

Dossier : Statistique et probabilités

Claudine Robert, présidente du Groupe Technique Disciplinaire de mathématiques (c'est-à-dire du groupe chargé d'écrire les nouveaux programmes), nous a fait parvenir, à notre demande, le texte suivant, où elle explicite ses choix en matière d'enseignement de la statistique. Qu'elle en soit remerciée.

Par cette publication, nous souhaitons contribuer à votre information : il ne s'agit pas ici de clore un débat mais bien d'alimenter celui qui a été amorcé dans le BGV « Spécial rentrée ». Notre Association se veut un lieu de libre parole, où chacun trouve place pour s'exprimer, et c'est l'expression des uns et des autres qui fait avancer la réflexion. Nous attendons donc vos réactions, à travers les divers moyens d'expression qui sont à votre disposition : réunions régionales, nationales, propositions d'articles, courrier postal ou électronique,

À propos de l'introduction de l'enseignement de la statistique dans les lycées

Claudine Robert

Un groupe (le Groupe Technique Disciplinaire de mathématiques) chargé d'écrire les nouveaux programmes de mathématiques des lycées d'enseignement général, avec notamment la demande d'adapter les programmes en statistique, a été mis en place en janvier 1999.

Les quelques lignes ci-dessous sont destinées à expliciter certaines des motivations qui nous ont conduit à ne pas nous limiter au traitement numérique ou graphique de l'information chiffrée, mais à introduire dans les programmes de mathématiques les premières éléments d'une démarche statistique.

Il y avait, jusqu'à présent, dans les programmes de Seconde, de Première et Terminale L et ES, un chapitre dont le titre était *statistique*. L'esprit de ces chapitres est celui « des statistiques » et non de « la statistique » et témoigne d'une époque où stocker un grand nombre de données était réservé aux instituts spécialisés. Dans le cadre de ce programme et avec le relais des manuels, s'est développée une statistique propre à l'enseignement secondaire et qui s'est peu à peu dissociée de celle que pratiquent les analystes et ingénieurs statisticiens (ainsi, dans de nombreux manuels, la médiane d'une série de données est calculée à partir d'une interpolation linéaire de la fonction de répartition, ce que les statisticiens ne font jamais).

L'esprit statistique naît lorsque l'on prend conscience de la fluctuation d'échantillonnage, et donc de la variabilité de la moyenne empirique et autres paramètres résumant une série. Il n'en était jusqu'à présent pas fait mention dans les programmes. La pratique ainsi induite par les anciens programmes et les manuels correspondants constitue à mon avis un réel barrage à la compréhension de la statistique, ce que la plupart des enseignants ont d'ailleurs fortement ressenti.

L'objectif du groupe qui compose les programmes n'est pas d'augmenter systématiquement le temps consacré à la statistique, mais par contre de poser les bases d'une statistique plus moderne ; c'est pourquoi nous avons volontairement choisi d'initialiser ce changement en mettant, dès la seconde, l'accent sur l'expérimentation avec des dés ou des tirages de boules dans des urnes, et la simulation aléatoire sur calculatrice de poche pour appréhender la notion de fluctuation d'échantillonnage.

Le programme que nous proposons est sans doute déroutant pour un corps professoral compétent, mais qui, dans son ensemble, n'a jamais fait de statistique, ou alors en annexe d'un cours de probabilité. Les enseignants de mathématiques devront se former dans un domaine qu'ils n'ont en général pas travaillé lors de leurs études. Certains auraient souhaité attendre encore quelques années afin notamment que les enseignants se forment ; mais comment les professeurs peuvent-ils se former et avoir simultanément une pratique enseignante qui, si elle est conforme aux programmes actuels, sera en franche contradiction avec ce qu'ils apprendront ? Nous pensons au contraire que l'enseignement de rudiments de la statistique les aidera à acquérir peu à peu des connaissances plus profondes dans ce domaine. Leurs premiers pas dans cet enseignement seront sans doute hésitants, mais cela nous paraît préférable à la situation actuelle. De plus, des formations seront mises en place pour lesquelles nous avons de l'aide de la communauté des statisticiens.

