

Les Conférences de la Famille Mathématique

(APMEP – IPR – IREM – IUFM)

Mercredi 14 juin 2006

17 h – 19 h

**Enseigner la statistique en seconde :
un problème de la profession**

par Yves Chevallard & Floriane Wozniak

Prologue 1

Qu'appelons-nous *problème de la profession* ? À quoi cela s'oppose-t-il ? Qu'est-ce que cette notion pousse en avant ?

Apologue 1

L'exemple de Pascale Pombourcq : ayant elle-même lancé une pièce un assez grand nombre de fois, elle a observé des résultats qui l'ont porté à penser – intuitivement – que cette pièce tendait à « privilégier les faces », alors même que le calcul des intervalles de confiance correspondant aux séries qu'elle a obtenues montre que, si l'on suit la recette qu'elle présente par ailleurs au lecteur dans sa brochure, l'hypothèse que la pièce est équilibrée ne peut jamais être rejetée (au seuil de 5 %).

Épilogue 1 // Prologue 2

Les questions dans la suite de cet exposé à deux voix, que l'on regarde généralement comme des difficultés avec lesquelles chacun se débrouille – plus ou moins bien ou plus ou moins mal –, seront regardées comme des problèmes de la profession.

Apologue 2

Choix et commentaire cursif de questions de PCL2 regardées comme symptômes de problèmes de la profession.

Épilogue 2 // Prologue 3

Pour situer les questions suscitées par l'enseignement de la statistique (en seconde), il convient de situer ce qu'est la statistique à une variable (univariée) : *questions génératrices et études statistiques*.

- Les *questions génératrices* de la statistique

Cet éléphant, il est gros, non ? C'est quoi un gros salaire dans cette profession ? Etc.

- La notion d'*étude statistique*

→ Une *question* Q à propos d'un *caractère* X des *individus* ω d'une *population* Ω : cet ω a-t-il un gros X ? Cela signifie quoi de dire que $\omega \in \Omega$ n'a pas un X trop grand ? Etc.

→ Ce qui importe : la *distribution* des valeurs de X sur Ω , que font connaître la *fonction de répartition* F_X définie par

$$F_X(x) = \frac{1}{N} \sum_{v \leq x} n_v ;$$

et son « inverse », la *fonction quantile* Q_X définie par

$$Q_X(u) = \inf \{ x / F_X(x) \geq u \}$$

→ Première conquête de l'étude de la statistique : on réfute l'assertion « ω a un gros X » en montrant par exemple que $X(\omega) < Me$: la médiane est première ; la moyenne, qui n'en est qu'une valeur approchée plus ou moins bonne, est seconde

Apologue 3

La notion d'étude statistique : exemple(s) illustratif(s).

Épilogue 3 // Prologue 4

La diffusion scolaire de la statistique ainsi entendue se heurte à divers obstacles, dont celui-ci : la statistique est-elle bien un domaine des mathématiques ? Son enseignement est-il du ressort du professeur de mathématiques ?

Le point de vue d'un « expert » :

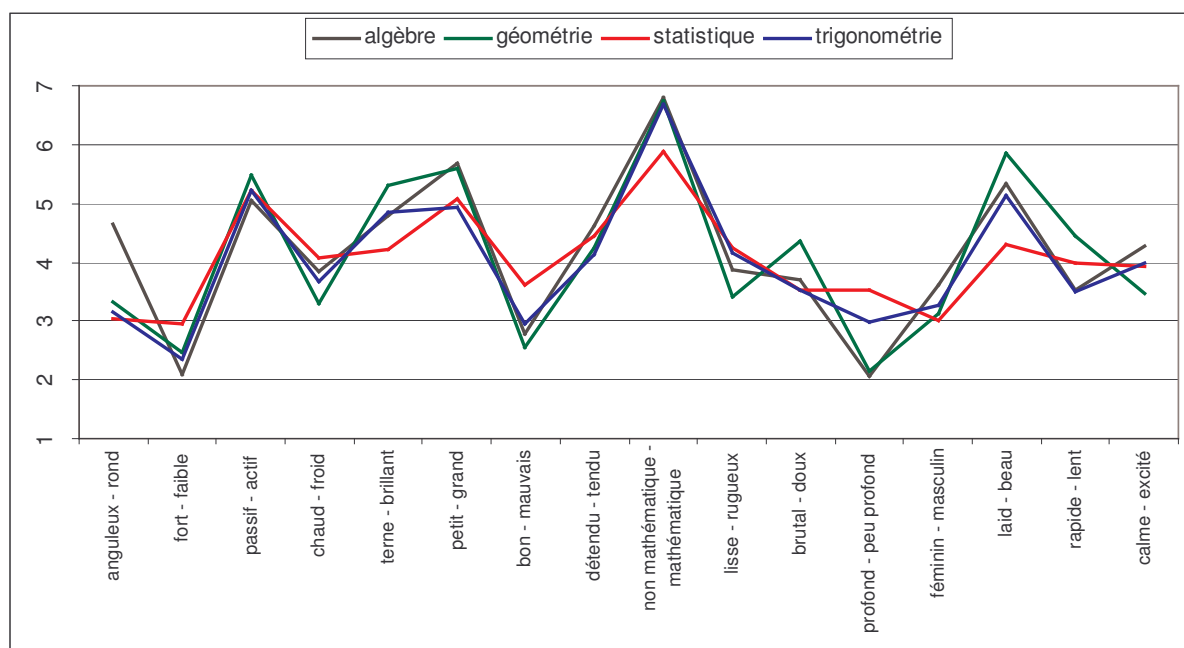
“The science of statistics is essentially a branch of Applied Mathematics and may be regarded as mathematics applied to observational data. As in other mathematical studies the same formula is equally relevant to widely different groups of subject matter. Consequently the unity of the different applications has usually been overlooked, the more naturally because the development of the underlying mathematical theory has been much neglected.”

