Commission de réflexion sur
l’enseignement des mathématiques

Rapport d’étape

Statistique et probabilités

Sommaire

La statistique traite de données expérimentales ou d'observation, à étudier dans leur contexte (« data with contexts ») : sa spécificité est d'établir des liens entre ces données et la théorie mathématique des probabilités, d'expliquer ainsi le passé et de prévoir l'avenir. L'objet de la statistique exploratoire ou descriptive est de représenter graphiquement, de résumer, de classer des données expérimentales ou d'observation. Confronter des données à des modèles probabilistes pour en expliquer la structure et faire de la prévision est l'objet de la statistique inférentielle. La modélisation ne peut se faire « en aveugle », c'est-à-dire sans observer, résumer, étudier la structure des données expérimentales : des allers et retours sont nécessaires entre leur exploration et leur modélisation stochastique. Cependant, si les deux composantes, exploratoire et inférentielle, sont au cœur de la pratique de nombreux statisticiens professionnels, celles-ci se sont développées au point que chacune a aussi ses domaines de recherche et ses champs d'applications propres et autonomes.

Ainsi, en statistique exploratoire, des outils tels la classification, l'analyse descriptive multivariée peuvent être employés pour eux-mêmes, sans modélisation stochastique. Le traitement de l'information chiffrée, c'est à dire le calcul d'indices à partir de données brutes (pourcentages divers, taux de natalité, etc.), qui est la partie la plus ancienne de la statistique descriptive, ne nécessite pas systématiquement des prolongements de nature probabiliste. Il ne faut pas pour autant oublier le lien essentiel de la statistique et des probabilités.

La statistique n'est par ailleurs pas la seule science ayant recours à des modèles probabilistes et ceux-ci sont au cœur de nombreuses disciplines. Les probabilités sont aujourd'hui une spécialité en interaction forte avec l'extérieur (de la physique à la finance, en passant par la biologie et l'économie), et avec l'intérieur des mathématiques (la théorie des nombres, la combinatoire, la géométrie, l'algèbre, l'analyse). La pratique des probabilités marie l'aspect ludique des questions et la rigueur dans l'application des méthodes. (cf. « En passant par hasard, les probabilités de tous les jours », Gilles Pagès et Claude Bouzitat, Vuibert-1999).

Les problématiques conduisant à des questions de nature statistique sont variées. La prise en compte de l'aléatoire a gagné presque tous les domaines : le contrôle de qualité en milieu industriel, la prévision des petits et des grands risques, l'élaboration de politiques de santé publique, les calculs financiers, etc. ; on trouvera une analyse des pratiques de la statistique actuelle dans « Les chemins de l'aléatoire » de Didier Dacunha-Castelle (Flammarion 1996). Enfin, loin de vouloir faire dire ce qu'on veut aux chiffres, la statistique revendique pleinement le rôle de dévoiler plusieurs aspects d'une même réalité, de prendre en charge des études dont la conclusion ne peut pas être affichée avec certitude.

Pour comprendre l'actualité, une formation à la statistique est aujourd'hui indispensable ; c'est une formation qui développe des capacités d'analyse et de synthèse et exerce le regard critique. Le langage élémentaire de la statistique (avec ses mots tels moyenne, dispersion, estimation, fourchette de sondage, différence significative, corrections saisonnières, espérance de vie, risque, etc.) est, dans tous les pays, nécessaire à la participation aux débats publics : il convient donc d'apprendre ce langage, ses règles, sa syntaxe, sa sémantique ; l'enseignement de la statistique étant, par nature, associé à celui des probabilités, il s'agit en fait d'une « formation à l'aléatoire ».

La question n'est plus « faut-il ou non se fier aux statistiques », mais « comment faire partager au plus grand nombre la connaissance des fondements de cette discipline, des questions qui la concernent, de la nature des preuves qu'elle apporte ». La réponse passe par l'intégration de l'aléatoire à tous les niveaux de l'enseignement.

Ce rapport s'inscrit en complément du 8-ème rapport de juillet 2000 sur la science et la technologie de l'Académie des Sciences, publié aux éditions TEC et DOC ; ce dernier répond à une commande du ministre de l'Education nationale de 1998 de procéder à une évaluation prospective de l'activité scientifique et universitaire française.
Pour ce qui concerne la statistique, le rapport de l'académie a été construit autour des questions suivantes :
-Qu'appelle-t-on statistique ?
-Quelles sont la nature et la qualité
-de la recherche en statistique en France ; quelle est sa place en Europe et dans le monde ?
- de la mise en œuvre des méthodes statistiques, dans les grands secteurs de l'économie et de la vie sociale,
- de la mise en œuvre des méthodes statistiques dans la recherche scientifique et technique,
- de la formation initiale de l'enseignement des statistiques, du primaire au supérieur, et de la formation continue .

Dans le rapport de l'académie s'expriment des gens de différents horizons qui donnent leur point de vue sur la statistique. Les visions personnelles des auteurs ne s'accordent pas toutes ; mais, s'il n'y a pas une pensée statistique unique (il ne peut en être autrement d'une discipline vivante), les zones de convergences sont vastes qui permettent d'envisager sereinement l'enseignement de la statistique.

Le présent rapport de la CREM a pour objectif de prolonger celui de l'académie par des pistes de réflexion pouvant influencer l'évolution future de l'enseignement des statistiques et des probabilités. Il ne s'agit pas ici de définir des curriculums, mais d'une part de rendre compte de questions qui animent vivement les débats à propos de ce chapitre de la formation scientifique, et d'autre part d'éclairer ⁻en illustrant parfois par des exemples didactiques simples⁻ des éléments susceptibles de guider des choix de contenus en différents temps et lieux d'enseignement et de formation.

I-La place de l'aléatoire dans l'enseignement des mathématiques

Une question récurrente à ce propos est : quelle est, sur ce sujet, l'implication des mathématiques, quelles sont les conséquences de l'introduction d'un enseignement de l'aléatoire dans la vision des mathématiques que l'on souhaite transmettre ? Nous proposons ci-dessous, au travers d'exemples, des éléments de réponse.

Le traitement élémentaire de l'information chiffrée repose sur la manipulation et la comparaison de nombres (indices, pourcentages, proportions) ; néanmoins, il convient dans un enseignement de statistique de ne pas s'en tenir à un strict point de vue du calcul numérique, mais de toujours replacer les données dans leur contexte et de chercher à quantifier les différents points de vue qui alimentent les débats.

- Dans le journal « Le Monde » du 28 novembre 2000, on trouve les informations suivantes :
« Le citoyen des Etats-Unis est le premier émetteur au Monde de gaz à effet de serre : 20 tonnes de CO2 par an, contre 10 pour un Allemand et 2,30 pour un Chinois. »
Chiffres éloquents à l'appui, on peut ainsi ranger ces trois pays du plus pollueur au moins pollueur : USA, Allemagne, Chine.
« La Chine produit 3,54 tonnes de CO2 pour la production d'un certain revenu (l'équivalent de 90$ de PNB) tandis que l'Allemagne ne produit que 0,46 tonnes de CO2 pour la même production de revenu et les Etats-Unis 0,77 tonnes. »
Chiffres éloquents à l'appui, on peut ainsi ranger ces trois pays du plus pollueur au moins pollueur : Chine, USA, Allemagne.

