MATHÉMATIQUES
Collège

Projet de document d'accompagnement

Organisation et gestion de données

Le programme de la classe de sixième entre en application à la rentrée 2005

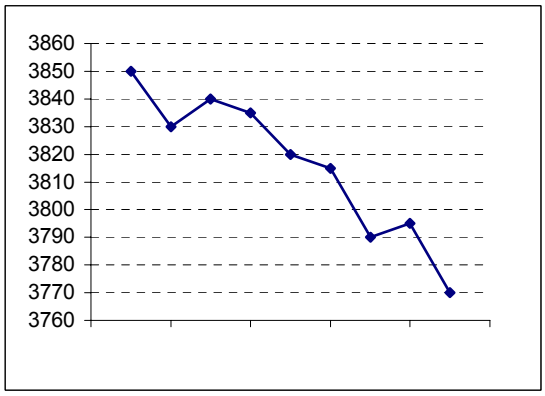
22 août 2005

Ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche
Direction de l'Enseignement scolaire
eduscol.education.fr

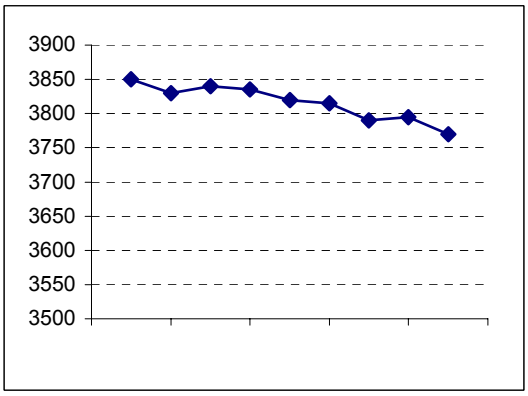
ORGANISATION ET TRAITEMENT DES DONNÉES

1- Objectifs généraux

La partie relative à l'organisation et la gestion de données a pour objectif principal de permettre aux élèves de construire et travailler des compétences nécessaires pour recevoir ou produire de l'information chiffrée. Il s'agit d'une part de continuer[1] à initier les élèves de collège à la lecture, à l'utilisation et à la production de tableaux, de représentations graphiques[2], d'autre part de mettre en place les premiers outils de la statistique descriptive, en particulier la notion de résumé statistique à partir de l'étude de quelques caractéristiques de position. Il s'agit aussi, à travers ces premiers contacts, d'aider les élèves à percevoir que la mise en forme de l'information proposée résulte de choix qui en accentuent ou en atténuent certains aspects et donc de contribuer ainsi au développement de l'esprit critique indispensable dans la vie de tout citoyen.
L'exemple ci-dessous permet de mettre en évidence le caractère subjectif de toute représentation graphique (souvent liée à la plage des données représentée sur les axes) et des interprétations qui pourraient en être tirées.



Un vendredi noir à la Bourse !



L'indice des valeurs est en repli de 2,5 %…

De même, tout citoyen devrait pouvoir décoder les slogans publicitaires comme par exemple celui, récent, d'un jeu de hasard : "Cent pour cent des gagnants ont tenté leur chance". Il est évident que dire "tous les gagnants ont joué" n'a pas le vernis "scientifique" qui est sensé lui donner sa crédibilité ! Dans un autre registre, celui des sondages d'opinion, il est indispensable de comprendre que 60 % d'avis favorables parmi 75 % de personnes ayant donné une réponse ne constituent pas une majorité absolue de la population tout entière.

Pour donner du sens aux notions étudiées et susciter l'intérêt, les travaux sont conduits à partir d'exemples et en liaison, chaque fois qu'il est possible, avec l'enseignement des autres disciplines. De fait, il est souvent plus pertinent de s'appuyer sur des situations réelles, par exemple sur des activités de relevés (enquêtes, mesurages…) réalisées par les élèves, en particulier dans leur environnement proche. Il est possible aussi d'utiliser des données réelles directement fournies. Les sources de données exploitables sont multiples. Il est ainsi très simple d'accéder à de nombreux aspects des résultats du recensement de 1999, sur le site internet de l'INSEE, ou à des données chiffrées concernant l'élevage ou la pêche sur le site du ministère de l'Agriculture (cf annexe).

