Enseigner la statistique : un problème de la profession

Yves Chevallard & Floriane Wozniak

IUFM de l'académie d'Aix-Marseille, Université de Provence, et UMR ADEF

IUFM de l'académie de Lyon et LEPS, université Lyon 1

Résumé :

Notre conférence se donnera pour objet de nourrir la réflexion sur les enjeux d'une formation des professeurs de mathématiques à l'enseignement de la statistique regardé comme un *problème de la profession*.
Dans un premier temps, nous présenterons certaines des contraintes auxquelles les professeurs sont soumis dans leur projet d'enseigner la statistique en classe de seconde. Après avoir défini un *modèle épistémologique de référence*, nous expliciterons alors un ensemble de conditions sous lesquelles un tel enseignement pourrait vivre en classe de mathématiques. Enfin, en prenant appui, notamment, sur le dispositif du *Forum des questions* mis en place à l'IUFM d'Aix-Marseille par Yves Chevallard, nous mettrons au jour certains besoins de formation, en présentant les lignes de force de ce que pourrait être une formation à l'enseignement de la statistique au collège et au lycée.

Écologie et économie didactiques

Dans des conditions et sous des contraintes données, tout fait didactique n'est pas possible. Il est des cas en effet où l'*écologie* didactique s'oppose à l'*économie* didactique souhaitée ; où le *projet d'enseignement* que l'on a formé ne trouve pas à se réaliser. S'il l'on veut alors que quelque chose advienne, on peut tenter
– soit de modifier le projet pour le rendre compatible avec l'écologie didactique régnante ;
– soit de modifier cette écologie ;
– soit de modifier l'un et l'autre de façon « bien combinée ».

Le premier cas est le plus fréquent : le projet s'adapte, de fait, aux conditions imposées, et en un sens les renforce. S'agissant de la *statistique*, ce processus d'adaptation conduit à sa *réduction arithmétique*, c'est-à-dire à la réduction du travail *statistique* à du travail *arithmétique*. En guise d'illustration de ce phénomène de réduction, voici un choix d'exercices proposés, pour la classe de 2de, par le serveur Wims ("WWW Interactive Multipurpose Server") [3].

[3] Voir http://wims.auto.u-psud.fr/wims/faq/fr/math.2.html.

Moyenne statistique

Exercice. Soit la série statistique suivante :

Valeurs	6.5	7	8	2.5	10.5	2	1	0.5	9.5	4.5	3	4	6	0	10.5	9.5	8.5	2	0

Calculer pour cette série statistique
- la moyenne :
- la valeur maximale :
- la valeur minimale :
- l'étendue statistique :

Effectif et fréquence

Exercice. Voici les âges recensés dans une population à la suite d'une étude :

39	40	42	39	42	41	43	39	40	39
42	43	40	43	43	40	41	41	40	43
42	41	43							

Compléter le tableau suivant

Ages	39	40	41	42	43
Effectif					
Fréquence					

Séries statistiques : taille

Exercice. Les tailles arrondies à un nombre entier de centimètres des 50 garçons d'un club de football sont les suivantes :

184	170	173	184	165	172	186	164	163	165
173	187	176	169	169	167	164	162	181	180
168	189	168	186	173	184	171	170	184	165
165	178	161	162	170	182	180	164	185	188
173	179	164	185	173	164	182	168	187	178

Compléter le tableau correspondant :

Tailles	[160, 165[	[165, 170[	[170, 175[	[175, 180[	[180, 185[	[185, 190]
Effectif						
Fréquence (%)						

On donnera les fréquences en arrondissant par excès à 0.1 près.

Modifier l'écologie didactique – c'est-à-dire le système des conditions et contraintes pesant sur le didactique – est une entreprise ambitieuse, mais souvent essentielle, comme ici. Car si la réduction arithmétique de la statistique est écologiquement possible, elle n'est guère *durable*, en ce sens que, se situant à un bas niveau mathématique, elle est peu significative d'une discipline spécifique supposée, « la statistique ». Si l'élève doit trouver la moyenne annuelle de Fanny, qui a eu 6,9 de moyenne au premier trimestre, 9,8 de moyenne au deuxième trimestre et 9,3 de moyenne au troisième trimestre, il peut effectuer mentalement le calcul suivant : 6,9 + 9,8 + 9,3 = (6 + 0,9 + 0,8 + 0,3) + 9 + 9 = (6 + 1,7 + 0,3) + 9 + 9 = 8 + 9 + 9. Or ce calcul, loin d'être emblématique d'une science « à forte personnalité », relève en vérité de l'arithmétique scolaire la plus banale qui soit ! N'ayant pas de spécificité, n'étant pas significatif de quelque chose de spécifique, le « travail mathématique » correspondant tendra à se réduire à la portion congrue, voire à être escamoté : le fait de donner ce travail « en DM » ou de renvoyer les élèves au manuel pour qu'ils s'aident eux-mêmes à ce propos sont les effets les plus communs de cette dépréciation scolaire.

Pour dépasser cette dégradation mathématique, une autre direction peut alors être spontanément tentée : *complexifier* le travail *arithmétique*, l'ennoblir en introduisant de façon plus ou moins clandestine des thèmes d'étude plus « sophistiqués ». C'est ce phénomène que l'on constate lorsqu'on observe le succès (illicite), en classe de seconde, du thème des *effets de structure* : nombre de professeurs ont en effet trouvé dans l'étude de ce phénomène une manière d'épicer un enseignement jugé sans doute bien fade. Voici d'abord comment la notion d'effet de structure se trouve présentée sur le site Internet de l'INSEE [4].

Effet de structure

Définition

Lorsqu'une population est répartie en sous-populations, il peut arriver qu'une grandeur évolue dans un sens sur chaque sous-population et dans le sens contraire sur l'ensemble de la population. Ce paradoxe s'explique parce que les effectifs de certaines sous-populations augmentent alors que d'autres régressent : c'est l'effet de structure.
Par exemple, le salaire de chaque profession peut stagner (ou augmenter faiblement) alors que le salaire moyen augmente fortement ; cela arrive si les professions très qualifiées, les mieux payées, sont de plus en plus nombreuses et, réciproquement, les emplois non qualifiés, les moins payés, de plus en plus rares.
A contrario, la variation à structure constante se calcule comme une moyenne pondérée des variations des moyennes de chaque sous-population, les pondérations étant les masses de la grandeur pour chaque sous-population.

On notera que le dernier paragraphe décrit un type de situations – celui de la moyenne pondérée – qui est normalement au cœur du travail sur les moyennes au collège et en 2de, dans le cas d'une *même* structure des populations que l'on compare ou d'une population dont la structure demeure *inchangée* dans le temps. Voici à présent un témoignage de ces pratiques tel qu'il apparaît sur le site « L'île aux mathématiques » [5].

posté le 05/02/2006 à 14:08

[4] Voir http://www.insee.fr/fr/nom_def_met/definitions/html/effet-structure.htm.
[5] Voir http://www.ilemaths.net/forum-sujet-68197.html.

effets de structure
posté par : **adilelgh**

Salut, j'ai besoin d'une assitance pour cet exercice
Dans les deux entreprise E1 et E2, les salariés sont classés en deux catégories: employés et cadres. Les deux tableaux qui suivent donnent la répartition des employés en fonction de leur catégorie professionnelle et de leur salaire annuel S en milliers d'euros. Dans les calculs qui suivent, les sommes seront exprimés en euros et arrondis à l'euro inférieur, et on considérera que tous les élements d'une classe sotn situés en son centre.
Le tableau est à la fin...
……
…

Entreprise E_1			
	Salaires		
Catégories	$10 \leq S < 20$	$20 \leq S < 30$	$30 \leq S < 40$
Employés	170	100	0
Cadres	0	10	20

Entreprise E_2			
	Salaires		
Catégories	$10 \leq S < 20$	$20 \leq S < 30$	$30 \leq S < 40$
Employés	280	140	0
Cadres	0	40	40

Quelle « manœuvre » tenter pour contrer cette banalisation dépréciative engendrée par la réduction arithmétique du travail statistique ? La réponse la plus authentique et la plus efficace nous paraît consister à faire vivre dans la classe de mathématiques des *questions « de haut niveau »*, où se montre véritablement quelque chose de la spécificité de la discipline à enseigner : la statistique.