Le mot variabilité peut être considéré comme le premier mot de la statistique ; d'autres mots viennent ensuite qui permettent de parler de cette variabilité : l'acquisition de leur sens et la maîtrise de leur usage font intervenir des raisonnements de nature mathématique. Par exemple :

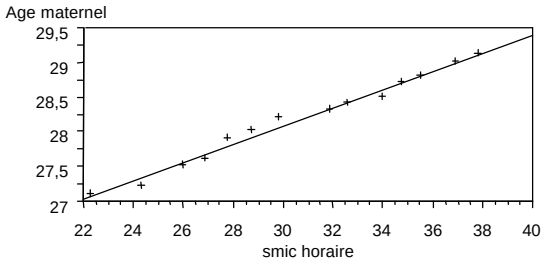
-Comment se transforment les paramètres statistiques élémentaires (moyenne et espérance, variance, pente et ordonnée à l'origine d'une droite d'ajustement par moindres carrés) par transformation affine des données, i.e. par changement d'unité et décalage de l'origine ? Est-il nécessaire, pour des mesures faites tous les 5 ans à partir de 1945, de garder ces dates à quatre chiffres, ou peut-on prendre comme unité 5 ans et comme origine l'année 1945 ? La compréhension de la linéarité passant aussi par la reconnaissance de ce qui ne l'est pas, il convient de comprendre pourquoi, si on a deux séries de données numériques x_i et y_i de même taille, la moyenne des x_iy_i n'est en général pas le produit de la moyenne des x_i par celle des y_i, notion qui au plan théorique devient « l'espérance d'un produit n'est en général pas le produit des espérances ». L'acquisition de réflexes vis a vis de la linéarité est un premier pas vers une pensée statistique autonome, et relève bien du champ des mathématiques.
Toutefois, pour comprendre la pertinence des raisonnements mathématiques en statistique, il convient de les pratiquer dans leur contexte. La statistique exploratoire ou inférentielle n'a pas pour fin ultime de décrire ou modéliser n'importe quelles données ; les études de statistique sont toujours motivées par un questionnement, qui n'est en général pas d'ordre mathématique : introduire ce questionnement dans un enseignement de mathématique et voir comment il s'articule avec des raisonnements mathématiques est considéré par certains comme une ouverture et un enrichissement ; pour d'autres, c'est un dévoiement. Ces deux opinions traduisent un fait indéniable : la statistique est à la fois dans et en dehors des mathématiques. Son enseignement relève des mathématiques et, nous y reviendrons ci-dessous, des autres disciplines.

-Une maxime statistique est « une corrélation forte n'implique pas nécessairement une causalité » ; ce qui s'énonce aussi : si un nuage de points est presque rectiligne, cela n'implique pas nécessairement de relation de cause à effet entre le phénomène mesuré par les abscisses et celui qui est mesuré par les ordonnées ; si cette maxime ne relève que d'un argument d'autorité, et n'est justifiée que par l'étude d'exemples choisis pour leur caractère absurde, elle risque d'être irrépressiblement récusée au premier graphique marquant (tel l'alignement de 10 points, dont l'abscisse donne le nombre d'entrées réalisées par les films étiquetés « film violent » ces dix dernières années et l'ordonnée le nombre d'agressions sur la voie publique pendant les mêmes années). Il appartient au professeur de mathématiques de faire comprendre que :

-si l'abscisse et l'ordonnée des points sont presque des fonctions affines d'une même variable t (le temps par exemple), alors « mathématiquement » le nuage des

points sera presque une droite ; les points n'ont pas une « tendance naturelle » à s'aligner, et un alignement appelle une explication : celle-ci est le plus souvent à rechercher dans l'existence d'une dépendance linéaire à un même facteur (appelé en médecine facteur de confusion).

-l'évolution de très nombreux phénomènes est en première approximation affine (même les phénomènes exponentiels à taux d'accroissements faibles en sont des cas particuliers) .



Dans le graphique ci-dessus, chaque point correspond à une année,
(de gauche à droite : de 1983 à 1996) ; l'abscisse est la valeur du SMIC horaire en début d'année
et l'ordonnée l'age moyen des femmes ayant eu un bébé cette année là.
Le coefficient de corrélation linéaire pour ces 14 points est 0,99.
Cet exemple illustre simplement qu'en première approximation, entre 1983 et 1996,
l'accroissement absolu annuel du SMIC d'une part,
et de l'age maternel d'autre part, sont approximativement constants.

- La vision géométrique est indispensable à la compréhension de l'analyse descriptive multivariée ; celle-ci consiste, pour « voir » un nuage de points dans un espace à n dimensions, n>3, à en regarder les projections sur ses plans d'inertie : l'analyse des représentations graphiques (qualité globale de la représentation, interprétation de la proximité entre points projetés) passe d'abord par la reconnaissance de propriétés de nature purement géométriques.
De plus, pour des étudiants qui manipuleraient un peu les espaces euclidiens, de nombreux calculs s'éclairent avec le point de vue suivant, où des facteurs mesurés sur n individus sont des vecteurs de R^n, muni du produit scalaire défini par $<x,y>= x_i y_i/n$. La moyenne est la projection sur le vecteur **1** (dont toutes les composantes valent 1), l'écart-type de x est la norme de sa projection sur le sous-espace orthogonal à **1** ; écrire que la variance est égale à la moyenne des carrés moins le carré de la moyenne, c'est écrire le théorème de Pythagore. Le coefficient de corrélation entre deux variables est le cosinus de l'angle que forment les variables centrées (si ce cosinus vaut 1 ou –1, les deux variables centrées sont linéairement dépendantes) ; la droite d'ajustement linéaire par moindre carrés de y sur x est la projection de y sur le sous-espace engendré par **1** et x, d'où le calcul des paramètres de la droite. Cette vision géométrique se prolonge naturellement au plan théorique avec les espaces L^2 et fait comprendre l'harmonie des calculs dans « le monde gaussien ». Si ce bagage mathématique n'est pas indispensable, il constitue pour ceux qui le possèdent un socle tout à fait consistant ; en retour ces considérations d'ordre statistique contribuent à faire vivre ces concepts mathématiques.

La théorie des probabilités est aujourd'hui une branche importante des mathématiques : cela implique-t-il pour autant qu'elle doit être prise en considération dans toute formation

mathématique ? Ce qui rend la théorie des probabilités aujourd'hui inévitable est son emprise sur le réel, son lien avec les autres branches des mathématiques, son usage en des lieux inattendus, que ce soit dans ou en dehors des mathématiques, comme ne témoignent les deux exemples ci-dessous :

- Situation 1 : Dix suspects d'un délit commis par une seule personne sont proposés à l'identification par quatre témoins. Chaque témoin désigne un suspect comme étant le coupable, sans connaître le choix des autres témoins. Un des suspects est désigné deux fois. Est-ce que cela constitue une lourde charge contre lui ? Ce problème n'a a priori rien à voir avec les probabilités, mais imaginons la situation la plus absurde qui soit : chaque témoin désigne un suspect au hasard, les choix étant indépendants. Dans ce cas, on a une probabilité 0,504 que les 4 témoins désignent des suspects différents, et donc une probabilité 0,496 qu'au moins un suspect soit désigné au moins deux fois : il paraît difficile qu'un événement que le « complet hasard » produirait presque une fois sur deux constitue une charge.

