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).

HUMEUR, par JEAN-MICHEL THENARD (Libération 30-31/12/2000).

LE MILLÉNAIRE, C'ÉTAIT HIER

Champs Elysées piétonniers, Concorde en lumière, tour Eiffel bleutée, tiens, voilà le millénaire nouveau. Et tout le monde s'en fiche comme des Tiberi. Car le millénaire, c'est l'an dernier qu'on l'a fêté. A l'hiver 1999, avec son cortège de bugs annoncés pour la première seconde de l'année 2000, chiffre rond à se pâmer que le *merchandising* a alors décliné sur tous les tons. 2001, en revanche, ça ne ressemble à rien. Ni à un changement de siècle, ni même à une page qui s'ouvre, depuis que les Arabes ont inventé le zéro. Alors, bien sûr, les esprits doctes continuent à râler. "*La minorité intelligente de cette planète célébrera le 1er janvier200l comme le vrai début du XXI^{e} siècle et du troisième millénaire*", vient encore de clamer Arthur C. Clarke, écrivain de science-fiction qui, du haut de ses 83 ans et de son titre de sir britannique, ne craint plus le ridicule. Il a une excuse : auteur de la nouvelle qui inspira *2001, odyssée de l'espace*, le bonhomme est un brin *deux-mille-un-centré*. L'an dernier déjà, il n'avait pas épargné de ses sarcasmes ceux qui avaient salué trop tôt à son goût l'entrée dans le nouveau siècle. D'autres avec lui avaient expliqué que le XXI^{e} siècle ne commençait pas là où le sens commun l'entendait, qu'il faudrait attendre un an pour pénétrer le troisième millénaire. Et de plaider le système métrique, la logique arithmétique et tout ce que l'on veut. Pourquoi pas? Mais le fait est là, indiscutable, indécrottable: c'est l'an dernier que la planète a dansé, chanté, enterré 2000 ans d'un coup, et aujourd'hui personne ne se trémousse à l'idée qu'il ne reste plus que 999 années jusqu'au prochain réveillon du millénaire. Comme quoi, belle leçon, la loi du plus grand nombre parfois s'impose à l'esprit prétendu savant. Faut-il voir dans cette modernité un péché contre l'esprit scientifique, un retour à l'obscurantisme? Pas même, puisque le calendrier chrétien n'est pas plus rond que plat. Et pourtant il tourne, en dépit des erreurs de calcul de Denis le Petit, son créateur, qui ignorait le zéro et bien d'autres choses... Alors, un an après, la grande querelle est tranchée: le troisième millénaire a bien débuté le 1er janvier 2000. N'en déplaise aux érudits et autres Trissotin des temps présents, rien ne se décrète contre la multitude, pas même le passage des siècles. Signe, sans doute, que celui qui s'ouvre sera démocratique ou ne sera pas.

QUELQUES NOUVELLES DU GROUPE “JEUX” REGIONAL

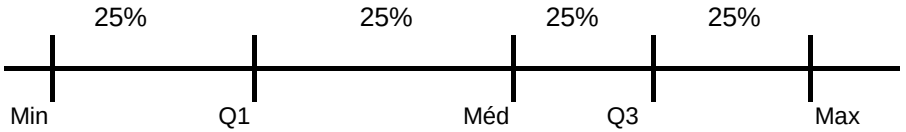
Le tome 2 de la brochure "Objets mathématiques" (dont la décision d’édition avait été prise lors du séminaire régional de Pierre-Percée en juin dernier) est en voie d'achèvement.
Vous y trouverez (ou retrouverez) le cube Soma, les pentaminos et les pentacubes plats, le tangram et un puzzle hexagonal, des développements de solides à colorier, des boites de dominos à ranger, ainsi que les 24 carrés de Mac Mahon. Nous avons essayé de privilégier des activités abordables par des élèves en difficulté, mais dans lesquelles les très bons élèves trouveront aussi leur intérêt.
Nous préparons des panneaux et des objets à manipuler pour compléter les 10 stands actuels de notre exposition régionale. La brochure sera terminée pour les journées nationales de Lille, pendant lesquelles nous présenterons la version agrandie de cette exposition "Objets mathématiques" .

Nous aimerions bien pouvoir faire circuler aussi ces nouveaux stands dans les quatre départements lorrains. Y aurait-il des adhérents bricolant le bois pour nous aider dans la réalisation des objets?
Contacter François DROUIN, Collège Les Avrils, 55300 Saint Mihiel (f.drouin@ac-nancy-metz.fr).