2- Les représentations graphiques de données : diagrammes et histogrammes

De nombreuses formes de représentations graphiques de données peuvent être rencontrées. Les élèves doivent être habitués à exploiter la plupart de ces formes. Les programmes font explicitement référence aux diagrammes en tuyaux d'orgue, en bandes, à secteurs pour les données relatives à un caractère qualitatif, aux diagrammes en bâtons pour les données relatives à un caractère quantitatif discret, aux histogrammes pour les données relatives à un caractère quantitatif continu.
Voici, par exemple, un tableau récapitulant l'évolution des tonnages et des chiffres d'affaire de la pêche dans le département des Côtes d'Armor de 1991 à 1998.

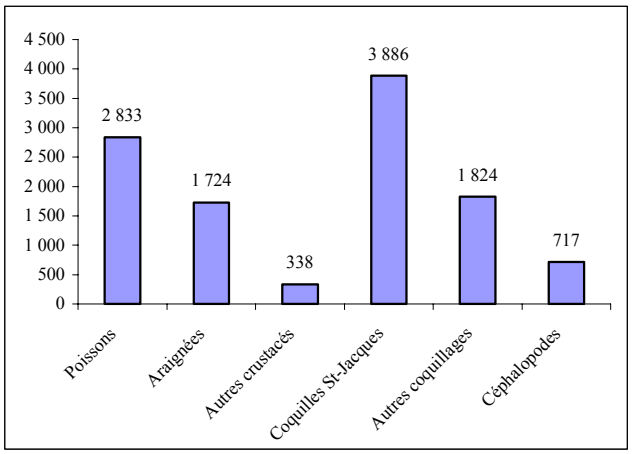
		1991	1992	1993	1994	1995	1996	1997	1998
Poissons	(tonnes)	2 823	2 833	2 938	3 565	3 903	4683	5 323	6 097
	(milliers d'euros)	9385	8905	8318	9449	9188	10651	12580	14625
Araignées	(tonnes)	1 588	1 724	1 510	1 332	898	770	831	867
	(milliers d'euros)	3801	4296	3232	3659	2120	1761	2144	2482
Autres crustacés	(tonnes)	307	338	339	429	344	411	435	489
	(milliers d'euros)	1591	1914	2013	2328	1759	2003	1817	2123
Coquilles St-Jacques	(tonnes)		3 886	4 764	5 479	4 501	4 330	4 011	3 180
	(milliers d'euros)	4844	6644	7593	9071	8225	8146	7300	6536
Autres coquillages	(tonnes)	2 321	1 824	1 607	1 212	866	1 701	2 278	5 935

[1] En effet, à l'école primaire, les élèves ont déjà été mis en situation de prendre de l'information à partir de tableaux, de diagrammes ou de graphiques.
[2] Le tableur grapheur fait l'objet d'une initiation dès la classe de cinquième et doit être largement utilisé.

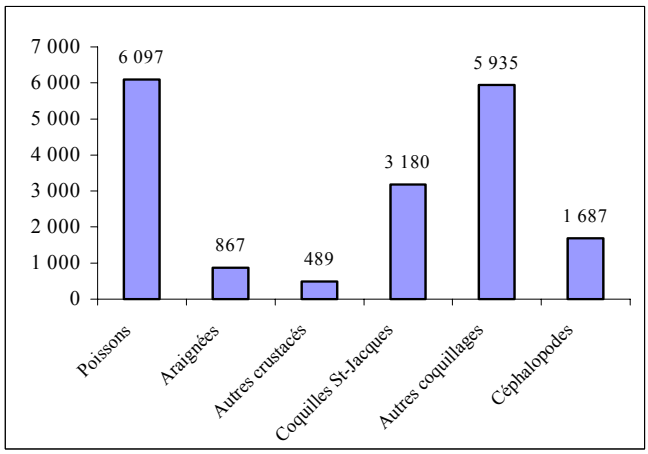
	(milliers d'euros)	2358	1824	1514	1176	1110	1700	2006	3628
Céphalopodes	(tonnes)	978	717	1 317	1 188	1 514	1 477	1 799	1 687
	(milliers d'euros)	1368	1624	2432	2506	2923	2884	5485	4005
Total	(tonnes)	10 024	11 322	12 475	13 205	12 027	13 372	14 677	18 257
	(milliers d'euros)	23347	25207	25102	28189	25325	27145	31332	33399

Pour l'étude des tonnages par catégorie pour des années données (1992 et 1998), l'utilisation d'un diagramme en tuyaux d'orgue est déjà exploitable.

