI) et le logiciel CHIC (Classification Hiérarchique Implicative et Cohésitive, logiciel développé par R. Couturier, A. Bodin, R. Gras).

2. Point de départ de l'analyse statistique implicative.

L'analyse implicative travaille à donner un sens statistique à a => b dans le cas où l'implication est non stricte. Ci-dessous un exemple (GRAS, 1996) d'implication entre deux variables binaires :

Sujet	1	2	3	4	5	6	7	8	9	10	
a	0	0	1	1	0	1	1	0	0	0	4
b	0	1	1	0	0	1	1	0	1	0	5

A partir de ce tableau, on calcule classiquement :

a / b	1	0	**Marge**
1	3	2	**5**
0	1	4	**5**
Marge	**4**	**6**	**10**

Soit en généralisant :

a / b	1	0	**Total**
1	$n_{a \cap b}$	$n_{\bar{a} \cap b}$	n_b
0	$n_{a \cap \bar{b}}$	$n_{\bar{a} \cap \bar{b}}$	$n_{\bar{b}}$
Total	n_a	$n_{\bar{a}}$	n

La quantité $n_{a \cap \bar{b}}$ est la réalisation de la variable aléatoire $Card\left(X \cap \overline{Y}\right)$ qui suit sous certaines conditions (Gras, 1996) la loi de Poisson de paramètre $n * p(a) * p\left(\bar{b}\right)$.
A l'origine, pour ce genre de modélisation, la loi pressentie était la loi binomiale. La méthode de maximum de vraisemblance consiste à proposer la valeur pour laquelle la probabilité de l'observation dans le modèle est la plus forte. En appliquant cette méthode, on a pu approximer cette loi par une loi de Poisson.

3. De l'indice d'implication à l'intensité d'implication

On peut alors construire (Gras, 1996) un indice d'implication et une intensité d'implication de la façon suivante :

$$Q(a, \bar{b}) = \frac{\mathrm{Card}\left(X \cap \overline{Y}\right) - np(a)\,p(\bar{b})}{\sqrt{np(a)\,p(\bar{b})}} = \frac{Card(X \cap \overline{Y}) - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}}$$

La variable aléatoire $Q(a, \bar{b})$ suit une loi normale centrée réduite. Elle permet de définir l'intensité d'implication $\varphi\left(a, \bar{b}\right)$ qui correspond à la qualité de l'admissibilité de a =>b.

$$\varphi(a,\bar{b}) = 1 - P\left[Q(a,\bar{b}) \le q(a,\bar{b})\right] = \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt$$

4. Le graphe implicatif, l'arbre hiérarchique ou cohésitif

Parmi les signifiants de ce concept statistique existent deux représentations graphiques : le graphe implicatif et l'arbre hiérarchique dit cohésitif
Tout d'abord, le graphe implicatif permet de voir les liens d'implication entre plusieurs variables. En effet, le principe du graphe est de calculer les indices d'implication. Puis on ne représente que les variables qui ont les liens les plus forts.
Ci-dessous un exemple de graphe implicatif à partir de données quelconques.

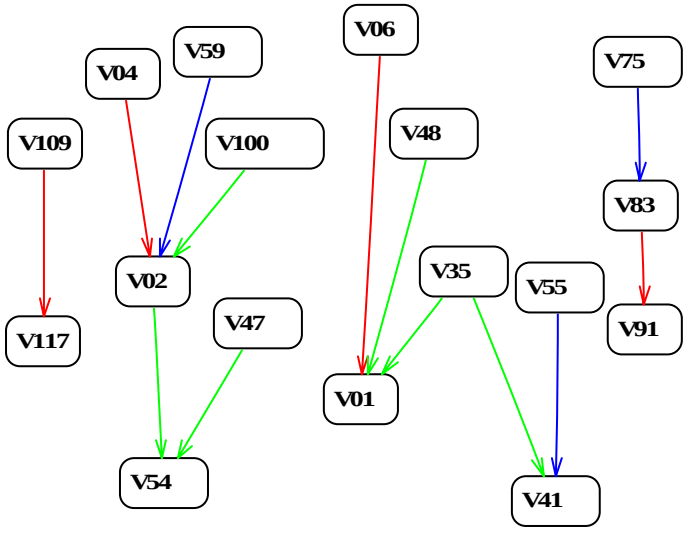

Figure 1 : Exemple de graphe implicatif

Par exemple, si V04 est vraie alors la probabilité pour que V02 soit vraie est égale à 0,99.
La deuxième représentation illustrant le concept d'implication statistique est l'arbre hiérarchique. Cette figure représente les variables avec leur significativité présentée sous forme de niveaux, à partir d'un fichier de données quelconques.