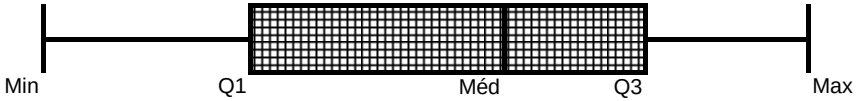
PLAN DE FORMATION ACADÉMIQUE

Le “ groupe technique mathématique ” s'est réuni le 2 février. L'ordre du jour avait pour but d'examiner les offres de formation répondant à l'appel d'offres envoyés dans les établissement au 1er trimestre
Les offres sont nettement plus nombreuses et variées que l’an passé :
44 stages proposés contre 26 l'année dernière (sans compter les Journées Régionale et Nationales de l'APMEP) ; 42 ont été retenus. Quelques aménagements vont être proposés : augmentation ou diminution de durée, regroupements de stages (c'est le cas en particulier en statistique pour la seconde, où deux formations étaient proposées : les deux formateurs vont réunir leur offre).
Quelques nouveautés intéressantes : liaison Term S- Deug MIAS, liaison Ecole-Collège sur des thèmes très précis...

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 que cela n’a de sens que pour un caractère numérique, et que

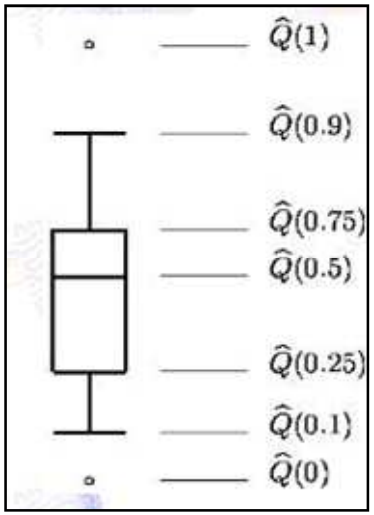


l’axe horizontal doit respecter la graduation. Le rectangle ici grisé correspond aux 50% “ centraux ” (ou mieux, “ médians ”). Des moustaches très courtes indiquent une très forte concentration d’individus sur un petit intervalle, au contraire de moustaches très longues (voir ci-après).
Les boites à 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 ”

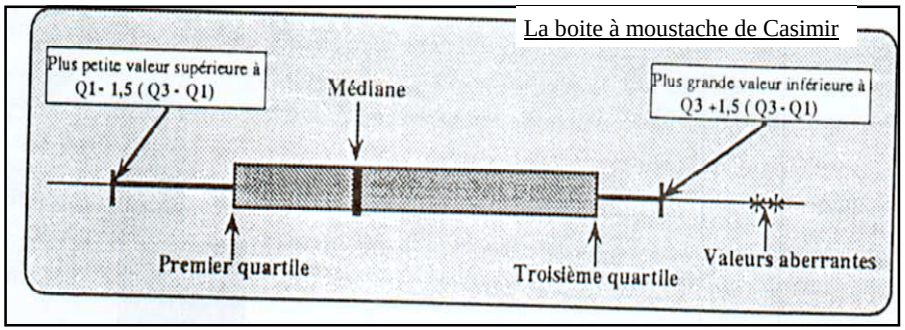
sont sensibles à une modification des valeurs extrêmes (ce qui importe, c'est donc bien le corps de la DAM, et pas sa moustache !).
C'est pourquoi les statisticiens ont décidé de tailler les " pointes " des moustaches. Les uns (c'est le cas du G.P.E.S., présidé par Claudine Robert, qui écrit les programmes 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



Le langage de cet algorithme est purement imaginaire, mais il s'inspire du C ou du Java.
Au bout d'un certain temps (500 fois 6 000 000, ça fait quand même 3 milliards de tirages), le résultat s'affiche :

```
Sur 500 expériences, écart moyen = 958,2
Maximum de l'écart : 4066 ; minimum de l'écart : 9
```

Et on est satisfait, car cela correspond au résultat théorique !

Si on avait recherché, par simulation, la fréquence d'un écart supérieur à $\sqrt{n}/2$ (ce qui correspond aux dires de B. PARIZOT, voir ci-dessous), on aurait trouvé environ 1/3 : ce n'est pas un événement si rare que ça.

