Les cafés de la statistique

"La statistique éclaire-t-elle les questions de société" ?

Soirée du 12 décembre 2006 :

« Enseigner la statistique dans le secondaire »

Dossier préparatoire

[*] pour les premières et terminales, la situation est plus complexe, à cause de la pluralité de filières et options : nous avons choisi de mettre ici le programme de Première L, qui est *a priori* la section dont les élèves auront le moins un usage professionnel de la statistique et qui marque donc mieux ce que serait une initiation pour la population en général.

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

Claudine ROBERT
(*Présidente du groupe technique disciplinaire en mathématiques*)

SMF – Gazette – 82, octobre 1999

Un groupe chargé d'écrire les nouveaux programmes de mathématiques des lycées d'enseignement général, avec commande institutionnelle d'introduire l'enseignement de la statistique, a été mis en place en janvier 1999.

Les quelques lignes ci-dessous sont destinées à donner des informations sur le vif débat que soulève l'introduction de la statistique dans l'enseignement secondaire ; schématiquement, ce débat oppose les spécialistes de l'enseignement des probabilités et statistiques tel qu'il se pratique actuellement dans les lycées et les praticiens ou les enseignants-chercheurs en statistique.

Il y avait, jusqu'à présent, dans l'enseignement secondaire français, quelques chapitres de mathématiques dont le titre était statistique. L'esprit de ces chapitres est bien celui des statistiques et non de la statistique et témoigne de l'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 scolaires s'est ainsi développée une pratique de la statistique propre à l'enseignement secondaire et qui s'est peu à peu dissociée de celle que pratiquent les statisticiens (qu'ils soient enseignants chercheurs ou analystes).

Sans entrer dans les détails, jusqu'à la terminale, il n'était jamais fait mention de la notion de fluctuation d'échantillonnage (ou même simplement de variabilité de la moyenne empirique pour des séries de données aléatoires). La pratique induite par ce programme et les manuels correspondants constituent à mon avis un réel barrage à la compréhension de la statistique, ne serait ce que par les maladresses considérables qui fleurissent à tous les niveaux et que les enseignants ressentent fortement.

L'optique du groupe qui compose les programmes n'est pas du tout de donner une place centrale à l'enseignement de la statistique dans toutes les sections mais par contre de poser les bases d'une statistique plus moderne. 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. Certains auraient souhaité attendre encore quelques années afin notamment que les enseignants se forment en statistique ; mais comment les professeurs peuvent-ils se former et avoir une pratique enseignante qui, si elle est conforme aux programmes actuels, sera en contradiction totale avec ce qu'ils apprendront ? Nous pensons qu'avoir à enseigner des rudiments de la statistique les aidera au contraire à acquérir peu à peu des connaissances plus profondes dans ce domaine.

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 qui a une longue tradition dans ce domaine. Mais malgré des traditions différentes, pourquoi, en ce qui concerne l'enseignement des statistiques au lycée, ne pourrait-on pas 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, 2002 respectivement pour les classes de seconde, première et terminale. Rappelons qu'au collège les élèves travaillent abondamment depuis plusieurs années sur la moyenne arithmétique d'une série et le langage graphique (histogrammes, diagrammes en bâtons, camemberts) et apprennent en technologie et en mathématiques à manipuler un tableur.

En seconde. Introduction de la fluctuation d'échantillonnage selon le schéma suivant :

— Réalisation effective par les élèves d'expériences de lancers de pièces ou de dés, de tirage de boules dans des urnes ; observation de la variabilité des séries de résultats : on introduit la notion de distribution empirique de fréquences et ce sont les variations de cette dernière

que l'on observe ; la moyenne, l'étendue et les paramètres qui s'en déduisent sont ainsi eux aussi fluctuants.

— Utilisation de simulation de la loi uniforme sur l'ensemble des chiffres (touche random des calculatrices pour les élèves, logiciels type excel 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 empirique 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.

Les programmes de première et terminale ne sont pas complètement 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 pour la statistique sont actuellement les suivantes (elles seront déclinées 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 à pattes. On pourra travailler sur des données classiques (courbes des tailles des carnets de santé des enfants par exemple) ou sur des données que les élèves recueillent eux-mêmes (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.) en insistant toujours 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 usuelles dans le cadre de la fluctuation d'échantillonnage) et comment pourrait se généraliser l'étude faite.

