L'E.D.A. AU SECOURS DE L'O.G.D.

ou

QUELQUES REMARQUES CONCERNANT L'ENSEIGNEMENT DE LA STATISTIQUE DANS LES COLLEGES

Michel JULLIEN
Gérard NIN
I.R.E.M. d'Aix-Marseille

L'article qui suit est le bilan (provisoire) d'un travail de recherche-formation-développement (Chevallard, 1986) qui a commencé au début de l'année 1987. La recherche entreprise alors a permis d'organiser un stage du Plan Académique de Formation pendant le premier trimestre de l'année scolaire 1987-1988. Le développement (en cours) consiste, pour les stagiaires, à dispenser un enseignement en classe de sixième qui utilise certains éléments du stage - c'est ce qui a été fait pendant le troisième trimestre - et, pour le formateur, à récupérer les informations issues de cet enseignement (au cours de séances de travail avec certains ex-stagiaires) pour approfondir sa recherche et modifier le contenu de son stage - ce qui devrait lui permettre d'être mieux à même de satisfaire d'éventuelles futures demandes de formation.

Précisons que, dans la référence à la rubrique Organisation et Gestion de Données (O.G.D.), l'interprétation des textes que nous avons faite consiste à penser que le désir des auteurs des programmes est précisément la mise en place d'un enseignement des statistiques. Un article, paru (petit x n° 13), signé par Henri Bareil vient d'ailleurs confirmer cette interprétation.

I - QUELQUES QUESTIONS QUI ONT JALONNE NOTRE TRAVAIL.

1. Qu'est-ce que la statistique ?

Réglons d'abord un problème de nombre (au sens grammatical du mot) : doit-on parler de la statistique ou des statistiques ? Pour éviter cette difficulté linguistique l'usage est de qualifier les noms : on parlera alors de **la** statistique mathématique (ou inférentielle) et **des** statistiques descriptives. Le passage du singulier au pluriel est à rapprocher de l'opposition, un temps soulignée, entre **la** mathématique et **les** mathématiques. D'un côté l'unité des structures et de l'autre la multiplicité de techniques «hétéroclites».

Cela dit, la question que nous posons ne concerne pas la seule statistique mathématique mais l'objet culturel «statistique» dans son ensemble, ce qui lui donne une ampleur écrasante et convaincra sans doute le lecteur qu'il y a bien là matière à interrogation, même si cette interrogation ne fait pas partie des habitudes de l'enseignant qui entretient souvent, inconsciemment, l'illusion de la transparence du savoir.

«petit x» n° 19 pp. 29 à 41, 1989

Au risque de décevoir nous laisserons de côté la question de la **nature** de la (des) statistique(s) (Chevallard, 1978) et nous tenterons seulement de répondre à la question «qu'est-ce qu'avoir une activité statistique ?». Pour cela nous **poserons** que l'activité statistique est essentiellement celle... du statisticien. Or il existe des statisticiens et certains d'entre eux (Volle, 1980) ont eu le bon goût de nous livrer de précieux témoignages sur leurs activités dont nous distinguerons trois types :

- **type 1** : la collecte, le regroupement d'informations et leurs stockages sous des formes diverses ;
- **type 2** : le résumé, l'illustration des informations recueillies et la recherche de nouvelles informations à partir de ces résumés et de ces illustrations ;
- **type 3** : la confirmation et l'extrapolation d'informations obtenues à partir d'un échantillon.

Il est alors facile de concevoir qu'une même personne puisse, à des moments différents, être concernée par chacun de ces trois types d'activités et donc faire tour à tour des statistiques et de la statistique.

Plus précisément on reconnaîtra dans l'activité de type 1 un travail qui est le plus souvent réservé aux statisticiens professionnels, c'est-à-dire à ceux qui travaillent au sein de l 'institution statistique au centre de laquelle, en France, se trouve l'I.N.S.E.E. Concevoir des nomenclatures, des questionnaires, des procédés de dépouillement et de vérification de ces questionnaires, constituer des fichiers «propres», les fusionner avec d'autres fichiers, etc., telle est une grande partie du travail de ces statisticiens.

L'activité de type 3 est celle que nous appellerons l'activité confirmatoire (Tuckey, 1977) ; elle englobe toutes les méthodes de la statistique mathématique traditionnelle : estimation ponctuelle, par intervalles de confiance, théorie des tests, analyse de la variance, régression..., ainsi que des méthodes statistiques plus récentes comme les méthodes issues de la théorie de la robustesse et celles qui constituent la statistique non paramétrique (Lecoutre et Tassy, 1977). L'outil commun à toutes ces méthodes est le calcul des probabilités et leur mise en œuvre nécessite toujours le choix **a priori** d'une loi ou d'une famille de lois de probabilité.

Les activités de type 2 sont classiquement abordées par des techniques de statistiques descriptives : calculs de paramètres de position, de dispersion, diagrammes divers et, plus récemment, lorsque les données ont un caractère multidimensionnel, par des méthodes dites d'analyse des données (analyse en composantes principales, analyse factorielle des corrrespondances, analyses typologiques,...) qui s'appuient, pour la plupart, sur la géométrie des espaces vectoriels réels de dimensions finies, le calcul matriciel et les algorithmes numériques qui s'y rattachent. Ces dernières méthodes sont particulièrement adaptées aux traitements de grands volumes de données et nécessitent l'utilisation de moyens de calculs informatiques.

Si l'on s'en tient à des corpus de données compatibles avec un traitement «à la main» la partie **exploratoire** de ces activités a vu se développer, depuis une quinzaine d'années environ, un domaine nouveau, celui de l'analyse exploratoire des données - en anglais, Exploratory Data Analysis (E.D.A.) - qui nous paraît intéressant dans la nouvelle perspective curriculaire du collège (O.G.D.) tant par son esprit que par certaines de ses méthodes. Bien sûr, l'intérêt que nous allons accorder à ce domaine ne saurait être interprété comme un rejet des techniques de statistiques descriptives classiques, lesquelles peuvent aussi être enseignées dans une perspective exploratoire.