Pour donner une idée de ce qui *pourrait être* à cet égard, voici un scénario de *fiction didactique* inspiré de l'article de Christine A. Franklin et Madhuri S. Mulekar intitulé "Is Central Park Warming?" et paru dans le numéro de mai 2006 de la revue *The Mathematics Teacher* (vol. 99, n° 9, p. 600-605).

1. La question

Une classe décide, sous la direction de son professeur de mathématiques, d'étudier la question suivante : « est-il vrai que la température a augmenté au cours du siècle écoulé ? »

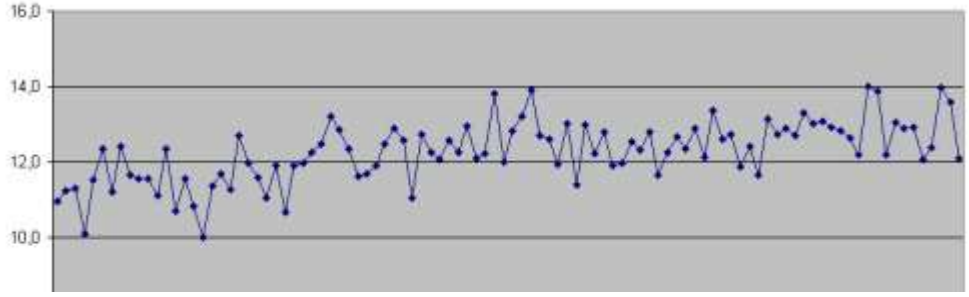
2. Les données

La classe recherche des données chiffrées relatives à la période 1901-2000. Elle les trouve : il s'agit des moyennes annuelles dans un lieu proche de l'établissement. (Dans chaque groupe de deux lignes successives, la première indique l'année, la seconde la température en degrés Celsius.)

1901	**1902**	**1903**	**1904**	**1905**	**1906**	**1907**	**1908**	**1909**	**1910**
11,0	11,3	11,3	10,1	11,5	12,4	11,2	12,4	11,7	11,6
1911	**1912**	**1913**	**1914**	**1915**	**1916**	**1917**	**1918**	**1919**	**1920**
11,6	11,1	12,4	10,7	11,6	10,9	10,0	11,4	11,7	11,3
1921	**1922**	**1923**	**1924**	**1925**	**1926**	**1927**	**1928**	**1929**	**1930**
12,7	12,0	11,6	11,1	11,9	10,7	11,9	12,0	12,3	12,5
1931	**1932**	**1933**	**1934**	**1935**	**1936**	**1937**	**1938**	**1939**	**1940**
13,2	12,9	12,4	11,6	11,7	11,9	12,5	12,9	12,6	11,0
1941	**1942**	**1943**	**1944**	**1945**	**1946**	**1947**	**1948**	**1949**	**1950**
12,7	12,3	12,1	12,6	12,3	13,0	12,1	12,2	13,8	12,0
1951	**1952**	**1953**	**1954**	**1955**	**1956**	**1957**	**1958**	**1959**	**1960**
12,8	13,2	13,9	12,7	12,6	11,9	13,0	11,4	13,0	12,2
1961	**1962**	**1963**	**1964**	**1965**	**1966**	**1967**	**1968**	**1969**	**1970**
12,8	11,9	12,0	12,5	12,3	12,8	11,7	12,3	12,7	12,4
1971	**1972**	**1973**	**1974**	**1975**	**1976**	**1977**	**1978**	**1979**	**1980**
12,9	12,1	13,4	12,6	12,7	11,9	12,4	11,7	13,2	12,8
1981	**1982**	**1983**	**1984**	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**
12,9	12,7	13,3	13,0	13,1	12,9	12,8	12,7	12,2	14,0
1991	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**
13,9	12,2	13,1	12,9	12,9	12,1	12,4	14,0	13,6	12,1

3. Étude des données

La classe représente graphiquement les données ci-dessus. À examiner le graphique obtenu, il semble bien, en effet, qu'il y ait une tendance à l'augmentation de la température annuelle moyenne (v. ci-après).



4. Situation par rapport à la médiane

La médiane des 100 températures est trouvée égale à 12,35 ºC. On observe que, dans les 50 premières années du siècle, les températures annuelles sont 15 fois supérieures à la médiane, alors qu'elles le sont 35 fois dans les 50 années suivantes. Cela confirme l'observation « qualitative » du graphique.

5. Un test

Sur les 30 dernières années du siècle (de 1971 à 2000), les températures annuelles moyennes ont été 23 fois supérieures à la moyenne. La classe imagine le test suivant : supposons que la température annuelle moyenne soit tirée à pile ou face : dans le premier cas (pile) elle est supérieure à la médiane, dans le second cas (face), inférieure. L'événement observé, à savoir 23 fois la sortie de pile en 30 lancers, est-il assez fréquent ou est-il rare ? La classe s'attend à ce qu'une telle observation soit rare, voire rarissime.

6. Une simulation

Faute de disposer des notions utiles de la théorie des probabilités, la classe procède par simulation. Elle réalise ainsi 1200 simulations de 30 tirages à pile ou face, avec les résultats suivants (dont seuls les 500 premiers ont été reproduits ci-après).

12 ; 18 ; 18 ; 10 ; 13 ; 12 ; 18 ; 18 ; 18 ; 18 ; 16 ; 20 ; 17 ; 15 ; 16 ; 15 ; 16 ; 15 ; 16 ; 18 ; 16 ; 16 ; 18 ; 16 ; 15 ; 16 ; 12 ; 13 ; 15 ; 13 ; 15 ; 11 ; 14 ; 13 ; 15 ; 15 ; 16 ; 13 ; 13 ; 22 ; 19 ; 13 ; 18 ; 16 ; 14 ; 10 ; 13 ; 18 ; 13 ; 18 ; 15 ; 15 ; 17 ; 12 ; 13 ; 18 ; 17 ; 12 ; 19 ; 11 ; 13 ; 19 ; 13 ; 15 ; 12 ; 13 ; 15 ; 13 ; 17 ; 11 ; 14 ; 20 ; 20 ; 12 ; 15 ; 10 ; 16 ; 15 ; 21 ; 11 ; 17 ; 20 ; 14 ; 14 ; 18 ; 20 ; 17 ; 16 ; 17 ; 17 ; 13 ; 15 ; 14 ; 18 ; 13 ; 15 ; 19 ; 17 ; 16 ; 17 ; 21 ; 17 ; 13 ; 14 ; 12 ; 18 ; 11 ; 14 ; 16 ; 14 ; 20 ; 11 ; 12 ; 16 ; 13 ; 18 ; 7 ; 17 ; 9 ; 16 ; 14 ; 11 ; 10 ; 16 ; 18 ; 14 ; 15 ; 17 ; 12 ; 12 ; 8 ; 16 ; 17 ; 19 ; 14 ; 15 ; 13 ; 11 ; 15 ; 16 ; 17 ; 16 ; 16 ; 14 ; 18 ; 12 ; 14 ; 20 ; 14 ; 20 ; 11 ; 14 ; 14 ; 17 ; 16 ; 13 ; 12 ; 14 ; 13 ; 16 ; 15 ; 16 ; 16 ; 13 ; 16 ; 10 ; 17 ; 14 ; 14 ; 14 ; 12 ; 13 ; 11 ; 13 ; 14 ; 16 ; 16 ; 18 ; 16 ; 12 ; 16 ; 16 ; 17 ; 16 ; 15 ; 11 ; 16 ; 16 ; 16 ; 15 ; 17 ; 14 ; 19 ; 16 ; 17 ; 17 ; 14 ; 15 ; 18 ; 20 ; 16 ; 18 ; 15 ; 12 ; 14 ; 20 ; 18 ; 11 ; 14 ; 15 ; 17 ; 12 ; 18 ; 16 ; 19 ; 18 ; 16 ; 17 ; 15 ; 13 ; 13 ; 18 ; 15 ; 15 ; 18 ; 19 ; 16 ; 17 ; 15 ; 14 ; 17 ; 9 ; 18 ; 20 ; 14 ; 14 ; 12 ; 16 ; 15 ; 18 ; 13 ; 17 ; 13 ; 21 ; 14 ; 11 ; 19 ; 15 ; 15 ; 12 ; 14 ; 16 ; 19 ; 12 ; 14 ; 13 ; 15 ; 18 ; 12 ; 15 ; 11 ; 14 ; 18 ; 13 ; 19 ; 13 ; 13 ; 16 ; 18 ; 17 ; 16 ; 14 ; 14 ; 16 ; 12 ; 12 ; 17 ; 15 ; 15 ; 11 ; 15 ; 15 ; 14 ; 11 ; 16 ; 16 ; 16 ; 14 ; 11 ; 18 ; 14 ; 14 ; 16 ; 16 ; 19 ; 19 ; 15 ; 14 ; 14 ; 14 ; 11 ; 12 ; 20 ; 13 ; 18 ; 16 ; 16 ; 16 ; 17 ; 16 ; 17 ; 16 ; 14 ; 19 ; 15 ; 10 ; 12 ; 12 ; 15 ; 12 ; 17 ; 20 ; 15 ; 16 ; 13 ; 15 ; 14 ; 18 ; 16 ; 15 ; 20 ; 10 ; 19 ; 18 ; 14 ; 11 ; 17 ; 13 ; 15 ; 18 ; 11 ; 14 ; 14 ; 18 ; 16 ; 14 ; 13 ; 17 ; 12 ; 11 ; 18 ; 14 ; 13 ; 17 ; 12 ; 12 ; 16 ; 12 ; 11 ; 15 ; 14 ; 14 ; 10 ; 23 ; 15 ; 13 ; 14 ; 18 ; 21 ; 15 ; 16 ; 14 ; 11 ; 13 ; 19 ; 15 ; 14 ; 19 ; 15 ; 13 ; 15 ; 17 ; 17 ; 15 ; 16 ; 14 ; 18 ; 13 ; 17 ; 9 ; 14 ; 13 ; 12 ; 13 ; 16 ; 11 ; 14 ; 17 ; 18 ; 17 ; 10 ; 13 ; 18 ; 13 ; 13 ; 15 ; 12 ; 14 ; 14 ; 15 ; 16 ; 14 ; 16 ; 14 ; 19 ; 10 ; 15 ; 19 ; 16 ; 17 ; 13 ; 17 ; 17 ; 17 ; 13 ; 17 ; 15 ; 20 ; 19 ; 16 ; 13 ; 17 ; 15 ; 18 ; 20 ; 16 ; 16 ; 15 ; 15 ; 17 ; 11 ; 11 ; 15 ; 16 ; 13 ; 14 ; 16 ; 18 ; 15 ; 13 ; 14 ; 17 ; 18 ; 19 ; 19 ; 15 ; 15 ; 12 ; 18 ; 16 ; 14 ; 13 ; 14 ; 17 ; 13 ; 9 ; 10 ; 15 ; 24 ; 14 ; 13 ; 13 ; 14 ; 10 ; 12 ; 21 ; 13 ; 16 ; 8 ; 20 ; 13 ; 16 ; 10 ; 16 ; 15 ; 12 ; 13 ; 15 ; 14 ; 17 ; 16 ; 12 ; 15 ; 15 ; 11 ; 17 ; 13 ; 15 ; 14 ; 8 ; …