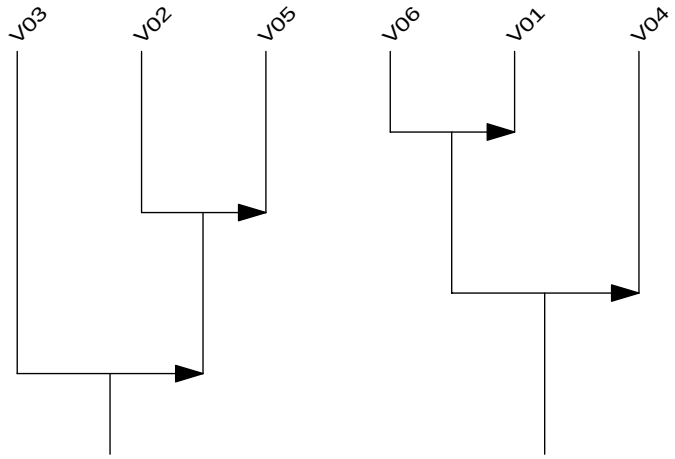

Figure 2 : Exemple d'arbre cohésitif

En lisant ce graphique, nous constatons que la variable V03 implique le couple de variables constitué de V02 et V05. En revanche c'est le couple de variables V06/V01 qui contribue à la réalisation de la variable V04.

5. L'analyse des similarités

Il existe un deuxième concept nécessaire pour comprendre le fonctionnement de la classification ascendante hiérarchique : l'analyse des similarités. Selon I.C. LERMAN, cette analyse consiste à étudier des classes de variables semblables, grâce à la création d'une typologie. Cette typologie est créée de la manière suivante : à partir de l'ensemble des variables, le logiciel regroupe celles-ci en classes de taille de plus en plus grande, dont l'effectif est de moins en moins important.
Plus le lien entre deux variables est fort, plus celles-ci sont semblables. En effet, la similarité

se définit de la sorte : soit deux variables quelconques A et B, plus la valeur de $A \cap B$ est élevée, plus ces deux variables sont similaires. L'indice de similarité est calculé par la probabilité que la valeur de $A \cap B$ soit supérieure à un nombre au hasard.
Le graphique de l'arbre des similarités se construit de la manière suivante : on réunit en une classe, au premier niveau, les deux variables qui se ressemblent le plus au sens de l'indice de similarité ; puis on fait de même pour deux autres variables ou une variable et une classe déjà formée et ainsi de suite.

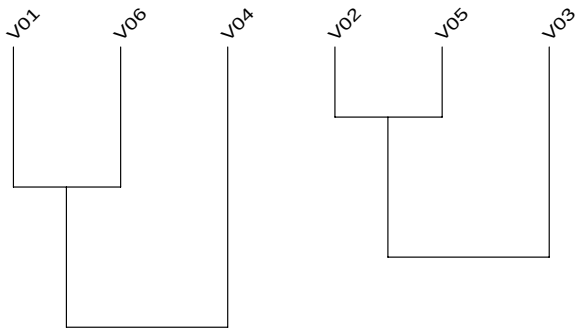

Figure 3 : Exemple d'un arbre des similarités

Ce concept est plus familier aux étudiants habitués, en raison de leurs enseignements de statistique, aux problématiques liées aux notions de distance. On note ici la distance implicative d'un individu x à la classe C est le nombre :

$$d(x,C) = \left[\frac{1}{n} \sum_{i=1}^{i=n} \frac{(\varphi_i - \varphi_{x,i})^2}{1-\varphi_i} \right]^{\frac{1}{2}}$$

où φ correspond à l'intensité d'implication définie précédemment.

6. Les difficultés immédiates rencontrées par les étudiants

Voici dans l'ordre d'apparition les difficultés rencontrées par les étudiants :

6.1. les cours de logique ayant presque disparus de l'enseignement (dans le DUT qu'ils suivent ils étaient faits en mathématique mais depuis le nouveau programme pédagogique national ils sont inclus dans les enseignements d'informatique qui, semble-t-il, se « passent » allègrement de la formalisation. En conséquence ils ne maîtrisent pas la logique des propositions.

6.2. On le sait la variabilité a une place centrale dans la statistique et joue un rôle primordial dans la conceptualisation des notions rencontrées dans ce champ. Devant ces nouveaux concepts les étudiants manquent de repères et « perdent pied ».