Une attention particulière sera portée à la variance qui sera utilisée pour des données gaussiennes notamment dans les domaines de la biologie et la médecine, en production industrielle et pour les erreurs de mesure. En physique, les élèves ont vu qu'il valait mieux utiliser la moyenne de plusieurs mesures plutôt qu'une seule : à travers l'observation de données réelles ou simulées on illustrera le fait que l'écart type de la moyenne est en $1 \triangleleft \sqrt{n}$.

En section scientifique, on peut dès la classe de première, justifier la définition de la variance — plutôt que la définition d'une mesure de dispersion que les élèves choisissent naturellement, à savoir la moyenne des valeurs absolue des écarts ; une « justification » classique de la variance est actuellement la facilité de calcul puisque les calculatrices de poche la proposent directement !

Tableau de contingence. 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 programmes actuels de l'enseignement secondaire ont dans le chapitre statistiques un paragraphe faisant une étrange 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 parle d'un poids idéal fonction affine de la taille (ce qui ne peut guère aider les élèves à comprendre la notion de fonction ou de modèle). On apprend de plus qu'il est légitime de faire un ajustement linéaire dès 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. La droite des moindres carrés sera maintenue dans les programmes (mais pas dans le cadre du chapitre statistique) et utilisée pour faire de l'interpolation et l'extrapolation linéaire sur un intervalle bien défini, notamment pour des données chronologiques.
— Le lien avec les probabilités sera traité dès la classe de première, en étant vigilant à ne pas mélanger comme cela se fait actuellement ce qui relève de la théorie et ce qui relève de l'expérience.

Statistiques en Classe de Seconde
Programmes et accompagnement

À titre indicatif, le temps à consacrer aux différents chapitres pourrait être de 1/8 pour les statistiques, le reste se répartissant équitablement entre les deux autres chapitres.

Rappel des programmes antérieurs :

Sixième	**Cinquième**
Exemples conduisant à lire et établir des relevés statistiques sous forme de tableaux ou de représentations graphiques, éventuellement en utilisant un ordinateur.	*Lecture, interprétation, représentations graphiques de séries statistiques. Diagrammes à barres, diagrammes circulaires. Classes, effectifs. Fréquences.*
Quatrième	**Troisième**
Effectifs cumulés, fréquences cumulées. Moyennes pondérées. Initiation à l'usage des tableurs-grapheurs. Valeur approchée de la moyenne d'une série statistique regroupée en classesd'intervalles.	*Caractéristiques de position d'une série statistique. Approche de caractéristiques de dispersion d'une série statistique. Initiation à l'utilisation des tableurs-grapheurs en statistique.*

En seconde le travail sera centré sur :

– la réflexion conduisant au choix de résumés numériques d'une série statistique quantitative ;
– la notion de fluctuation d'échantillonnage vue ici sous l'aspect élémentaire de la variabilité de la distribution des fréquences ;
– la simulation à l'aide du générateur aléatoire d'une calculatrice. La simulation remplaçant l'expérimentation permet, avec une grande économie de moyens, d'observer des résultats associés à la réalisation d'un très grand nombre d'expériences. On verra ici la diversité des situations simulables à partir d'une liste de chiffres.

L'enseignant traitera des données en nombre suffisant pour que cela justifie une étude statistique ; il proposera des sujets d'étude et des simulations en fonction de l'intérêt des élèves, de l'actualité et de ses goûts.

La notion de fluctuation d'échantillonnage et de simulation ne doit pas faire l'objet d'un cours. L'élève pourra se faire un « cahier de statistique » où il consignera une grande partie des traitements de données et des expériences de simulation qu'il fait, des raisons qui conduisent à faire des simulations ou traiter des données, l'observation et la synthèse de ses propres expériences et de celles de sa classe. Ce cahier sera complété en première et terminale et pourra faire partie des procédures d'évaluation annuelle.

En classe de première et de terminale, dans toutes les filières, on réfléchira sur la synthèse des données à l'aide du couple moyenne, écart-type qui sera vu à propos de phénomènes aléatoires gaussiens et par moyenne ou médiane et intervalle inter-quartile sinon. On amorcera une réflexion sur le problème de recueil des données et la notion de preuve statistique ; on fera un lien entre statistique et probabilité. L'enseignement de la statistique sera présent dans toutes les filières mais sous des formes diverses.

