HONNEUR AUX DAM ! (OU LES AVATARS DE LA FEMME À BARBE)

Jacques VERDIER
Lycée Varoquaux
54-TOMBLAINE

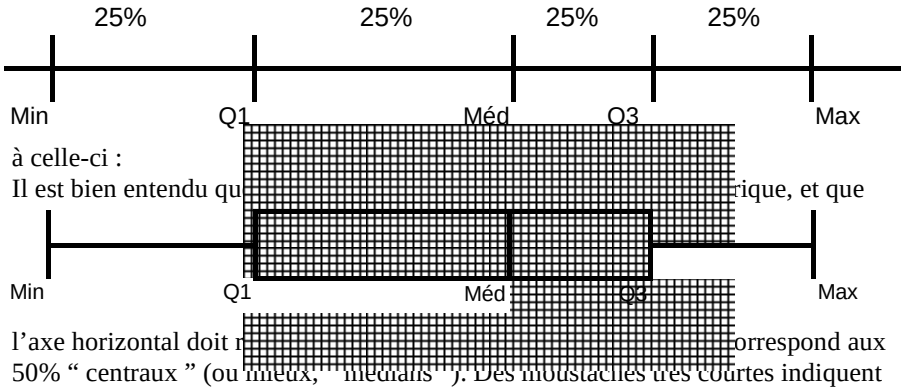
Pour ceux qui ne le savent pas, une DAM est une DEB qui a vieilli d'un an… (dans les instructions de l'an passé, on parlait de Diagrammes en Boites ; depuis on a vu apparaître les Diagrammes à Moustaches).
Cet article fait suite à celui que j'ai publié dans le bulletin n°430 de l'A.P.M.E.P. sous le titre " Deux ou trois choses que je sais de la médiane ". Je vous invite à vous y reporter avant de lire celui-ci, que j'aurais pu intituler " deux ou trois choses que je sais des boites à moustaches ". Celui-là est disponible sur le Web, à l'adresse URL suivante : http://www.ac-nancy-metz.fr/enseign/maths/APMEP/mediane.htm

Quartiles, déciles…
Tout ce qui a été dit sur les problèmes de " définition rigoureuse " de la médiane dans l'article cité reste vrai pour les quartiles, et les déciles. Ce que l'élève doit retenir, c'est que 25% de la population se situe en dessous du premier quartile (Q1), 25% au-dessus du troisième quartile (Q3), et 50% entre les deux. De même pour les déciles : on utilise surtout D1 et D9, et les 80% " centraux " se trouvent entre D1 et D9 ; D9 – D1 s'appelle d'ailleurs l'intervalle interdécile.

Excel et les quartiles
Sur le tableur Excel, la syntaxe en est =QUARTILE(plage_de_valeurs ; k) où k = 1 pour Q1 et k = 3 pour Q3. Bien entendu, k = 2 redonne la médiane. Le résultat n'est pas nécessairement celui auquel on pourrait s'attendre. Un exemple : soit une série (triée dans l'ordre croissant) de 100 valeurs, la 25ème valeur étant 39, la 26ème valeur étant 40. On a Q1 = 39,5 (médiane des 50 premières valeurs). Excel donne Q1 = 39,75…
Quant aux déciles, ils n'existent pas en tant que tels ; il faut utiliser les centiles, avec la syntaxe =CENTILE(plage_de_valeurs ; k) où cette fois k est un nombre compris entre 0 et 1. D1 s'obtient en prenant k = 0.1 et D9 en prenant k = 0.9. Avec toujours les mêmes " surprises " : si le 10ème nombre d'une série de 100 vaut 22 et que le 11ème vaut 23, Excel annoncera D1 = 22.9 (et pas 22.5).

Les boites à moustaches

On passe de cette représentation (qui correspond aux définitions de la médiane et des quartiles) :



à celle-ci :

Il est bien entendu qu[illegible]tique, et que

l'axe horizontal doit r[illegible]orrespond aux 50% " centraux " (ou mieux, " médians "). Des moustaches très courtes indiquent une très forte concentration d'individus sur un petit intervalle, au contraire de moustaches très longues (voir ci-après).

Les boites à moustaches permettent de comparer très facilement des échantillons correspondant au même caractère statistique, en les plaçant parallèlement les unes aux autres, **relativement au même axe gradué**. Selon les auteurs, ces boites sont placées horizontalement ou verticalement : les élèves doivent avoir rencontré les deux, pour ne pas être 'surpris' le jour de l'examen.