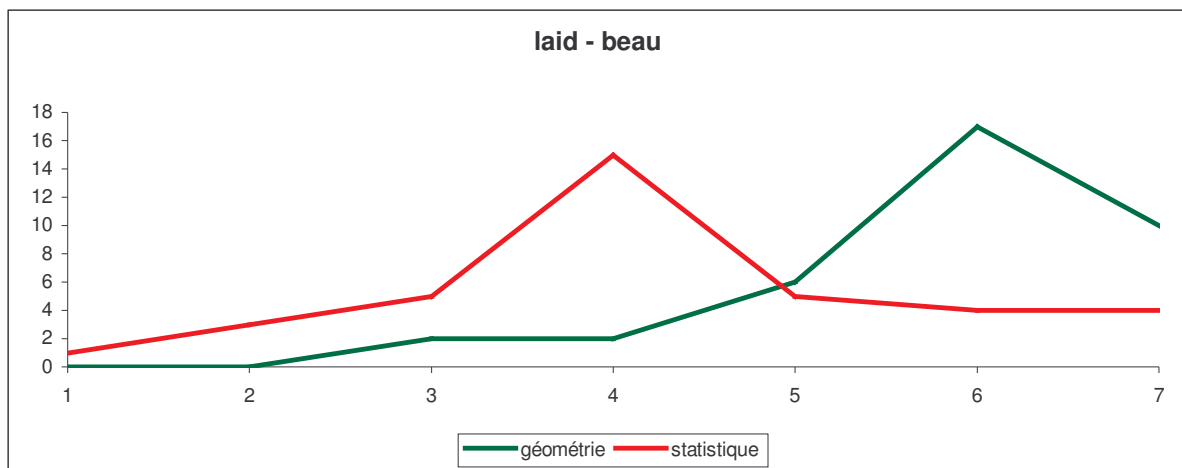
R. A. Fischer, *Statistical Methods for Research Workers* (1925)

Apologue 4

Le point de vue des professeurs stagiaires saisi à travers un différenciateur sémantique et à travers un questionnaire :

- Une enquête par questionnaire
- Une enquête par différenciateur sémantique





Épilogue 4 // Prologue 5

Un savoir citoyen : le problème des obstacles dans la culture et la société.

- Herbert George Wells (1866-1946) : *Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.*
- Recommandations de l'*American Statistical Association* et de la *Mathematical Association of America*.

1. *Emphasize the elements of statistical thinking*: (a) the need for data, (b) the importance of data production, (c) the omnipresence of variability, (d) the measuring and modeling of variability.

2. *Incorporate more data and concepts, fewer recipes and derivations. Wherever possible, automate computations and graphics.* An introductory course should: (a) rely heavily on *real* (not merely realistic) data, (b) emphasize *statistical* concepts [...] (c) rely on computers rather than computational recipes, (d) treat formal derivations as secondary in importance.

3. *Foster active learning*, through the following alternatives to lecturing: (a) group problem solving and discussion, (b) laboratory exercises, (c) demonstrations based on class-generated data (d) written and oral presentations, (e) projects, either group or individual.

Apologue 5

Illustration de la pénétration différentielle de la culture statistique dans des journaux français et des journaux américains.

Épilogue 5 // Prologue 6

La statistique, science de la variabilité :

– l’objectif premier de l’éducation statistique à l’école, au collège et au lycée est de faire reconnaître que chaque variable statistique que nous voyons autour de nous a une certaine distribution de fréquences qui, sauf exception, n’est ni celle d’une variable *constante*, ni celle

d'une variable *équirépartie* : en d'autres termes, ce qui doit se mettre à vivre pour les élèves (et pour les citoyens qu'ils deviendront), c'est la *variété* des distributions de fréquences existant « dans la nature » ;

– tout le monde sait en principe que, s'il est vrai qu'on meurt à tout âge, on ne meurt pas *également* à tout âge ; mais la chose est souvent vite perdue de vue dès lors qu'on évoque d'autres phénomènes – l'affirmation que des maris qui battent leur femme se rencontrent dans toutes les classes de la société se transforme subrepticement en la croyance qu'on bat autant sa femme dans telle classe qu'en toute autre, par exemple ;

– le phénomène est parfois utilisé à des fins on ne peut plus honteusement intéressées ; dans le numéro de septembre 2005 du *Harper's Magazine*, en un article sous-titré *America's educational apartheid*, l'auteur dénonce ainsi l'emploi hypocrite et trompeur de l'adjectif *diverse* appliqué à des écoles pourtant très fortement ségréguées, écrivant par exemple (art. cit., p. 43) :

Schools in which as few as 3 or 4 percent of students may be white or Southeast Asian or of Middle Eastern origin, for instance—and where *every other child* in the building is black or Hispanic—are referred to as “diverse”.

– dans le cas d'une variable quantitative X , il semble que le déni de la variabilité se traduise, non par la croyance spontanée (et manipulable) en l'équirépartition, mais par un ensemble de pratiques culturellement établies qui tendent à accréditer l'idée qu'il existerait une valeur a dont on pourrait dire sans autre correctif que $X \approx a$. Répondre à la question « C'est grand comment un ... ? » par « Ça a une taille de ... environ » concourt à conforter ce déni.

Apologue 6

Le « refoulement » (ou l'évitement), dans les manuels, de l'obstacle culturel de la variabilité.

Épilogue 6 // Prologue 7

Une « pathologie » classique de l'enseignement de la statistique : la réduction arithmétique

Apologue 7

La réduction arithmétique dans les manuels et les activités de classes.

Épilogue 7 // Prologue 8

La réduction arithmétique est-elle dans le programme 2000 ? Le cas du secteur des « résumés numériques ».

Apologue 8

• Le point de vue « authentique » (non arithmétiquement réduit) est en fait présent déjà dans le programme de 3^e.

La statistique en 3^e : programme

Contenus 1

Caractéristiques de position d'une série statistique

Compétences exigibles 1

Une série statistique étant donnée (sous forme de liste ou de tableau, ou par une représentation graphique), proposer une valeur médiane de cette série et en donner la signification.

Commentaires 1

Il s'agit essentiellement d'une part, de faire acquérir aux élèves les premiers outils de comparaison de séries statistiques, d'autre part de les habituer à avoir une attitude de lecteurs responsables face aux informations de nature statistique.

On repère, en utilisant effectifs ou fréquences cumulées, à partir de quelle valeur du caractère on peut être assuré que la moitié de l'effectif est englobée. Les exemples ne devront soulever aucune difficulté au sujet de la détermination de la valeur de la médiane.