Contenu	Capacités attendues	Commentaire
Résumé numérique par une ou plusieurs mesures de tendance	Utiliser les propriétés de linéarité de la moyenne d'une série	L'objectif est de faire réfléchir les élèves sur la nature des données traitées, et de s'appuyer sur

centrale (moyenne, médiane, classe modale, moyenne élaguée) et une mesure de dispersion (on se restreindra en classe de seconde à l'étendue).	statistique. Calculer la moyenne d'une série à partir des moyennes de sous-groupes. Calcul de la moyenne à partir de la distribution des fréquences.	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.*
Définition de la distribution des fréquences d'une série prenant un petit nombre de valeurs et de la fréquence d'un événement.	Concevoir et mettre en œuvre des simulations simples à partir d'échantillons de chiffres au hasard.	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.
Simulation et fluctuation d'échantillonnage		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 donnera alors éventuellement les résultats de simulations de même taille N préparées à l'avance et obtenues à partir de simulations sur ordinateurs.

Thèmes d'étude *Pour chacun des chapitres, le professeur choisira, pour l'ensemble des élèves ou pour certains seulement en fonction de leurs centres d'intérêt, un ou plusieurs thèmes d'étude dans la liste ci-dessous.*

– Simulations d'un sondage ; à l'issue de nombreuses simulations, pour des échantillons de taille variable, on pourra introduire la notion de fourchette de sondage, sans justification théorique. La notion de niveau de confiance 0,95 de la fourchette peut être introduite en terme de « chances » (il y a 95 chances sur 100 pour que la fourchette contienne la proportion que l'on cherche à estimer) ; on pourra utiliser les formules des fourchettes aux niveaux 0,95, 0,90 et 0,99 pour une proportion observée voisine de 0,5 afin de voir qu'on perd en précision ce qu'on gagne en niveau de confiance. On incitera les élèves à connaître l'approximation usuelle de la fourchette au niveau de confiance 0,95, issue d'un sondage sur n individus (n>30) dans le cas où la proportion observée est comprise entre 0,3 et 0,7, à savoir $[\hat{p}-1/\sqrt{n}\,;\hat{p}+1/\sqrt{n}\,]$.

– Simulations de jeux de pile ou face distribution de fréquences du nombre maximum de coups consécutifs égaux dans une simulation de 100 ou 200 lancers de pièce équilibrée ; distribution de fréquences du gain sur un jeu d'au plus dix parties où on joue en doublant la mise (ou en la triplant) tant qu'on n'a pas gagné. On pourra aussi faire directement l'expérience avec des pièces pour bien faire sentir la notion de simulation...

– Simulation du lancer de deux dés identiques et distribution de la somme des faces. On pourra aussi faire directement l'expérience avec des dés pour bien faire sentir la notion de simulation...

– Simulations de promenades aléatoires sur des solides ou des lignes polygonales, fluctuation du temps et estimation du temps mis pour traverser un cube, ou pour aller d'un sommet donné à un autre sommet donné d'une ligne polygonale.

– Simulation de naissances : distribution du nombre d'enfants par famille d'au plus quatre enfants lorsqu'on s'arrête au premier garçon, en admettant que pour chaque naissance, il y a autant de chances que ce soit un garçon qu'une fille.

Document d'accompagnement du programme de la classe de seconde (extrait)

On trouvera à la suite de ce document d'accompagnement des fiches sur le programme de statistique de la classe de seconde et sur les thèmes associés.

Les choix, traduits en termes de programme pour la classe de seconde, sont guidés par les perspectives suivantes pour le lycée :

- acquérir une expérience de l'aléatoire et ouvrir le champ du questionnement statistique ;
- voir dans un cas simple ce qu'est un modèle probabiliste et aborder le calcul des probabilités.

Au collège, les élèves se sont familiarisés avec les phénomènes variables et ont appris des éléments du langage graphique (représentations diverses, "camemberts", diagrammes en bâtons) qui permettent de visualiser une série de données expérimentales ; par ailleurs, ils ont travaillé sur la notion de moyenne arithmétique.

En seconde, différents éléments apparaissent au programme :

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 XVIIe 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 coté 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 fluctuation 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.

Simulation

Formellement, simuler une expérience, c'est choisir un modèle de cette expérience puis simuler ce modèle : cet aspect sera introduit ultérieurement en première. Dans le cadre du programme de seconde, simuler une expérience consistera à produire une liste de résultats que l'on pourra assimiler à un échantillon de cette expérience (voir plus loin la fiche *listes de chiffres au hasard*).
On se contentera de simuler des situations très simples, reposant le plus souvent sur la simulation d'expériences de références où toutes les issues ont des chances égales d'apparaître.