2. Qu'est-ce que l'E.D.A. ?

L'E.D.A. est née aux Etats-Unis sous l'impulsion du statisticien J.W. Tuckey (op. cit.). Comme l'analyse des données elle est **anti-probabiliste**, c'est-à-dire qu'elle ne suppose pas que les données soient distribuées suivant une loi de probabilité

classique (Normale le plus souvent), elle n'utilise que des notions mathématiques très élémentaires et des procédés graphiques faciles à mettre en œuvre. Jusque-là, elle est donc assez proche des statistiques descriptives traditionnelles, mais elle s'en éloigne par ses intentions.

La représentation ou le calcul ne sont pas en E.D.A. une fin en soi mais plutôt un moyen de **découvrir** une information cachée dans les données. Comme le dit I.J. Good (1983) «l'analyste qui fait de l'E.D.A. sait qu'il veut formuler des hypothèses tandis que celui qui fait des statistiques descriptives classiques en est moins conscient». Pour l'aider dans son **travail de détective** celui qui fait de l'E.D.A. mise sur la multiplicité des représentations et sur l'utilisation de grandeurs statistiques peu sensibles aux valeurs aberrantes (qui entachent souvent les données réelles) ; c'est ce qui explique que la médiane sera préférée à la moyenne et un intervalle interquartile à un écart-type.

Présentons maintenant quelques techniques d'E.D.A. à partir d'un exemple.
Le tableau suivant donne les pourcentages du P.I.B. consacrés à la protection sociale par 9 pays de la communauté européenne (d'après «Tableaux de l'économie française», 1987).

	1970	1981	1982	1984
Pays-Bas	20,8	31,4	33,3	34,0
Belgique	18,7	30,0	31,4	39,6
R.F.A.	21,5	29,4	29,4	28,9
Danemark	19,6	30,1	30,3	27,6
Luxembourg	15,9	27,8	28,9	29,3
France	19,2	27,4	28,5	28,8
Italie	17,4	25,3	25,8	27,3
Royaume-Uni	15,9	23,4	23,0	23,8
Irlande	13,2	21,9	23,8	23,5

Un même principe de représentation appelé «stem and leaf» (branche et feuille) dans la littérature statistique américaine va permettre de comparer entre elles les différentes années et suggérer une typologie des différents pays concernés. Pour cela on prend pour «branches» les unités des nombres qui mesurent les pourcentages et pour «feuille» les dixièmes d'unités de ces mêmes nombres. Ce qui donne :

1984	1970		1970	1981	1982	1984
	IRL	13	2			
		14				
	GB ; LUX	15	9 ; 9			
		16				
	ITA	17	4			
	BEL	18	7			
	FRA ; DK	19	2 ; 6			
	PB	20	8			
	RFA	21	5	9		
		22				
IRL ; GB		23		4	0 ; 8	8 ; 5
		24				
		25		3	8	
		26				
ITA ; DK		27		8 ; 4		6 ; 3
FRA ; RFA		28			9 ; 5	9 ; 8
LUX ; BEL		29		4	4	6 ; 3
		30		0 ; 1	3	
		31		4	4	
		32				
		33			3	
PB		34				0

Dans ce tableau, les deux «9» de la colonne de droite, notée 1970, signifient que deux pays affectaient 15,9 % de leurs P.I.B. à la protection sociale en 1970.

Dans les colonnes de gauche ce sont les différents pays qui sont les feuilles de la représentation, les branches restant les mêmes. Bien sûr, il est tout à fait possible, et même fréquent, de ne représenter que les colonnes de droite, ou encore, si on ne cherche qu'à comparer deux séries de données, de les représenter «dos à dos» : l'une à droite et l'autre à gauche. Il faut bien comprendre qu'ici aucune disposition n'est imposée, c'est celle qui **parle** le plus qui est la meilleure !

Comme nous l'avions annoncé, le graphisme associé à ce type de représentation est particulièrement frustre, mais il n'en constitue pas moins pour autant un **enrichissement** d'une des représentations les plus utilisées en statistiques descriptives classiques, à savoir l'histogramme. En effet, ici les données sont individualisées et leurs répartitions au sein de chaque intervalle de valeurs restent visibles (elles peuvent même être ordonnées). La cardinalité de chacune des classes peut aussi être mentionnée mais elle est en général lisible sans qu'il soit nécessaire de la calculer et cette lisibilité est parfaitement suffisante pour apporter un renseignement qualitatif sur la distribution des valeurs dans les différentes classes : symétrie, unimodalité ou plurimodalité, etc. Sur ce point il faut bien observer que l'utilisation qui est habituellement faite d'un histogramme ne va pas plus loin, sauf, et c'est d'ailleurs son utilisation historique première, à vouloir comparer la distribution empirique à une distribution théorique donnée (pour davantage d'indications à ce propos, voir Delecroix 1983), mais, ce faisant, nous introduisons l'outil probabiliste, ce qui ne nous paraît pas être le but visé par cette partie du programme.