Mais… car il y a un mais !
Bernard PARIZOT a écrit « *Par le simple fait du hasard, c'est-à-dire dans le cas où aucune opinion ne guiderait les votes, on devrait donc obtenir un écart de voix de l'ordre de racine carrée de 6 millions, c'est-à-dire environ 2500 voix* ».
Ce qui ne correspond ni à ce qu'on a calculé, ni à ce qu'on a simulé.
On a certainement mal interprété son texte : il parlait certainement de l'écart entre les deux candidats, et non de l'écart entre le nombre de voix de l'un des deux et 3 000 000. Il faut donc doubler tous les résultats trouvés auparavant.

Mais cela ne donne toujours qu'un espérance de $\sqrt{\frac{2n}{\pi}}$, soit environ 1 954 voix.
L'explication tient peut-être dans cette phrase : « *les fluctuations statistiques sont de type gaussien, avec une largeur relative typique en racine de 1/N* ». Ce qui correspond à la « plage de normalité » dont il est question dans le programme de seconde à propos des fluctuations d'échantillonnage : quand ou joue à 'pile ou face', on a en effet 95% de chances d'observer, dans l'échantillon tiré, que la proportion de 'PILE' est dans l'intervalle

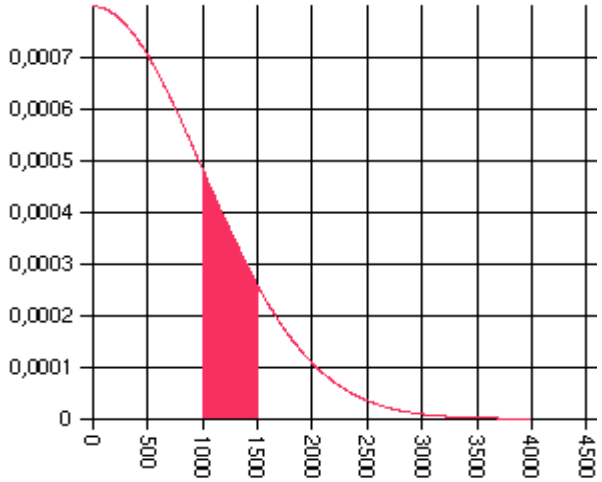
$$\left]\frac{1}{2}-\frac{1}{\sqrt{n}};\frac{1}{2}+\frac{1}{\sqrt{n}}\right[$$

.
L'auteur pose inconsciemment le problème de la vulgarisation scientifique, qui oblige parfois à une trop grande simplification du propos pour qu'il puisse être compris par tous. Mais le cœur du message est présent : l'écart est dans l'ordre de grandeur de $\sqrt{n}$…

Jacques VERDIER

P.S. Dans un message électronique du 27/01/01, Michel BRISSAUD m'écrivait ceci : « *En conclusion, c'est très compliqué, et je continue à penser qu'un approche correcte de la fluctuation d'échantillonnage est impossible en classe de seconde : il faut rester à un niveau très élémentaire.* »

L'image ci-dessous est la densité de 2D (différence en valeur absolue entre le nombre d PILE et le nombre de FACE (pour n = 1 000 000).
Nous l'avons trouvée sur le site personnel de Michel Brissaud, qui a lui aussi traité ce problème de l'écart des voix (avec l'aide d'un tableur). Nous l'en remercions et nous vous conseillons de consulter sa page personnelle :
<http://www.perso.wanadoo.fr/michel.brissaud>



Le point de vue du simulateur (!)
Le simulateur est l'auteur de ces lignes : il ne sait ni calculer des intégrales (fussent-elles de Wallis, de Futuna ou d'ailleurs), ni manier correctement la combinatoire. Il a donc réalisé un petit programme pour simuler 6 000 000 de tirages à 'pile ou face', et regardé l'écart (en valeur absolue) entre le nombre de 'PILE' et 3 000 000. Et il a répété 500 fois cette expérience aléatoire.
Voici l'algorithme :

```
nbvotants=6000000 ; nbA=0, écart=0, totalécart=0;
écartmax=0, écartmin=nbvotants; écartmoy=0;
entrer nbrepet; (équivalent d'un input)
for(j=0;j<nbrepet;j=j+1)
     { nbA=0; écart=0;
     for (i=0;i<nbvotants;i=i+1) if (random()<0.5) nbA=nbA+1;
     écart=abs(nbvotants/2-nbA); totalécart=totalécart+écart;
     if (écart>écartmax) écartmax=écart;
     if (écart<écartmin) écartmin=écart;
     } (fin de la boucle for)
écartmoy=totalécart/nbrepet;
print("Sur "+nbrepet+" expériences, écart moyen =
"+écartmoy);
print("Maximum de l'écart : "+écartmax+" ; minimum de
l'écart : "+écartmin);
```


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 1ère 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.