Contenus 2

Approche de caractéristiques de dispersion d'une série statistique

Compétences exigibles 2

Une série statistique étant donnée, déterminer son étendue ou celle d'une partie donnée de cette série.

Commentaires 2

L'étude des séries statistiques ayant même moyenne permettra l'approche de la notion de dispersion avant toute introduction d'indice de dispersion.

On introduira l'étendue de la série ou de la partie de la série obtenue après élimination de valeurs extrêmes.

On pourra ainsi aborder la comparaison de deux séries en calculant quelques caractéristiques de position et de dispersion, ou en interprétant des représentations graphiques données.

Contenus 3

Initiation à l'utilisation de tableurs-grapheurs en statistique

Compétences exigibles 3

Ø

Commentaires 3

Les tableurs que l'on peut utiliser sur tous les types d'ordinateurs permettent, notamment en liaison avec l'enseignement de la technologie, d'appliquer de manière rapide à des données statistiques les traitements étudiés.

Accompagnement du programme de 3^e (extrait)

En classe de 3^e, il s'agit d'aider les élèves à franchir une nouvelle étape dans le développement de leur autonomie de jugement à propos d'informations qui peuvent être nombreuses. Dans le cas d'un regroupement en classes, les choix effectués peuvent avoir des effets sur les résultats numériques ou les représentations graphiques et leurs interprétations.

En classe de 4^e, on a pu observer que « la moyenne d'une population dont les éléments sont rangés par ordre croissant ne sépare pas ceux-ci, en général, en deux parties de même effectif », ce qui justifie l'introduction de la médiane en classe de 3^e. Les élèves disposent alors de deux indicateurs de la tendance centrale d'une population, leur position relative pouvant faire l'objet d'une interprétation dans des situations appropriées.

La nécessité de distinguer deux séries statistiques de même tendance centrale justifie l'intérêt de la notion de dispersion. Dans ce premier contact, le programme se limite à l'étendue d'une série statistique ou à l'étendue d'une partie donnée de celle-ci ; cela permet, sans difficulté technique, de familiariser les élèves avec une démarche habituelle en statistique : procéder à une synthèse de l'information sous la forme de nombres mesurant respectivement la position et la dispersion de la série étudiée.

- Si le programme de 2^{de} met en garde contre la dérivation arithmétique et fournit les points d'appui de l'étude d'une statistique authentique, il le fait de façon bien timide, en sorte que l'on peut aisément ignorer ses prescriptions (voire ses proscriptions !).

La statistique en 2^{de} : secteur 1, résumés numériques

Contenus 1

Résumé numérique par une ou plusieurs mesures de tendance centrale (moyenne, médiane, classe modale, moyenne élaguée) et une mesure de dispersion (on se restreindra en classe de seconde à l'étendue).

Capacités attendues 1

Utiliser les propriétés de linéarité de la moyenne d'une série statistique.
Calculer la moyenne d'une série à partir des moyennes de sous-groupes.

Commentaires 1

L'objectif est de faire réfléchir les élèves sur la nature des données traitées, et de s'appuyer sur des représentations graphiques pour justifier un choix de résumé.

On peut commencer à utiliser le symbole Σ .

On commentera quelques cas où la médiane et la moyenne diffèrent sensiblement.

On remarquera que la médiane d'une série ne peut se déduire de la médiane de sous-séries. Le calcul de la médiane nécessite de trier les données, ce qui pose des problèmes de nature algorithmique.

Contenus 2

Définition de la distribution des fréquences d'une série prenant un petit nombre de valeurs et de la fréquence d'un événement.

Capacités attendues 2

Calcul de la moyenne à partir de la distribution des fréquences.

Commentaires 2

Ø

Accompagnement du programme de 2^{de} (extrait)

• Statistique descriptive

Le programme comporte quelques éléments sur les résumés numériques de séries statistiques, déjà travaillés au collège ; il s'agit essentiellement d'entretenir les acquis, de les réinvestir dans certains thèmes et/ou à l'occasion de certains événements que pourrait offrir l'actualité.

La statistique donne lieu à de nombreuses activités numériques et favorise la maîtrise du calcul ; cependant, de tels calculs ne doivent être demandés que dans la mesure où ils permettent aux élèves de mieux comprendre la spécificité de la série statistique en jeu. Estimer la moyenne de séries de données quantitatives en les regroupant par classe n'est plus une pratique utile en statistique depuis que des ordinateurs calculent la moyenne de milliers de données en une fraction de seconde ; par contre savoir calculer une moyenne à partir de moyennes des sous-groupes ou comprendre la linéarité de la moyenne peut donner lieu à des exercices pertinents au regard de la pratique de la statistique. Calculer simplement, à partir de la moyenne, la moyenne élaguée d'une ou plusieurs valeurs extrêmes montre l'influence d'éventuelles valeurs aberrantes.

Épilogue 8 // Prologue 9

Les problèmes du secteur « Simulation et fluctuation d'échantillonnage » :

1. L'importance annoncée de la familiarisation avec les fluctuations d'échantillonnage.

2. Le phénomène est observée à l'aide de la simulation (plutôt qu'avec des données « réelles »).
3. Le travail sur la fluctuation d'échantillonnage est écrasé par le passage à la limite et la « dérive probabiliste ».

Apologue 9

- Le programme et son accompagnement

La statistique en 2^{de} : secteur 2, simulation et fluctuation d'échantillonnage

Contenus 3

Simulation et fluctuation d'échantillonnage.

Capacités attendues 3

Concevoir et mettre en œuvre des simulations simples à partir d'échantillons de chiffres au hasard.

Commentaires 1

La touche « random » d'une calculatrice pourra être présentée comme une procédure qui, chaque fois qu'on l'actionne, fournit une liste de n chiffres (composant la partie décimale du nombre affiché). Si on appelle la procédure un très grand nombre de fois, la suite produite sera sans ordre ni périodicité et les fréquences des dix chiffres seront sensiblement égales.