Attention, il y a des exemples où médiane et boite à moustache ne sont pas du tout des indicateurs (résumés) pertinents :

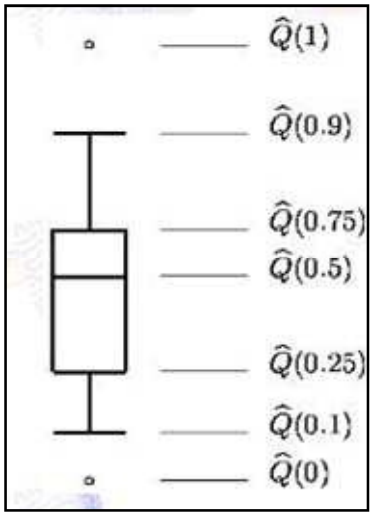
1er exemple : calculer des déciles sur une série de 25 notes d'élèves !

2ème exemple : dans ma commune il y a 174 familles sans enfant, 265 familles de 1 enfant, 207 familles de 2 enfants, 88 familles de 3 enfants, 17 familles de 4 enfants, etc. Le meilleur "résumé" que l'on puisse faire est un petit tableau reprenant l'intégralité de l'information, ou de le remplacer par un diagramme en bâtons. A la rigueur on pourrait donner le nombre moyen d'enfants par famille. Mais surtout pas une boite à moustache !!!

Faut-il tronquer les extrémités des moustaches ?

Prenons encore un exemple : dans une série statistique A de 100 valeurs (ordonnées), les 25 dernières valeurs sont 66, 66, 67, 67 ….. 85, 86, 88, 91, 91, 92, 94, 95, 96 et 97 ; dans une autre série, B, les 25 dernières valeurs sont 66, 66, 67, 67 ….. 85, 86, 88, 91, 91, 92, 94, 95, 96 et 124. L'étendue du dernier quartile vaut 31 dans la série A, et 58 dans la série B ; de même l'étendue du dernier décile vaut 12 dans la série A, et 39 dans l'autre. On se rend compte combien un seul élément (ici le maximum) augmente la longueur de la moustache : autant la médiane et les quartiles (et donc l'intervalle inter-quartile) sont des indicateurs **stables** pour des variations des valeurs du caractère, autant les " moustaches "

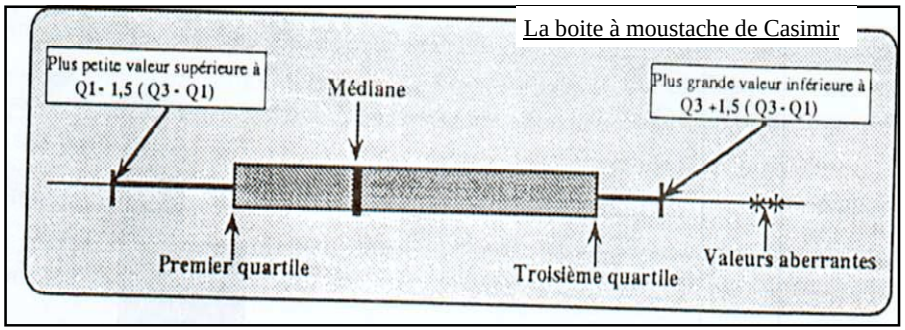
sont sensibles à une modification des valeurs extrêmes (ce qui importe, c'est donc bien le corps de la DAM, et pas sa moustache !).
C'est pourquoi les statisticiens ont décidé de tailler les " pointes " des moustaches. Les uns (c'est le cas du G.P.E.S., présidé par Claudine Robert, qui écrit les programmes actuels de lycée) coupent 10% de chaque côté (ils arrêtent donc les moustaches à D1 et à D9, *voir schéma ci-contre, issu de http://www.inrialpes.fr/sel/*), les autres (c'était le cas de Tukey, 'inventeur' de ces diagrammes, et de la nouvelle version de 'Casimir', logiciel d'exploitation de l'évaluation à l'entrée en sixième, *voir schéma ci-dessous*) ôtent tout ce qui dépasse 1,5 fois l'intervalle inter-quartile (c'est la méthode reprise par certaines calculatrices graphiques), d'autres – moins nombreux – retirent 5% en tout (soit 2,5% de chaque côté). Nous ne prendrons parti ni pour les uns, ni pour les autres, ni pour les 'intégristes' qui ne veulent rien couper…



Bien sûr, pour que l'on se rende compte de ce qui a disparu de la série, on signale sur l'axe les points où se trouvaient le maximum et le minimum, et - le plus souvent - les points correspondant aux abscisses des valeurs ainsi " supprimées ".
En outre, nommer "valeurs aberrantes" les valeurs que l'on a fait disparaître peut être admissible quand il s'agit de relevés de mesures ou de contrôles de fabrication, mais pose des problèmes d'éthique quand il s'agit des poids ou des scores des élèves.