  Situation 2 : Dans une procédure d'identification, parmi les 10 personnes que voient les quatre témoins, un seul est un vrai suspect. Si deux témoins désignent le vrai suspect, la situation est très différente : en reprenant la situation de choix au hasard et indépendants des témoins, la probabilité qu'exactement deux d'entre eux (resp. au moins deux) désignent ce vrai suspect est 0,0486 (resp. 0,0523).
- Un théorème de Ramsay dit que pour tout entier k>0, il existe un entier n tel que si on trace n points et les n(n-1)/2 segments qui les relient, en coloriant n'importe comment tous ces segments soit en rouge soit en bleu, il existe nécessairement k points parmi les n tels que les k(k-1)/2 segments qui les relient soient tous de la même couleur. On sait que $N<2^{2k}$. Minorer N est longtemps resté un problème ouvert, pour lequel le mathématicien Paul Erdös a proposé le raisonnement suivant ; pour r>k on considère tous les coloriages possibles des r(r-1)/2 segments qui joignent ces r points. Soit X_r la variable aléatoire qui au choix d'un de ces coloriages au hasard associe le nombre de sous-ensembles à k points tels que tous les segments qui les joignent soient de la même couleur. L'espérance de X_r est $2^{1-k(k-1)/2}\binom{r}{k}$; pour $r\leq 2^{k/2}$, cette espérance est inférieure à 1, ce qui démontre l'existence d'un coloriage tel que $X_r=0$. D'ou : $n>2^{k/2}$.

Le calcul des probabilités fait intervenir des objets qui trouvent naturellement leur place dans tout enseignement de mathématiques et il conviendrait inversement de situer ces objets dans un des contextes où ils sont très souvent utilisés. Voici deux exemples simples parmi d'autres :

- la fonction définie par $f(t)=\exp(-t^2/2)$ est une fonction riche à étudier dès qu'on commence à manipuler les fonctions exponentielles (symétrie, point d'inflexion, décroissance rapide à l'infini, aire sous la courbe représentative non seulement finie mais valant $\sqrt{2\delta\pi}$-résultat étonnant lorsqu'on le voit pour la première fois) ; elle est centrale en théorie des probabilités ; $f/\sqrt{2\pi}$ est la densité de la loi de Gauss centrée réduite et intervient dans les calculs d'erreurs pour toutes les sciences expérimentales : elle mérite d'être mieux connue que par le seul nom de « courbe en cloche » qui lui est souvent donné en sciences humaines et parfois aussi en biologie ! Sa primitive $\Phi(x)=(1/\sqrt{2\pi})\int_0^x \exp(-t^2/2)dt$, dont Laplace préconisait la tabulation en 1778, est une fonction de référence, au même titre que les fonctions trigonométriques, dans les grands logiciels de calculs (Mathematica, maple, matlab etc.).
- le temps aléatoire de vie X d'un système sans mémoire vérifie l'équation : $P(X> t+h/X>t)=P(X> h)$, soit $F(t+h)-F(t)=F(h)(1-F(t))$, où F est la fonction de répartition de X (soit $F(h)=P(X\leq h)$) ; en divisant par h et en passant à la limite, on tombe sur l'équation différentielle $F'(t)=F'(0)(1-F(t))$, pour $x\geq 0$ et $F(0)=0$, caractérisant les lois de probabilités continues de durée de vie des phénomènes sans mémoire ; on trouve alors que la densité de X est de la forme $f(t)=ae^{-at}$ pour t>0 ; de telles lois sont à la base des processus de Poisson dont un exemple classique est la désintégration radioactive.

Enfin le calcul des probabilités fait intervenir des objets fascinants. L'exemple du mouvement brownien est caractéristique (voir dans « leçons de mathématiques d'aujourd'hui », Cassini, 2000 la leçon « le théorème de Pythagore et l'analyse multifractale, le mouvement brownien » de J.P. Kahane). C'est d'abord le mouvement désordonné du pollen en suspension dans un liquide, observé par le botaniste Brown ; puis, avec Einstein, le mouvement causé par des chocs moléculaires. Avec Wiener, c'est un objet mathématique parfaitement défini, dont la place en mathématiques est aujourd'hui centrale. Le mouvement brownien a en effet envahi l'analyse (extrema de fonctions d'un grand nombre de variables), la géométrie (exploration des surfaces et des variétés), la théorie des nombres (tests de primalité). L'intuition des mathématiciens s'est exercée sur cet objet (avec l'image de l'ivrogne brownien ou celle de promenade aléatoire) avec tant de succès que le mouvement brownien désigne maintenant (suivant Paul Levy) l'objet mathématique et non l'objet étudié tour à tour par les biologistes et les physiciens. C'est même l'objet mathématique qui alimente aujourd'hui l'intuition des physiciens et structure en partie leurs images mentales sur ce sujet.

En 1947, dans un livre qui est encore d'actualité, intitulé « Les méthodes statistiques adaptées à la recherche scientifique », Sir R. Fisher, (un des fondateurs de la statistique moderne) introduit ainsi son ouvrage :
« La statistique peut-être considérée comme une branche des mathématiques appliquées, concernant des données d'observation. Comme toujours, les mêmes formules s'adressent également à des groupes très différents de sujets ; mais l'unité des diverses applications est perdue si la théorie mathématique de base est négligée. »
Le cours de mathématiques est un lieu où peut se tisser un lien entre les divers champs d'application de la statistique, où se créent des éléments d'une culture et d'une pratique commune. La simulation aléatoire est de plus aujourd'hui une composante non négligeable de cette pratique commune, et elle change complètement le mode d'accès à l'aléatoire ; elle met en œuvre des concepts et des résultats récents des mathématiques qu'il n'est pas obligatoire de comprendre pour la pratiquer ; elle permet de déterminer des propriétés des expériences dont on simule un modèle, mais n'est compréhensible et efficace qu'accompagnée d'une réflexion mathématique.

Dans différents domaines, dont la biostatistique, les publications des résultats d'études statistiques se font selon un certain standard. Dans une première partie, appelée « matériel et méthode », on présente l'objectif de l'étude, la définition précise des variables prises en compte, le protocole détaillé de recueil des données, les traitements statistiques qui seront utilisés et souvent le logiciel utilisé. Dans une deuxième partie figurent les résultats : représentations graphiques, estimations de divers paramètres, tests d'hypothèses, ainsi que d'éventuels commentaires mathématiques de ces résultats ; la troisième partie, souvent intitulée discussion, est celle de l'interprétation : les résultats mathématiques sont réinterprétés dans le contexte et, à l'aide d'éléments extérieurs à l'étude, des explications sont proposées. La partie centrale et sa compréhension relève ainsi complètement du champ des mathématiques, ce qui milite pour l'exposition de quelques fondements des probabilités et de la statistique dans des enseignements de mathématiques à différents niveaux.

La théorie des probabilités fait partie aujourd'hui de la formation scientifique tous les jeunes reçoivent dans l'enseignement secondaire, et cependant il y a cinquante ans encore, la

place de cette discipline en mathématiques était l'objet de vives controverses. C'est aujourd'hui la statistique qui pose problème à une partie de la communauté des chercheurs en mathématiques. En effet, c'est à travers la statistique que les mathématiques sont aujourd'hui les plus visibles dans la vie quotidienne ; les courbes de poids suivant l'âge, dans les carnets de santé de tous les enfants, donnent pour chaque âge des intervalles de dispersion ; les données du chômage corrigées de variations saisonnières, les prévisions sur le calcul des retraites, les facteurs de confiance en météo font intervenir la statistique. Cependant, la statistique n'est pas partie prenante dans les grands problèmes de recherche en mathématiques actuellement posés et la démarche proprement statistique ne semble pas de nature à intervenir dans leur résolution : certains pensent que l'apprentissage de la démarche statistique ne devrait pas interférer avec la formation « classique » en mathématiques et qu'un enseignement un peu consistant de cette discipline pourrait être réservé aux filières de sciences expérimentales.