La simulation permettra de disposer d'échantillons de grande taille et d'observer des phénomènes appelant une explication dans le champ des mathématiques. Pour bien comprendre les mathématiques, il est utile d'apprendre quel type de questions sont à adresser à cette discipline et aussi d'apprendre à reformuler ces questions dans le langage propre des mathématiques ; le langage des probabilités présenté en première S, ES et en option de première L, formalisera le langage naïf des *chances* et du *hasard* employé en seconde ; le calcul des probabilités permettra ensuite d'expliquer certains phénomènes observés.

En seconde, on approche dans le cadre d'un langage simple et familier les techniques de simulation ; pour que l'élève ne soit pas écrasé par la puissance des outils modernes de simulation, il convient qu'il ait établi un lien concret entre l'expérience et sa simulation : certaines expériences simples pourront être réalisées par une partie de la classe et simulées par le reste de la classe ; il n'est pas nécessaire, dans un premier temps, de lier les premiers pas vers la simulation de l'aléatoire à l'introduction de concepts théoriques difficiles tel celui de modèle.

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.

Cahier de statistique

Les élèves pourraient commencer en seconde un cahier de statistique rendant compte des expériences faites ou simulées, en classe ou chez eux, à la demande de l'enseignant ou de leur propre initiative. La rédaction d'un tel document individuel leur permettrait d'organiser et de planifier les expériences et les simulations, de donner forme à la conclusion qu'ils en tirent, aux questions théoriques qui se sont posées et qu'ils pourront reprendre ultérieurement. La tenue de ce cahier pourrait contribuer efficacement à structurer le travail expérimental proposé et aider ultérieurement chaque élève à mieux expliciter le lien entre l'expérience et la théorie ; cela permettrait à l'enseignant de contrôler la qualité des travaux réalisés, de vérifier que ne s'installe pas des perceptions erronées sur les phénomènes aléatoires, de faire des évaluations sur la partie statistique du programme. Ce cahier pourrait être continué en première et

terminale : l'enseignant de première pourrait ainsi savoir quels thèmes ont été travaillés par ses élèves en seconde.

La production d'un texte écrit est en soi un élément formateur ; un tel cahier, où se mêlent texte écrit et représentations graphiques, présentant des éléments narratifs et des argumentations, s'inscrit de plus dans le cadre du nouveau programme de français des élèves de seconde.

Classe de Première L
Mathématiques – informatique ; enseignement obligatoire
Programme - Extrait : 2- Statistique

BO HS n°7 du 31 août 2000

En seconde, les élèves ont abordé les notions de fluctuation d'échantillonnage et de simulation. On va maintenant définir de nouveaux paramètres à associer à une série de données numériques ; pour l'interprétation des valeurs de ces paramètres, on gardera à l'esprit qu'ils fluctuent d'une série de données à une autre.

L'objectif de ce chapitre est :

- de familiariser les élèves avec des questions de nature statistique ;
- de montrer, à travers la notion de phénomènes gaussiens, la nature de l'information prévisionnelle apportée par un écart-type ;
- d'étudier des tableaux de pourcentages.

Contenus	Commentaires
Diagrammes en boîtes Intervalle inter-quartile Définition de l'intervalle interquartile. Construction de diagrammes en boîtes (aussi appelés *boîtes à moustaches* ou *boîtes à pattes*).	On étudiera des données recueillies par les élèves, tout en choisissant des situations permettant de limiter le temps de recueil de ces données. À cette occasion, on s'attachera à : - définir une problématique ou une question précise motivant un recueil de données expérimentales, - définir les données à recueillir, leur codage et les traitements statistiques qu'on appliquera pour avoir des éléments de réponses à la question posée, - élaborer un protocole de recueil et aborder les problèmes que cela pose. Proposition d'exemples : battements cardiaques, estimation de longueurs, durée des repas du soir, nombre et durée de conversations téléphoniques, temps de passage en caisse dans une grande surface, etc.
Variance, écart-type Introduction de l'écart-type pour des données gaussiennes	L'objectif est ici de rendre les élèves capables de comprendre l'information apportée par la valeur de l'écart-type lors de mesures issues de la biologie ou du contrôle industriel. On pourra prendre comme exemple de référence l'étude des courbes de taille et/ou de poids dans les carnets de santé des enfants, en se limitant éventuellement à des âges inférieurs à quatre ou six ans.
Définition de la plage de normalité pour un niveau de confiance donné.	On se limitera ici aux exemples de résultats fournis par les laboratoires biologiques lors de certains examens. Pour l'interprétation lorsque le niveau de confiance est 0,95, on notera que le choix de ce dernier résulte d'un consensus pour avoir des formules simples et implique qu'environ une personne sur vingt sorte de cette plage.
Tableaux croisés Analyse d'un tableau de grands effectifs ; Construction et interprétation : - des marges ; - du tableau des pourcentages en divisant chaque cellule par la somme de toutes les cellules ; - du tableau des pourcentages par ligne en divisant chaque cellule par la somme des cellules de la même ligne ; - du tableau des pourcentages par colonnes en divisant chaque cellule par la somme des cellules de la même colonne.	On ne parlera pas des tableaux théoriques ou dits de proportionnalité ; les commentaires sur les pourcentages des lignes (resp. des colonnes) se feront simplement à partir des distributions de fréquences associées aux marges horizontales (resp. verticales). On pourra prendre comme exemple de référence l'étude de résultats d'élection (classification selon les régions ou les classes d'âge des votes à une élection où plusieurs candidats sont en présence).