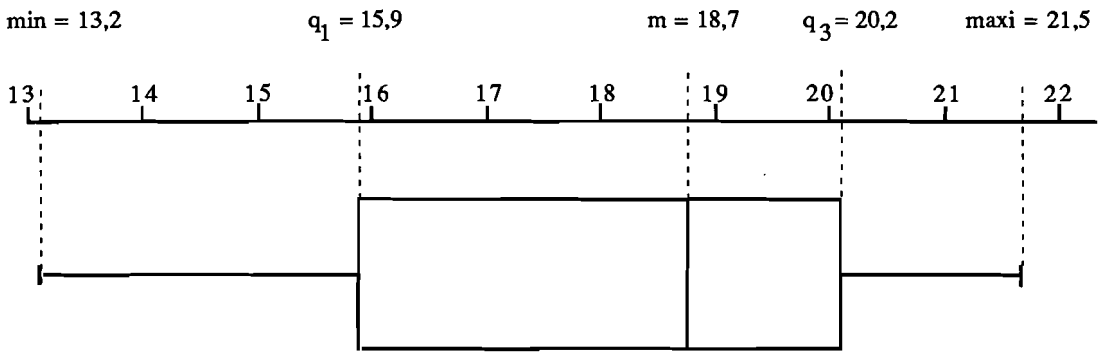
Revenons à l'essentiel de l'E.D.A. : que peut-on découvrir grâce à ce type de représentation (compte tenu de notre manque de culture socio-économique) ?

Tout d'abord on **voit** un glissement vers le bas de la représentation au fil des années et on en déduit que les pourcentages des P.I.B. consacrés à la protection sociale ont augmenté, c'était évident. Mais ce qui l'était moins c'est que l'homogénéité des comportements de ces 9 pays, évaluée visuellement par l'amplitude de la distribution, n'a pas augmenté pendant plus d'une décennie de concertation des politiques économiques et sociales. On peut préciser cela en examinant l'année 1984 : un groupe de pays, 6 sur 9, semble bien avoir un comportement qui s'est homogénéisé mais les autres ont fait «bande à part» dans les deux directions opposées. Il est possible, à partir des colonnes de gauche, d'identifier facilement les pays concernés et, à partir de là, d'engager une analyse socio-économique dont nous ferons grâce au lecteur.

Toujours dans le seul but d'illustrer l'esprit et les techniques de l'E.D.A., soulignons un autre intérêt de cette représentation. Les colonnes de gauche font apparaître entre 1970 et 1984 des «inversions de rangs» : le Luxembourg passe du 7ème au 2ème rang tandis que la République Fédérale Allemande passe du 1er au 4ème rang. Quelle(s) explication(s) peut-on donner de ce fait ? Simple coïncidence des valeurs qui, à quelques dixièmes de pourcentage près, modifient les rangs, ou changements notables des politiques sociales des pays concernés ? Ou encore modifications des nomenclatures nationales qui n'englobent plus les mêmes postes de dépenses relatifs à la protection sociale ?

Un autre mode de représentation utilisé est celui des «box-plot». Il consiste à associer à chaque distribution d'une variable X observée un résumé en cinq valeurs : la valeur minimale, le quartile inférieur[1], la médiane (ou le 2ème quartile), le quartile supérieur et la valeur maximale. Le résumé en cinq points correspondant à la variable «1970» est représenté par la «boîte» de la figure suivante :

1. C'est-à-dire la valeur Q_1 qui partage la population P étudiée en deux parties P_1 et P_2 telles que card P_1 = (card P)/4, card P_2 = 3(card P)/4 et pour tout individu i de P_1, $X(i) <= Q1$ tandis que pour tout individu de P_2 $X(i) >= Q_1$.



La longueur de cette «boîte» rectangulaire est proportionnelle à la longueur de l'intervalle interquartile et renferme donc la moitié la plus centrale de la population. La ligne horizontale figure la valeur de la médiane et sa position reflète la symétrie ou la non symétrie de la distribution. Dans notre exemple, la distribution présente une dissymétrie à droite qui n'est pas nettement mise en évidence par un diagramme en bâtons classique.

Remarquons, pour terminer, que le calcul de la médiane d'une variable ne prenant qu'un nombre fini de valeurs est particulièrement élémentaire : on écrit les valeurs de la variable dans l'ordre croissant (par exemple) et on «barre» tour à tour une valeur à chaque extrémité ; si la série comprend un nombre pair de termes on termine en prenant le milieu de l'intervalle médian (il n'y a rien là qui soit susceptible de paralyser un élève de collège, même s'il est maladroit avec sa calculette).

3. Pourquoi la statistique apparaît-elle dans les programmes ?

Nous proposons dans ce paragraphe une tentative d'analyse des raisons pour lesquelles est apparue dans les nouveaux programmes de collège (donc dans l'enseignement obligatoire) une rubrique dont l'intitulé même - «Organisation et gestion de données» - la désigne comme ressortissant au domaine de la statistique.

En aval de cet enseignement, la rubrique «Statistique» existait dans les programmes des classes de lycée depuis longtemps déjà. Depuis 1971, au moins, en première et terminale, et depuis 1981 en seconde. En amont, par contre, les programmes du Cycle Moyen de 1980 ne font pas de référence directe à un tel enseignement. Cependant on peut y lire un signe avant-coureur : dans la rubrique intitulée «Situations-problèmes» il est notamment écrit que «dans des situations, vécues ou décrites» il faut que l'élève sache «organiser et exploiter» l'information contenue dans la question posée. S'il est clair que cette «organisation» et cette «exploitation» mentionnées ci-dessus concernent plus largement des données et des informations qui ne sont pas nécessairement à caractère numérique, on ne peut s'empêcher de penser que, dans l'esprit des concepteurs des programmes au moins, il y a un lien étroit entre apprendre à bien organiser et gérer des données numériques et être capable d'une bonne organisation et d'une bonne gestion d'informations de nature quelconque contenues dans un énoncé. Il y aurait donc là une sorte de filiation dans laquelle la statistique apparaîtrait comme une propédeutique à la capacité de bien gérer des informations de différentes natures.