Année 1992

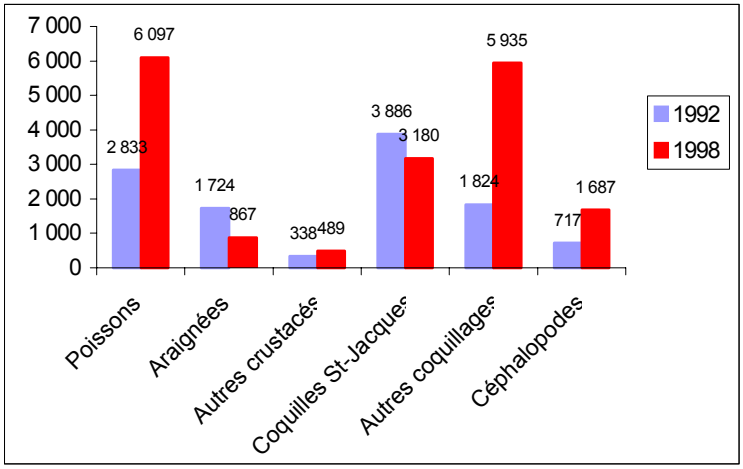


Année 1998

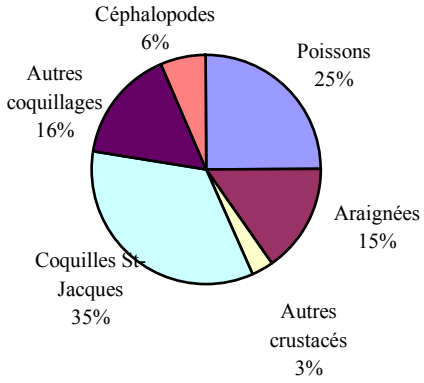


Des réponses simples aux questions suivantes, nécessitant des lectures directes, peuvent être sollicitées dès la sixième :
- *Quel est le tonnage de poissons pêchés en 1998 ?*
- *Quelle est la catégorie la plus pêchée en 1992, en 1998 ?*
- *Quelles sont les différences les plus significatives entre les deux années ?*

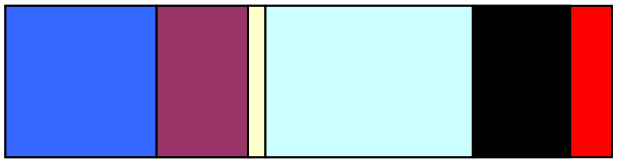
Il est à noter que ces réponses peuvent être données aussi à partir du tableau mais l'utilisation du graphique facilite le travail à condition que l'élève soit en mesure d'estimer visuellement certains rapports entre les hauteurs des « tuyaux ».
En outre, les questions posées, lors d'une étude de population, peuvent induire le type de graphique retenu. Ainsi, pour la troisième des questions ci-dessus, la représentation des deux années « sur le même graphique » facilite les comparaisons.



Pour réaliser des études relatives, par exemple la place d'une catégorie parmi un tout, le recours aux diagrammes en bandes ou à secteurs est plus adapté. Ils sont souvent associés aux fréquences des différentes catégories. Ainsi la pêche, par catégorie, en 1992, peut se résumer par le diagramme circulaire suivant :

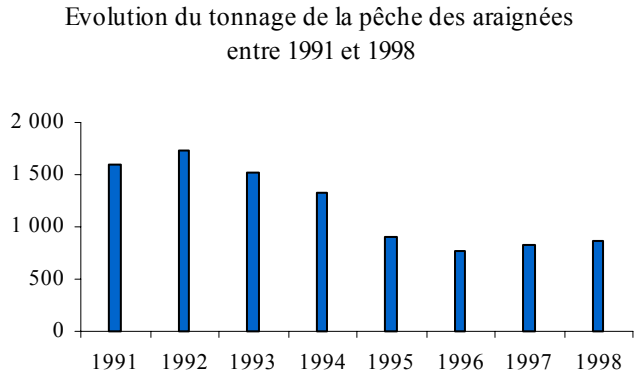


La représentation par un diagramme en bandes obéit à la même démarche. Il s'agit de découper la surface d'un rectangle en sous-surfaces dont les aires sont proportionnelles aux effectifs de chaque catégorie (c'est-à-dire dont les longueurs sont proportionnelles aux effectifs ou aux fréquences de chaque catégorie).