Document d'accompagnement Programme de la classe de Première L

Extraits : Le chapitre statistique

Ce chapitre se divise en trois parties. Dans la première, l'objectif est de voir sur des exemples qu'une question peut trouver une réponse dans le champ de la statistique sous réserve, éventuellement, de transformer la question initiale. À partir de cette nouvelle question, on réfléchira simultanément sur les données à recueillir et sur le traitement statistique que l'on peut envisager pour ces données.

Exemples

Considérons la question : Quel est le nombre de battements cardiaques à la minute ? La question est trop imprécise et il convient au moins de spécifier si c'est au repos ou après un effort clairement défini. Il peut se poser alors de nouvelles questions : sur la comparaison des données au repos ou après effort, par exemple. Les élèves peuvent aussi proposer que chacun étudie son propre pouls en faisant plusieurs mesures (il y a alors à la fois la variabilité individuelle de la fréquence cardiaque et les erreurs de mesure qui s'ajoutent) ou faire une étude sur une classe entière.

Considérons une autre question : Sait-on estimer à l'oeil une longueur ? Cette question est, elle aussi, trop imprécise. Au niveau de la population visée, s'adresse-t-on à des gens de tous âges ? De tous métiers ? Par ailleurs, que signifie «estimer une longueur» ? S'agit-il de petites longueurs ou de grandes distances ? Lorsque cette situation a été expérimentée dans des classes, la question initiale s'est transformée pour devenir par exemple : Si on demande à un élève de première de couper 20 cm d'une ficelle sans appareil de mesure, que se passet-il ? On peut alors mettre en place un protocole expérimental qui permettra d'observer des données liées à cette question.

L'objectif est donc ici de montrer la diversité des questions qui se posent ainsi que le soin nécessaire à la définition et au recueil des données. Il s'agit aussi de montrer aux élèves qu'une première expérience permet de préciser et de reformuler la question initiale et que, si l'on veut apporter des réponses, généraliser ce qui est fait et interpréter des différences, il faut faire un traitement statistique plus sophistiqué et tenir compte en particulier de la fluctuation d'échantillonnage. Le résumé des données observées pourra se faire à l'aide de diagrammes en boîtes (souvent appelés boîtes à moustaches ou boîtes à pattes), éventuellement accompagnés de la moyenne ou d'une moyenne élaguée.

On trouvera à la fin de ce document une annexe relative aux diagrammes en boîtes donnant toutes les indications nécessaires sur les paramètres utiles (médianes, quartiles), sur les modes de construction de ces diagrammes, ainsi que sur leur utilisation. L'essentiel, dans un tel diagramme, est la construction de la boîte contenant la moitié des valeurs de la série ; pour les « moustaches », on pourra choisir les premier et neuvième déciles ou les valeurs extrêmes, comme l'indique l'annexe citée : en première L, on privilégiera l'utilisation des valeurs extrêmes ; dans tous les cas, les élèves devront légender leur schéma. Le tableur servira avant tout ici à ordonner les valeurs de la série observée et éventuellement à les numéroter : les élèves effectueront ensuite «à la main» le calcul de la médiane et des quartiles ainsi que la construction de la boîte.

Dans la seconde partie, on pourra d'abord définir l'écart type d'une série. On remarquera que l'écart type et la moyenne sont sensibles aux valeurs extrêmes alors que la médiane et l'écart interquartile ne le sont pas. On travaillera ensuite selon l'esprit décrit dans l'annexe ci-dessous relative aux données gaussiennes.