Cette remarque faite, le programme du cours moyen de 1985, comme son prédécesseur, ne fait nulle mention d'un tel enseignement. Pourtant, certains manuels (nous avons consulté le manuel du Cours Moyen intitulé «Mathématique contemporaine», Magnard, 1985, pp. 278-287), selon une coutume bien ancrée dans l'enseignement français qui veut qu'au niveau N, sous prétexte de préparer l'avenir, on traite le niveau N + 1, consacrent plusieurs chapitres à l'étude de tableaux et de graphiques.

C'est donc à partir de la sixième et depuis la rentrée 1986 que la statistique, ou au moins quelque chose qui y fait implicitement référence, doit être enseignée. En dehors du fait que telle ou telle personne ou groupe de personnes appartenant à la **noosphère**[2] a pu jouer un rôle essentiel dans ce choix, la théorie didactique nous suggère qu'il existe des éléments objectifs qui ont déterminé l'action de ces personnes et c'est sur ces éléments-là que porte notre interrogation.

Un premier élément concerne, en fait, l'ensemble du cursus. Après la commotion de la réforme des mathématiques modernes, les deux changements successifs des programmes ont orienté les mathématiques au collège dans la même direction : le **retour aux situations concrètes**. Plus nettement encore dans les programmes de 1986, le concret doit être le point de départ obligé de toute activité mathématique. C'est dans ce cadre-là et à côté des rubriques «Travaux géométriques» et «Travaux numériques» qu'il est fait mention de la rubrique «Organisation et gestion de données. Fonctions». Si on ajoute à cela le parti pris de l'interdisciplinarité avec l'apparition des «thèmes transversaux» on peut analyser la venue dans le curriculum de l'organisation et de la gestion de données comme une solution à un problème que l'on pourrait formuler ainsi :

que peut-on faire de simple - voire de très simple - avec des nombres qui puisse satisfaire l'exigence de ***concret*** *et qui ait un rapport avec* ***d'autres disciplines*** *enseignées au collège et avec les «thèmes transversaux» ?*

La réponse surgit alors : faire des tableaux et des graphiques. Elle résout un problème - en réalité, jamais réellement posé, impensé - qui apparaît dès lors que l'ensemble des programmes du collège est imprégné d'un numérisme et d'un concrétisme qui sont les marques de l'avènement d'un **empirisme** - dont nous n'entendons pas rester les prisonniers - dans l'enseignement à ce niveau[3].

Le second élément qui nous paraît, en partie au moins, expliquer l'émergence d'une initiation aux statistiques dès le collège se trouve explicité dans le libellé des programmes de seconde de 1987. En effet, on peut lire dans la deuxième rubrique intitulée «Statistique» :

Ce chapitre présente un quadruple intérêt : d'abord la ***lecture pertinente des tableaux statistiques*** *est maintenant nécessaire à la compréhension du fonctionnement de la société (...).*

L'aveu, si l'on peut dire, est sans détour. L'un des éléments mis en avant pour justifier la pertinence de l'enseignement de la statistique est son aspect **utilitaire** : pour vivre dans la société d'aujourd'hui, il faut apprendre à, et savoir, lire des tableaux et des graphiques car cela sert à l'individu dans sa vie quotidienne. Initier à la statistique est donc la réponse (positive) apportée par le système d'enseignement à une pression sociale qui exige que l'utilité des apprentissages faits à l'école soit immédiatement **visible**. On devine alors que, si la contrainte sociale était la seule, l'initiation à l'activité du statisticien passerait vite au second plan pour ne laisser la place qu'à un apprentissage de lecture (et de traduction) de tableaux et de graphiques. Transformer cette exigence sociale en y reconnaissant une activité mathématique (la statistique) c'est déjà, pour le système d'enseignement, adopter une position de compromis devant les différentes contraintes qui pèsent sur lui.

Les rapides analyses qui précèdent montrent que la venue dans l'enseignement d'un nouvel élément de savoir n'est pas seulement le fait d'un choix, ou d'une politique à laquelle est attaché le nom de tel ou tel ministre, mais provient plutôt du produit d'un ensemble de conditions qui forment le cadre plus ou moins restreint dans lequel ce choix

2. Pour plus de précision sur ce concept et son rôle dans la théorie didactique, voir Chevallard 1985.
3. Pour une analyse plus complète de l'entrée de l'empirisme dans l'enseignement du collège, voir Chevallard 1988, article paru dans ces mêmes colonnes.

peut se faire. Elles montrent de plus que, sous l'étiquette «Statistique» (qui réfère au savoir savant), on peut enseigner des savoirs bien différents selon que l'on mettra en avant, de manière sans doute spontanée, telle ou telle idée du rôle que l'on veut faire jouer à cet enseignement.

4. Qu'est-ce qu'enseigner la statistique ?

Si l'on fait abstraction du niveau d'enseignement trois tendances sont perceptibles.

La première, classique au niveau du second cycle des universités, consiste à rabattre l'enseignement de la statistique sur celui des mathématiques et plus précisément, nous l'avons déjà noté, sur celui du calcul des probabilités. Dans un tel enseignement les théorèmes succèdent aux définitions et les corpus de données étudiés ne sont là que pour constituer des applications aux théorèmes du cours : c'est la tendance mathématique.

La deuxième concerne les enseignements de disciplines qui utilisent de la statistique, comme l'Economie ou la Médecine. Il s'agit ici de connaître le vocabulaire et les techniques de statistiques descriptives mais aussi d'utiliser des résultats qui s'appuient sur le calcul des probabilités (sondages, tests, estimations diverses,...) sans étudier la théorie des probabilités : c'est la tendance utilisatrice. Dans ce cas, les données sont prétexte à l'exercice de savoir-faire et la connaissance qui résulte de leur traitement statistique est factice, ce sont souvent ce qu'on appelle des données «d'école».