Dans les 1200 échantillons de taille 30 obtenus, 3 seulement comportent 23 sorties « pile », tandis qu'un seul comporte 24 sorties « pile » : la proportion d'échantillons comportant 23 sorties « pile » ou plus est donc de 4/1200, soit un peu plus de 0,003. (Le professeur indiquera que, selon la théorie des probabilités, la « fréquence théorique » est de 0,0026 environ.)

7. La conclusion

L'idée que le hasard pourrait expliquer à lui seul le réchauffement observé perd ainsi beaucoup de son crédit.

Enseigner avec la société

À l'instar de ce que suggère le scénario didactique précédent, un projet d'enseignement de la statistique doit faire apparaître nettement l'objet de cette science : l'étude de la *variabilité*. Y renoncer implique, à terme, sinon la disparition complète de l'enseignement correspondant, du moins sa dégradation et son régime indéfiniment végétatif, tant du moins que le corpus de savoir rassemblé sous l'étiquette de statistique n'aura pas conquis une identité claire, qui le distingue des autres parties du corpus mathématique enseigné tout en permettant de le situer par rapport à elles. Il s'agit là d'une « loi » curriculaire qui n'est pas propre aux mathématiques. Dans un ouvrage récent (*The Fight for English*, Oxford University Press, 2006), un spécialiste britannique réputé de la langue anglaise, David Crystal, consacre de longues pages à conter (et à analyser) la disparition de l'enseignement traditionnel de la grammaire anglaise, dont il constate le décès en 1965. Mais il note que, dès 1921, dans un rapport établi par une commission de travail du "Board of Education" (ministère de l'Éducation), on pouvait lire que, aux yeux des rédacteurs du rapport, il était devenu *"impossible at the present juncture to teach English grammar in the schools for the simple reason that no-one knows exactly what it is."* Semblablement, quelle définition pourrait-on donner, aujourd'hui, de la statistique à partir de la seule observation de son enseignement scolaire ? Gageons que les phénomènes transpositifs intenses, et dénaturants, que laissent voir tant les manuels que les classes ne permettraient guère de répondre à cette question cruciale de façon satisfaisante.

L'exigence d'identifier la statistique comme une discipline *sui generis*, irréductible à toute autre, ne doit pas être limitée à ce que nous nommons *la profession*, c'est-à-dire le collectif plus ou moins intégré constitué, en l'espèce, des professeurs de mathématiques, des formateurs de ces professeurs, des chercheurs sur l'enseignement des mathématiques et sur le métier de professeur de mathématiques ou sur la formation à ce métier, ainsi que des responsables officiels ou associatifs de l'enseignement, de la formation et de la recherche en question. Cette exigence doit en effet être partagée *avec les parents* ; ou, plus exactement, la reconnaissance de la spécificité de la discipline enseignée doit être partagée par la profession et *avec les élèves*, soit la génération qui se forme, *et avec les générations précédentes*.

Ce partage doit être pris en charge par la profession, en tant que *problème de la profession*. En d'autres termes, la profession doit chercher à faire évoluer sur ce point essentiel le système des conditions et contraintes à l'intérieur duquel les professeurs doivent opérer. La géométrie, l'arithmétique ou l'algèbre ont une image *spécifique* pour tous ceux qui ont eu une scolarité obligatoire complète : pour eux, la géométrie, qui est la science de la *« spatialité »*, ce sera au moins « des figures » ; l'arithmétique, science de la *« numérosité »*, ce sera des chiffres et des opérations ; l'algèbre, science de la *« calculabilité »*, ce sera des *x* et des *y*. Ces exemples suggèrent au reste que l'*identité* ne suffit pas : ces disciplines mathématiques gagnent à avoir aussi une réelle *attractivité*, qui suscite une véritable curiosité, certains parents (ou grands-parents) allant même jusqu'à *envier* leurs enfants (ou petits-enfants) d'avoir à étudier ce dont eux-mêmes, le cas échéant, auront été privés.

La science de la variabilité

L'un des grands obstacles à l'ambition d'un pacte sociétal et scolaire autour de l'enseignement de la statistique tient à ce fait que l'idée d'une *science de la variabilité* n'existe pas intensément dans la culture de la société française « éternelle » : fait massif, qui, aussi bien, touche les professeurs de mathématiques eux-mêmes. La *dénégation de la variabilité* est en effet la loi plutôt que l'exception. On tend ainsi à regarder toute grandeur comme *constante*, attitude dont le langage courant porte témoignage (on demandera facilement « combien ça pèse un lion », « combien il y a de pétales dans une rose », etc.). Or la première conquête collective à accomplir se trouve là : dans le fait *d'assumer la variation*. Cela consiste par exemple à ne pas dire que les conjoints violents (ou les violeurs) se recrutent dans tous les milieux sociaux en laissant croire (voire en croyant soi-même) qu'ils s'y recrutent *également*, pas davantage qu'on ne dirait « qu'on meurt (également) à tout âge ». Dès les

débuts de la *Political Arithmetic* anglaise, dont John Graunt (1620-1674) et William Petty (1623-1687) sont les noms les plus connus, la science statistique s'inscrit en faux contre le « tout se vaut », qui conduit à croire, par exemple, qu'on peut mourir de ceci « aussi bien » que de cela. De quoi mourait-on ainsi dans la ville de Londres autour de 1650 ? Dans ses *Natural and Political Observations Made upon the Bills of Mortality* (1662), John Graunt présente les causes de mortalité à Londres pour diverses années. L'un des buts qu'il poursuit est de donner aux personnes des indications sur ce qui les conduira à quitter ce monde et, ce faisant, à se libérer de certaines inquiétudes répandues liées au « on peut mourir de tout » populaire. Il écrit à ce propos ceci :
In the next place, whereas many persons live in great fear, and apprehension of some of the more formidable, and notorious diseases following; I shall only set down how many died of each: that the respective numbers, being compared with the Total 229250, those persons may the better understand the hazard they are in.
Graunt dispose de données sur une période où la ville de Londres a vu mourir 229 250 personnes. Il présente alors à son lecteur la table des *notorious deseases* – des maux dont les gens craignaient le plus de mourir [6].