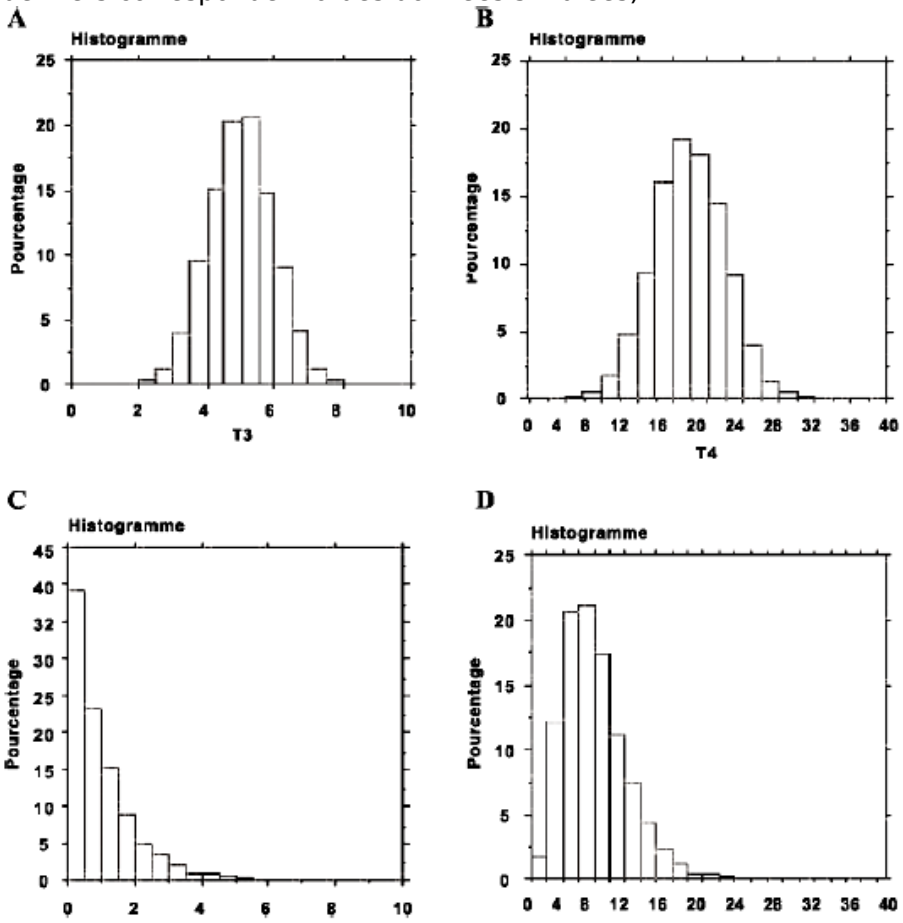
La troisième partie est consacrée à l'étude de tableaux croisés. On trouvera dans le document d'accompagnement de première ES (page 12 et suivantes) quelques exemples d'études de tableaux à double entrée. On s'intéressera aussi à des situations pour lesquelles l'enseignant sait qu'il n'y a pas indépendance entre les deux caractères qualitatifs étudiés sur la population (tableaux présentés lors d'élections, par exemple). Cette partie pourrait aussi bien figurer dans le chapitre « Information chiffrée» : elle a été mise dans le chapitre « Statistique » afin que l'enseignant puisse indiquer (sans le justifier) que les fluctuations des distributions des fréquences d'une ligne à l'autre (resp. d'une colonne à l'autre) sont éventuellement d'une ampleur que la fluctuation d'échantillonnage ne peut seule expliquer.

Le travail réalisé dans ce chapitre sera rédigé dans le cahier de statistique commencé en seconde.

Annexe : à propos des données gaussiennes

Exemple 1 - Bilan de santé

Pour faire le bilan de l'activité thyroïdienne d'un individu, on mesure les quantités de «T3 libre» (tri-iodo thyronine libre) et «T4 libre» (thyroxine libre); ces mesures sont exprimées en pmol/litre (pmol signifie pico-mole, soit 10^{-12} mole, une mole étant composée d'environ $6{,}02.10^{23}$ molécules). Les résultats de deux séries de 5 000 mesures chez des individus dont la thyroïde fonctionne normalement sont résumés par les histogrammes A et B; ces histogrammes ont la même allure dite en cloche, nettement différente de celle des histogrammes C et D (ces deux derniers correspondent à des données simulées).