La troisième tendance est celle qui vise à former des statisticiens. On ne la rencontre guère que dans les troisièmes cycles des universités et dans les écoles supérieures de statistique. C'est en fait la seule qui va permettre de rencontrer des données réelles et qui va poser le problème de la découverte et de la confirmation (ou de l'infirmation) de faits remarquables issus de ces données.

L'enseignement secondaire, - en fait le lycée jusqu'à ces dernières années - s'est orienté vers la deuxième tendance. Orientation timide avec les programmes de 1971 (1973 pour la classe de seconde où la statistique ne figure pas) et beaucoup plus nette avec les programmes de 1981 où deux évolutions méritent notre attention. La disparition de toute allusion statistique dans les programmes de terminale C (jusque et y compris le vocabulaire probabiliste utilisé en statistique élémentaire comme, par exemple le mot de variable aléatoire) d'une part, et d'autre part, en classe de seconde, l'invitation (explicitée dans les programmes) à une démarche exploratoire.

Cet inventaire, sans doute incomplet, des différentes tendances de l'enseignement de la statistique a, à nos yeux, le mérite de pointer ce qui n'est pas possible pour l'enseignement au collège, mais il ne saurait encore tracer une voie possible puisque nous ne nous sommes pas encore penchés sur les difficultés particulières qui pèsent sur cet enseignement.

5. Quels problèmes pose l'enseignement de la statistique dans le secondaire ?

Une constatation s'impose : les professeurs n'aiment pas, en général, enseigner cette partie du programme. Une explication spontanée d'un tel fait peu se formuler de la façon suivante : les professeurs n'aiment pas la statistique **donc** ils n'aiment pas l'enseigner. A partir de là, une analyse plus approfondie consiste à comprendre pourquoi ils n'aiment pas la statistique. Dans l'état actuel de nos connaissances de la question nous pouvons avancer une autre explication. Les professeurs de mathématiques n'aiment pas enseigner la statistique parce qu'ils n'identifient pas cet enseignement à un authentique enseignement de mathématique. La «paternité» scientifique de la statistique enseignée au lycée s'est diluée dans le temps ; choisir parmi trois nombres qui résultent d'une même mesure fait réfléchir Laplace et ses contemporains

mais ne mérite plus de faire réfléchir, aujourd'hui, un élève de seconde à qui l'on proposera la méthode «naturelle» qui consiste à en faire la moyenne arithmétique. Mais ce problème de la **perte de sens** du savoir n'est pas le seul problème que rencontre l'enseignement de la statistique.

Devant l'impossibilité «d'accrocher» la statistique enseignée à la statistique savante la démarche officielle consiste - comme nous l'avons déjà signalé - à jouer la carte de **l'efficacité pratique** : «apprenez leur au moins à lire un tableau» pourrait être, en schématisant un peu, la consigne que le professeur est invité à suivre. On peut voir deux raisons à cela. Tout d'abord le poids de l'idéologie scientiste ambiante qui, s'appuyant sur l'utilisation sans cesse accrue du **langage** statistique dans certains domaines de l'actualité, donnera à entendre que «l'on ne peut plus rien faire sans le recours aux statistiques»,. Ensuite et en liaison avec ce qui vient d'être souligné, le rôle du paradigme de l'enseignement primaire qui peut être explicité ainsi : puisque les maîtres de l'enseignement primaire peuvent enseigner sous l'étiquette mathématique les quatre opérations et les problèmes arithmétiques qui s'y rattachent sans que cela pose problème, compte tenu de l'efficacité pratique indiscutable de ce type de connaissances (et malgré la distance évidente qui les sépare d'une activité relevant directement de la science mathématique), les professeurs de mathématiques du lycée, et aujourd'hui ceux du collège, doivent pouvoir enseigner la statistique comme un objet de savoir nécessaire à tout citoyen de cette fin du vingtième siècle.

Quelle que soit la générosité de ces intentions elle se heurte à une double difficulté. D'une part, les professeurs de mathématiques des collèges et des lycées sont bien plus sensibles que les instituteurs à la distance qui sépare ce qu'on leur demande d'enseigner des mathématiques savantes. D'autre part, la prétendue nécessité culturelle et pratique de la statistique n'est pas suffisante pour lui assurer une vie stable à l'intérieur du curriculum du collège **si elle se limite à** la dialectique lecture-représentation entre tableaux et graphiques. On touche là un problème d'écologie didactique (Rajoson, 1988), problème particulièrement aigu quand il s'agit d'introduire un nouvel objet d'enseignement.

Un troisième problème auquel se trouve confronté l'enseignement de la statistique est d'ordre culturel. La statistique n'est pas un objet culturel «au dessus de tout soupçon» et, bien que magnifié par les uns (et peut-être parce que magnifié par les uns), il est aussi largement déprécié par les autres. Il est courant d'entendre dire ou de lire que la statistique n'est que la forme scientifique du mensonge, que l'on peut tout faire dire à la statistique, etc. Dans ces conditions on peut comprendre que certains enseignants aient des scrupules à influencer des esprits en formation avec un objet aussi peu recommandable.

II - UNE SOLUTION POSSIBLE.

1. La voie ouverte par l'E.D.A.

Tout d'abord, et c'est essentiel, l'E.D.A., en ajoutant la dimension exploratoire aux activités de statistiques descriptives traditionnelles, permet un enrichissement considérable de l'enseignement de la statistique en lui donnant une **finalité**. (Que penserait-on du calcul algébrique s'il ne devait jamais servir qu'à développer et à factoriser des expressions littérales ?).