Apoplex: 1306	Lunatique: 0158
Bleeding: 069	Murthered: 0086
Cut of the Stone: 0038	Overlaid [Infants suffocated when
Burnt, and Scalded: 125	their mother or nurse rolls over on
Falling Sickness [= Epilepsy]:	them in bed], and Starved: 0529
0074	Poysoned: 014
Drowned: 829	Palsy: 0423
Dead in the Streets: 0243	Smothered: 026
Excessive drinking: 002	Rupture: 0201
Gowt: 0134	Shot: 007
Frighted: 022	Stone and Strangury [restricted
Head-Ach: 0051	urine flow. A difficulty of urine
Grief: 279	attended with pain]: 0863
Jaundice: 0998	Starved: 051
Hanged themselves: 222	Sciatica: 0005
Lethargy: 0067	Vomiting: 136
Kil'd by several accidents: 1021	Sodainly: 0454
Leprosy: 0006	

On voit ainsi que, par exemple, la fréquence des décès du fait d'un *accident*, relativement élevée, est seulement de $\frac{1021}{229250} \approx 4{,}45$ ‰. À l'opposé des conclusions de Graunt, fondées sur des données chiffrées, l'assomption « théorique » *a priori*, souvent implicite, qu'il y aurait distribution à peu près *uniforme* (des causes de décès, etc.) apparaît souvent comme l'indice *d'un refus paradoxal de la variabilité*. Qui n'a entendu dire, par exemple, que les maris violents « se recrutent dans toutes les classes sociales » ? Le quotidien *Le Monde* écrivait ainsi, dans son édition datée du 9 août 2003 : « L'image […] de la femme battue dans un foyer pauvre par un mari alcoolique a vécu : la violence touche tous les milieux sociaux (8,9 % des femmes concernées sont des cadres, 3,3 % des ouvrières). L'article où figure ce passage se faisait ainsi l'écho d'un ouvrage alors récemment paru, et qui faisait grand bruit : *Les violences envers les femmes en France* (Maryse Jaspard *et al.*, La Documentation française, Paris, 2003). Une saine culture statistique aurait voulu que l'on fasse connaître au lecteur du quotidien, non pas seulement deux valeurs isolées, supposées emblématiques, mais *l'ensemble de la distribution des fréquences* à travers les « classes sociales », telle du moins que l'enquête de référence la faisait connaître. Or, lorsqu'on examine l'ouvrage cité, trois points soulèvent des

[6] Voir http://www.ac.wwu.edu/~stephan/Graunt/2.html.

interrogations sur l'information chiffrée apportée au lecteur du *Monde*. Tout d'abord, les chiffres indiqués (8,9 % et 3,3 %) semblent être, non des pourcentages de femmes cadres ou de femmes ouvrières dans l'ensemble des femmes victimes de violences conjugales mais des pourcentages de femmes victimes de violences conjugales parmi les femmes cadres, ou parmi les ouvrières, etc. Ensuite, on ne voit pas bien comment le premier chiffre (8,9 %) a été fabriqué : il semble qu'il provienne de l'addition erronée des pourcentages des « violences graves » (6,1 %) et des « violences très graves » (2,6 %) déclarées par les femmes cadres de l'échantillon. Enfin, on découvre que le pourcentage de 3,3 % pour les ouvrières (qui correspond aux « violences très graves ») doit être rapproché, non du chiffre de 8,7 % (ou 8,9 %), mais de 2,6 % : la conclusion mise en avant par l'article du *Monde* s'en trouve alors affaiblie.

Par contraste avec ces formes de refus de la variabilité, la vision statistique conduit à regarder les objets du monde naturel ou social, non comme le siège de grandeurs *fixes*, mais de grandeurs *variables*. À cet égard, la *connaissance du monde* s'égale, à un premier niveau, à la connaissance, pour chaque type d'« objets » et pour chaque grandeur qui lui est attachée, à la *distribution de fréquences* de cette grandeur. Dans un livre intitulé *Les conduites déviantes des lycéens* (Hachette Éducation, 2000), le sociologue Robert Ballion rend compte d'une enquête par questionnaire auprès de lycéens (9919 de ces questionnaires, recueillis entre avril et novembre 1997, ont été analysés). Les enquêtés étaient interrogés sur leur propre estimation de leur « valeur scolaire ». Si l'on interroge un individu pris au hasard, le fait qu'il estime avoir une bonne réussite scolaire ou non dépendra sans doute de beaucoup de facteurs. Mais voici ce qu'on trouve, à l'instar de John Graunt examinant les causes de mortalité à Londres, en réponse à la question que voici : qu'un lycéen déclare avoir des résultats « bons ou excellents », est-ce rare ? En fait, les élèves disant avoir des résultats bons ou excellents représentent 10,6 % de l'échantillon étudié ; ceux qui s'attribuent des résultats assez bons ou moyens sont en revanche 74,5 %, tandis que ceux qui jugent leurs résultats médiocres ou faibles sont 13,8 % (il y a 1,1 % de non-réponses). Il s'agit donc là d'un comportement relativement peu fréquent. Mais ce qui est remarquable encore, c'est la distribution des réponses des *parents* à la question de la réussite scolaire de leur enfant. Une enquête conduite par ailleurs, citée par le même auteur, montre que la part des parents jugeant leur enfant « excellent » n'est que de 7,3 % alors qu'ils sont 82,1 % à le juger « bon » ou « moyen ». Ceux qui, à l'opposé, estiment que leur enfant a « des difficultés » ou « de grosses difficultés » sont 8,2 % (il y a 2,4 % de non-réponses). Les auteurs de l'enquête concluent que les parents manifestent « une certaine réticence à placer leurs enfants aux extrémités de l'échelle scolaire, parmi les élèves excellents, ou parmi ceux qui ont de grosses difficultés ». Les deux distributions apparaissent en effet différentes.

On peut envisager aussi de distinguer, au sein de l'échantillon des lycéens interrogés par Robert Ballion, ceux qui fréquentent un lycée d'enseignement général et technologique (LEGT) et ceux qui fréquentent un lycée professionnel (LT). Les distributions correspondantes sont-elles semblables, voire superposables ? Voici.

	LEGT	LP
Disent avoir des résultats bons ou excellents	9 %	14 %
Disent avoir des résultats moyens	75 %	78 %
Disent avoir des résultats médiocres ou faibles	16 %	8 %

La comparaison de ces distributions sera peut-être une surprise pour le profane : la distribution est translatée *vers le haut* quand on passe des élèves de LEGT aux élèves de LP ! Le phénomène est connu des spécialistes, et l'auteur cité écrit à ce propos :

… les lycées d'enseignement professionnel proposent à leurs élèves des situations d'apprentissage qui favorisent le sentiment de réussite mieux que ne le font les lycées d'enseignement général et technologique. Comme l'écrit Bernard Charlot : « Le lycée professionnel, lycée de relégation au départ, devient en cours de route un lieu de reconstruction d'élèves en échec. »

On notera la différence entre appréciation subjective et réalité objective de la réussite scolaire : alors que (seulement) 39,7 % des élèves de LEGT ont redoublé durant leur scolarité, ce pourcentage passe à 82,3 % en LP – il fait plus que doubler.

Il est intéressant aussi de comparer filles et garçons. Les premières ont *objectivement* une meilleure réussite scolaire que les seconds : alors que 55,9 % des garçons ont redoublé au cours de leur scolarité, ce pourcentage tombe à 46,9 % pour les filles, soit 9 *points de moins*. Les distributions de fréquences sont les suivantes.