Chaque élève produira des simulations de taille n (n allant de 10 à 100 suivant les cas) à partir de sa calculatrice ; ces simulations pourront être regroupées en une simulation ou plusieurs simulations de taille N , après avoir constaté la variabilité des résultats de chacune d'elles. L'enseignant pourra alors éventuellement donner les résultats de simulation de même taille N préparées à l'avance et obtenues à partir de simulations sur ordinateurs.

Accompagnement du programme de 2^{de} (extrait)

• La fluctuation d'échantillonnage

Nous appellerons échantillon de taille n d'une expérience la série des résultats obtenus en réalisant n fois cette expérience ; on dira aussi qu'un échantillon est une liste de résultats de n expériences identiques et indépendantes ; on se limite en seconde aux échantillons d'expériences ayant un nombre fini d'issues possibles. La distribution des fréquences associée à un échantillon est le *vecteur* dont les composantes sont les fréquences des issues dans l'échantillon ; on ne donnera pas de définition générale de la notion de distribution des fréquences, on se contentera de la définir comme liste des fréquences dans chacune des situations que l'on traitera. Les distributions des fréquences varient d'un échantillon à l'autre d'une même expérience : c'est ce qu'on appellera en classe de seconde la fluctuation d'échantillonnage.

Aborder la notion de fluctuation d'échantillonnage se fera en premier lieu dans des cas simples (lancers de dés, de pièces), où la notion d'expériences identiques et indépendantes est intuitive et ne pose pas de problème ; l'élève reprendra ainsi contact avec des expériences aléatoires familières (lancer de dés équilibrés) et les enrichira. Historiquement, l'honnête homme du XVII^e siècle s'est familiarisé à l'aléatoire en pratiquant les jeux de hasard ; maintenant, les calculatrices et les ordinateurs permettent la production aisée de listes de chiffres au hasard ; la production de telles listes fera partie, à côté des lancers de dés ou de pièces équilibrés, à côté de tirage de boules dans des urnes, du bagage d'expériences de référence de l'élève. L'étude de ces expériences de référence sera ainsi à la base de la formation sur l'aléatoire des élèves.

L'esprit statistique naît lorsqu'on prend conscience de l'existence de fluctuations d'échantillonnage ; en seconde, l'élève constatera expérimentalement qu'entre deux échantillons, de même taille ou non, les distributions des fréquences fluctuent ; la moyenne étant la moyenne pondérée des composantes de la distribution des fréquences est, elle aussi, soumise à fluctuation d'échantillonnage ; il en est de même de la médiane. On observera aussi que l'ampleur des fluctuations des distributions de fréquences calculées sur des échantillons de taille n diminue lorsque n augmente. Par ailleurs, on n'hésitera pas à parler de la fréquence d'un événement (« le nombre observé est pair », « le nombre est un multiple de

trois », etc.) sans pour autant définir formellement ce qu'est un événement, ni donner de formules permettant le calcul automatique de la fréquence de la réunion ou de l'intersection de deux événements. Le choix pédagogique est ici d'aller de l'observation vers la conceptualisation et non d'introduire d'abord le langage probabiliste pour constater ensuite que tout se passe comme le prévoit cette théorie.

- Dans les manuels

Q. Un fabricant de chaussures pour hommes ne veut pas investir dans la fabrication de chaussures dont la pointure ne dépasse pas 5 % de la demande. Un sondage, sur 250 hommes adultes choisis au hasard donne la répartition des pointures suivantes :

Pointure	38	39	40	41	42	43	44	45	46	47	48
Nombre d'hommes	1	4	21	34	48	55	42	37	7	0	1

Si l'on se fie à cet échantillon, quelles pointures seront fabriquées ? Quel pourcentage de clients potentiels ne trouveront pas chaussures à leurs pieds ?

R. 38 ; 39 ; 46 ; 47 ; 48 → pas fabriquer
5,2 % des personnes ne pourront pas trouver de chaussures à leurs pieds.

- Les 11 fiches de statistique.

Les « fiches de statistique » introduisent subrepticement (voire ouvertement) les notions clés d'une théorie probabiliste de la variabilité.

1) Ces fiches sont précédées d'un texte « doctrinal » qui reprend pourtant les points de vue explicités dans le programme et son document d'accompagnement.

- Ce texte rappelle d'abord que le but de l'étude de la statistique en seconde est l'initiation à la variabilité et à la « fluctuation d'échantillonnage » :

L'objectif du chapitre de statistique, en classe de seconde, est de sensibiliser les élèves à l'aléatoire et plus particulièrement de leur faire prendre conscience de l'existence de la fluctuation d'échantillonnage dans des cas simples : cette fluctuation est ici celle de distributions des fréquences entre des séries obtenues par répétition d'expériences identiques. La moyenne, la médiane, le maximum, le minimum se calculant à partir de cette distribution des fréquences sont aussi soumis à la fluctuation d'échantillonnage.

- À nouveau se trouve précisé qu'il s'agit d'aborder les faits de variabilité par l'observation et l'expérimentation, et non par la théorie mathématique qui en rendra raison plus tard. (On notera la notation « psychologique » finale à propos « des enfants ou des adolescents », là où on attendrait une notation *épistémologique*, valable aussi bien pour l'adulte profane en la matière...)

Le choix pédagogique de ce programme est d'aller de l'observation vers la conceptualisation et non d'introduire d'abord le langage probabiliste pour constater ensuite que tout se passe comme le prévoit cette théorie ; pour des enfants ou des adolescents, le langage des probabilités ne préexiste pas à la perception de l'aléatoire.

- La seule conceptualisation admissible, rappelle le texte que nous suivons, est celle de la « protothéorie » du hasard, combinée à la notion de série statistique associée à une certaine expérience aléatoire.

Le langage utilisé est celui de la vie courante : égalité des chances, pièce ou dé équilibré, choix au hasard dans un ensemble fini, etc. Une série statistique dont les termes sont les résultats issus de la réplication d'une même expérience sera appelée un échantillon de cette expérience (ou plus simplement échantillon si cela ne prête pas à confusion) ; par exemple : les résultats de n lancers d'un dé, les couleurs de n boules tirées au hasard dans une urne, l'opinion de n personnes choisies au hasard dans une population bien définie (l'expérience que l'on répète est ici le recueil de l'opinion d'une personne choisie au hasard), la taille de n personnes choisies au hasard dans une population bien définie (l'expérience que l'on répète est ici la mesure de la taille d'une personne choisie au hasard), etc.