De plus cet enrichissement ne s'accompagne pas d'une augmentation du niveau mathématique des outils requis, ce qui permet d'aborder un véritable travail statistique dès la classe de sixième et de le prolonger à travers toutes les classes du collège. Prolongement qui est d'ailleurs tout à fait conforme à l'esprit dans lequel le programme officiel envisage l'enseignement de la statistique en classe de seconde, comme nous l'avons déjà noté.

Enfin l'E.D.A. redonne aux objets d'enseignement une paternité scientifique en l'incluant dans un domaine de mathématiques appliquées qui, pour n'être pas né en France, n'en n'est pas moins très actif ailleurs.

2. Les contraintes didactiques qui en découlent.

Le problème qui se pose alors et celui de la mise en place d'un tel enseignement dans les classes de collège et, pour commencer, en sixième. Nous allons pour ce faire devoir identifier, en une analyse a priori, les différentes contraintes que l'on va rencontrer.

a) Dans la mesure où notre souci est de faire rencontrer à l'élève une problématique, il est nécessaire qu'il s'affronte à des problèmes du champ considéré. Il va donc y avoir une phase durant laquelle l'élève sera face à un corpus de données - dont le degré d'organisation sera plus ou moins grand[4] -, et il faudra faire en sorte qu'il se pose des questions à propos de ces données. On voit tout de suite que le choix du domaine dans lequel ces données seront prises est essentiel pour parvenir à ce but. Et c'est une première difficulté à prendre en compte :

dans quel domaine de réalité choisir le corpus de données ?

Ce domaine doit présenter diverses propriétés : il doit être relativement **familier** à l'élève et suffisamment **riche** pour que des questions apparaissent. Par exemple, le corpus de données que présente un ticket de caisse de supermarché (exemple emprunté au bulletin inter-IREM «Suivi scientifique 1985-1986», p. 165), s'il possède la première qualité, nous paraît d'une possibilité de questionnement à son propos assez réduite. Inversement, un sujet comme la répartition des dépenses des ménages français par postes (alimentation, etc.) (exemple tiré du manuel «Mathématique 6ème», Didier, Paris, 1986, p. 197) risque de laisser coi plus d'un élève de collège (et de sixième en particulier).

De plus, les informations qui sont incluses dans ces corpus - et que l'on va donc découvrir - gagneraient à être susceptibles de créer une certaine **surprise** chez les élèves (en mettant en défaut une idée reçue, par exemple), et, partant, de susciter davantage leur intérêt. Il est d'ailleurs à noter que cette propriété se trouvera plus aisément dans un corpus de données qui réunit les deux qualités de richesse et de familiarité.

b) Dans un tel type de travail avec des élèves, nous savons qu'une phase importante et qui va conditionner toute la suite est la **dévolution du problème** (Brousseau, 1987). Il faudra donc se donner les moyens de mener à bien cette dévolution. En particulier, on devra vraisemblablement consacrer beaucoup de temps d'horloge à la conduite de débats, surtout dans les classes de sixième où elle est toujours un peu problématique. Or il semble, compte tenu de l'ensemble du programme de ces classes, qu'on ne puisse consacrer que trois ou quatre semaines à cette partie. En dehors de cette question de temps, le problème est donc le suivant : quelles questions poser aux élèves qui les mettent en situation de s'interroger sur le corpus de données présenté, puis de produire (et donc, déjà, d'éprouver le besoin de produire) un traitement de données adéquat qui fournisse des éléments de réponse aux questions posées ? C'est en répondant à cette question que l'on pourra espérer construire une séquence didactique dans le droit fil de ce qui a été écrit au paragraphe précédent. Dans la réalisation d'une telle séquence - en classe de sixième - que nous présentons dans le paragraphe suivant, nous ne prétendons pas avoir résolu ce problème, et la dévolution, si elle a eu lieu, a été obtenue davantage par la force persuasive qu'ont dû développer les professeurs que grâce à une situation didactique bien adaptée. Ceci pose d'ailleurs le problème de la **robustesse** des situations didactiques.

4. Nous ne discuterons pas ici de l'intérêt de la collecte des données, qui peut être grand, surtout lorsque cette collecte a été suscitée par un questionnement des élèves eux-mêmes.

c) Un apprentissage se fait d'autant plus facilement qu'il est identifié par l'élève comme une réponse plus performante que celles qu'il peut fournir à une situation donnée. Autrement dit, si l'on admet l'hypothèse que la connaissance se construit comme une ancienne connaissance, la question est alors la suivante : dans l'apprentissage des statistiques quel est l'**«ancien»** sur lequel on va s'appuyer pour créer du **«nouveau»** ? Sur quelle(s) connaissance(s) ancienne(s) bâtirons-nous l'exploration de données ? La réponse à cette question nécessite sans doute un examen approfondi de ce qui se fait, dans la réalité de la classe, en matière de tableaux et graphiques, à l'école primaire.

d) Dans tout enseignement il arrive un moment où le professeur doit dire : «parmi toutes les choses que nous avons vues ensemble, c'est celle-ci que vous devez connaître et sur lesquelles vous serez, à un moment ou à un autre, évalué». On appelle cette situation une situation **d'institutionnalisation** (Brousseau 1987). Dans la forme traditionnelle, c'est le cours qui remplit ce rôle, aidé en cela par les exercices et les interrogations écrites. En général, ce que doit connaître les élèves est identifié par une expression qui permet aux professeurs d'en parler. Ainsi, on dira que les élèves doivent connaître les identités remarquables ou la formule donnant l'aire d'un rectangle ou savoir trouver l'équation d'une droite, etc. C'est la tradition de l'enseignement qui a permis l'identification de ces éléments de savoir ainsi que les expressions qui servent à les désigner[5].