La prise d'information à partir de ces représentations s'appuie essentiellement sur la capacité à estimer visuellement les rapports partie/tout ou partie *a*/partie *b* et inversement, de telles constructions contribuent à développer cette capacité chez les élèves.

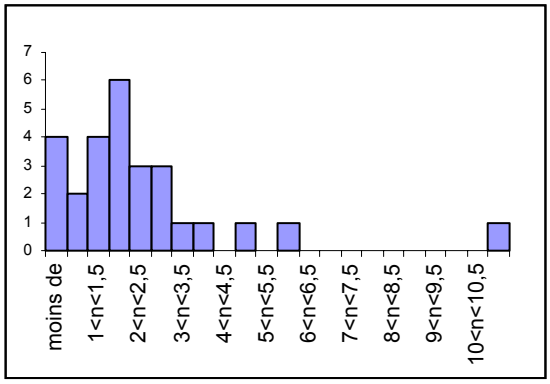
L'évolution du tonnage de la pêche à l'araignée au cours de la période 1991-1998 est une donnée à caractère quantitatif discret, comme la plupart des séries chronologiques. L'utilisation d'un diagramme en bâtons[3] est ici appropriée.
Il faut remarquer que des représentations graphiques voisines sont souvent utilisées : par exemple, un nuage de points ou une courbe représentative. Il importe alors d'indiquer aux élèves le lien avec la représentation standard et de leur faire comprendre les différences d'implicites qu'elles introduisent[4].

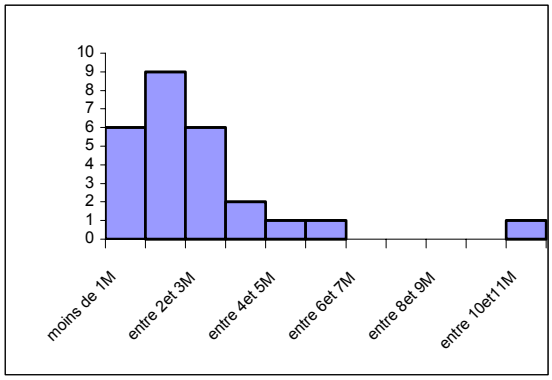


Toutes les représentations précédentes sont rencontrées dès la classe de sixième[5], voire au niveau de l'école primaire, dans de nombreuses disciplines. C'est à partir de la classe de cinquième que sont introduits les histogrammes dans le programme de mathématiques. Lorsque les valeurs possibles pour un caractère quantitatif discret sont très nombreuses, par exemple les notes moyennes (arrondies au dixième) d'un groupe d'élèves, il devient difficile voire impossible d'utiliser un diagramme en bâtons. Il convient alors de réaliser des regroupements par classes et d'avoir recours à un histogramme pour représenter les données. Il en est de même pour les données qui peuvent prendre toutes les valeurs d'un intervalle réel, par exemple la taille des nouveaux-nés de l'année 2004 ou plus généralement tout caractère faisant l'objet d'une mesure physique. Le choix des classes dépend de la nature du problème étudié. Pour étudier les irrégularités d'une distribution, les intervalles doivent être assez petits. Si c'est la forme générale qui présente de l'intérêt, les intervalles sont plus grands. Ainsi, l'exemple ci-dessous, construit à partir de tableaux de l'INSEE (recensement 1999), décrit le nombre de régions de France suivant le nombre d'habitants. L'amplitude des classes est 0,5

[3] Les logiciels les plus fréquemment utilisés ne permettent pas de réaliser des diagrammes en bâtons. Il faut faire un diagramme en tuyaux d'orgue et réduire la largeur des barres en choisissant un écartement maximum entre deux barres.
[4] Dans la représentation de droite, le fait de relier les points ne correspond à aucune réalité concrète ou théorique.
Il s'agit simplement d'avoir un aperçu plus "parlant" de l'évolution étudiée.
[5] Il s'agit en sixième essentiellement de lecture et d'interprétation. La construction de telles représentations par les élèves relève de la cinquième.

million pour l'histogramme de gauche, de 1 million pour celui de droite. Il est possible de constater l'existence de quatre régions "peu peuplées" sur l'histogramme de gauche, ce que ne révèle pas celui de droite. Les programmes précisent que les exemples étudiés se limitent au cas de classes d'égales amplitudes[6]. L'histogramme se lit alors comme un diagramme en bâtons.