Les deux séries obtenues lors de ces bilans thyroïdiens correspondent à des «données gaussiennes»; pour de telles données on peut déterminer un modèle (modèle gaussien) à partir de deux paramètres *m* et σ calculés sur une série de référence aussi longue que possible :

$m = \frac{1}{n}\sum x_i$ – la moyenne *m* de la série de référence :

$\sigma = \sqrt{\frac{1}{n}\sum (x_i - m)^2}$ – l'écart type de la série de référence :

Pour les mesures de T4L, on trouve sur la série de référence (ici de taille 5 000) :
m = 16,7 et σ = 4,0 (les unités étant des pmol/litre).
Pour les mesures de T3L, on trouve sur la série de référence (ici de taille 5 000) :
m = 4,9 et σ = 0,9 (les unités étant des pmol/litre).

À partir de ce modèle, on montre que pour des mesures ultérieures de données analogues:
– environ 68 % des mesures sont dans l'intervalle [*m* – σ ; *m* + σ] ; environ 16 % seront inférieures à (*m* – σ) et environ 16 % seront supérieures à (*m* + σ) ;

– environ 95 % des mesures sont dans l'intervalle [$m - 2\sigma$; $m + 2\sigma$] ; environ 2,5 % seront inférieures à ($m - 2\sigma$) et environ 2,5 % seront supérieures à ($m + 2\sigma$) ;
– environ 99,8 % des mesures sont dans l'intervalle [$m - 3\sigma$; $m + 3\sigma$]; environ 0,1% seront inférieures à ($m - 3\sigma$) et environ 0,1% seront supérieures à ($m + 3\sigma$).

Dans l'exemple des analyses biologiques de dosages de T3L et T4L, et dans de nombreux examens, l'intervalle [$m - 2\sigma$; $m + 2\sigma$] est appelé « plage de normalité » : il contient environ 95 % des valeurs observées chez des individus non-malades. Les plages de normalité sont ici :

– [3,1; 6,7] pour le dosage de T3L;
– [9,7; 25,7] pour le dosage de T4L.

(Les plages de normalité sont indiquées par les laboratoires sur les comptes rendus d'analyse biologique. Ces plages de normalité sont voisines mais pas nécessairement identiques d'un laboratoire à l'autre; elles sont en effet calculées à partir de séries de référence traitées par l'appareil de mesure du laboratoire.)

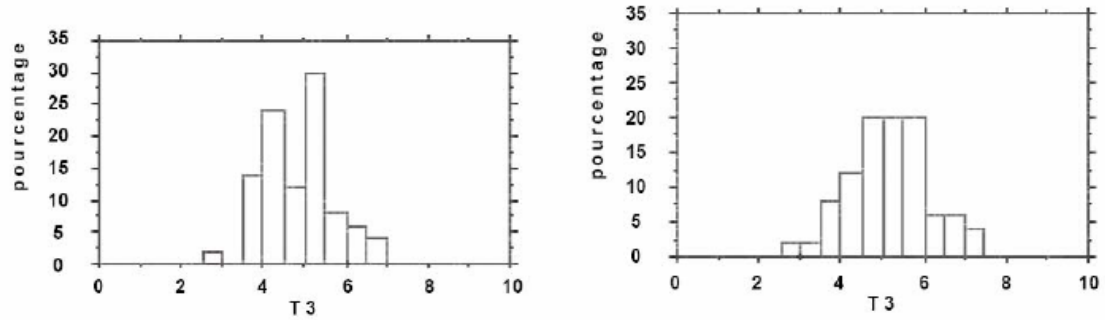
Si on faisait des dosages de T3L chez des personnes choisies au hasard dans une population donnée, environ une sur vingt aurait une valeur sortant de la plage de normalité. Cela dit, les personnes à qui l'on fait de tels dosages (les individus de la série de référence mis à part) ne sont pas choisies au hasard et présentent en général des symptômes justifiant cet examen; sortir de la plage de normalité constitue un symptôme de plus en faveur d'une maladie de la thyroïde (symptôme d'autant plus marqué que l'on s'éloigne beaucoup de la moyenne : il est classique pour certaines pathologies de dépasser la moyenne de dix écarts types).

Exemple 2 - Taille

Les tailles de garçons (resp. de filles) nés la même année constituent des données gaussiennes, et les courbes situées à la fin du carnet de santé des enfants donnent, entre autres, pour chaque âge une plage de normalité égale à [$m - 2\sigma$; $m + 2\sigma$], où les paramètres m et σ pour un âge donné sont calculés sur des séries de référence (malheureusement, ces séries sont anciennes et ne sont plus vraiment des séries de référence pour les enfants qui naissent aujourd'hui). On notera que dans la population concernée par la série de référence, pour chaque âge, environ un individu sur vingt sort de la plage de normalité.