Mais pour un enseignement nouveau, ce travail reste à faire. Ainsi, lorsqu'on aura montré aux élèves la démarche exploratoire à propos d'un corpus de données bien choisi, quel sera le contenu du cours qui s'ensuivra ? Quelles connaissances devrons-nous institutionnaliser ? Il faudra peut-être dire que ce qu'il faut retenir ce n'est pas telle ou telle connaissance qu'on aura développée dans le domaine duquel les données sont tirées (ce domaine étant contingent par rapport au domaine de connaissances que l'on veut mettre en œuvre). Par exemple, si l'on propose un corpus de données concernant les tailles des Français selon leur âge, quel pourra être le **statut des connaissances** élaborées en classe à propos de ce corpus ?

On voit que les problèmes posés ci-dessus ne sont pas tous spécifiques du domaine de connaissance abordé et qu'ils peuvent souvent se rattacher à des situations générales bien étudiées en didactique. Cependant, les solutions apportées à ces problèmes (qu'elles soient bonnes ou mauvaises d'ailleurs) seront, elles, complètement spécifiques du domaine étudié. Nous proposons, dans le paragraphe qui suit, un exemple d'enseignement réalisé dans cinq classes de sixième au cours du troisième trimestre de l'année scolaire écoulée. Les professeurs de ces classes sont ceux qui ont suivi le stage de formation mentionné au début de cet article. Ces professeurs travaillent en équipe depuis trois ans au niveau des classes de sixième, en s'interrogeant plus particulièrement sur la manière de mettre en place des groupes de niveau constitués à propos d'un contenu précis de savoir et dans lesquels l'enseignement de ce contenu précis tiendrait le plus grand compte des difficultés spécifiques des élèves du groupe.

III - UN EXEMPLE EN CLASSE DE SIXIEME.

L'enseignement dispensé dans ces classes de sixième se prétend inspiré des différentes remarques qui précèdent et tente de tenir compte des contraintes didactiques que le travail dans le stage de formation a permis de mettre à jour (contraintes explicitées dans le paragraphe précédent). Il présente deux grandes parties :

5. Il est à noter que la vigilance, épistémologique et didactique, du didacticien doit s'exercer à plein à propos de ces expressions qui désignent couramment un savoir à enseigner. Ainsi, les professeurs peuvent déclarer qu'ils ont fait les «identités remarquables» à leurs élèves. Mais le didacticien se demandera ce que c'est que «faire les identités remarquables», aujourd'hui dans telle classe.

- la première doit permettre de rencontrer le schéma données-questionnement-traitement dans des exemples à la fois simples et significatifs ;
- la seconde doit montrer, dans l'esprit développé lors de la première partie, quelques traitements de données plus particulièrement visés par le programme (représentations graphiques, diagrammes en bâtons) et quelques cas où les données sont caractérisées par le fait d'existe, entre elles, une relation de proportionnalité.

Ici nous présenterons seulement la première partie.

Le travail proposé aux élèves possède la structure ternaire désormais classique : une première phase au cours de laquelle les élèves sont mis en situation de construire des connaissances (phase que nous avons nommée Activité) ; la deuxième phase leur précise quelles sont les connaissances construites (en principe...) dans l'Activité et qui devront être connues d'eux (nous l'avons nommée Leçon) ; enfin les Exercices complètent ce travail en fournissant l'occasion d'une reprise et d'un entraînement. C'est donc sur ce schéma de principe que se déroulera le cours.

L'Activité est donc le moment où l'on va faire rencontrer aux élèves la problématique décrite dans les paragraphes précédents. Confrontés à un corpus de données, ils devront «jouer au détective» et «découvrir» des renseignements cachés dans ces domaines. Comme il a été dit plus haut, la première difficulté réside dans le choix du domaine dans lequel prendre des données. Nous avons choisi le tableau qui suit, emprunté à une publication de l'I.N.S.E.E., que nous avons fait précéder d'un texte de présentation et au sujet duquel nous avons posé cinq questions :

Problème *: Voici un tableau donnant la taille moyenne (en centimètres) des adultes selon leur âge, en France.*

Classes d'âges	*Hommes*		*Femmes*	
	en 1970	*en 1980*	*en 1970*	*en 1980*
entre 20 et 30 ans	*172,5*	*174,0*	*161,5*	*162,0*
entre 30 et 40 ans	*171,0*	*173,0*	*160,5*	*161,5*
entre 40 et 50 ans	*170,0*	*171,0*	*160,5*	*160,5*
entre 50 et 60 ans	*169,0*	*170,5*	*160,5*	*160,5*
entre 60 et 70 ans	*168,0*	*169,0*	*160,0*	*160,0*
70 ans et plus	*168,0*	*168,0*	*159,0*	*158,5*

1. Expliquez ce que signifie, dans le tableau, le nombre 173,0 ?
2. Expliquez ce que signifie chacun des quatre nombres de la ligne «entre 50 et 60 ans» ?
3. Expliquez ce que signifie chacun des six nombres de la colonne «Femmes en 1970» ?
4. Quelles réflexions vous suggère l'examen attentif de ce tableau ?
5. Imaginez des moyens pour montrer ce que vous prétendez voir dans le tableau.

Les trois première questions sont évidemment destinées à s'assurer que les données ont pris du sens pour les élèves. En réalité, ce sont elles qui vont **permettre** cette prise de sens et, par suite, l'entrée dans le problème. C'est dire leur importance et l'attention toute particulière que devra leur consacrer le professeur.

La quatrième question n'est pas vraiment une question que l'on pose aux élèves en espérant qu'ils y répondent. Elle est simplement destinée à lancer le débat dans la classe, débat qui, avec l'élan donné par les premières questions, devrait pouvoir s'enclencher. On attend deux remarques essentielles à propos de ce tableau que l'on peut formuler de la façon suivante :

- «les hommes sont plus grand que les femmes» ;
- «les jeunes sont plus grands que les vieux».