Le choix de l'amplitude des classes joue donc un rôle fondamental car certaines caractéristiques importantes peuvent être gommées par le choix d'une amplitude trop grande.

3- Les premières notions de résume statistique

3.1• Effectifs et fréquences

La notion de fréquence est introduite en classe de cinquième. Son utilisation peut être légitimée par des questions de comparaison de sous-populations ayant un caractère donné dans des populations d'effectifs différents. A partir de là, le premier objectif est de savoir calculer des fréquences dans un contexte donné. Diverses écritures peuvent être utilisées pour désigner une fréquence mais, le plus souvent, une valeur décimale exacte ou approchée ou un pourcentage permettent de mieux fixer les idées (par exemple, dans l'exemple précédent de la pêche, la fréquence de la catégorie "poissons" dans les prises de 1992 s'exploite plus facilement sous la forme 0,25 ou 25 % que sous la forme $\frac{2833}{11322}$).

Dans un deuxième temps, la notion de fréquence peut être utilisée dans quelques exemples de comparaison de deux distributions d'une même variable qualitative. Il importe alors de choisir des contextes dans lesquels cette comparaison a un sens. Pour reprendre l'exemple précédent, s'il est cohérent de comparer les fréquences du tonnage de poisson en 1992 (25 %) et en 1998 (33 %), il est illusoire de vouloir comparer cette fréquence avec celle des prises de poisson dans un port donné du département des Côtes d'Armor.

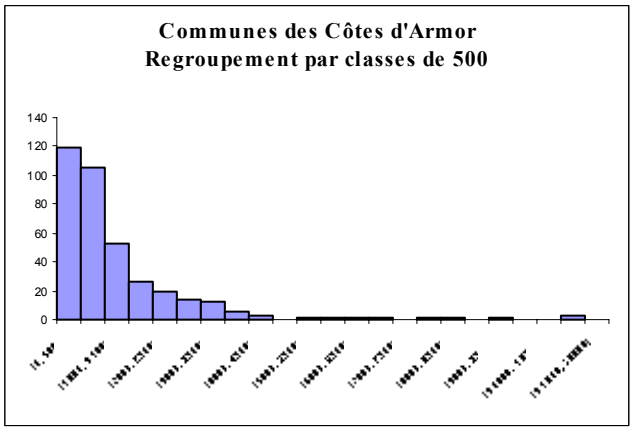
3.2• Moyenne, moyenne pondérée

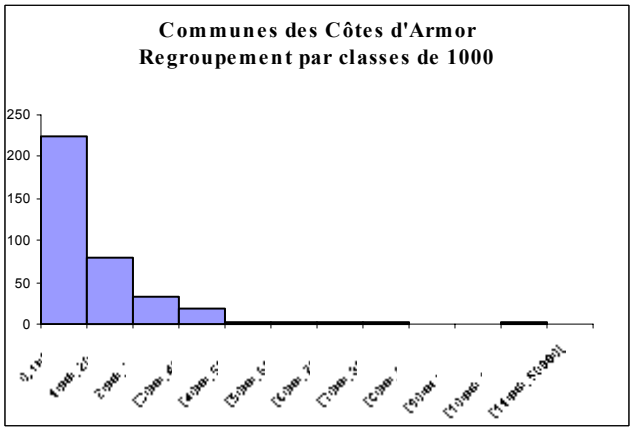
Les diagrammes et les histogrammes sont des outils de description des données assez complets mais leur mise en œuvre est parfois lourde. Pour effectuer des travaux plus globaux ou des études comparatives, il s'avère nécessaire de synthétiser davantage l'information.

La moyenne est la première caractéristique de position étudiée. Les élèves connaissent cette notion dans la mesure où elle est très présente dans leur scolarité[7]. Ils doivent approfondir leurs connaissances à son sujet : différents procédés de calcul, compréhension des effets de regroupements (la moyenne des moyennes partielles n'est pas forcément égale à la moyenne), dépendance des valeurs extrêmes, fait qu'une même moyenne peut résumer des ensembles de données très différents (par exemple : forte concentration autour de la moyenne ou forte dispersion par rapport à celle-ci), à partir de situations variées et significatives au delà du classique travail sur les notes obtenues par les élèves.

Le recours à des regroupements en classes pour l'estimation d'une moyenne[8] n'est pas un objectif des programmes. Il peut être entrepris sur un exemple pour faire constater la perte d'information et le confort et la fiabilité qu'apporte l'utilisation d'un tableur pour effectuer un tel calcul.