Cependant, il faut bien comprendre que lorsqu'on procède ainsi, un modèle de répartition gaussienne est implicitement sous-jacent. Enlever 10% de chaque côté correspond (approximativement) à s'arrêter à la moyenne plus ou moins 1,3 fois

La boite à moustache de Casimir



l'écart-type. Mais arrêter la moustache aux valeurs déterminées par le graphique précédent (*Casimir*) correspond, sur une distribution " normale ", à ne considérer comme " aberrantes " que 0,7% des valeurs (0,35% de chaque côté) ; ce qui me paraît plus raisonnable que d'en enlever 20% en tout…

Or toutes les séries statistiques ne sont pas gaussiennes, loin de là. Prenons par exemple la population de la Meurthe et Moselle : la plus petit commune (Leménil-Mitry) a 2 habitants ; la plus grosse (Nancy) en a 103 606. La médiane vaut 263,5 (donc la moitié des 594 communes ont 263 habitants ou moins) ; 50% des communes ont entre 130 et 659 habitants (intervalle inter-quartile) ; 10% des communes ont plus de 2404 habitants. Ces 10% des communes les plus peuplées représentent, à elles seules, près de 71% de la population du département. On voit à quoi conduirait, sur de telles séries, le " taillage " des pointes de moustaches !!!

La polémique

A l'occasion de la création de la matière " Mathématique-Informatique " en 1 ère L cette année, l'APMEP a été à l'initiative une liste de diffusion d'activités dans cette classe, liste où les uns et les autres peuvent s'exprimer. Voici quelques extraits de messages relatifs à la médiane et aux boites à moustaches.

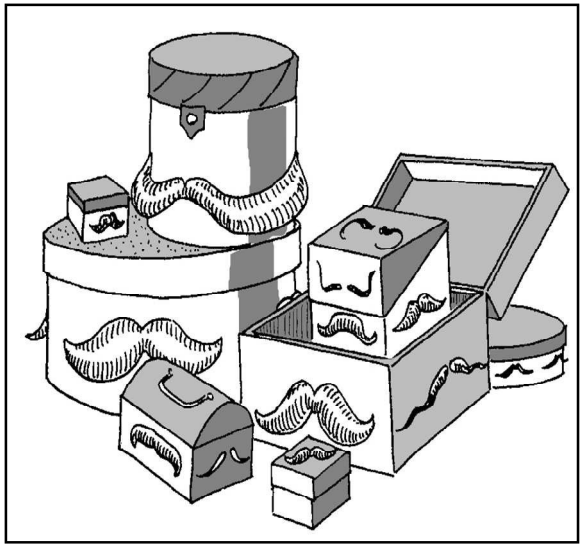
Les BAM ont un intérêt qui justifie leur présence dans le programme : elles permettent de comparer 2 séries statistiques en un coup d'œil. L'exemple classique est la comparaison des suites obtenues en lançant un dé plusieurs fois ; la comparaison des BAM des séries de 100 lancers et de 500 lancers illustre de façon spectaculaire la fluctuation d'échantillonnage ; pour ma part je regrette qu'elles ne soient pas dans le programme de seconde. (Rémy Coste, 10/01/01, en réponse à un " détracteur ").

Je reviens sur les problèmes de définition pour médiane, quartiles et déciles.
J'ai considéré, presque inconsciemment, que le texte du GEPS (ex-GTD) du 1/12/00 induisait, avec sa définition des quantiles, que nous n'avions pas à proposer aux élèves de 1ère L de calculs sur des séries statistiques dont les données sont regroupées en classes puisque celles-ci supposent, pour les calculs, l'utilisation d'interpolations...
Ni le programme officiel de 1ère L, ni le document d'accompagnement du même programme n'abordent explicitement le cas des séries statistiques dont les données sont regroupées en classes.
(…) J'ai quand même relu quelques textes officiels :
- Dans le programme de 4ème, en compétences exigibles, on peut lire : " Calculer une valeur approchée de la moyenne d'une série statistique regroupée en classes d'intervalles ".
- Dans le document d'accompagnement du programme de seconde il est écrit : " Estimer la moyenne de séries de données quantitatives en les regroupant par classes n'est plus une pratique utile en statistique depuis que les ordinateurs calculent la moyenne de milliers de données en une fraction de seconde ".
Vérité en 4ème... Archaïsme en 2de... Il est vrai que ce programme de 4ème est entré en application …en septembre 1998... Il est temps de le mettre au rayon des soldes ! (Michel