- C'est alors que s'introduit une ambiguïté qui, en effet, on va le voir, aura des effets sensibles sur le contenu des études statistiques proposés dans les « fiches » : il semble, à lire le passage suivant, que la stabilisation des fréquences arrive très vite, et soit la chose à observer et à expliquer, comme si ce phénomène *annulait* celui des fluctuations d'échantillonnage, comme si l'on ne savait pas résister au passage à la limite – et... en classe de première !

Les situations traitées mènent à l'observation de résultats qui appellent une explication ; par exemple, si on lance deux pièces équilibrées et qu'on regarde le nombre de *pile* obtenus, la distribution des fréquences de 0, 1, 2 semble voisine de (1/4, 1/2, 1/4) ; « pourquoi en est-il ainsi ? » est une question à adresser aux mathématiques, qui justifie de formaliser le langage utilisé et d'introduire à la théorie des probabilités. Le langage probabiliste permettra de se dégager de la polysémie des termes employés dans la vie courante (tels ceux de *chance* ou de *hasard*) ; des éléments simples de théorie des probabilités (introduits en classe de première) rendront intelligibles de nombreuses observations liées aux phénomènes aléatoires.

- La tension entre cette « tentation de l'infini », associée à demi-mot à la puissance des « outils de simulation », et le travail sur des échantillons de taille limitée se voit notamment dans le passage ci-après.

On simulera les expériences de référence et on pourra ainsi étudier de longues séries ; on simulera aussi de nouvelles situations à l'aide de listes de chiffres au hasard. Les outils de simulation sont puissants, mais pour les élèves une première étape importante est d'établir un lien entre l'expérience et une simulation de cette expérience : quand les thèmes s'y prêtent (et c'est le plus souvent le cas), la classe pourra être partagée en deux groupes : dans un groupe, les élèves feront effectivement des expériences, dans l'autre les élèves simuleront l'expérience.

2) L'examen des fiches montre cette tension à l'œuvre : on en recense quelques témoignages, sans prétention d'exhaustivité.

- La 2^e fiche, intitulée *De plus en plus de lancers d'un dé*, comporte le résumé thématique suivant, où l'on saisit notamment la tension entre le 3^e et le 4^e points :

- Échantillon de taille n d'une expérience.
- Notion de distribution de fréquences et de fluctuation d'échantillonnage.

- La distribution des fréquences de lancers d'un dé **équilibré**, calculée sur une série de taille n varie d'une série à l'autre ; on observe que cette fluctuation est d'autant plus faible que n est grand ; quand n augmente, la distribution des fréquences se rapproche de $(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$.
- La distribution des fréquences de lancers d'un dé **non équilibré**, calculée sur une série de taille n , varie d'une série à l'autre ; on observe que cette fluctuation est d'autant plus faible que n est grand.

• La 3^e fiche, intitulée *Le lièvre et la tortue*, ne fait pas qu'apparaître en quelque sorte hypnotisée par le comportement « à l'infini » : elle évoque explicitement la théorie probabiliste et la notion de probabilité. Observant que, dans une suite de 10 000 parties, le lièvre bat la tortue dans environ 67 % des cas, le rédacteur de la fiche s'interroge :

D'où sort ce 67 % ? On conçoit que ce pourcentage ne dépend que du fait que le dé est équilibré, donc calculable à partir de cette hypothèse, mais comment faire ce calcul ?

Voici la réponse :

Expliquer le phénomène observé, c'est démontrer dans un cadre théorique que le lièvre gagne et calculer ses chances théoriques de gagner : mais pour cela, il faut donner une définition mathématique des hypothèses faites, à savoir que le dé est équilibré et les lancers indépendants les uns des autres. La théorie qui permet de démontrer tout cela est la théorie des probabilités. On montre que la « chance théorique » du lièvre est $1 - (5/6)^6$, soit à peu près 0,665. Voici de nouvelles simulations ; on pourra observer que quand n augmente, les fréquences se rapprochent des valeurs théoriques.

	f	$1 - f$	f_1	f_2	f_3	f_4	f_5	f_6
10	0.6100	0.3900	0.1356	0.1504	0.2116	0.1788	0.1713	0.1524
1000	0.6701	0.3299	0.1661	0.1662	0.1635	0.1656	0.1696	0.1689
10000	0.6667	0.3333	0.1667	0.1671	0.1660	0.1659	0.1668	0.1675
100000	0.6651	0.3349	0.1666	0.1667	0.1667	0.1669	0.1664	0.1671

S'interrogeant ensuite sur la durée des parties, le même auteur indique que, pour 10 000 parties, la durée moyenne est 3,958 et note :

Là aussi, la seule hypothèse faite concerne les lancers de dés et il devrait y avoir une formule qui permette de calculer une valeur théorique μ pour cette moyenne. Cette formule existe bien et se démontre par le calcul des probabilités : $\mu = \sum_{i=1}^{i=5} i \times \frac{5^{i-1}}{6^i} + 6 \times \frac{5^5}{6^5} \approx 3,99$.

• La 5^e fiche, *Politique nataliste*, est pareillement fascinée par le passage à la limite et à la théorie. Elle se conclut ainsi :

Les nombres moyens d'enfants observés dans cette simulation sont respectivement $m_4 = 1,90$, $m_5 = 1,84$, $m_6 = 1,95$. On pourrait s'attendre à ce que $m_6 > m_5 > m_4$. S'agit-il là d'un tour joué par la fluctuation d'échantillonnage ? On pourrait refaire d'autres simulations. Mais on peut démontrer, dans le cadre de la théorie des probabilités, une formule donnant la valeur théorique du nombre moyen d'enfants pour les trois cas envisagés :

$$\mu_4 = \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{4}{8} \quad \mu_5 = \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{4}{16} + \frac{5}{16} \quad \mu_6 = \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{4}{16} + \frac{5}{32} + \frac{6}{32}$$

On pourra comparer ces valeurs théoriques entre elles.