On trouvera en annexe les grandes lignes du programme sur lequel nous travaillons actuellement. Ce programme s'apparente à ce qui se fait et va se faire en Angleterre, pays qui a une longue et riche tradition dans ce domaine. Mais pourquoi, en ce qui concerne l'enseignement des statistiques au lycée, ne pourrait-on pas avoir l'ambition de faire au moins aussi bien que nos voisins d'Outre-Manche ?

ANNEXE

Voici quelles sont les grandes lignes que nous envisageons pour les nouveaux programmes du lycée en statistique. Ces nouveaux programmes entreront en vigueur en septembre 2000, 2001 et 2002 respectivement pour les classes de Seconde, Première et Terminale.

Rappelons qu'au collège les élèves ont travaillé sur la moyenne arithmétique d'une série et le langage graphique (histogrammes, diagrammes en bâtons, camemberts) et apprennent en technologie et en mathématiques à se servir d'un tableur.

En seconde :

– **Introduction de la fluctuation d'échantillonnage.** Cette partie est insuffisamment détaillée dans la rédaction actuelle du programme, mais voici ce dont il s'agit :

– réalisation effective, par les élèves, de lancers de pièces ou de dés et de tirages de boules dans une urne (nous espérons que chaque lycée aura à terme un matériel pédagogique adapté pour cela) ; observation de la variabilité des séries de résultats. Mais l'expérimentation et l'observation ne peuvent se transformer en connaissance que lorsqu'on a des mots et des concepts pour réfléchir, analyser et comprendre ; c'est pourquoi, on introduit la notion de distribution de fréquences qui nous paraît fondamentale. Ce sont les variations de cette dernière que l'on observe ; on montrera ensuite que la moyenne, l'étendue s'en déduisent et sont par conséquent eux aussi fluctuants.

– utilisation de simulation de la loi uniforme sur l'ensemble des chiffres (touche *random* des calculatrices pour les élèves, logiciels pour les enseignants) ; il s'agit en premier lieu d'appréhender ce que signifie simuler une expérience aléatoire (sans disposer du concept de probabilité) ; ensuite, grâce à la simulation, on pourra observer à grande

échelle et ainsi expérimenter que l'ampleur de la fluctuation de la distribution des fréquences diminue quand le nombre de simulations augmente. La quantification de cette diminution pourra être approfondie dans l'un des thèmes facultatifs de ce nouveau programme et qui concerne la notion de fourchette de sondage.

Ce travail prépare bien sûr au concept de probabilité, mais ce n'est cependant pas là le but principal. L'objectif est avant tout de faire sentir et vivre la notion de distribution de fréquences et de fluctuation d'échantillonnage ; les distributions des fréquences, les moyennes et médianes empiriques et leurs fluctuations sont des objets mentaux qu'il convient d'observer et d'étudier d'une part pour comprendre ce qu'est une question dans le champ de la statistique, d'autre part pour comprendre le rôle et la fonction d'un modèle probabiliste.

On pourra consulter à partir du mois de novembre le site WEB du GTD pour plus de précisions sur des activités en seconde à ce sujet.

- **Travail pratique sur des séries de données.**

La mesure de la dispersion retenue est l'étendue. Signalons que cette mesure, bien que grossière, figure sur toutes les cartes de contrôle industriel et est systématiquement calculée par tous les grands logiciels de statistique. Enfin, cette mesure intervient pour l'estimation d'un écart-type théorique lorsqu'on a moins de 10 données (et cela arrive aussi bien dans l'industrie pour des expériences coûteuses qu'en médecine pour des résultats d'examen invasifs ou concernant des maladies rares).

Et puis, n'est-il pas plus parlant de résumer une petite série de notes d'un élève ou d'une classe par l'étendue plutôt que par l'écart-type ? En fait, en dehors d'un ordre de grandeur de référence ou de la connaissance du caractère gaussien des données, l'écart-type est un paramètre peu interprétable.