Ces réticences s'expliquent en partie par une mauvaise perception de la cohérence que ses fondements mathématiques confèrent à la statistique. Dans « Mathematics : frontiers and perspectives », publié en 2000 par l'American Mathematical Society, David Mumford, spécialiste de géométrie algébrique, se souvient avoir dit à ses étudiants en 1970 : « mon dieu, ne gaspillez pas votre temps à étudier la statistique, ce n'est qu'un recueil absurde de recettes» . Cette image du livre de recettes n'a pas complètement disparue et les doutes de certains se trouvent renforcés par une pratique sociale manifestement excessive de sondages sur tout et n'importe quoi.

Nous sommes cependant d'accord avec David Mumford qui, trente ans après ces considérations dont la tonalité ne semble pas vraiment positive, écrit dans l'ouvrage cité ci-dessus : « la théorie des probabilités et l'inférence statistique émergent comme éléments majeurs de la modélisation scientifique et vont profondément influencer les mathématiques à venir ». Et cette idée rejoint ce que le physicien James Clerk Maxwell écrivait vers 1860 :

« La vraie logique de ce monde se trouve dans le calcul des probabilités »

II-Statistique et outils logiciels

On pourrait dire, pour relier ce paragraphe au précédent, que le matériau brut travaillé par la statistique est constitué de données expérimentales, les outils théoriques utilisés sont essentiellement la géométrie et l'algèbre linéaire pour la statistique exploratoire et les probabilités pour la statistique inférentielle et l'outil matériel est l'ordinateur. La mise en œuvre des méthodes de la statistique demande de gros moyens de calculs ; la statistique a pu se développer à grande échelle parce que les ordinateurs mettent à la disposition d'un large public des possibilités dépassant celles des centres de calculs d'il y a 30 ans.

L'utilisation des outils logiciels dans les milieux professionnels ou de recherche est double :

- mise en œuvre de calculs longs ou complexes (calculs de moyenne, de variance, inversions de matrices, calculs d'extrema sous contraintes, etc.), calculs des valeurs approchées (fonctions de répartition de lois de probabilités par exemple), calculs à distance finie, mise en œuvre de méthodes non paramétriques, de ré-échantillonnage.
- estimation des résultats à partir de simulations : par exemple, le volume de certains ensembles, l'estimation de probabilités d'événements pour lesquels on n'a pas établi de formules qui en permettrait une approximation numérique directe.

Dans une perspective de formation, l'usage de l'informatique est aussi double :

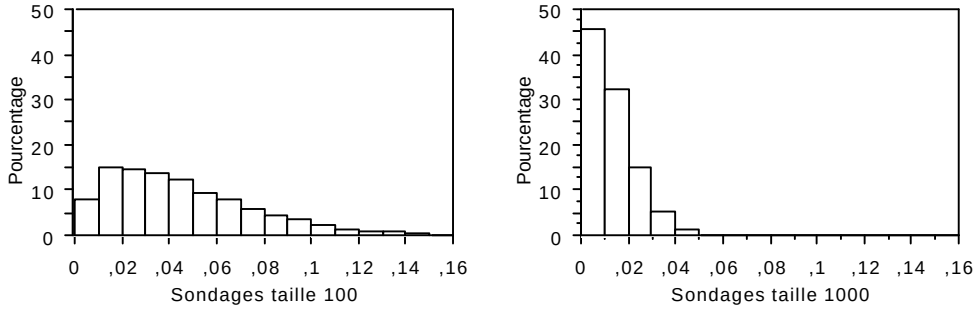
- accès à des données diverses et de qualité et possibilité de mettre en œuvre sur ces données les traitements statistiques pertinents.
- la simulation est un outil privilégié pour acquérir une expérience des phénomènes aléatoires, comprendre des théorèmes de convergence, voir où se situent les questions et appréhender la nature des preuves statistiques.

Dans l'exemple ci-dessous, on s'interroge sur diverses notions de parité des sexes dans une assemblée de 2n personnes choisies dans une population Π :
(i)une notion déterministe : « la parité, c'est lorsque le nombre de femmes est égal à celui des hommes »
(ii)deux notions statistiques :

-« la parité, c'est lorsque l'écart absolu à ½ de la proportion f de femmes est du même ordre de grandeur que celui qu'on obtiendrait en faisant un sondage de taille 2n dans une population où les hommes et les femmes sont en proportions égales ».

- la parité, c'est lorsque l'écart absolu de f à p, où p est la proportion de femmes dans Π est du même ordre de grandeur que celui qu'on obtiendrait en faisant un sondage de taille 2n dans Π.

La notion déterministe est simple à comprendre, les autres sont difficiles à mettre en œuvre si on n'a jamais fait de statistique : comment quantifier cette notion de « même ordre de grandeur» ?
Pour se faire une idée, par exemple pour la première notion statistique, on peut faire des simulations ; on trouvera ci-dessous les histogrammes correspondant d'une part à 10 000 simulations de sondages de taille 100, d'autre part à 10 000 simulations de sondages de taille 1000 (dans une population où hommes et femmes sont en proportions égales). Ci-dessous, on trouve les résumés graphiques des valeurs de d=|f-0,5|, où f est le pourcentage de femmes. Pour les sondages de taille 100, 63% des valeurs de d sont supérieures à 0,2 contre12% pour les sondages de taille 1000 : on voit tout de suite que la notion du « même ordre de grandeur » mentionnée ci-dessus devra prendre en compte la taille de l'assemblée ; pour une assemblée de taille 100 par exemple, la graphique de gauche peut aider à fixer une borne δ pour d, au delà de laquelle on dira que la parité statistique n'est pas respectée ; en fait, en modélisant cette situation, on peut rendre ici le choix de δ indépendant de toute simulation et établir des formules mathématiques sur lesquelles fonder ce choix. On pourra aussi, par cette modélisation, établir un lien avec la notion déterministe de parité ; la probabilité qu'un choix de 2n personnes mène à n hommes et n femmes est approximativement $1/\sqrt{\pi n}$: la probabilité d'avoir 50 hommes parmi 100 (resp. 500 parmi 1000) sur une population où hommes et femmes sont équirépartis est 0,056 (resp0,018).



Enfin, les outils logiciels créent des communautés. Il y a en effet quelques grands logiciels dédiés à la statistique (S+, SAS, SPSS, Spad, Stata, etc). Ils sont d'utilisation conviviale, mais il faut du temps pour en devenir un utilisateur expérimenté et averti ; acquérir une expérience en statistique implique et la manipulation de concepts et la pratique d'un logiciel ; chaque entreprise a un ou deux logiciels privilégiés : c'est souvent à travers l'usage de ces derniers que se crée une culture statistique commune. Dans de nombreuses entreprises, la formation interne en statistique est d'ailleurs souvent une formation à un logiciel.
Pour décrire l'activité des statisticiens on peut ainsi envisager plusieurs grilles de classement :
- selon le champ disciplinaire (statistiques industrielles, économétrie, médecine, sondages, etc.).