6.3. Trouver un lien avec les concepts déjà étudiés dans le cours de statistique.

En termes de réponse à la première difficulté nous avons dispensé aux étudiants un cours de logique, et pour les autres l'idée de faire construire des simulations par les étudiants est apparue comme « naturelle », c'est ce que nous développons dans les paragraphes suivants.

7. Simulations concernant l'analyse statistique implicative

7.1. La place de la simulation

C'est la notion de champ conceptuel en tant qu'ensemble des situations renvoyant à l'idée de procédure (Vergnaud, 1990) qui permet de situer la simulation en statistique dans sa finalité qui est de fournir à l'apprenant un éclairage sur le signifié (invariants opératoires).

Par ailleurs l'approche développée par Jean-Claude Régnier (1988) intégrant l'apprentissage fondé sur le tâtonnement expérimental de l'apprenant offre une perspective pour analyse didactique de la simulation et relation à la résolution de problème.

7.2. La simulation outil d'appropriation de concepts statistiques

Nous avons débattu de l'utilité et de l'utilisation de la simulation dans diverses situations rencontrées dans l'enseignement de la statistique entre autres concernant la corrélation (Oriol et Régnier 2003a), ou d'autres concepts (Oriol, 2007).

C'est sans doute un des points spécifiques de l'enseignement de la statistique obligé de développer un enseignement s'appuyant sur les mathématiques mais hétérodoxe par rapport aux outils traditionnels construits dans l'exclusion entre le vrai et le faux. Le propre du raisonnement scientifique est que des mêmes conditions vont produire des effets identiques. En statistique il n'en est rien et c'est cela sans doute une des difficultés que rencontrent les étudiants. La simulation permet de distinguer les invariants dans la variabilité.

Notons d'ailleurs que dans notre contexte de l'analyse statistique implicative la pensée développée est « doublement » hétérodoxe : d'une part comme toute pensée statistique et d'autre part comme s'intéressant à des énoncés a =>b « partiellement » vrais.

7.3. La simulation proposée

Le choix d'Excel : notre pratique pédagogique vise à intégrer le plus tôt possible l'outil informatique comme instrument canonique d'une pratique de la statistique.

La construction de la simulation : les étudiants doivent construire une feuille Excel tirant 100 fois au hasard les valeurs binaires de a et b, évaluer l'indépendance des variables a et b, calculer pour ces 100 valeurs l'indice d'implication et l'intensité d'implication entre a et b. Et recommencer… Voici les 10 premières lignes d'un tableau (le tableau comporte en réalité 100 lignes) dans lequel les valeurs de a et de b sont 0 ou 1 obtenues aléatoirement avec le générateur de nombres aléatoires d'Excel :

	A	b					Vérif.
							100
			16	35	26	23	100
1	0	1	0	1	0	0	1
2	1	1	0	0	0	1	1
3	1	0	0	0	1	0	1
---	---	---	---	---	---	---	---

Le tableau récapitulatif correspondant obtenu :

a →	0	1	
↓b			
0	16	26	42
1	35	23	58
	51	49	100

On en déduit le tableau de contingence et le calcul de la valeur du Khi deux

Tableau d'indépendance			
	21,42	20,58	
	29,58	28,42	
Calcul du Khi²			
	1,371	1,427	
	0,993	1,034	
Valeur du Khi²			4,826

Et également le calcul de l'indice d'implication et de l'intensité d'implication

Calcul de l'indice d'implication	
q(a,non(b)) =	1,195

Calcul de l'intensité d'implication

Phi(a,non(b)) =	0,116

Figure 4 : Comparaison des valeurs de q et des valeurs de Phi correspondantes

Voici une série de résultats obtenus concernant des indices d'implication et les valeurs d'intensités correspondantes

Valeurs de k	Valeurs de q	Valeurs de Phi
1	-1,74207716	0,95925257
2	-0,48564293	0,68638982
3	-0,79259392	0,78599284
4	1,46026115	0,0721092
5	-0,06428571	0,52562871
6	-0,00769231	0,50306877
7	-0,48107024	0,68476671
8	0,20225996	0,41985677
9	-0,38392627	0,6494834
10	-0,32433749	0,62715866

Et une représentation graphique :