Commentaires

L'objectif essentiel du paragraphe relatif aux données gaussiennes est de faire comprendre, à partir d'exemples, d'une part, le type d'information qu'apporte l'écart type et, d'autre part, la notion de plage de normalité, en particulier pour une lecture correcte de certains examens biologiques ou des courbes de croissance présentes dans les carnets de santé. On s'est longtemps demandé pourquoi de nombreuses données (mesures biologiques, erreurs de mesure) pouvaient être qualifiées de «gaussiennes»; un théorème de mathématiques appelé «théorème central limite» en propose une explication. Ce théorème est totalement hors programme. De même, est totalement hors programme la reconnaissance du caractère gaussien de données : on dira aux élèves que des études statistiques ont prouvé qu'il en était ainsi et on leur apprendra simplement à comprendre et utiliser cette information.



On sera attentif à la formulation des conclusions : la normalité évoquée ici est une normalité statistique et il y a une chance sur vingt pour qu'un individu « normal » choisi au hasard soit en dehors de la plage de normalité! De même, un échantillon de petite taille pris au hasard peut s'écarter sensiblement de la forme «en cloche», comme le montrent les deux histogrammes ci-dessous relatifs à deux échantillons de taille cinquante, pris au hasard dans la série de l'exemple 1.

On pourra prolonger la réflexion en simulant, par exemple, la situation suivante où *n* personnes choisies au hasard dans une population de gens en parfaite santé subissent quatre examens médicaux indépendants : on constate alors qu'environ une personne sur cinq a au moins un examen qui sort de la plage de normalité ! Pour faire cette simulation avec une calculatrice, il suffit de disposer de huit chiffres au hasard, de les prendre deux par deux pour fabriquer quatre nombres entre 00 et 99, puis de compter 1 si l'un au moins de ces quatre nombres est supérieur ou égal à 95, 0 sinon; la proportion de 1 est de l'ordre de 1/5 !

Extraits du rapport d'étape Statistique et probabilités de la Commission de réflexion sur l'enseignement des mathématiques

24 mars 2006

introduction

« 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'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 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. »

« Le traitement de l'information chiffrée, c'est à dire le calcul de certains 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 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 » et la réponse passe par l'intégration de l'aléatoire à tous les niveaux de l'enseignement. »

chapitre 4

« des concepts, des formules ou des pratiques que l'on peut acquérir dans l'enseignement secondaire et qui facilitent les formations ultérieures de statistique. »

formations professionnelles

« 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
- il y a souvent plusieurs origines de la dispersion
- l'esprit de la démarche statistique, c'est réfléchir avant de collecter les données
- 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. »

« 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. »

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 du 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 ou….
- 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. »

La formation des professeurs

« une formation initiale et continue des enseignants ayant en charge l'enseignement de cette « statistique du citoyen » est un vaste chantier. »

« Elle concerne pour les mathématiques environ 40 000 personnes (professeurs de collège et des lycées) ; 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. »

« 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 ».

Extrait d'une enquête faite en 2005 par l'APMEP auprès de professeurs :

Ce qu'ils pensent des contenus du programme

Certains professeurs ne se lancent pas dans une réponse à la question à réponse ouverte : "quelle partie du programme aimeraient-ils supprimer?", cela demande une réflexion globale non seulement sur le programme de première S mais sur l'architecture générale de l'enseignement des mathématiques au lycée. Cependant, 61% des professeurs (78 sur 128) jugent nécessaire de donner un avis. Les uns disent leur perplexité ("je ne sais pas"), ou ne font qu'une demande quantitative ("au moins deux chapitres"), les autres mentionnent les autres niveaux du lycée, soit pour demander que des modifications éventuelles se fassent en cohérence avec le programme de Terminale, soit pour proposer de "muscler" le programme de seconde pour alléger celui de première S. 15% des professeurs qui rédigent une réponse ne souhaitent supprimer aucun chapitre, mais cette réponse s'accompagne presque toujours d'une demande d'augmentation de l'horaire("j'aimerais surtout avoir plus d'heures car c'est la course pour terminer le programme et nous n'avons pas le temps de faire beaucoup d'exercices").

(...)

57 professeurs sur les 78 qui répondent à cette question citent expressément des chapitres à supprimer. Les chapitres écartés sont nombreux et variés, mais les réponses révèlent quelques régularités : le chapitre le plus souvent rejeté est le chapitre "statistiques" : 35 fois. Le chapitre sur les probabilités n'est cité que 4 fois, avec en outre une mention spécifique de la notion de variable aléatoire, citée 2 fois."