- selon la nature des données qui sont traitées (séries de petites tailles ou grandes bases de données, données temporelles, données censurées, etc.), ou, ce qui revient un peu au même, selon les modèles employés, ou encore selon le logiciel dont ils ont une bonne maîtrise.

Signalons cependant une dérive liée à la qualité des logiciels de statistique : s'ils permettent à de nombreux utilisateurs non spécialisés de réaliser des études statistiques une fois les données produites, ils sont muets pour ce qui est de la réflexion fondamentale qui précède le recueil des données (ainsi, en contrôle de qualité, si les sources de variabilité sont mal identifiées, si les données sont entachées d'un biais, aucun usage rigoureux d'un logiciel ne mènera à des conclusions correctes). Or, on embauche de moins en moins de statisticiens et la fuite en avant vers l'usage de techniques statistiques évoluées et conceptuellement mal maîtrisées par l'utilisateur est à craindre. Parallèlement, la statistique a parfois tendance à être gommée en tant que telle, i.e. à ne plus être repérée et nommée. Des concepteurs de logiciels de statistique proposent des formations d'une journée à de tels logiciels où on ne prononce pas le terme de statistique (on y parle de modèle, de degré de confiance, de traitement de l'information, d'incertitude, de « data mining ») ; il sera alors difficile à l'utilisateur de savoir où chercher pour compléter sa formation et résoudre des problèmes non prévus dans le mode d'emploi.

III- L'aléatoire dans quelques champs disciplinaires

L'aléatoire se retrouve dans de nombreux champs disciplinaires et même dans la description de phénomènes pour lesquels on connaît des équations déterministes. Citons à ce propos David Ruelle dans *Le hasard aujourd'hui* , éditions du Seuil, collection Points sciences, 1991 :
« *Je préfère ne pas considérer le hasard comme une partie du monde physique, mais comme une partie de sa description. … Le chaos permet de comprendre comment le hasard s'introduit malgré des descriptions déterministes. En fait, ce qui se passe, c'est que les descriptions d'évolution sont bien déterministes, mais on n'a jamais une connaissance parfaite de l'état initial du système et, par conséquent, au bout d'un certain temps, on ne sait plus où on en est : on a , comme on dit alors, une situation de hasard ».*

Les probabilités interviennent en physique, non seulement en mécanique statistique, mais aussi dans la théorie des objets plus simples. En effet, à l'échelle atomique, les particules sont régies par la physique quantique, qui présente un caractère probabiliste irréductible : à cette échelle, on démontre qu'il est impossible de rendre compte des phénomènes observés à l'aide de variables cachées qui obéiraient à des lois déterministes. Ainsi, les probabilités sont essentielles à la théorie la plus fondamentale sur laquelle repose aujourd'hui tout l'édifice de la physique, et qui permet d'expliquer les propriétés les plus variées, de la radioactivité à la liaison chimique, du magnétisme au rayonnement solaire.
La définition même de nombreuses grandeurs de la physique est de nature statistique (par exemple, à l'échelle atomique la température s'interprète comme le paramètre d'une loi de probabilité exponentielle).
A propos de l'usage des probabilités en physique, citons un extrait de l'introduction de l'ouvrage de Roger Balian, (From microphysics to macrophysics : methods and applications of statistical physics, Springer Verlag, 1991) :
« *La plupart des quantités d'intérêt physique, accessibles à l'expérience et nécessaires aux applications sont macroscopiques : volume pression, température, capacité calorifique,*

viscosité, indice de réfraction, susceptibilité magnétique, résistivité, etc. Si on cherche à les calculer, à partir de propriétés microscopiques, on doit les rattacher à des moyennes statistiques sur l'ensemble des particules, dont les caractéristiques individuelles sont à la fois inaccessibles et sans intérêt. L'explication des propriétés macroscopiques à partir des constituants microscopiques nécessite donc l'emploi de concepts et méthodes probabilistes, même si les lois élémentaires sont parfaitement connues et même si la théorie microscopique sous-jacente était parfaitement déterministe. Ainsi, comme les prévisions de la mécanique quantique, mais pour des raisons entièrement différentes, les prévisions de la mécanique statistique seront de nature probabiliste et porteront sur des quantités moyennes.
Cependant, le fait même que les systèmes étudiés soient très grands, ce qui oblige à des traitements incomplets de nature statistique, a une contrepartie heureuse: les grandeurs physiques macroscopiques ont des écarts quadratiques moyens négligeables devant leur valeur moyenne et des prévisions exactes deviennent possibles ».

L'exemple simple suivant illustre l'idée importante en physique statistique que la prévisibilité d'une fonction peut être d'autant plus grande que le problème a un grand nombre de degrés de libertés : choisissons un point au hasard x dans la boule unité de R^n (n est ici le degré de liberté du problème). Pour $n>10^6$ (si on parle de particules, 10^6 est un nombre petit), la probabilité que la norme de x soit inférieure à 0,9999 est inférieure à 10^{-22}; en effet, cette probabilité est égale au volume de la sphère de rayon 0,9999 divisé par le volume de la sphère unité et vaut donc $0{,}9999^n$; on peut ainsi « prévoir » que la norme d'un point choisi au hasard dans la boule unité de R^n, pour n grand, vaut 1 avec une bonne précision (c'est en fait une autre manière de dire que le volume de la boule « se concentre » en bordure de la sphère).

La statistique propose, pour la biologie, des outils puissants d'aide à la décision, des techniques permettant d'optimiser les protocoles expérimentaux, des instruments visant à assurer au mieux l'objectivité des analyses.
Le « hasard » préside par ailleurs aux mutations génétiques et permet l'évolution des espèces : le titre du livre de Jacques Monod, paru aux éditions du Seuil en 1970, où sont exposés des fondements de la biologie est « Le hasard et la nécessité ». Le bref extrait ci-dessous illustre le titre et témoigne aussi du fait que les réticences à enseigner l'aléatoire ont été précédées de réticences à en accepter le pouvoir explicatif :
« *Beaucoup d'esprits distingués, aujourd'hui encore, paraissent ne pas pouvoir accepter ni même comprendre que d'une source de bruit la sélection ait pu, à elle seule, tirer toutes les musiques de la biosphère. La sélection opère en effet sur les produits du hasard, et ne peut s'alimenter ailleurs ; mais elle opère dans un domaine d'exigences rigoureuses dont le hasard est banni* ».
La variabilité biologique est par ailleurs une des manifestations de l'aléatoire à laquelle l'être humain est confronté dès son plus jeune âge. Il appartient à tout enseignement de biologie, dans l'enseignement secondaire en particulier, de rendre familier l'aléatoire, notamment pour des phénomènes continus, de faire prendre conscience de la fluctuation d'échantillonnage, d'initier à la problématique de la prévision, de travailler sur la différence entre prédictibilité et causalité, etc.

