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 passe-t-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

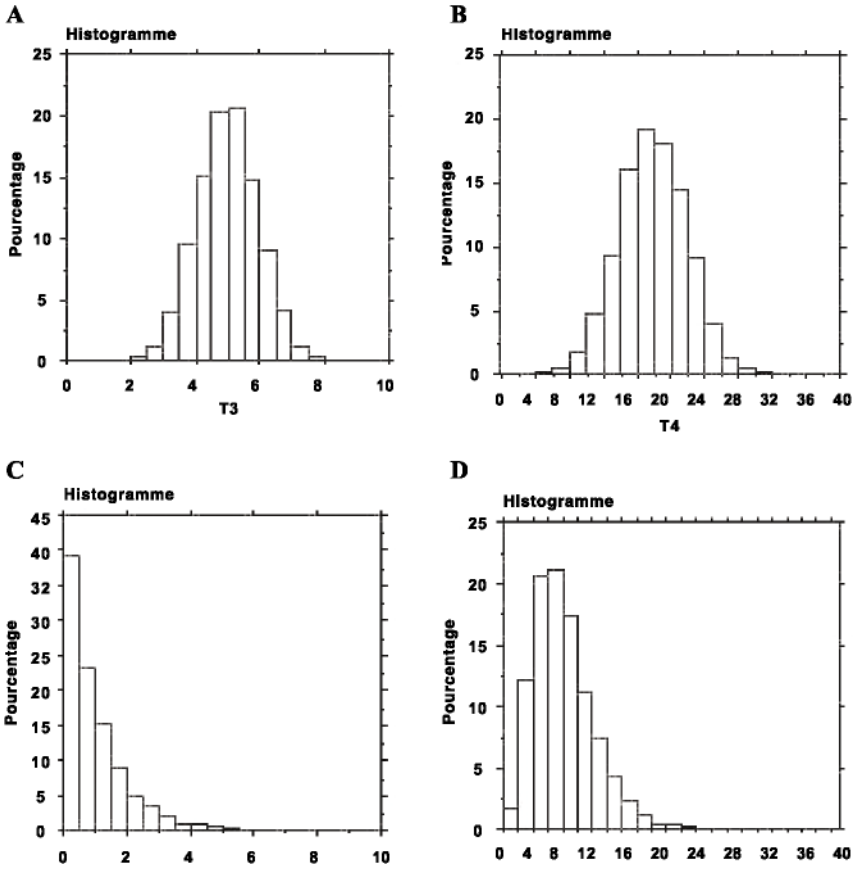
aussi bien figurer dans le chapitre « Information chiffrée» : elle a été mise dans le chapitre « Statistique » afin que l'enseignant puisse indiquer (sans le justifier) que les fluctuations des distributions des fréquences d'une ligne à l'autre (resp. d'une colonne à l'autre) sont éventuellement d'une 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 :

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

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

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 [$m - \sigma$; $m + \sigma$] ; environ 16 % seront inférieures à ($m - \sigma$) et environ 16 % seront supérieures à ($m + \sigma$) ;

– 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).

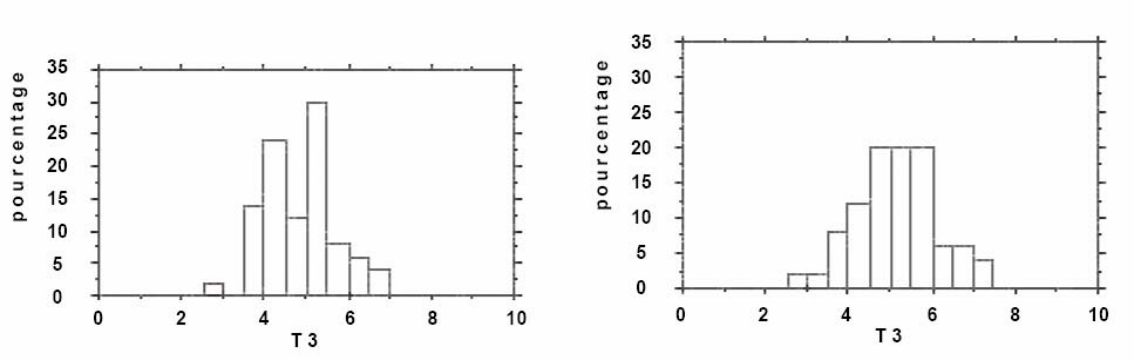
Exemple 2 - Taille

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 !