Evidemment ces remarques sont à préciser. En particulier la seconde, dans laquelle il faudra attirer l'attention des élèves sur le fait que l'on ne mesure pas les mêmes personnes et que, par conséquent, une meilleure formulation serait «les individus des nouvelles générations sont plus grands que ceux des anciennes». En tout état de cause, les élèves ont des remarques à faire, et des remarques le plus souvent pertinentes. Comme pour les questions précédentes, il est nécessaire de consacrer du temps à cette phase qui est censée opérer la dévolution du problème. En particulier, il nous a paru didactiquement nécessaire de se donner les moyens de permettre aux élèves de **s'engager** dans leurs réponses (par exemple, en exigeant des réponses écrites, même au brouillon).

La dernière question, comme la précédente, devra se traiter en classe sous la direction du professeur. Certes, on peut penser que les élèves auront des idées. Mais il s'agit ici de leur montrer un type de **traitement de données** très simple qui permet de mettre en évidence les remarques faites à la question précédente. Il consiste à refaire un tableau dans lequel on réécrit les tailles des femmes (aussi bien en 1970 qu'en 1980) et on leur associe les **écarts** correspondants avec les tailles des hommes. Ce travail effectué montre, non seulement que les hommes, d'après ces données, sont effectivement plus grands que les femmes - ce pourquoi le traitement était fait -, mais encore que l'écart est d'environ 10 cm et qu'il diminue légèrement lorsque l'âge des personnes augmente. Ces deux dernières informations ne sont pas directement visibles sur le tableau de départ. Le traitement de données envisagé a donc permis de rencontrer une information contenue dans les données mais dont la saisie n'est pas immédiate.

Cette phase du travail étant terminée, il s'agit d'indiquer aux élèves les connaissances qu'ils ont à retenir. Voici la leçon qui leur a été proposée :

*Un ensemble de **données** peut être organisé dans un tableau. L'examen d'un tel ensemble de données peut nous suggérer des **questions** à propos de ces données.*

Par exemple, dans l'activité précédente, les données sont les tailles moyennes des Françaises et des Français selon leur âge. L'examen de ces données nous a fait nous poser deux questions :
- pourquoi peut-on affirmer, à partir de ces données, que les hommes sont en moyenne plus grands que les femmes ?
- pourquoi peut-on affirmer, à partir de ces données, que les nouvelles générations sont en moyenne plus grandes que les anciennes générations ?

*On peut alors essayer de répondre à ces questions en réorganisant les données : on dit que l'on fait un **traitement de données**.*

Soit un ensemble de données, on ne peut en général pas répondre à toutes les questions que l'on se pose. Par exemple, dans l'exemple précédent, on ne peut certai- nement pas répondre à la question : «qu'est-ce qui fait que les hommes sont en moyenne plus grands que les femmes ?».

Ce texte est lu et commenté en classe. Il fixe le «jeu» auquel on va jouer par la suite. Les élèves peuvent alors s'attendre à devoir découvrir des informations cachées dans un corpus de données à l'aide d'un traitement de ces données - même si, pour l'instant, ils n'en connaissent qu'un seul type, le calcul des écarts[6].

6. Ce type de traitement, qui peut paraître quelque peu trivial, a, en réalité, un écho dans les pratiques statistiques puisqu'on peut le rattacher à des techniques comme l'analyse des résidus.

En comparaison avec les techniques propres de l'E.D.A., le lecteur pourra penser que les outils enseignés dans ce cours sont encore trop rudimentaires. Nous en convenons, tout en faisant observer que nous n'avons qu'une expérience limitée de cet objet d'enseignement qui est conçu pour pouvoir s'étendre à tous les niveaux du cursus du collège. On peut imaginer que, si l'esprit exploratoire est mis en œuvre dès la classe de sixième - et c'est ce que nous prétendons faire par ce cours -, ce n'est, par exemple, qu'en troisième qu'il sera possible d'enseigner une technique comme celle des «box-plot». Pour notre part, c'est dans cette direction que nous entendons poursuivre notre travail.

REFERENCES BIBLIOGRAPHIQUES.

BROUSSEAU G., 1980 et 1981, Problèmes de didactique des décimaux, *Recherches en Didactique des Mathématiques, Vol. 1.1 pp. 11-75 et 2.1 pp. 37-127.*

BROUSSEAU G., 1987, Théorisation des phénomènes d'enseignement des mathématiques, *Thèse d'état, Université de Bordeaux 1.*

CHEVALLARD Y., 1978, Notes pour la didactique de la statistique, *IREM. Université d'Aix-Marseille.*

CHEVALLARD Y., 1985, *La transposition didactique,* La Pensée Sauvage, Grenoble.

CHEVALLARD Y., 1986, La formation au carrefour de la recherche et du développement, Exposé à la «Nuit des formateurs», *IREM. Université d'Aix-Marseille,* (à paraître).

CHEVALLARD Y., 1989, Le passage de l'arithmétique à l'algébrique dans l'enseignement des mathématiques au collège : perspectives curriculaires, *Petit x n° 19.*

DELECROIX M., 1983, Histogrammes et estimation de la densité, *Que sais-je ?, n° 2055, PUF.*

GOOD I.J., 1983, The philosophy of E.D.A., *Philosophy of Sciences,* pp. 283-295.

LECOUTRE J.P. et PASSY P., 1987, *Statistique non paramétrique et robustesse,* Economica.

RAJOSON L., 1988, L'analyse écologique des conditions et des contraintes dans l'étude des phénomènes de transposition didactique : trois études de cas, *Thèse de 3ème cycle, Université d'Aix-Marseille II.*

TUCKEY J.W., 1977, *Exploratory Data Analysis,* Addison Wesley.

VOLLE M., 1980, *Le métier de statisticien,* Hachette.