- La 6^e fiche, *Faites vos jeux*, manifeste la même impatience « théoriste », comme si la théorie – et non l’expérience – pouvait seule combler le désir de « certitude » du rédacteur.

La question posée au début de cette fiche concerne l’événement $\{ m \geq 6 \}$. Or sur les 2000 simulations, environ 82 % réalisent cet événement : on ne peut pas dire qu’un tel événement est exceptionnel ! La fréquence observée est bien sûr soumise à la fluctuation d’échantillonnage ; le calcul des probabilités permet de montrer que pour n expériences, lorsque n augmente, la fréquence observée se rapproche d’un nombre p , où $p \approx 0,8$.

En l’espèce, la « théorie » est ici celle des *chaînes de Markov*...

Épilogue 9 // Prologue 10

En seconde, organiser la rencontre avec la variabilité et ses lois expérimentales, sans céder à l’attraction du modèle probabiliste : un débat...

Apologue 10

- Céder à l’attraction ?
- Y résister.

Dans l’enseignement prodigué en classe de seconde, on se défendra contre les tentations précédentes, moins peut-être pour ne pas empiéter sur ce qui viendra après – en première – que pour donner toute sa place à la fluctuation d’échantillonnage, et cela pour deux ordres de raisons, liés entre eux.

1) La première grande raison, c’est qu’il convient de se persuader que la distribution des fréquences d’un échantillon ne permet pas en général d’accéder à une connaissance « sûre » de la structure de la *population toute entière*, c’est-à-dire qu’un « sondage » – en particulier sur un échantillon qui n’a que le mérite d’être « disponible » – ne permet pas d’avoir une connaissance sûre de la population totale. On ne passe pas facilement d’échantillons en règle générale de taille limitée à une population de très grande taille.

- Pour aboutir à un tel enseignement, il n’est *nullement besoin* de connaître les « valeurs *théoriques* » : il suffit de constater que les fréquences *empiriques* diffèrent, et qu’elles ne « convergent » parfois que bien lentement !

- Voici par exemple un diagramme en bâtons donnant les fréquences d’apparition du numéro 2 lors de 20 séries de 20 lancers d’un dé à 10 faces :

0	xx
0,05	xxxx
0,1	xxxxxxxx
0,15	xxxxx
0,2	x
0,25	x

L'étendue est ici de 0,25. Voici maintenant les fréquences d'apparition du 2 lors de 10 séries de 40 lancers du même dé. Cette fois, l'étendue n'est plus que 0,125 : elle est réduite de moitié. Mais, comme on le verra, la fréquence peut encore varier, d'un échantillon de taille 40 à l'autre, de 5 % à 17,5 %.

0,05	xx
0,075	xx
0,1	x
0,125	xxx
0,15	x
0,175	x

Même si l'on dispose d'un échantillon de taille « assez grande », on ne peut pas davantage faire de conclusion tout à fait assurée : sur un échantillon de 100 lancers, la fréquence du 2 sera de 7 % ; sur un autre de même taille, elle sera de 13 % ; etc.

- Un résultat général, à propos d'un événement A de probabilité $p \in]0 ; 1[$, quant à la taille $n_0(\epsilon, \delta)$, où $0 < \epsilon \leq p(1 - p)$ et $\delta \in]0 ; 1]$, telle que l'on ait

$$n \geq n_0(\epsilon, \delta) \Rightarrow P^{(n)}(|f_{(n, \Sigma)}(A) - p| \geq \epsilon) \leq \delta$$

est celui-ci (A. Rényi, *Calcul des probabilités*, Dunod, Paris, 1966, p. 366) : $n_0(\epsilon, \delta) = \frac{9 \ln \frac{2}{\delta}}{8 \epsilon^2}$.

Si l'on admet que la sortie du 2 lorsqu'on lance le dé à dix faces de l'exemple précédent a une probabilité $p = 0,1$, si l'on prend alors $\epsilon = 1 \% = 0,01$ et si l'on souhaite que la fréquence du 2 ne s'éloigne de $\epsilon = 0,01$ ou plus de la fréquence théorique $p = 0,1$ sur en moyenne au plus un échantillon de taille n sur 50, c'est-à-dire si l'on prend $\delta = \frac{1}{50}$, il faut prendre $n \geq \frac{9 \ln 100}{8 \times 10^{-4}}$, soit

$n \geq 51\,809$. Si l'on accepte de s'éloigner de 2 % de p sur en moyenne au plus un échantillon sur 10 ($\epsilon = 2 \times 10^{-2}$ et $\delta = \frac{1}{10}$), il vient : $n_0(\epsilon, \delta) = \frac{9 \ln 20}{32 \times 10^{-4}} \approx 8425,49\dots$, ce qui suppose donc

$n \geq 8426$, nombre qui reste considérable. Pour les « petits » échantillons que l'on obtient « à la main » ou que l'on peut observer dans le monde naturel ou social, le point de vue du passage « à la limite » est inadéquat.

- L'adéquation de ce point de vue avec certaines situations où le « grand nombre » l'emporte sera mise en valeur en classe de 1^{re}, où l'on pourra s'inspirer notamment des considérations suivantes (Rényi, *op. cit.*, p. 366-367) :

Dans beaucoup de phénomènes aléatoires de masse, il apparaît que certaines grandeurs, bien que dépendant du hasard, peuvent pratiquement être considérées comme constantes. Comme exemple d'une telle grandeur, considérons la pression d'un gaz enfermé dans un récipient. Cette pression est le résultat des chocs des molécules du gaz contre les parois du récipient. Le nombre des chocs et la vitesse des molécules dépendent du hasard, la pression résultante est cependant pratiquement constante quand le nombre de molécules est très grand. De tels phénomènes s'expliquent comme une réalisation de la loi des grands nombres.

2) Il existe une seconde grande raison, qui compte dans l'éducation d'un citoyen : si l'on extrait un échantillon d'une population de structure connue (ou supposée), à quoi peut-on

s'attendre quant à cet échantillon ? Si je tire au hasard les noms de trois élèves d'une classe de 28 élèves composée de 16 filles et 12 garçons, puis-je parier que cet échantillon de taille 3 contiendra au moins un garçon et au moins une fille ?...