	Garçons	Filles
Disent avoir de bons résultats	11 %	10 %
Disent avoir des résultats moyens	74 %	76 %
Disent avoir des résultats faibles	15 %	14 %

Les distributions sont donc très voisines : objectivement, les filles se sous-estiment ou les garçons se surestiment. Ajoutons encore une touche à ce tableau, en distinguant, non entre garçons et filles, ou entre élèves de LEGT et élèves de LP, mais entre élèves *en fonction de l'âge*. Les différentes distributions de fréquences sont les suivantes.

Âge	*≤ 15*	*16*	*17*	*18*	*19*	*≥ 20*
Déclarent de bons résultats (%)	15	12	10	9	9	8
Déclarent des résultats faibles (%)	13	14	13	14	17	21

L'auteur commente ces résultats dans les termes que voici :
Plus on avance en âge, et donc dans le cursus, plus la valeur de l'auto-estimation baisse : le taux des élèves qui estiment avoir de bons résultats faiblit, tandis qu'au contraire augmente celui des élèves à résultats faibles. On peut voir dans ce phénomène un indicateur de dégradation dans le temps de l'expérience scolaire, le fait d'éprouver un sentiment de réussite devenant de moins en moins fréquent au fur et à mesure que se déroule la scolarité.

Études statistiques

On situera la suite de cette présentation par rapport à un *modèle épistémologique de référence*, en sorte que nous pourrons rapporter à ce modèle les pratiques d'enseignement observables et interroger les conditions de son « implémentation didactique ». Le modèle de référence se rapporte, ici, au seul cas de la statistique *à une variable* (avec, pour théâtre principal de sa réalisation didactique, les classes de 3^{e} et de 2de). De façon volontairement restrictive, on appelle ici *étude statistique* l'étude d'une question Q d'un type dont les questions suivantes, à la formulation volontairement naïve, sont des spécimens significatifs (le *type* de ces questions sera formalisé un peu plus loin) :

- Un bébé qui pèse 3,4 kg à la naissance, c'est un gros bébé ?
- Un éléphant de deux tonnes, c'est un gros éléphant ou c'est un petit éléphant, ou c'est ni l'un ni l'autre ?
- Quand on dit qu'un joueur de foot a marqué beaucoup de buts dans une saison, ça veut dire qu'il en a marqué au moins combien ?
- Est-il exact que les trois derniers jours ont été exceptionnellement froids ?
- Cet article est-il cher pour ce que c'est ?

La conduite de l'étude suppose alors que l'on précise un certain nombre de données de base : il faut, en principe, indiquer la *population* Ω (les bébés, les éléphants, etc.) à laquelle on se réfère ; le *caractère* X (le poids, le prix, etc.) défini sur Ω ; enfin l'*échantillon* $E \subseteq \Omega$ sur lequel on va étudier la

question *Q*. Voici, en guise d'illustration, une partie de la synthèse réalisée dans une classe de 3e mi-réelle, mi-imaginaire, sur laquelle nous reviendrons.

Collège Georges Bouligand
3e7 – Mathématiques
Synthèse : conduire une étude statistique
II. Les grands problèmes du travail statistique
a) Définir la population. Dans une l'étude statistique d'un caractère, on doit préciser ce qu'est la *population* sur laquelle on étudie le caractère : à quelle population d'élèves, ou de bébés, ou d'éléphants, ou de joueurs de football s'intéresse-t-on exactement ? Répondre à ce type de questions suppose une enquête dont l'objet n'est pas, en général, mathématique, mais à laquelle il faut tout de même procéder – de façon plus ou moins approfondie.
b) Définir l'échantillon. En général, on ne dispose pas de la valeur du caractère étudié pour chacun des « individus » de la population : on s'en tiendra à étudier ce caractère sur un ou plusieurs échantillons de cette population. L'idéal serait de disposer d'échantillons dits « représentatifs » de la population, c'est-à-dire qui « ressemblent » à la population. Mais on est souvent loin de cet idéal : on devra se contenter en général d'échantillons *disponibles*, c'est-à-dire pour lesquels on dispose des données utiles. Pour cette raison, une étude statistique est presque toujours partielle et provisoire.
c) Définir le caractère. De la même façon que la population est souvent indiquée de façon approximative (que ce soit celle des bébés, des footballeurs ou des éléphants), le caractère que l'on est censé étudier sur cette population est lui-même souvent *mal défini*. Qu'est-ce, par exemple, que la « longueur » d'une phrase ? Si on étudie le caractère « nombre d'occurrences de la lettre *a* » dans la population des « phrases françaises », va-t-on compter les occurrences de *à* et de *â* ? Comptera-t-on aussi les occurrences de *æ* ? Etc. Dans tous les cas, il conviendra de préciser le choix, en général *conventionnel*, que l'on aura fait.

Le type de tâches fondamental va porter sur la *série statistique* $\{ x_1, x_2, \ldots, x_n \}$, où $x_i = X(\omega_i)$, qui correspond à l'échantillon étudié, $E = \{ \omega_1, \omega_2, \ldots, \omega_n \}$. Ce qui importe, ici, n'est pas la valeur de X sur tel « individu » $\omega \in E$, mais la *distribution des fréquences* sur E. Pour l'obtenir, on considère d'abord la *distribution des effectifs*, n_j $(1 \leq j \leq p)$, des valeurs $v_1, \ldots, v_p$ prises par X sur E ; puis on considère la distribution des fréquences $f_j = n_j/n$. Notons que, si on connaît $n = \sum_{j=1}^{p} n_j$, on peut reconstituer la distribution des effectifs, puisque $n_j = N \times f_j$. Nous pouvons alors formaliser ce qui nous intéresse en définissant la *fonction de répartition* de X sur E, définie par

$$F_X(x) = \frac{1}{n} \text{Card} \{ i / x_i \leq x \} = \frac{1}{n} \sum_{v_j \leq x} n_j.$$

Inversement, la connaissance de F_X entraîne la connaissance de la distribution des fréquences de X : si on suppose que l'on a $v_1 < \ldots < v_p$, il vient en effet $n_1 = F_X(v_1)$, $n_j = F_X(v_j) - F_X(v_{j-1})$ pour $j = 2, \ldots, p$. La fonction de répartition F_X a ainsi l'intérêt de condenser tout ce qu'on veut savoir à propos de X sur E : en un sens, la connaissance statistique (univariée) du monde consiste, pour chaque couple (E, X), à connaître la distribution des fréquences de X sur E ou, de façon équivalente, la fonction de répartition de X sur E.

Supposons ainsi qu'on veuille répondre à la question suivante : « Il est gros, cet éléphant, non ? » Si l'éléphant en question a un poids p tel que, disons, $F_X(p) = 0{,}32$, on pourra répondre *par la négative* : car 68 % des éléphants sont alors *plus « gros »* que l'éléphant considéré, qui ne peut donc guère être dit « parmi les plus gros ». Supposons maintenant que l'on veuille répondre à la question : « C'est quoi un gros éléphant ? C'est gros comment ? » La réponse n'est bien sûr *pas univoque* ; si par exemple, pour un certain poids p, $F_X(p) = 0{,}49$, on pourra répondre qu'un gros éléphant est

certainement un éléphant dont le poids est supérieur à p. Si l'on décide qu'un éléphant sera dit « gros » s'il appartient à l'ensemble des 20 % les plus lourds, et si l'on sait que l'on a $F_X(p^*) = 0{,}80$, alors on dira qu'un gros éléphant est un éléphant de poids supérieur à p^*.

Considérons à présent les notes trimestrielles des élèves d'une classe de 3e : 8,9 ; 10 ; 8,5 ; 13,7 ; 15,6 ; 6,9 ; 12,3 ; 16,9 ; 8,7 ; 8,2 ; 15,3 ; 13,3 ; **14,1** ; 7,4 ; 15,7 ; 11,1 ; 9,5 ; 17,0 ; 10,0 ; 7,3 ; 14,6 ; 10,4 ; 7,7 ; 18,9 ; 8,5 ; 15 ; 14,6 ; 9,3. La note de 14,1 sur 20 obtenue par l'un des élèves est-elle une note « élevée » dans cette classe et pour ce trimestre ? En ordonnant la série donnée par ordre croissant, on obtient ceci : 6,9 ; 7,3 ; 7,4 ; 7,7 ; 8,2 ; 8,5 ; 8,5 ; 8,7 ; 8,9 ; 9,3 ; 9,5 ; 10 ; 10 ; 10,4 ; 11,1 ; 12,3 ; 13,3 ; 13,7 ; **14,1** ; 14,6 ; 14,6 ; 15 ; 15,3 ; 15,6 ; 15,7 ; 16,9 ; 17,0 ; 18,9. On peut dire d'abord que *ce n'est pas une note basse*, puisqu'il n'y a que dix notes qui lui soient strictement supérieures (ce qui représente 35,7 % des notes environ). Mais pour faire partie des 20 % de notes les plus élevées, il faudrait ici (puisque 20 % × 28 = 5,6) que la note examinée fasse partie des 5 notes les plus élevées, c'est-à-dire qu'elle soit supérieure ou égale à 15,6 – ce qui n'est pas le cas.