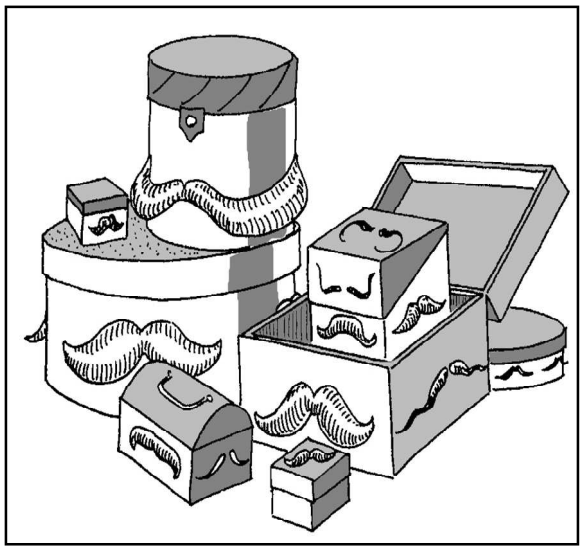
Les BAM ont un intérêt qui justifie leur présence dans le programme : elles permettent de comparer 2 séries statistiques en un coup d'œil. L'exemple classique est la comparaison des suites obtenues en lançant un dé plusieurs fois ; la comparaison des BAM des séries de 100 lancers et de 500 lancers illustre 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 2nde... 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).



D'où E(D) = $\frac{1}{2^{n-1}}\sum_{k=1}^{n/2} k.C_n^{n/2+k}$

Et par un astucieux calcul (si l'on sait bien manier la combinatoire), on trouve finalement que E(D) = $\frac{n.C_n^{n/2}}{2^{n+1}}$

L'inconvénient est qu'on ne sait pas calculer effectivement des combinaisons portant sur de si grandes valeurs (n'oublions pas que, dans notre exemple, n = 6 000 000). L'analyste va donc chercher la limite de cette expression lorsque n→∞. Et cela en utilisant les intégrales de Wallis $\omega_k = \int_0^{\pi/2} \sin^k(x).dx$, sachant que $\omega_n = \frac{\pi}{2} \times \frac{C_n^{n/2}}{2^n}$. On trouve donc, in fine, que E(D) est de l'ordre de grandeur de $\sqrt{\frac{n}{2\pi}}$ lorsque n est 'infiniment grand'.

A propos des intégrales de Wallis, consultez l'excellent ouvrage de B. et A. PARZYSZ *Fonctions d'une variable* dans la collection TD chez Dunod. Le calcul ci-dessus a d'ailleurs été fait par Bernard PARZYSZ, que nous remercions.

Le point de vue du probabilitste-statisticien

Le statisticien, lui, sait que la loi binomiale B(n ; ½) de X (nombre de 'PILE') est approchée, dès que n est suffisamment grand, par une loi normale N(n/2 ; √n/2).

La fonction de répartition de D est définie par $F_D(x) = p(|X - n/2| < x) = p(n/2 - x < X < n/2 + x)$ pour x≥0.

Soit $F_D(x) = F_X(n/2 + x) - F_X(n/2-x)$. En dérivant, il vient :

$f_D(x) = f_X(n/2 + x) - f_X(n/2-x)$, où f_X est la densité de X, dont on disait qu'on l'avait assimilée à une loi normale.

On trouve donc $f_D(x) = \frac{2\sqrt{2}}{\sqrt{n\pi}} \times e^{\frac{-2x^2}{n}}$

.D'où $\int_0^{\infty} x.f_D(x).dx = \frac{\sqrt{n}}{\sqrt{2\pi}} \times \left[-e^{\frac{-2x^2}{n}} \right]_0^{\infty}$.

Ce qui, in fine, redonne bien E(D) = $\sqrt{\frac{n}{2\pi}}$. Merci à Daniel Vagost et Franck Gaüzère, du département S.T.I.D. de l'I.U.T. de Metz, d'avoir ment à bien ce calcul.