Enfin, la statistique concerne d'entrée de jeu toutes les disciplines expérimentales : en effet, un résultat de mesure n'a de sens que si on l'assortit d'une précision. Toutes les sciences expérimentales sont ainsi utilisatrices de calculs d'intervalles de confiance faisant le plus souvent intervenir des lois de Gauss.
Pendant très longtemps, la prise en considération des incertitudes de mesures consistait, en physique, dans l'enseignement secondaire, à envisager comment se propageaient numériquement les erreurs de mesures, ou plutôt des majorants de ces erreurs : l'erreur absolue d'une somme de mesures est la somme des erreurs absolues, l'erreur relative pour un produit est la somme des erreurs relatives et si n mesures $x_1,\ldots x_n$, sont entachées d'une erreur

ε, leur moyenne arithmétique est aussi entachée de l'erreur ε (alors pourquoi faire plusieurs mesures ?). On ne s'intéressait ainsi qu'à la propagation de l'incertitude, comme s'il s'agissait de calculs numériques (si on appuie sur la touche π d'une calculatrice, on obtient toujours le même résultat ; c'est une valeur numérique approchée de π et la précision des calculs utilisant cette valeur approchée obéit aux règles de calculs d'incertitude ci-dessus). Aujourd'hui, on commence à introduire au niveau du lycée l'idée que les erreurs de mesures sont aléatoires ; l'erreur, ou dispersion, est le plus souvent quantifiée par un écart-type ; la dispersion du résultat sur une moyenne arithmétique de n mesures indépendantes, faites dans les mêmes conditions, est égale à la dispersion initiale des mesures divisée par $\sqrt{n}$ (et c'est là tout l'intérêt de reproduire des mesures).

Par ailleurs, certaines expériences doivent être conçues pour mettre en évidence des phénomènes très rares susceptibles d'être cachés par un bruit aléatoire : il convient alors d'introduire des biais contrôlés et d'utiliser des procédures statistiques non classiques. Enfin, citons l'analyse de données expérimentales incomplètes, par exemple lors de la détermination de la structure d'une protéine à partir de la diffusion du rayonnement synchrotron dont on ne détecte que le module de l'onde, pas la phase : il est nécessaire d'employer des méthodes probabilistes pertinentes.

En économie, la statistique fait partie du langage et est constitutive d'un mode de pensée de certains secteurs économiques : la consommation des ménages, l'évolution du chômage, la pyramide des âges, etc. Nicolas Bouleau, dans « Philosophie des mathématiques et de la modélisation » (éditions l'Harmattan) suggère même de considérer les modèles (dont les modèles aléatoires) comme des récits symboliques : on conçoit dès lors que plusieurs modèles, ne se déduisant pas les uns des autres, coexistent pour appréhender une même réalité.

Dans un premier temps, au niveau de l'enseignement secondaire, l'enseignement de l'économie pourrait prendre en charge la réflexion sur la nature de données qu'elle emploie, la difficulté à définir des indicateurs pertinents pour étudier la réalité économique ; par exemple une réflexion peut être menée en même temps en mathématiques et en économie sur l'élaboration des tables de mortalité et les calculs d'espérance de vie à tous les âges (pourquoi faut-il définir une cohorte fictive pour dresser les tables de mortalité, pourquoi l'espérance de vie d'un professeur est-elle supérieure à celle d'un individu pris au hasard dans la population?). Ce travail sur la définition d'indices pertinents et leur modes de calculs possibles est un élément fondamental de la formation en statistique.

L'économie est aussi un lieu où apprendre à confronter les données entre elles pour mieux les comprendre et les intégrer dans un environnement riche ; par exemple à partir des données telles que la taille des principaux cheptels bovins -215 millions de têtes en Inde, 163 au Brésil, 107 en Chine, 99 aux Etats-Unis, 20 en France, certains, pour les intégrer à leur univers, rapporteront au nombre d'habitants, d'autres à la surface du pays, d'autres en classant les pays suivant l'état de leur développement. Un tel travail mobilise des données dont seul l'ordre de grandeur importe ; dans cette confrontation des données, la connaissance des ordres de grandeur est capitale pour se repérer dans le magma des données chiffrées, et pour tout débat argumenté.

De nombreuses disciplines développent leurs outils de statistique descriptive. Ainsi les géographes ont des représentations particulières où les mois de l'année sont les rayons du cercle unité d'angles polaires $k\pi/6$, k=1,..,12 ; il appartiendrait en fait à chaque discipline de développer et de justifier ses modes spécifiques de représentation.

IV- Différents temps et lieux de formation

Ce paragraphe a pour objectif d'expliciter la nature des acquis de l'enseignement secondaire qui sont nécessaires à des pousuites d'études en statistique.

1- Les formations professionnelles en entreprise

Les stages internes de formation de techniciens par les ingénieurs des entreprises sont souvent très appréciés ; l'enjeu est clair : les stagiaires devront tout de suite mettre en oeuvre ce qu'ils ont acquis. Nous insistons ici sur ce type de formation car elle est souvent méconnue ; elle fait moins l'objet de rapports publics que les autres. Et surtout, elle présente des caractéristiques très spécifiques, notamment au niveau du recours aux mathématiques et au calcul.

Les considérations qui fondent la démarche statistique que les stagiaires auront à mettre en œuvre peuvent être résumées par quelques assertions :
- un chiffre statistique est toujours entaché d'incertitude,
- en prenant une décision à l'aide de ce chiffre, on prend des risques,
- la dispersion a souvent plusieurs origines,
- l'esprit de la démarche statistique, c'est réfléchir avant de collecter les données (et après aussi bien sûr),
- on ne doit pas rajouter, au niveau de la preuve statistique, des jugements ou des appréciations externes à l'étude.

Les principes de ces formations professionnelles peuvent être décrits ainsi :
- partir de la pratique des participants et de leurs questions ; souvent des outils "simples" permettent de les résoudre (le fait de travailler toujours avec le même type de métiers aide).
- introduire les notions en s'affranchissant au maximum du formalisme mathématique, afin de concentrer l'attention sur le raisonnement statistique et non sur les difficultés mathématiques (pas d'équation de la loi de Gauss, seulement des dessins).
- faire comprendre « avec les mains » et par l'exemple quelques méthodes statistiques, en insistant sur les conditions dans lesquelles on peut les utiliser, et sur celles où on ne peut pas.
- aller jusqu'au bout des calculs sur des exemples numériques très simples.
- montrer la nécessité d'aller jusqu'au bout de l'interprétation des résultats.
- montrer comment les outils présentés permettent de répondre aux questions posées, mais aussi d'aller plus loin (par exemple : déterminer le nombres d'essais à faire).
- faire réfléchir aux "pièges" de mise en œuvre à l'aide d'exemples classiques.
- montrer ce que peuvent apporter des méthodes plus compliquées, même si elles nécessitent l'intervention de spécialistes.
- signaler que, même si on ne les donne pas, il existe des justifications théoriques très rigoureuses sous-jacentes aux méthodes présentées : on n'invente pas soi-même des outils statistiques.

Si on part de la pratique des participants, dans leur domaine, on peut arriver à faire sentir simplement de nombreuses notions, et former aux traitements statistiques simples. La pratique professionnelle aide à formuler les questions statistiques et à introduire la nécessité d'une

démarche rationnelle. Il est remarquable que, dans ce cadre, les formateurs ont peu de difficultés à faire passer les notions de « population de référence » et de « valeur théorique » : c'est si on veut en donner un formalisme mathématique que cela devient très difficile, voire infaisable.