[6] Sauf éventuellement pour les classes extrêmes.
[7] Moyenne des notes trimestrielle calculée le plus souvent comme moyenne arithmétique des notes aux contrôles.
[8] En utilisant les milieux des classes, affectés de l'effectif de la classe correspondante.





Estimation de la moyenne : $m_{500} = 1409$ Estimation de la moyenne : $m_{1000} = 1444$

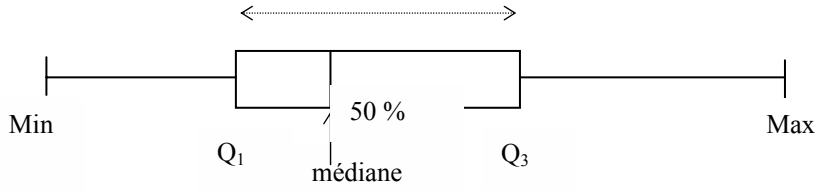
La moyenne donnée par un calcul direct au tableur est 1458[9].

3.3• Médiane, quartiles

La médiane est, comme la moyenne, un indicateur de tendance centrale.
La définition qui est retenue en collège pour la médiane d'une série est celle qui est adoptée dans le programme de seconde. Elle s'appuie sur la pratique :
Médiane (empirique) *: La série des données est ordonnée par ordre croissant. Si la série est de taille impaire (2n+1), la médiane est la valeur du terme de rang n+1. Si la série est de taille paire (2n), la médiane est la demi-somme des valeurs des termes de rang n et n+1.*
D'autres définitions sont parfois utilisées ; par exemple, la médiane est le deuxième quartile[10]. Dans la pratique statistique, ces différences n'ont pas d'importance. Pour les élèves, connaître la signification de la médiane en terme de position est l'objectif principal. La détermination de la médiane nécessite le classement des données, ce qui n'est pas le cas pour le calcul de la moyenne. De plus, contrairement à la moyenne, la médiane n'est pas sensible aux valeurs extrêmes, ce qui est mis en évidence sur des exemples. La position relative de la médiane et de la moyenne d'une série peut être interprétée quand cela est significatif. Ainsi des expressions comme « la moyenne des salaires est… » et « la médiane des salaires est … » doivent pouvoir être traduites par les élèves sous d'autres formes, par exemple : « Avec la masse des salaires distribués, si chacun recevait le même salaire, celui-ci serait de … », « La moitié de la population gagne plus de … et l'autre moitié moins de … ».
Pour mieux comprendre la notion de médiane, il est utile de mettre en évidence, sur quelques exemples, et sans en faire des connaissances exigibles, d'autres caractéristiques de position : les premier et troisième quartiles.
Pour mémoire, les définitions concernant les quartiles sont les suivantes :
Premier quartile (empirique) *: c'est le plus petit élément q des valeurs des termes de la série, tel qu'au moins 25 % des données sont inférieures ou égales à q.*
Troisième quartile (empirique) *: c'est le plus petit élément q' des valeurs des termes de la série, tel qu'au moins 75 % des données sont inférieures ou égales à q'.*

3.4• Etendue

Le seul paramètre relatif à la dispersion[11] d'une série de données dans les programmes de collège est l'étendue. Ce premier élément concernant la notion de dispersion est rudimentaire. Il présente un inconvénient : sa très grande sensibilité aux extrêmes. La détermination des quartiles peut alors compléter la connaissance de la distribution, en considérant l'intervalle interquartile (voir schéma ci-dessous).
Une fois déterminés ces différents paramètres, il est possible de donner un premier résumé statistique d'une série.
Remarque : Dans de nombreuses disciplines, il est d'usage de présenter ce résumé sous une forme graphique : le diagramme en boîte (ou à moustaches ou de Tuckey). Comme les histogrammes, les diagrammes en boîte représentent graphiquement une série de données. Au lieu de partager l'ensemble des valeurs possibles en intervalles d'amplitude constante, on le partage en segments qui contiennent une proportion fixée des valeurs de la série. La configuration la plus classique s'appuie simplement sur les quartiles.