• On sait que le calcul des probabilités est né en partie de problèmes posés par les jeux d'argent. Dans une lettre fameuse, datée du 29 juillet 1654, Pascal écrit ainsi à Fermat :

Je n'ai pas le temps de vous envoyer la démonstration d'une difficulté qui étonnait fort M. de Méré ; car il a très bon esprit, mais il n'est pas géomètre. C'est, comme vous savez, un grand défaut. Il me disait donc qu'il avait trouvé difficulté sur les nombres pour cette raison : Si on entreprend de faire un 6 avec un dé, il y a avantage de l'entreprendre par quatre coups. Si on entreprend de faire « sonnez » avec deux dés, il y a désavantage de l'entreprendre en vingt-quatre coups, et néanmoins 24 est à 36, qui est le nombre des faces de deux dés, comme 4 est à 6, qui est le nombre des faces d'un dé. Voilà qui était son grand scandale qui lui faisait dire hautement que les propositions n'étaient pas constantes et que l'Arithmétique se dément.

Si l'on fait rouler un dé équilibré, on peut « avoir avantage » à parier qu'on amènera un 6 en quatre lancers, c'est-à-dire que, plus d'une fois sur deux en moyenne, il se trouvera un 6 parmi les quatre sorties. La première chose à faire, *aujourd'hui*, c'est de tenter de vérifier cette « croyance » *par une simulation* : c'est ce qu'on fera en 2^{de}. Voici par exemple 50 séries de quatre lancers d'un dé.

3-5-3-6 // 4-3-1-4 // 5-6-3-3 // 2-3-1-5 // 1-5-5-5 // 2-1-6-6 // 6-4-5-4
 // 1-5-4-1 // 4-1-4-5 // 2-1-2-1 // 6-2-2-3 // 4-4-4-4 // 3-4-2-6 // 5-4-2-2
 // 3-4-3-1 // 5-3-6-5 // 1-1-5-3 // 2-2-1-2 // 1-4-3-4 // 3-4-1-6 // 3-2-4-2
 // 3-6-2-1 // 4-2-3-3 // 2-6-6-4 // 4-4-5-4 // 2-1-4-3 // 6-6-6-4 // 5-5-5-6
 // 2-1-4-6 // 6-1-3-2 // 3-2-5-2 // 2-1-1-1 // 5-6-4-1 // 1-2-4-6 // 3-5-6-5
 // 1-5-3-4 // 6-3-1-4 // 4-6-3-4 // 4-6-6-4 // 3-6-5-6 // 6-5-4-1 // 1-2-5-1
 // 2-5-3-2 // 2-3-2-6 // 3-4-5-4 // 4-6-3-6 // 4-6-4-5 // 6-6-1-2 // 2-2-5-3
 // 5-2-4-4

Parmi ces 50 séries, 24 contiennent un 6, ce qui est donc moins que « prévu » (ou qu'espéré) ! L'examen d'autres séries de quatre lancers permettrait de voir s'il s'agit là d'un fait un peu exceptionnel, ou bien s'il faut conclure par exemple que, s'il y a « avantage », celui-ci est fort limité... Si l'on avait parié d'obtenir un 6 en cinq lancers – mais il aurait fallu trouver un parieur qui relève le défi ! –, les 40 séries de 5 lancers ci-après semblent suggérer que l'affaire aurait été sensiblement plus avantageuse, puisqu'on y compte 23 séries (sur 40) contenant un 6.

3-5-3-6-4 // 3-1-4-5-6 // 3-3-2-3-1 // 5-1-5-5-5 // 2-1-6-6-6 // 4-5-4-1
 // 5-4-1-4 // 5-2-1-2-1 // 6-2-2-3-4 // 4-4-4-3-4 // 2-6-5-4-2 // 2-3-4-3-1
 // 5-3-6-5-1 // 1-5-3-2-2 // 1-2-1-4-3 // 4-3-4-1-6 // 3-2-4-2-3 // 6-2-1-4-2
 // 3-3-2-6-6 // 4-4-4-5-4 // 2-1-4-3-6 // 6-6-4-5-5 // 5-6-2-1-4 // 6-6-1-3-2
 // 3-2-5-2-2 // 1-1-1-5-6 // 4-1-1-2-4 // 6-3-5-6-5 // 1-5-3-4-6 // 3-1-4-4-6
 // 3-4-4-6-6 // 4-3-6-5-6 // 6-5-4-1-1 // 2-5-1-2-5 // 3-2-2-3-2 // 6-3-4-5-4
 // 4-6-3-6-4 // 6-4-5-6-6 // 1-2-2-2-5 // 3-5-2-4-4

En sens inverse, parier sur la sortie de 6 en trois coups apparaîtra nettement plus risqué... Bien entendu, on pourra aussi explorer par simulation la seconde « croyance » du chevalier de Méré – le fait qu'il y aurait avantage à parier sur la sortie de deux 6 (« sonnez ») en 24 lancers de deux dés.

• Examinons ici le problème évoqué plus haut : si l'on tire au hasard les noms de n élèves d'une classe de 28 élèves composée de 16 filles et 12 garçons, à partir de quelle valeur de n peut-on être « sûr » que l'échantillon tiré contiendra au moins un garçon et au moins une fille ? Traduisons « être sûr » par le fait que la non-réalisation de l'événement ne devrait se produire en moyenne qu'au plus une fois sur 20 (par exemple). On numérote les élèves de 1 à 28, les numéros de 1 à 12 correspondant aux garçons, les suivants aux filles. On verra rapidement qu'un échantillon de taille 3 n'a que bien peu de chances de satisfaire la condition attendue. Pour des échantillons de taille 4, on obtiendra par exemple ceci, qui montre qu'on n'y est pas encore :

6 – 14 – 15 – 3 // 2 – 25 – 5 – 16 // 24 – 18 – 26 – 27 // 27 – 4 – 26 – 18 // 13 – 10 – 5 – 3 // 5 – 23 – 4 – 16 // 11 – 8 – 7 – 2 // 9 – 23 – 16 – 21 // 28 – 4 – 20 – 18 // 21 – 28 – 10 – 15 // 19 – 18 – 20 – 3 // 9 – 10 – 6 – 27 // 7 – 19 – 25 – 14 // 19 – 4 – 2 – 12 // 3 – 5 – 9 – 12 // 10 – 9 – 15 – 8 // 5 – 9 – 18 – 4 // 11 – 22 – 27 – 20 // 26 – 22 – 23 – 27 // 22 – 7 – 19 – 21 // 26 – 13 – 16 – 24 // 27 – 5 – 19 – 10 // 19 – 13 – 20 – 27 // ...