Les considérations qui précèdent sont à l'origine de l'introduction de la fonction *quantile*, qui est en quelque sorte la fonction « inverse » de la fonction de répartition, et qui permet de répondre aux questions de type « au-dessus de quelle valeur x du caractère X ne trouve-t-on plus que 20 % de la population ? » ou « au-dessous de quelle valeur n'y a-t-il plus que 15 % de la population ? ». Généralisant le cas où il existe q tel que $F_X(q) = u$, on considère en effet la plus petite valeur q telle que $F_X(q) \geq u$; d'une façon générale, on définit alors la fonction quantile par $Q(u) = \inf \{ x / F_X(x) \geq u \}$, où $u \in [0; 1]$. On a ainsi $F_X(Q(u)) \geq u$, avec en outre, si $x < Q(u)$, $F_X(x) < u$. Le nombre $Q(u)$ est appelé u-quantile (ou u-fractile) ou quantile (ou fractile) d'ordre u. Comment déterminer un u-quantile ? Reprenons la série déjà rencontrée des notes trimestrielles d'une classe de 3e ; trions-les par ordre croissant et déterminons la valeur de

$$F(x) = \frac{1}{n} \text{Card} \{ i / x_i \leq x \}$$

pour chacune des valeurs prises $x = v$; on obtient le tableau ci-après, qui fait apparaître notamment ceci : $Q(10\ \%) = 7{,}4$; $Q(20\ \%) = Q(25\ \%) = 8{,}5$; $Q(30\ \%) = 8{,}9$; $Q(40\ \%) = 10$; $Q(50\ \%) = 10{,}4$; $Q(60\ \%) = 13{,}3$; $Q(70\ \%) = Q(75\ \%) = 14{,}6$; $Q(80\ \%) = 15{,}3$; $Q(90\ \%) = 16{,}9$; $Q(95\ \%) = 17$; $Q(100\ \%) = 18{,}9$.

v	*6,9*	*7,3*	*7,4*	*7,7*	*8,2*	*8,5*	*8,7*	*8,9*	*9,3*
100 F_X	3,6	7,1	10,7	14,3	17,9	25	28,6	32,1	35,7
v	*9,5*	*10*	*10,4*	*11,1*	*12,3*	*13,3*	*13,7*	*14,1*	*14,6*
100 F_X	39,3	46,4	50	53,6	57,1	60,7	64,3	67,9	75
v	*15*	*15,3*	*15,6*	*15,7*	*16,9*	*17*	*18,9*		
100 F_X	78,6	82,1	85,7	89,3	92,9	96,4	100		

Les quantiles $Q(k/10)$, où l'entier k varie de 1 à 9, sont les *déciles* : $Q(10\ \%)$ est le *premier* décile, …, $Q(90\ \%)$ le *neuvième* décile. Les trois quantiles $Q(25\ \%)$, $Q(50\ \%)$, $Q(75\ \%)$ sont les *quartiles*, respectivement le premier quartile, le deuxième quartile et le troisième quartile. On aura noté que le *deuxième quartile* ne correspond pas exactement (dans le cas précédent) à la *médiane* (qui, selon la convention adoptée en 2de, vaut ici 10,75).

On peut aussi procéder ainsi : n étant l'effectif de la série (dans le cas précédent, $n = 28$), $Q(u)$ est la valeur du terme de la série (supposée rangée par ordre croissant) dont l'indice est le plus petit entier supérieur ou égal à nu. Pour $u = 0{,}85$, par exemple, $nu = 23{,}8$: $Q(85\ \%)$ est donc la valeur du terme de rang 24, à savoir 15,6.

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*

6,9	7,3	7,4	7,7	8,2	8,5	8,5	8,7	8,9	9,3	9,5	10	10	10,4
15	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	***24***	*25*	*26*	*27*	*28*
11,1	12,3	13,3	13,7	14,1	14,6	14,6	15	15,3	**15,6**	15,7	16,9	17	18,9

Conditions de possibilité

L'implémentation didactique du modèle épistémologique de référence est-elle possible ? La réponse est certainement positive. En voici un exemple observée dans une classe de 3e où a été conduite une séance d'initiation à l'étude statistique d'une question *Q* donnée. Voici d'abord le scénario de l'étude.

2. Étude des notes d'un devoir en classe

2.1. L'étude

Dans une classe de 3e de 28 élèves a eu lieu un devoir surveillé. Les 28 notes attribuées sont reproduites ci-après :

15 ; 13 ; 11 ; 9 ; 4 ; 11 ; 13 ; 18 ; 9 ; 10 ; 14 ; 17 ; 13 ; 16 ; 11 ; 16 ; 12 ; 16 ; 9 ; 5 ; 13 ; 19 ; 16 ; 0 ; 9 ; 16 ; 11 ; 7.

On dira qu'une note de cette série est *satisfaisante* si elle est supérieure ou égale à au moins 50 % des notes, soit à 14 notes de la série au moins.

Il revient au même de dire qu'elle est inférieure strictement à au plus 50 % des notes de la série, c'est-à-dire à 14 notes au plus.

Question 1. Lors du devoir, quatre élèves ont obtenu la note 11. Cette note est-elle satisfaisante ? Comment faire pour le savoir ?

Réponse 1. Un décompte à la main permet de constater que la note 11…

– … est supérieure ou égale à 13 notes, ce qui ne représente que $\frac{13}{28} = \frac{1300}{28}$ % ≈ 46,4 % des notes ;

– … est strictement inférieure à 15 notes de la série, ce qui représente $\frac{15}{28} = \frac{1500}{28}$ % ≈ 53,6 % des notes.

En conséquence, la note 11 *n'est pas satisfaisante*.

Question 2. Pour savoir si la note 11 est satisfaisante ou non dans la série des 28 notes observées, un élève a eu l'idée de ranger ces 28 notes par ordre croissant et de les numéroter. Comment le faire ? Comment cela permet-il de décider si la note 11 est satisfaisante ou pas ?

Réponse 2. On peut saisir ces notes dans la colonne A d'un fichier du Classeur, puis utiliser l'icône de tri croissant. On obtient alors ceci.

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	***14***
0	4	5	7	9	9	9	9	10	11	11	11	11	**12**
15	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*	*27*	*28*
13	13	13	13	14	15	16	16	16	16	16	17	18	19

Pour que 11 soit une note satisfaisante, il faut et il suffit que la note numérotée 14 soit inférieure ou égale à 11. Ici, cette note est 12 : on retrouve donc que 11 n'est pas une note satisfaisante.

Question 3. Préciser la plus petite note satisfaisante dans la série des 28 notes.

Réponse 3. La plus petite note satisfaisante est celle qui a le numéro 14 : on a vu qu'il s'agit de la note 12.

Question 4. On dit qu'une note est *excellente* si elle est supérieure ou égale à 90 % des notes au moins. En utilisant à nouveau la série croissante et numérotée des 28 notes, déterminer les notes excellentes de la série.

Réponse 4. On a : 90 % × 28 = 25,2. Une note est donc excellente si et seulement si elle est supérieure ou égale à la note numérotée 26. D'après le tableau obtenu, la note numérotée 26 est 17. Les notes excellentes de la série sont donc 17, 18 et 19.

Question 5. On dit qu'une note est *remarquable* si elle est supérieure ou égale à au moins 70 % des notes. Déterminer les notes remarquables de la série.
Réponse 5. On a : 70 % × 28 = 19,6. Une note est donc remarquable si et seulement si elle est supérieure ou égale à la note numérotée 20. D'après le tableau, les notes remarquables sont donc les notes au moins égales à 15.
2.2. Le bilan de l'étude
Question. Qu'a-t-on appris, au fond, dans ce qui précède ?
Réponse. a) Étant donné une série statistique de longueur n et un pourcentage k %, on a appris à déterminer la plus petite valeur de cette série supérieure ou égale à au moins k % des valeurs de la série.
b) Pour cela, on a dû apprendre à mettre, à l'aide du Classeur, la série donnée sous la forme d'une série croissante.
c) Étant donné un pourcentage k %, on a calculé le premier entier supérieur ou égal k % × n : la plus petite valeur cherchée est celle qui a pour numéro cet entier.