[9] L'exemple montre aussi que, contrairement à une idée répandue, une amplitude plus petite ne garantit pas une meilleure estimation.
[10] En théorie, on définit, pour toute série numérique de données à valeur dans un intervalle I, la fonction quantile Q, de [0,1] dans I, par : $Q(t) = \inf \{x, F(x) \geq t\}$, où $F(x)$ désigne la fréquence des éléments de la série inférieurs ou égaux à x. La médiane est alors Q(0,50).
[11] L'intérêt de la notion de dispersion peut se dégager de la nécessité de distinguer deux séries de même tendance centrale.

De même qu'aucune compétence n'est exigible à propos des quartiles, les diagrammes en boîtes ne font l'objet d'aucune étude spécifique au collège. Cependant, pour quelques exemples, il peut être intéressant de les faire matérialiser pour mieux visualiser la distribution des valeurs et notamment comparer plusieurs répartitions comme dans l'exemple de l'étude de la population en Bretagne (cf. annexe).

4- Les intéractions avec d'autres domaines d'étude

4.1 • Liens avec le calcul

A l'évidence, le premier domaine des programmes de mathématiques qui entre en interaction avec la représentation et le traitement des données est celui des nombres et du calcul. Le classement des données, la détermination des différentes caractéristiques, des fréquences…sont des occasions de manipuler les nombres entiers et décimaux, de travailler sur les diverses représentations d'un même nombre et d'organiser des calculs. Les calculs induits par les problèmes posés sont élémentaires (quatre opérations) et restent ainsi accessibles à la plupart des élèves. C'est aussi l'occasion, quand la situation s'y prête, de pratiquer le calcul mental en lui donnant un objectif précis (ordre de grandeur, anticipation des résultats…). D'un autre point de vue, la nécessité (voulue) de traiter des situations réelles présentant des grands effectifs de données limite la pratique, dans ce domaine, du calcul "à la main". Le recours au calcul instrumenté est alors naturellement mis en œuvre. Il n'est évidemment pas inutile, cependant, d'aborder, sans instrument de calcul, des situations plus simples lors de la mise en place des notions pour en renforcer la compréhension.
L'utilisation du tableur-grapheur est intimement liée au travail de traitement des données dans la mesure où il permet non seulement d'exécuter les différents calculs nécessaires dans des conditions favorables mais aussi d'obtenir directement les représentations graphiques souhaitées. C'est donc une nécessité de débuter l'apprentissage de cet outil dès la cinquième (le niveau auquel on commence à construire de l'information) !

4.2• Liens avec la proportionnalité

La proportionnalité intervient très souvent dans le travail de traitement des données. C'est ainsi que la graduation d'un axe prend appui sur la proportionnalité des distances entre deux points et des écarts entre les deux valeurs qu'ils représentent. De même, la détermination des fréquences de différents caractères d'une même série fait intervenir la proportionnalité à leurs effectifs. L'élaboration de diagrammes circulaires ou en bande est une occasion particulière de faire fonctionner la proportionnalité dans la détermination des angles ou des longueurs. Il peut être aussi utile de faire remarquer, dans un histogramme, la proportionnalité de l'aire[12] d'un rectangle et de la fréquence correspondante. Il convient donc, à chaque occasion, d'expliciter la présence et l'utilisation de la proportionnalité et de travailler ainsi à consolider la cohérence interne des programmes.

4.3• Liens avec les autres disciplines

L'interaction de la partie "Organisation et gestion de données" avec les autres disciplines enseignées au collège se fait essentiellement sous deux formes. Dans un sens, les savoirs et savoir-faire construits dans le cadre de l'enseignement des mathématiques sont opérationnalisés dans des études spécifiques à la discipline concernée. La plupart des compétences travaillées en mathématique sont mobilisables : lecture de tableaux et lecture graphique, traitements calculatoires et graphiques divers, élaboration d'un résumé statistique… C'est le cas en particulier des disciplines expérimentales : Sciences de la vie et Physique-Chimie dans lesquelles les relevés d'observations ou de mesures nécessitent un traitement statistique, mais aussi de la géographie qui fait un usage important des représentations graphiques et des statistiques ou de la technologie…
Dans l'autre sens, les disciplines peuvent fournir aux mathématiques les exemples sur lesquels les notions à mettre en évidence sont dégagées ou exploitées.
L'étude de certains thèmes de convergence propres à l'ensemble des disciplines scientifiques : météorologie, santé, sécurité, environnement offre une occasion particulièrement pertinente de mettre en œuvre ces implications mutuelles.

[12] Même si les histogrammes en classes inégales ne sont pas abordés au collège.