Les programmes de Première et Terminale ne sont pas déterminés et dépendront du résultat des expérimentations qui vont se faire chaque année dans les classes (le programme de Seconde sera expérimenté dès cette année dans une cinquantaine de classes).

Néanmoins, les idées directrices sont actuellement les suivantes (elles seront déclinaées différemment suivant les sections).

- **Recueil de données** ; résumé de ces données soit à l'aide du couple **moyenne, écart-type**, soit par un **diagramme en boîte**.

On pourra travailler sur des données classiques (courbes des tailles des carnets de santé des enfants par exemple) ou sur des données en petit nombre que les élèves recueillent eux-mêmes sans que cela leur prenne trop de temps (pouls, durée des coups de téléphone ou du temps d'attente à une caisse d'une grande surface, poids des cartables, appréciation de longueurs, etc.). On se centrera sur la question qui motive le choix de l'étude et le lien avec les données recueillies, le mode de recueil de ces données et les problèmes que cela pose, le traitement statistique que l'on pourra en faire pour apporter des éléments de réponse ; en conclusion de telles études, on posera clairement la question du sens de certaines différences (i.e. on indiquera que la statistique donne des moyens de comparer des différences à celles qui sont le fait de la fluctuation d'échantillonnage) et comment pourrait se généraliser l'étude faite.

La statistique donne lieu à des activités numériques, mais celles-ci ne doivent être développées que dans la mesure où elles permettent aux élèves de mieux comprendre la nature des objets qu'ils manipulent. Ainsi, estimer des moyennes de séries de données en les regroupant par classes 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 estimer vite l'influence d'une donnée aberrante sur la valeur d'une moyenne, comprendre la linéarité de celle-ci sont des exercices pertinents au regard de la pratique de la statistique. De même, en section scientifique, on justifiera l'association (moyenne, écart-type) plutôt que (moyenne, moyenne des valeurs absolues des écarts à la moyenne) : on travaille ainsi sur le minimum d'un polynôme du second degré et sur celui d'une fonction affine par morceaux et ces calculs sont précieux pour la compréhension de ce qu'est un écart-type.

Une attention particulière sera d'ailleurs portée à la variance qui est une notion difficile à interpréter. On évitera son usage systématique et on l'introduira pour des données *gaussiennes* qui concernent les domaines de la biologie et de la médecine, la production industrielle et les erreurs de mesure (on notera que la notion de données *gaussiennes* est implicite dans les programmes actuels et dans les exercices où on demande de « vérifier » que le pourcentage des données en dehors de l'intervalle $[m-2\sigma ; m+2\sigma]$ est environ 5%).

En physique, dès la seconde, les élèves ont vu qu'il valait mieux utiliser la moyenne de plusieurs mesures plutôt qu'une seule. On pourra leur expliquer une raison de cette pratique à travers l'observation de données réelles ou

simulées illustrant que l'écart-type des moyennes de séries de n données d'écart-type σ est de l'ordre de σ/\sqrt{n} .

– **Tableaux croisés.**

Analyse de tableaux croisés portant sur un grand nombre de données. Interprétation des marges, construction du tableau des pourcentages associés, par ligne et par colonne.

Test du khi-deux pour des tableaux (2,2) (il s'agit ici de faire comprendre l'esprit du test en se référant aux résultats de simulations de tirages de boules dans des urnes).

– **Les probabilités** seront introduites et travaillées en Première ou Terminale de certaines sections et apparaîtront comme des « distributions théoriques » de fréquences. De même qu'on fait manipuler aux enfants des groupes de n objets avant d'introduire les entiers naturels, de même les élèves auront manipulé les distributions de fréquence avant qu'on leur parle de probabilité. Les lois de probabilité seront clairement présentées comme des objets du monde mathématique permettant de modéliser les situations aléatoires.