Moriceau, 14/01/01)

Moi je trouve qu'on commence à en faire (à nous en faire faire...) un peu trop de ces BAM que (presque) personne ne connaissait, il y a seulement un an... Un peu trop avec les statistiques en général, oh ! excusez-moi, avec LA statistique ! (comme il y a eu LA mathématique à l'époque bourbakiste ; autre temps... autre mode...). (Michel Moriceau, 17/01/01)

Pardon de jouer les rabat-joie (ou les social-traîtres comme vous voudrez), mais je ne partage pas du tout (et pour tout dire je ne comprends pas) cette animosité persistante contre l'introduction d'une part somme toute bien modeste de la statistique en lycée. Je ne veux pas faire un plaidoyer qui risque de ne pas servir à grand chose (…). Je voudrais simplement dire ceci : dans la communauté scientifique au sens large (physiciens, biologistes, chimistes, médecins, géographes, historiens, économistes, etc.), tous, et à tous les niveaux, se sont chaudement félicités de cette décision. J'ai vraiment le sentiment que nous pêchons par orgueil (nous les matheux), au point que nous ne nous daignons même pas regarder ce qui se passe ailleurs. TOUS les scientifiques utilisent et ont besoin des statistiques. Allons-nous continuer à nous draper dans notre dignité et condamner un pan entier des mathématiques au motif d'être utile, voire d'être des maths appliquées (quelle horreur !). Les mathématiques sont-elles devenues un dogme intouchable et sacré ? Qu'il y ait une partie (réduite) de maths appliquées dans l'enseignement des maths d'un lycéen me semble indispensable. C'est une composante essentielle de la formation scientifique. Nous pouvons bien sûr décréter que ce n'est pas aux profs de maths de le faire... au risque de nous isoler définitivement de tous les autres scientifiques, et, à terme, de la société.
Voilà pour ma réaction épidermique que l'on voudra bien me pardonner !
P.S. La "vraie" définition de la médiane ou les boîtes à moustaches ne sont vraiment que des points de détails dans les programmes. Qui en fait tout un fromage ? Il me semble qu'il vaut mieux dépenser notre énergie pour réclamer une formation solide, tant théorique que pédagogique, pour tous les profs de maths. (Rémy Coste, 22 janvier, en réponse à Michel et à d'autres).



ANNEXE
Un exemple de contrôle sur les médianes, boites à moustaches, etc.

Exercice 1

Voici une série de 160 valeurs. Ces valeurs sont triées :

26	27	28	28	28	28	29	29	30	30	31	32	32	32	32	33	33	33	34	34
34	34	34	34	34	34	35	35	35	35	35	35	35	35	35	35	36	36	36	36
36	36	36	36	36	36	36	36	37	37	37	37	37	37	37	37	37	37	38	38
38	38	38	38	38	38	38	38	39	39	39	39	39	39	39	39	40	40	40	40
40	40	40	40	40	40	41	41	41	41	41	41	41	41	41	41	41	41	41	41
41	41	41	42	42	42	42	42	42	42	42	42	42	42	42	42	42	43	43	43
43	43	43	43	43	43	43	43	43	43	43	44	44	44	44	44	44	45	45	45
45	45	45	45	45	45	46	46	46	46	47	47	48	48	49	49	50	50	50	52

1°) Calculer la médiane, les quartiles et les déciles (le 1 ^{er} et le 9ème) de cette série, en expliquant clairement comment vous procédez.
2°) Dessiner les 2 boites à moustache (l'une non élaguée, l'autre élaguée aux déciles) correspondant à cette série de valeurs.

Exercice 2

Une association de consommateurs relève les poids des baguettes dans trois boulangeries. En principe, les baguettes devraient peser 250 grammes.

Relevé de la boulangerie A :
Relevé portant sur 1248 baguettes. Poids minimum : 215 g. Poids maximum : 285 g. Poids médian : 250 g. Quartiles : 235 g et 265 g.

Relevé de la boulangerie B :
Relevé portant sur 908 baguettes. Poids minimum : 215 grammes. Poids maximum : 285 g. Poids médian : 255 g. Quartiles : respectivement 250 et 260 g.

Relevé de la boulangerie C :
Relevé portant sur 1035 baguettes. Poids minimum : 230 grammes. Poids maximum : 270 g. Poids médian : 250 g. Quartiles : respectivement 245 et 257 g.

1°) Réaliser les trois boites à moustache (non élaguées) sur le même graphique, l'une au-dessus de l'autre.