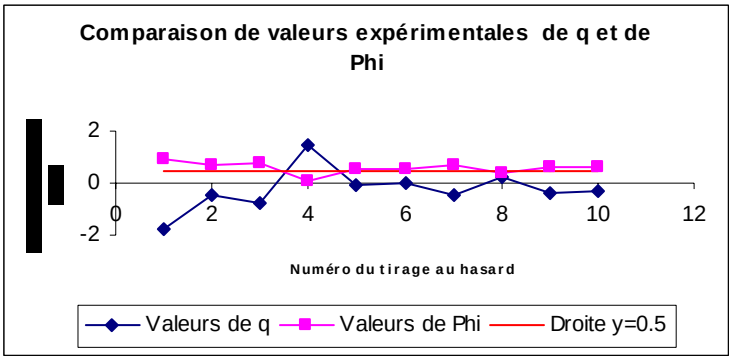

Figure 5 : Indices d'implication et intensités d'implication

La lecture des dizaines de graphiques obtenus, indique que si l'indice q est inférieur à 0 alors l'intensité est supérieure à 0,5 et permettra de passer aux démonstrations de ces propriétés puis à poursuivre et à aborder les concepts suivants du champ de l'ASI.

7.4. Recherche de bénéfices collatéraux : les valeurs du Khi², le test de Mc Némar

Comme nous pouvons « faire tourner » cette simulation il est intéressant de stocker les résultats afin de les comparer à des résultats prévisibles par la théorie statistique. Puisque on génère des valeurs du Khi² on peut les comparer aux seuils théoriques à 5% et à 1% par exemple. Ainsi dans le tableau ci-dessous sur les 640 valeurs du Khi2 obtenues nous en avons 6 au dessus du seuil de 1% et 32 au dessus du seuil de 5% ce qui correspond aux valeurs attendues.

Valeur du Khi2	2,793491

Seuils théoriques	
A 5%	A 1 %
3,841	6,635

Valeurs observées			
634	6	640	0,9%
608	32	640	5,0%

Figure 6 : Comparaison des valeurs simulées et théoriques

D'une façon similaire nous avons construit le test de Mc Nemar que les étudiants venaient d'étudier dans le cours sur les méthodes consacrées aux tests non paramétriques.

7.5. Recherche de bénéfices directs concernant l'ASI : q est symétrique, il existe une relation entre le Khi² et q(a,non(b)). Nous avons recherché à vérifier les relations entre q(a,non(b)) et q(non(b),a). Cela permet aux étudiants d'être plus à l'aise avec les formules

Calcul de l'indice d'implication	
q(a,non(b))=	0,909

Calcul de l'indice d'implication	
q(non(b),a)=	0,909

Et également à vérifier la relation entre le Khi² et q(a,non(b))

Phi(a,non(b))=	0,18

Khi²/q(a,non(b))=	3,92
n²/nb*nnon(a)=	3,92

7.6. Comparaison du calcul de Phi avec la loi de Poisson et avec la loi normale

Il nous a semblé intéressant de comparer le calcul de Phi d'une part à l'aide de la loi de Poisson et d'autre part avec la loi normale

Calcul de l'intensité d'implication (loi normale)		avec la loi de Poisson	Lambda=	26,3
Phi(a,non(b))=	0,674443		0,62776283	

Figure 7 : Comparaison des valeurs de l'intensité d'implication (loi de Gauss et loi de Poisson)

D'une expérience à l'autre l'écart varie peu et il est de l'ordre de 0.05.

7.7. Observation de 100 valeurs de q, de 500 valeurs de Phi

Nous avons ensuite représenté et observé 100 valeurs au hasard de q

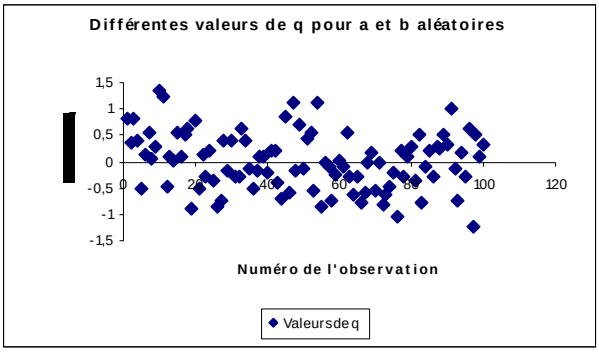

Figure 8 : 100 valeurs « au hasard » de l'indice d'implication

Puis 500 valeurs de l'intensité d'implication :