Actuellement, on introduit la probabilité d'un événement comme « la » fréquence limite d'un nombre infini d'expériences en mélangeant ce qui relève d'une loi empirique (diminution de la variabilité de la fréquence avec l'accroissement de la taille des échantillons) et ce qui relève d'un théorème de mathématique (loi forte des grands nombres). Une conséquence implicite de la présentation actuelle est qu'il n'existerait qu'un seul modèle relatif à une série nécessairement finie de données expérimentales, modèle déterminé par passage à « la » limite. Cette idée d'unicité du modèle est pour beaucoup un obstacle à la compréhension de la statistique et des probabilités (avec cette idée d'unicité de la probabilité associée à une expérience aléatoire, pour conclure qu'un traitement A d'une maladie est plus efficace qu'un traitement B, il suffirait de calculer la probabilité p de guérison par le traitement A et de regarder si elle est supérieure à la probabilité p' de guérison par B ; en pratique, on est amené à travailler sur l'ensemble des modèles « compatibles » avec les données expérimentales dont on dispose).

Dans les programmes actuels, une **variable aléatoire** est « une grandeur numérique associée à une expérience et pour laquelle on peut affecter une probabilité pour chaque valeur prise ». Cela occulte le rôle fondamental de la notion de variable aléatoire (au niveau du lycée, c'est même le seul rôle) qui est celui de transporteur de loi de probabilité ; en particulier, lorsqu'on a choisi un modèle (c'est-à-dire une loi de probabilité sur un ensemble fini), la

loi de la variable aléatoire relève de calculs théoriques et non plus d'estimations ou d'affectations subjectives.

Les schémas expérimentaux, type tirages de boules ou de cartes, lancers de pièces ou de dés, auront le statut d'expériences de référence ; le travail sur de telles expériences de référence permettra d'acquérir un certain sens de l'aléatoire. Au niveau du lycée, un modèle associé à une expérience aléatoire sera toujours une loi de probabilité sur un ensemble fini. On insistera sur le fait qu'en général, un modèle doit être estimé, puis validé avant d'être utilisé pour expliquer et prévoir, sans pour autant détailler les techniques et théorèmes utilisés pour cela. Pour ce qui est des expériences de référence, on admettra que le modèle de l'équiprobabilité, naturel à comprendre à l'aide d'arguments de symétrie, a été validé.

La notion de fonction de répartition empirique (c'est à dire calculée sur une série de données) ou de fonction de répartition d'une loi de probabilité disparaîtra des programmes. Cette notion est utile pour des lois sur des ensembles infinis, mais, au niveau des lycées, l'usage de diagrammes en boîtes est suffisant.

Les programmes actuels de la section ES ont dans le chapitre statistique un paragraphe faisant référence à la régression linéaire. Cela donne trop souvent lieu à des études surprenantes où la note en mathématiques au bac devient directement une fonction de la note en physique, où on définit un modèle dans lequel le poids d'un individu est une fonction affine de sa taille (ce qui ne saurait aider les élèves à comprendre la notion de fonction ou ce qu'est un modèle). On apprend de plus qu'il est légitime de faire un ajustement linéaire des que le coefficient de corrélation linéaire empirique est supérieur à 0,8 ou 0,9 suivant les ouvrages, et ceci indépendamment du nombre de données observées, ce qui n'a pas de sens. La droite des moindres carrés sera sans doute maintenue dans les programmes (mais pas dans le cadre du chapitre statistique) et utilisée pour faire de l'interpolation et de l'extrapolation linéaires sur un intervalle bien défini, notamment pour des données chronologiques.

Les quelques lignes ci-dessus ne constituent pas un programme à proprement parler et ne sont là que pour éclairer l'état d'esprit dans lequel le travail sera fait et en quoi il y a changement. Notons aussi un dernier changement : jusqu'à présent, l'enseignement en section S ne comportait pas de statistique ; ce ne sera plus le cas, sans toutefois, bien sûr, que ce sujet soit aussi central qu'en section ES.