On verra ensuite qu'on se rapproche de la « certitude » que l'on souhaite en passant à $n = 5$:

6 – 14 – 15 – 3 – 2 // 25 – 5 – 16 – 24 – 18 // 26 – 27 – 4 – 18 – 13 // 10 – 5 – 3 – 23 – 4 // 16 – 11 – 8 – 7 – 2 // 9 – 23 – 16 – 21 – 28 // 4 – 28 – 20 – 18 – 21 // 28 – 10 – 15 – 19 – 18 // 20 – 3 – 9 – 10 – 6 // 27 – 7 – 19 – 25 – 14 // 19 – 9 – 4 – 2 – 12 // 3 – 5 – 9 – 12 – 10 // 9 – 15 – 8 – 5 – 18 // 4 – 11 – 22 – 27 – 20 // 26 – 22 – 23 – 27 – 7 // 19 – 21 – 26 – 13 – 16 // 13 – 24 – 27 – 5 – 19 // 10 – 19 – 13 – 20 – 27 // ...

g) Le travail demandé en 2^{de} préfigure des développements à plus longs termes, qui relève de l'*inférence statistique* : d'une part, le fait d'inférer, de ce qu'on *observe* sur un ou plusieurs échantillons, ce qu'il en *est* de la *population* ; d'autre part, et comme on vient de le voir, le fait de « prévoir », lorsqu'on tire un échantillon à partir d'une population connue ou supposée telle, ce qu'il en sera de l'échantillon.

1) Dans son livre *Probabilités et statistique* (Masson, Paris, 2001), le biostatisticien Alain-Jacques Valleron note ceci au début du chapitre intitulé « Les tests statistiques : calcul du nombre de sujets nécessaires » (*op. cit.*, p. 81) : « La question la plus souvent posée au biostatisticien est “Combien faut-il de sujets dans mon étude pour répondre à ma question ?” » Sur la notion corrélative d'*intervalle de pari* (qui fait pendant à la notion d'intervalle de confiance), on pourra, outre la référence citée, se reporter par exemple à cette adresse : <http://www.chups.jussieu.fr/polys/biostats/poly/POLY.Chp.9.4.html>. D'une façon plus générale, on pourra également se pencher sur le contenu de l'article que l'on trouvera à l'adresse suivante : www.spim.jussieu.fr/IMG/doc/lecturestat-2.doc.

2) Si l'on choisit d'extraire un échantillon de taille 5 de la classe prise en référence plus haut (12 garçons, 16 filles), est-il raisonnable de parier qu'il y aura parmi les cinq au moins une fille et au moins un garçon, c'est-à-dire que l'échantillon ne sera pas composé que de garçons ou que de filles ? Plus généralement, dans quel intervalle situera-t-on, disons, la proportion des filles dans l'échantillon de taille 5, sachant qu'elle est de 57 % environ dans la population totale (la classe) ? Les connaissances que l'auteur cité ci-dessus voudrait voir se répandre dans la culture biomédicale commune sont aussi à diffuser dans la culture du citoyen : cette diffusion devrait avoir son camp de base en seconde.

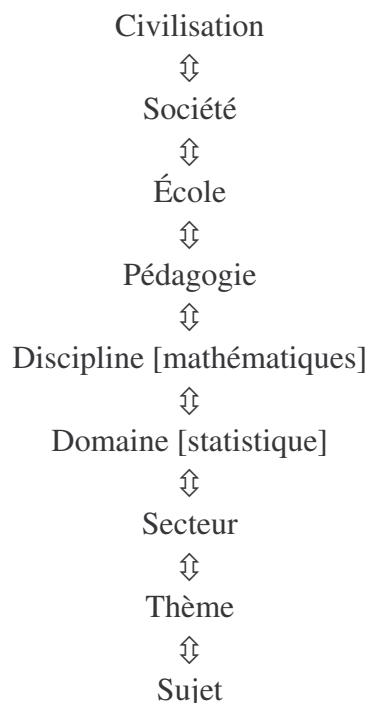
Débat

Matériaux possibles

- La dialectique entre l'école et la société
- Les contraintes pédagogiques
 - « Liberté pédagogique » et illusion idionomique
 - Le format de la séance de classe
 - L'enfermement dans la discipline
- Les contraintes de la discipline
 - Doutes identitaires face à l'altérité épistémologique
 - Entre repliement et métissage
- Les collectifs réformateurs
 - Entre-soi obstiné et collègues invisibles
 - La notion de parcours d'étude et de recherche : « *C'est petit ou c'est gros ? C'est beaucoup ou c'est pas beaucoup ?* »

Quelques restes...

L'échelle des niveaux de co-détermination didactique



La notion de fait social total

« Le fait social total n'est sûrement pas la totalité exhibée d'une société, car qui a jamais pu saisir pareille totalité, même dans la meilleure des monographies ? Il est cette propriété unique

des faits qui sont l'objet même de la science de l'homme, d'être signifiants, c'est-à-dire partiels, contingents, arbitraires et cependant reliés, rattachés, dépendants toujours de quelque chose qui est en eux, à la fois manifesté et caché par eux. »

Clément Tarot, *De Durkheim à Mauss, l'invention du symbolique* (Paris, Éditions La Découverte & Syros, 1999), p. 663

Un point de vue spontané de la profession

« Étant donné que les futurs S n'en feront plus (et seront sans doute aptes à s'y mettre...) et que de toute façon les ES en refont en 1^{re}, certains ne voyaient guère l'intérêt d'en faire vu qu'actuellement c'est la plupart du temps bâclé en fin d'année si c'est fait... Mais nous sommes gênés de demander qu'on supprime les stats alors qu'on revendique des maths citoyennes... »

APMEP [Jean-Pierre Richeton], *BGV*, 87 (juin 1999)