Voici maintenant la ***synthèse*** issue de ce travail et de sa suite (portant sur d'autres questions).

Collège Georges Bouligand
3e7 – Mathématiques
Synthèse : conduire une étude statistique
III. Le type de tâches fondamental
a) Le type de tâches. Étant donné une série statistique de longueur n et un pourcentage k %, déterminer la plus petite valeur de cette série qui est supérieure ou égale à au moins k % des valeurs de la série.
b) La technique. On range la série par ordre croissant ; on calcule alors l'entier N qui est le plus petit entier supérieur ou égal au nombre k % × n (lequel, en général, n'est pas entier) ; la valeur cherchée est alors la valeur de rang N dans la série rangée par ordre croissant.
c) Exemple. On cherche la plus petite valeur x d'une série de longueur n = 162 telle qu'au moins 80 % des valeurs de la série soit inférieure ou égale à x. On a : 80 % × 162 = 129,6. Le plus petit entier N supérieur ou égal à 129,6 est 130. La valeur cherchée est donc celle qui occupe le rang 130 dans la série rangée par ordre croissant.
d) Justification & remarques. 1) Supposons que la série que l'on vient d'évoquer se présente ainsi :

Rang :	129	130	131	132	133
Valeur :		014	015	016	017	017

La valeur donnée par la technique est 15. Cette valeur est supérieure ou égale à 130 valeurs de la série, soit à $\frac{130}{162} = \frac{13000}{162}$ % ≈ 80,25 % de l'effectif de la série. La valeur immédiatement plus petite, 14, ne convient pas : elle n'est supérieure ou égale qu'à 129 valeurs de la série, c'est-à-dire à $\frac{129}{162} = \frac{12900}{162}$ % ≈ 79,63 % de l'effectif de la série.
2) Supposons que la série se présente ainsi :

Rang : 128	129	130	131	132
Valeur :	014	015	015	016	017

La valeur cherchée est toujours 15 ; mais on voit que la valeur classée au 129e rang convient aussi, puisque c'est la même ! Le rang qu'on a calculé (130) n'est donc pas le plus petit rang possible. Mais la *valeur* ayant ce rang est bien la plus petite possible. La valeur immédiatement plus petite, 14, est supérieure ou égale à 128 valeurs, ce qui ne représente que $\frac{128}{162} = \frac{12800}{162}$ % ≈ 79 % de l'effectif de la série.
3) Supposons enfin que la série se présente ainsi :

Rang : 129 130 131 132 133
Valeur : 14 015 015 015 016
La valeur cherchée est toujours 15 ; mais, ici, cette valeur, qui, dans les cas précédents, était la plus petite possible qui soit supérieure ou égale à « juste un peu plus » que 80 % des valeurs de la série, est en fait supérieure ou égale à « sensiblement plus » de 80 % des valeurs de la série, puisqu'on a en effet : $\frac{132}{162} = \frac{13200}{162}\ \% \approx 81{,}5\ \%$.

4) On peut généraliser la justification précédente en utilisant des lettres : la valeur x cherchée occupe le rang N : elle est donc supérieure ou égale à $\frac{100\,N}{n}$ % de l'effectif de la population. Comme $N \geq k$ % $\times\, n$, on a $100\,N \geq k \times n$ et il vient donc : $\frac{100\,N}{n}\ \% \geq \frac{k \times n}{n}\ \% = k\ \%$.

La diffusion scolaire de ce modèle épistémologique (et son implémentation corrélative dans les classes) se heurte à divers obstacles, dont l'élimination est un problème que *la profession* doit contribuer à résoudre notamment – mais pas seulement – à travers la formation de ses membres (même si, bien entendu, elle doit rechercher des coopérations afin de résoudre les problèmes qui se posent à elles). Si l'on suppose acquise (ce qui n'est pas le cas aujourd'hui) l'acceptation large par la profession du point de vue selon lequel les études statistiques menées par les classes doivent partir de questions « effectives » et viser à leur apporter des réponses « significatives », même si elles demeurent partielles et provisoires, on peut ramener les autres problèmes à une question fondamentale : quelles sont les questions Q dont l'étude statistique 1) donne à voir la spécificité de l'apport de la science statistique à la compréhension du monde, et 2) se réfère à des populations Ω et à des variables X pour lesquelles il est possible et réaliste d'obtenir des échantillons E se traduisant par des séries statistiques $X(E)$ appropriées à l'étude à conduire ?

On s'arrêtera ici sur le second point de difficulté : le choix de Q de façon que des *données* appropriées soient *disponibles*. En nombre de cas, dans les études extrascolaires, une *enquête ad hoc* permet de recueillir un échantillon de données que l'on veut « représentatif » de la population Ω. Par exemple, dans l'étude précédente sur les lycéens, l'auteur a exploité 9919 questionnaires, recueillis dans six académies. Dans chacune d'elles, « les services statistiques ont constitué un échantillon d'une quinzaine de lycées et, dans chaque lycée, 100 à 150 élèves ont été soumis à la passation du questionnaire », cela en avril 1997 pour quatre des six académies, en novembre 1997 pour les deux autres. Dans l'étude sur les violences à l'encontre des femmes, l'enquête a été réalisée « par téléphone de mars à juillet 2000, auprès d'un échantillon représentatif de 6 970 femmes âgées de 20 à 59 ans, résidant en métropole et vivant hors institutions ». Cette technique est évidemment disponible dans les classes, mais son usage, souvent « coûteux », reste limité (même si l'on tente, au sein d'un établissement ou d'un groupe d'établissements, de mutualiser des données). Une autre technique consiste alors à rechercher des données disponibles *par ailleurs*, données que l'on utilisera au mieux dans la classe : en ce cas, la question Q étudiée sera choisie en fonction des données disponibles, telles par exemple qu'on peut en trouver sur le site Internet de l'INSEE [7]. Bien entendu, dans certains cas, il n'est pas entièrement déraisonnable d'engendrer – selon une loi déterminée – des données numériques sur lesquelles la classe travaillera.

Fluctuations et probabilités

[7] Voir http://www.insee.fr/fr/ppp/fichiers_detail/edufa03/telechargement.htm.

L'un des objectifs d'une éducation statistique « citoyenne » d'aujourd'hui est d'apprendre aux générations montantes à porter un regard juste sur la composition des « petits échantillons » : si l'on tire au sort 8 citoyens, et s'il en résulte un groupe de 6 femmes et de deux hommes, doit-on s'étonner en disant qu'il s'agit là d'un fait *rare* ? Doit-on penser que, « normalement », il aurait dû « sortir » une des compositions (4, 4), (5, 3), (3, 5) ? Comme on le voit, on retrouve ici la problématique illustrée par les études précédentes : contenir 6 femmes pour un groupe de 8 personnes tirées au sort dans une population dans laquelle on suppose réalisée la parité, est-ce beaucoup ? Pour répondre, la procédure consiste à créer des échantillons de taille 8 en grand nombre. Ainsi, pour 400 échantillons au hasard de taille 8, le tableau ci-après donne les effectifs des échantillons selon le nombre de femmes.

0	1	2	3	4	5	6	7	8
2	14	43	103	102	79	46	11	0

On voit que, *sur ces 400 échantillons* de taille 8, on a plus d'une chance sur 10 – exactement 11,5 chances sur 100 – de tirer un groupe contenant exactement 6 femmes, et même 14,25 chances sur 100 de tirer un groupe comportant *au moins* 6 femmes. Mais on a en même temps 14,75 chances sur 100 de tirer un groupe comportant *au plus* 2 femmes. Et on a 71 chances sur 100, donc environ 7 chances sur 10, de tirer un groupe contenant entre 3 ou 5 femmes.