2°) Commenter les affirmations suivantes :

a) Comme le poids " normal " est 250 g, seul le boulanger B est dans la norme, c'est à dire qu'il est en règle vis à vis de la répression des fraudes.

b) Si vous achetez votre baguette chez le boulanger A, il y a une chance sur deux qu'elle ne fasse pas les 250 g attendus.

c) Si vous achetez votre baguette chez le boulanger C, il y a une chance sur deux qu'elle ne fasse pas les 250 g attendus.

d) Celui qui a le moins de variation dans ses poids est le boulanger C.

e) A la boulangerie B, les trois quarts des baguettes font au moins le poids réglementaire.

f) Il vaut mieux acheter son pain chez les boulangers A ou B, car le poids maximum relevé est 285 g (contre 270 g chez C).

g) Chez le boulanger C, les baguettes pèsent en moyenne 256 grammes.

QUELQUES COMMENTAIRES

On pourra discuter sur l'opportunité d'un contrôle entièrement consacré aux médianes, quartiles, boites à moustaches... Mais j'ai été pris par le temps : il me fallait rendre mes notes du trimestre pour le 28/02, et j'ai annoncé aux élèves le 27 qu'ils auraient un contrôle le lendemain.

En ce qui concerne les **boites à moustaches**, elles sont bien réussies par 14 élèves sur 20. Deux élèves ont dessiné des boites en quatre parties de longueurs égales (puisqu'il y a quatre fois 25%) : leurs boites sont toujours superposables, quelle que soit la série donnée. Pour les autres, l'erreur la plus fréquente consiste en une échelle assez " élastique " des valeurs de la variable.

En ce qui concerne le **premier exercice**, 15 élèves sur les 20 ont calculé de la façon suivante : il y a 160 valeurs, la moitié cela fait 80, le quart cela fait 40 ; la médiane est donc la 80ème valeur, le premier quartile la 40ème valeur, etc. Alors qu'il aurait fallu prendre " entre " la 80ème et la 81ème, entre la 40ème et la 41ème, etc. Il est vrai que cela n'avait aucune incidence sur le résultat, puisque ces valeurs étaient ex æquo. Le premier exercice que l'on avait fait en classe (puis en TD info) portait pourtant sur une série de 100 valeurs, où j'avais vu les élèves faire les choses correctement.

En ce qui concerne le **second exercice**, j'ai été surpris par la longueur des justifications et explications apportées, parfois jusqu'à 5 ou 6 lignes à chaque item. Mais ces explications sont souvent confuses, et sans aucun rapport avec ce qui est demandé

Question a : Une seule élève a répondu correctement : aucun boulanger n'est en règle, puisque tous vendent des baguettes de moins de 250 g. Beaucoup disent que "tout va bien", puisqu'il y a plutôt plus de baguettes au-dessus du poids

réglementaire...

Question b et c : Il n'y avait aucune ambiguïté, la moitié des baguettes étant au-dessus de 250 g et l'autre moitié au-dessous. Beaucoup de réponses font intervenir, dans une savante alchimie, les quartiles, les extrêmes, le nombre de baguettes testées ; en somme, " l'âge du capitaine (cf. Stella Baruk). Six élèves sur les 20 ont trouvé qu'une des deux affirmations était vraie et l'autre fausse.

Question d : Selon que l'on s'intéressait à l'étendue ou à l'intervalle inter-quartile, la réponse différait. J'ai bien sûr considéré comme correcte toute réponse clairement argumentée.

Question e : Il n'y a aucune ambiguïté possible ici (c'est la définition même du quartile), et je m'attendais à beaucoup de réponses exactes. Je n'en ai eu que 7 (je ne compte pas comme bonnes les réponses du type " C'est vrai, car l'intervalle inter-quartile est de 10 grammes " ou autres).

Question f : Je m'attendais à une variété de réponses plus ou moins floues ou ambiguës, je n'ai pas été déçu. Certaines ont même fait intervenir le prix de vente dans leur argumentation...

Question g : Je considère seulement 5 réponsescomme correctes. La majorité a répondu "Non, c'est 250 g" (erreur attendue). Je cite une des réponses : *" C'est faux : les baguettes ne sont pas proportionnelles les unes aux autres à chaque fois, à cause des quartiles ".*

Les notes de mes 20 élèves s'échelonnent de 1 à 18, avec une médiane de 9,75 (entre 9,5 et 10 !), et 1/3 de la classe entre 13,5 et 16.
Celui qui a 1 n'a écrit que ces trois lignes sur sa copie :

La médiane = 40
Les quartiles = 25%
Les déciles = 10%.