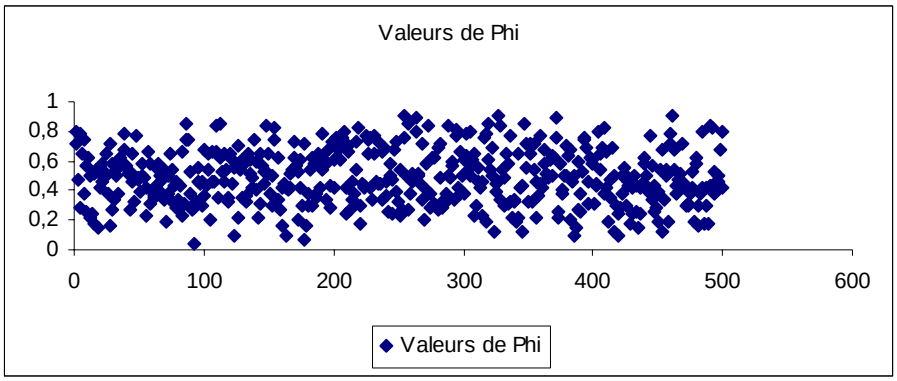

Figure 9 : 500 valeurs de l'intensité d'implication

8. Conclusion sur les simulations concernant l'ASI

Ici l'exigence d'utiliser une théorie nouvelle et un outil inconnu oblige les étudiants à un aller retour entre le réel et ses représentations, entre signifié et signifiant, à construire des invariants opératoires, à construire le sens des situations, bref à conceptualiser l'analyse statistique implicative. Il nous apparaît donc, une fois encore, que l'organisation du couple situation schème permet une meilleure appropriation des concepts statistiques par les étudiants.

Dans le cas présenté, comme dans beaucoup d'autres (Oriol, 2007), la construction par les apprenants de l'outil de simulation leur permet de dégager des invariants de la variabilité de phénomènes non déterministes. Les simulations «prêtes à être admirées» emmènent le plus souvent les étudiants à une «bovarysation» des notions étudiées. Ici, en revanche, la construction par les apprenants, eux-mêmes, de l'outil s'apparente aux méthodes d'apprentissage où les premiers travaux des apprentis consistaient à fabriquer leurs propres outils (l'équerre pour un outilleur par exemple, ou encore une brouette pour un menuisier) et à la mise en scène du tâtonnement expérimental producteur d'apprentissages.

La théorie des champs conceptuels (Vergnaud, 1990) et le modèle d'apprentissage fondé sur le tâtonnement expérimental (Régnier, 1988) permettent de situer ces activités de simulation dans l'apprentissage de la statistique lieu où elles ont une place mais bien entendu pas toute la place. Elles permettent comme nous l'avons dit de dégager les invariants de la variabilité des situations statistiques.

Bibliographie

[1] Gras, R (1996), *L'implication statistique – Nouvelle méthode exploratoire de données*, La pensée sauvage.

[2] Oriol, J-C. Régnier, J-C. (2003a) « Fonctionnement didactique de la simulation en statistique : Exemple de l'enseignement du concept d'intervalle de confiance », *35èmes Journées de Statistique*, SFDS Lyon, tome 2.

[3] Oriol, J-C. Régnier J-C., (2003b) « Fonctionnement didactique de la simulation en statistique dans l'enseignement du concept de corrélation », *Espace Mathématique Francophone 2003*, Tozeur, Tunisie.

[4] Oriol, J-C. Régnier, J-C. (2006) Formation en statistique en DUT STID et Transposition didactique, *38èmes Journées de Statistique*, SFDS Clamart.

[5] Oriol, J-C. Régnier, J-C. (2007a) Conceptualisation de l'analyse statistique implicative, *39èmes Journées de Statistique*, SFDS Angers.

[6] Oriol, J-C. Régnier, J-C. (2007b) Enseignement - apprentissage de l'ASI en 1er cycle universitaire, *Actes ASI 4, Castellon* (Espagne).

[7] Oriol, J-C, (2007) *Formation à la statistique par la pratique d'enquêtes par questionnaires et la simulation : étude didactique d'une expérience d'enseignement dans un département d'IUT*, Thèse Lyon2.

[8] Régnier, J-C. (1988), Étude didactique d'une méthode d'apprentissage fondé sur le tâtonnement expérimental de l'apprenant, *Annales de Didactique et de Sciences Cognitives, séminaire de Didactique des Mathématiques de Strasbourg*, pp 255-279.

[9] Régnier, JC, (2006) Formation de l'esprit statistique et raisonnement statistique. Que peut-on attendre de la didactique de la statistique ? in C.Castela et C.Houdement (Dir.) *Actes du séminaire national de Didactique des Mathématiques*. Année 2005. Editeurs: ARDM & IREM de Paris 7 (pp.13-37)

[10] Vergnaud, G. (1990), La théorie des champs conceptuels. *Recherches en Didactique des Mathématiques*. 10(2-3).