Les obstacles psychologiques auxquels sont confrontés les formateurs lors de ces stages sont essentiellement liés aux compétences, aux goûts, aux expériences passées des stagiaires et aux rumeurs. Ainsi, ceux qui n'aiment pas les mathématiques croient que la statistique, c'est des maths, et qu'on va les bombarder de démonstrations : il convient de calmer leurs appréhensions et montrer que la démarche statistique n'est pas uniquement de nature mathématique, loin s'en faut. A l'inverse, les "esprit matheux" ont du mal à accepter d'utiliser des méthodes sans tout savoir des justifications théoriques, à simplifier les problèmes pour pouvoir les résoudre (par exemple : utiliser une loi normale ou log- normale même si elle est à la limite de ce qu'on peut accepter, car c'est avec elle que l'on sait résoudre le problème). Enfin, certains croient que faire des statistiques, c'est cliquer sur des cases d'un écran d'ordinateur : ceux là doivent être convaincus que c'est un sujet où la réflexion est nécessaire.

2- Les formations en école d'ingénieurs

Il semble que les ingénieurs français ne soient pas assez formés en statistique. Pour beaucoup d'entre eux, ce domaine n'est vraiment abordé qu'à la fin de leurs études : s'ils en gardent un « vernis », cela arrive vraiment trop tard pour être intégré à leur personnalité intellectuelle.

On notera que la formation d'ingénieurs dans des écoles plus spécialisées en statistique ne peut ni mettre en sourdine les soubassements conceptuels, ni reculer le temps de l'exercice d'un mode de pensée : il convient d'enseigner les deux ensemble, et d'arriver à établir une continuité. Ainsi, à l'ENSAE (école nationale de la statistique appliquée à l'économie) qui forme entre autre une partie des cadres de l'INSEE, la part de la formation aux pures techniques des sondage est faible : on dispense essentiellement une formation au traitement de l'information, aux techniques d'observation, à la représentation, à la modélisation.

3-Enseignement supérieur

Dans l'enseignement supérieur, l'étudiant se concentre sur un petit nombre de champs disciplinaires, son orientation professionnelle se précise et la réalisation de projets où interviennent les probabilités et la statistique va lui permettre à la fois d'avancer en statistique et de préciser ses choix professionnels. De plus, l'étudiant peut avoir accès à des logiciels plus spécifiquement dédiés à la statistique ainsi qu' à de nombreuses bases de données, régulièrement mises à jour.

Si le temps de collège et du lycée est pour l'aléatoire celui de l'apprentissage d'un langage élémentaire et l'ouverture à un mode de pensée, l'enseignement post-baccalauréat est celui de l'approfondissement. Dans de nombreux cursus, la statistique est le seul exemple d'apprentissage d'une discipline interprétative et c'est là une des difficultés majeures de son enseignement. C'est un apprentissage long et difficile - pendant un certain temps, les étudiants ne comprennent pas ce qu'on attend d'eux ; cet inconfort s'ajoutant à une vision souvent à court terme (il faut comprendre, non pour être professionnellement compétent, mais pour avoir une bonne note à l'examen) rend le rôle des enseignants ardu - et parfois décourageant.

Aux extrêmes des formations post-baccalauréat, on peut distinguer :

-les maîtrises de sciences, où le repli vers l'enseignement des seuls outils mathématiques est rassurant, tant pour l'enseignant que pour l'étudiant. Néanmoins, pour que les étudiants acceptent d'articuler concepts et données d'observation (ce qui est encore embryonnaire aujourd'hui), il serait particulièrement utile que l'initiation lors de l'enseignement secondaire ait montré comment des questions sont susceptibles d'être reformulées avec la langage de la statistique.

-l'enseignement de la statistique pour des non scientifiques : il s'agit dans ce cas de mettre complètement en sourdine la référence aux mathématiques, et de se centrer sur la démarche, l'interprétation. Mais pour dépasser la collection de recettes, il convient de pouvoir trouver quelques appuis théoriques venant de l'enseignement secondaire.

-un cas particulier : la formation en statistique des médecins ; c'est une situation assez à part du fait que la statistique y a sa place depuis plus longtemps que dans les autres cursus ; elle comporte une formation initiale indispensable pour comprendre les bases de la médecine actuelle (la place que l'on accordait au début du 20ème siècle à « la leçon de nos maîtres » doit maintenant être partagée avec la formation en statistique) ; elle est sérieusement complétée pour les médecins hospitaliers qui s'engagent dans des recherches cliniques.

L'exemple suivant montre l'utilité pour tous les médecins d'avoir une formation suffisante pour appréhender l'efficacité des formules issues du calcul des probabilités.
Le fabriquant d'un test de diagnostic T d'une maladie M fournit les caractéristiques suivantes :
La probabilité qu'un individu malade ait un test positif est s=0,99.
La probabilité qu'un individu non malade ait un test négatif est s'=0,99.
On suppose que ce test est fait systématiquement chez tous les sujets d'une population contenant une proportion p d'individus atteints de la maladie M.
Désignons par V la valeur diagnostique du test dans cette population, où V est la probabilité qu'un individu de cette population, dont le test est positif, soit malade. Une formule simple donne, dans tous les cas, la valeur diagnostique du test :

$$V = sp / (sp+(1-s')(1-p))$$

Pour les valeurs de s et s' considérées , V=99p / (98p+1) ; le tableau suivant montre alors la différence entre un dépistage systématique d'une maladie rare (p<0,001) et le dépistage dans un sous-groupes à risque (p>0,1).

p	0,001	0,010	0,100
V	0,090	0,500	0,912

4-L'enseignement au collège et au lycée

L'objectif d'une initiation aux probabilités et à la statistique au niveau collège et lycée est d'enrichir le langage, de repérer les questions de nature statistique, de définir des concepts qui fonderont un mode de pensée pertinent, rassurant, remarquablement efficace. Les modes de représentation graphiques usuels (histogrammes, diagrammes en bâtons notamment), c'est à dire les éléments de base du langage graphique de la statistique sont aujourd'hui enseignés en collège et une introduction à l'aléatoire, appuyée sur le calcul des probabilités et la simulation est proposée dans les nouveaux programmes de lycée.
Nous décrivons ci-dessus quelques unes de rencontres les plus usuelles des enfants avec l'aléatoire, sur lesquelles un enseignement pourra s'appuyer.

- les jeux de hasard

Pour les lancers de dés, l'aléatoire peut être relatif à un nombre fini de lancers, et de nombreux calculs sont accessibles au niveau de l'enseignement secondaire ; mais il peut aussi s'agir d'expériences du type « tant que le six n'est pas sorti » : là se forgent des intuitions pour lesquels l'élève ne dispose pas des concepts et du vocabulaire qui permettrait une

formulation « juste ». La simulation serait un moyen d'aborder ce sujet avant d'avoir les moyens théoriques de le traiter.

Par exemple de nombreux jeunes (et moins jeunes) pensent qu'en répétant 1000 fois une expérience pour laquelle le succès a une chance sur 1000 d'arriver, elle « arrivera forcément »; cette image mentale n'est en général pas démentie par l'expérience ne serait-ce que parce qu'il est presque impossible de refaire 1000 fois la même. Si ce « arrivera forcément » est faux, cette intuition accorde à juste titre à l'inverse de la probabilité de succès une place particulière : c'est l'espérance du rang du premier succès.

- les phénomènes biologiques, par exemple la croissance : la variabilité des vitesses de croissance et de l'état final ; il s'agit ici de phénomènes continus, et l'imprévisibilité n'est pas totale (la taille d'un adulte est inférieure à 3 mètres),
- les attentes (d'un autobus, à un guichet),
- les coïncidences : chacun dans son histoire est l'objet de coïncidences et on peut parfois essayer d'en apprécier le caractère exceptionnel et non reproductible,
- les collections d'images et la recherche d'une série complète d'images par achat de suffisamment de boites de céréales ou de plaques de chocolat,
- le temps (au sens météorologique),
- le risque.