Le type de travail statistique que l'on vient d'évoquer à propos des « petits échantillons » est en vérité très fragile au plan de l'écologie mathématico-didactique *actuelle*. En nombre de cas, en effet, la considération de petits effectifs est phagocytée par l'obsession du « passage à la limite » et de la « stabilisation des fréquences ». En d'autres termes, partant d'un échantillon de taille 8 (extrait de la population mentionnée plus haut), où il y aurait par exemple 5 femmes et 3 hommes, on va faire croître la taille comme pour montrer que, *en fait*, ce ne serait là qu'une « anomalie », laquelle s'estompe et disparaît peu à peu (non sans une certaine « résistance » cependant) quand on augmente la taille de l'échantillon. Ainsi, partant d'un échantillon de taille 8 où la proportion de femmes est de 5/8 = 0,625, on a obtenu les fréquences suivantes en augmentant la taille de l'échantillon comme le tableau ci-après l'indique.

8	20	50	100	200	250	400	500
0,625	0,55	0,58	0,49	0,5	0,516	0,525	0,524

En d'autres termes, on ne va pas « rester » sur un échantillon de petite taille, mais regarder celui-ci comme le début d'un échantillonnage indéfiniment recommencé. Le phénomène est connu et Guy Brousseau [8] n'a pas manqué de le souligner encore récemment dans les termes que voici :
L'introduction de la temporalité, de l'ordre, de l'évolution des données, préfigure la sortie des situations de description statistique et l'entrée dans les jeux de la prévision, auxquels est attachée une autre famille d'obstacles, ceux attachés à la pensée probabiliste mais qui vont largement déborder sur la statistique.
À titre d'illustration d'un tel phénomène dans les classes, voici un extrait d'un compte rendu d'observation d'une séance qui a eu lieu en mars 2002 dans une classe de seconde.

« Allez, vous sortez vos cahiers de statistique ! Vous aviez à faire le premier exercice pour aujourd'hui ! » L'énoncé est le suivant :
Un joueur lance un dé. S'il obtient un numéro différent de 6, il reçoit une somme égale au numéro obtenu (en francs). S'il obtient 6, il doit verser 6 F. Ce joueur joue cent fois de suite. Il obtient les résultats suivants :

[8] « Situations fondamentales et processus génétiques de la statistique », *Balises en didactique des mathématiques*. Grenoble, La Pensée sauvage, Grenoble, p. 165-193. (Le passage cité se trouve p. 184.)

4, 6, 1, 2, 5, 2, 1, 2, 3, 5, 4, 3, 1, 6, 6, 4, 2, 2, 1, 4
6, 4, 2, 6, 4, 3, 6, 4, 6, 3, 3, 4, 4, 6, 2, 3, 5, 6, 5, 5
3, 3, 5, 4, 1, 3, 1, 3, 2, 4, 5, 6, 4, 4, 3, 6, 2, 6, 5, 6
3, 4, 6, 3, 4, 5, 3, 1, 5, 1, 3, 5, 6, 1, 5, 3, 4, 2, 2, 4
1, 5, 1, 4, 6, 1, 2, 3, 2, 1, 3, 3, 1, 2, 6, 1, 2, 3, 3, 3

1° Ordonner les données brutes ci-dessus. Préciser l'effectif et la fréquence de chaque modalité.
2° Quel est le mode de cette série ?
3° On s'intéresse maintenant au gain du joueur (noté négativement s'il s'agit d'une perte).
a. Construire un tableau montrant les gains et leurs effectifs.
b. Calculer le gain moyen par partie du joueur.

P appelle une élève au tableau et rappelle la consigne : ordonner les données brutes, etc. L'élève dessine un tableau. Précisant ce que c'est qu'ordonner les valeurs, P écrit :

$$\overbrace{1\,1\,\ldots\,1}^{15}$$

$$\overbrace{2\,2\,\ldots\,2}^{15}$$

$$\overbrace{3\,3\,\ldots\,3}^{22}$$

$$\overbrace{4\,4\,\ldots\,4}^{18}$$

$$\overbrace{5\,5\,\ldots\,5}^{13}$$

$$\overbrace{6\,6\,\ldots\,6}^{17}$$

P vérifie si le travail demandé a été fait. Pendant ce temps, l'élève a progressé, remplissant le tableau qu'elle avait dressé et avançant dans le travail demandé :

1°)

Valeur dé	1	2	3	4	5	6
Effectif	15	15	22	18	13	17
Fréquence	$\frac{15}{100} \simeq 0,15$	$\frac{15}{100} \simeq 0,15$	$\frac{22}{100} \simeq 0,22$	$\frac{18}{100} \simeq 0,18$	$\frac{13}{100} \simeq 0,13$	$\frac{17}{100} \simeq 0,17$

2) le mode de cette série est 3

P contrôle avec la classe que le mode indiqué est exact. Puis on passe à la question 3 : P rappelle la règle du jeu. L'élève écrit :

3) les valeurs sont –6, 1, 2, 3, 4, 5

P : « On fait un tableau montrant les gains et les effectifs ». L'élève s'exécute :

Gain	– 6	1	2	3	4	5
Effectif	17	15	152	22	18	13

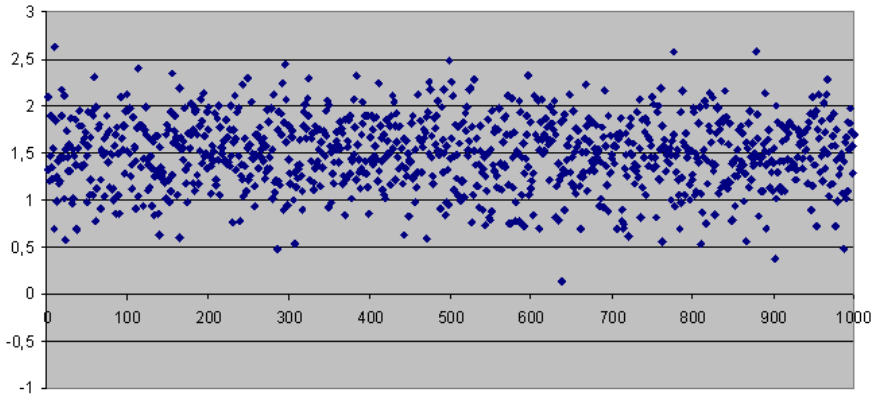
P : « Très bien !… Petit b, calculer le gain moyen, c'est-à-dire calculer la moyenne. Je sais qu'ici il y a des experts en moyenne !… » L'élève écrit :

Erreur ! = **Erreur !** = **Erreur !** = 1,46 Frs

P confirme que le gain moyen est « appelé g barre » ; puis il conclut : « Si vous jouez, vous gagnez. Pas beaucoup, mais vous gagnez ! Y a-t-il des questions ? »

On retrouve ici, en apparence, la croyance en un monde « constant » : le résultat obtenu pour *cette* série de 100 parties est tenu pour *universel* : « Si vous jouez, vous gagnez. Pas beaucoup, mais vous gagnez ! » On saisit ici le rôle que joue la stabilisation des fréquences « fantasmée » : elle permet de chasser de façon imaginaire la variabilité d'un monde où la variation serait une pure et simple anomalie inessentielle.

Ici, il aurait été pertinent de simuler un grand nombre de fois une partie et d'en noter le gain : le graphique suivant montre un échantillon de 1000 parties de cent lancers de dé chacune tout à fait suggestif.



Le danger que fait courir la proximité « culturelle » (dans l'enseignement des mathématiques) de la notion de probabilité à l'existence d'un enseignement de la statistique comme science de la variabilité est liée à deux contraintes massives, l'une propre à la statistique, l'autre beaucoup plus large. La première est le *refoulement de la variation*, qui pousse à mettre en avant le « passage à la limite ». Le second est au moins aussi profondément enraciné dans la culture de la profession : de même que la « géométrie théorique » (dans le style euclidien), à travers ses rejetons contemporains, étouffe dans l'œuf la géométrie « expérimentale » *au lieu de naître d'elle*, de même la statistique théorique, fondée sur le concept de probabilité, asphyxie le travail statistique empirique et expérimental au lieu de le servir en proposant, à travers le fait central de la stabilisation des fréquences, une « algèbre des fréquences ». Le « socle commun de connaissances et de compétences » indiquant que « les élèves doivent connaître », « pour ce qui concerne l'organisation et la gestion de données et les fonctions », « les notions de chance ou de probabilité », le programme de 3^{e} qui entrera en vigueur en septembre 2007 comporte une initiation à la notion de probabilité. Il y a là, pour la profession, un grand combat à livrer.