De telles rencontres ne sont pas toujours reconnues comme pouvant être pensées en terme d'aléatoire ou d'imprévisibilité : le seul fait de le reconnaître, d'envisager pour certaines quelques modèles simplifiés, la possibilité de les simuler, favoriserait l'élaboration d'images mentales variées à partir desquelles pourront s'élaborer des concepts théoriques.

V. La formation des professeurs

La formation des enseignants des lycées et collèges actuellement en poste est aujourd'hui un problème clé en ce qui concerne la formation citoyenne à l'aléatoire. Quelles sont les spécificités d'une telle formation ?

-elle concerne pour les mathématiques environ 40 000 personnes (professeurs de collège et de lycée) ; étant donné le caractère nécessairement réparti d'un enseignement de la statistique, il concerne aussi les autres disciplines. Il n'y a aucune raison de cloisonner les formations ; il conviendrait donc de former ensemble les enseignants de mathématiques, de physique, de biologie et de sciences économiques et sociales, soit un public d'environ 60 000 personnes.

-cette formation est à créer ; elle doit être construite en collaboration avec les enseignants, comme a été créé l'enseignement de la statistique aux médecins, en s'appuyant pour une très large part sur les médecins eux-mêmes.

- une telle formation doit comporter des éléments théoriques, et elle doit aussi confronter les enseignants aux traitements de données réelles. Elle doit intégrer et utiliser la simulation aléatoire. Comme toutes les formations, sa qualité dépend de la pertinence des exemples traités, des questions posées, et bien que ce soit une formation professionnelle, les questions les plus motivantes ou éclairantes ne sont pas pour la plupart issues de la pratique professionnelle d'enseignant.

-pour monter des dispositifs de formation, il convient de faire participer les professeurs de l'enseignement secondaire connaissant bien la statistique, des universitaires de différentes disciplines, mais aussi des professionnels d'institutions tels l'INSEE, l'INRA, l'INED, des enseignants des écoles d'ingénieurs, des médecins etc.

-de nombreux enseignants de mathématiques se trouvent pour la première fois de leur carrière devoir enseigner un domaine non abordé dans leurs études, et qui fait intervenir un mode de pensée nouveau. On peut espérer que ceux qui reprendront à leur compte la notion de « bonne fortune » décrite il y a 100 ans par l'entomologiste Henri Fabre dans ses souvenirs entraîneront dans leur sillage leurs collègues :
« Une racine carrée à extraire, la surface de la sphère à évaluer avec démonstration étaient pour moi les points culminants de la science. Le terrible logarithme, lorsque par hasard j'en ouvrais une table, me donnait le vertige, avec son amoncellement de nombres ; certaine frayeur, mêlée de respect, me prenait rien que sur le seuil de cette caverne à calculs. De l'algèbre, aucune notion. J'en savais le nom et sous ce vocable tourbillonnait en ma pauvre cervelle la cohue de l'abstrus…..La bonne fortune me valut la première leçon d'algèbre, leçon donnée et non reçue, cela va de soi ».

Il convient par ailleurs que l'aléatoire fasse partie de la formation initiale des enseignants. Certains points notés ci-dessus concernent aussi cette formation.
Des épreuves spécifiques de probabilités et statistiques devraient devenir obligatoires dans les concours d'enseignement, pas seulement en mathématiques. La statistique étant une discipline d'interprétation il n'est pas facile d'en faire des épreuves d'examens et de concours. Néanmoins, le succès de l'épreuve de modélisation à l'agrégation de mathématiques indique des voies possibles (cf. « Mathématiques en situation", n°11 collection SCOPOS, Springer, 2000).

En guise de conclusion

Définir les grandes lignes de ce qu'on pourrait appeler la formation citoyenne en probabilité et statistique, préciser les éléments d'une formation initiale et continue des enseignants ayant en charge l'enseignement de cette « statistique du citoyen » est un vaste chantier. La France a comme atout d'être un pays où existe une école mathématique forte, un pays de tradition mathématique ; pour l'enseignement secondaire, il y a aujourd'hui un large consensus pour ancrer la formation à l'aléatoire en mathématiques, sans pour autant, bien au contraire, que cette discipline en fasse une chasse gardée. Dans ce cadre là, il ne s'agira pas de remplacer des chapitres classiquement enseignés par des probabilités et de la statistique mais plutôt, prenant l'aléatoire comme l'un des fils conducteurs d'un parcours de formation, de se promener à travers différents chapitres des mathématiques.

L'expérience de cette dernière décennie montre que travailler sur l'aléatoire après le baccalauréat, sans une première approche dans l'enseignement secondaire, est très difficile. Pour des raisons diverses suivant les filières, il convient que l'enseignement post-bac puisse prendre un véritable appui sur ce qui est fait dans l'enseignement secondaire.

A travers la modélisation stochastique se forge pour toutes les disciplines l'expérience de la modélisation, avec la nécessité parfois douloureuse d'adapter le modèle au données et non l'inverse, avec ses règles de simplicité et de parcimonie, la prise en compte de contraintes matérielles (à coté des « grands modèles » susceptibles de lever un coin du voile des mystères de l'univers, on a aujourd'hui grand besoin de modèles à élaborer rapidement et dont la durée de vie est limitée). La formation à l'aléatoire est en fait exemplaire d'un enseignement de « sciences mathématiques ».

Enfin, rappelons les deuxième et neuvième constats et recommandations du rapport de l'Académie :
Constat 2 : La recherche statistique est insuffisante en France si on la compare à celle des autres pays de même niveau technologique. Le nombre de chercheurs en statistique en France devrait être multiplié par un facteur minimal de 5 pour aboutir à un niveau relatif d'activité comparable à celui des Etats-Unis.
Recommandation 2 : Le groupe recommande un effort à long terme pour rattraper le retard de développement de la France en statistique. L'action doit porter sur le recrutement et la formation des enseignants-chercheurs et des chercheurs en statistique.

Constat 9 : En France, à la différence d'autres pays européens, les citoyens n'ont pas une formation suffisante à la prise en compte du mode de pensée statistique. Pour améliorer cette situation, des initiatives récentes ont été prises dans le cadre d'une réforme de l'apprentissage des mathématiques dans l'enseignement primaire et secondaire.
Recommandation 9 : Le groupe souligne l'opportunité de ces réformes et encourage les responsables de l'enseignement à en assurer la mise en œuvre, en particulier par un effort de formation initiale et continue des professeurs des lycées et collèges.

Pour établir ce rapport, la commission s'est mise en contact avec :

Anestis Antoniadis Professeur à l'université Joseph-Fourier, Grenoble,
Nicolas Bouleau, Professeur à l'école nationale des Ponts et Chaussées
Yves Escoufier, Professeur à l'université de Montpellier II
Georges Oppenheim, Professeur à l'université de Paris XI
Alain Trognon, directeur de l'ENSAE
Annie Uhry, Ingénieur statisticien Statistique et Productique Industrielles Péchiney Centre de Recherches de Voreppe.
Vidal Cohen, Professeur à l'école nationale des Ponts et Chaussées
Marc Yor , Professeur à l'université de Paris VI