DÉFINITION DES QUARTILES D'UNE SÉRIE STATISTIQUE

27 février 2009

On peut sauter le texte en regard d'un trait épais dans la marge.

Cette note est suivie des articles :

Quartiles et dispersion, Incertitude liée à une mesure et Boîte à moustache.

1 Définition des quartiles

Est donnée une suite de n nombres réels :

$$(\mathcal{S}) = (x_i \ \ 1 \leq i \leq n) = (x_1, \cdots, x_n)$$

qu'on appelle une série statistique [1] ; n est sa taille.

Définition 1.1 [Premier quartile (empirique)] *C'est le plus petit élément q [2] des valeurs des termes de la série tel qu'au moins 25 % des données soient inférieures ou égales à q.* □

Il est courant de le noter q_1. Par définition, q_1 est un élément de $\mathcal{S}$, ce qui n'est pas le cas de la médiane qui peut appartenir ou non à $\mathcal{S}$.

Définition 1.2 [Troisième quartile (empirique)] *C'est le plus petit élément q [3] des valeurs des termes de la série tel qu'au moins 75 % des données soient inférieures ou égales à q.* □

Il est courant de le noter q_3. Par définition, q_3 est aussi un élément de $\mathcal{S}$.

Évidemment, tout le monde se demande où est passé q_2, qui aurait dû être défini comme suit :

Définition 1.3 [Deuxième quartile (empirique)] *C'est le plus petit élément q [4] des valeurs des termes de la série tel qu'au moins 50 % des données soient inférieures ou égales à q.* □

[1]Le vocabulaire utilisé dans cette note est conforme, sauf erreur, au Lexique du fascicule *Mathématiques classe de première des séries générales, Documents d'accompagnement des programmes*, Classe de première S, Annexe 1 Statistique et Probabilités, pages 73-74.

http ://www.cndp.fr/archivage/valid/86906/86906-13718-17372.pdf

[2]Bien sûr, il faudrait démontrer qu'un tel élément existe : c'est facile

[3]idem

[4]idem

En fait, la définition 1.3 aurait dû être choisie comme définition de la médiane dans un système cohérent de définitions. Comme celle-ci a été définie d'une manière différente, on perd le plaisir de dire que la médiane est le second quartile. On aurait de la même façon pu faire apparaître le minimum de $(\mathcal{S})$ comme le $0^{\text{ème}}$ quartile et son maximum comme le $4^{\text{ème}}$. À oublier.

L'important est de retenir la signification des premier et troisième quartiles :

> il y a en gros un quart des valeurs de la série étudiée qui sont $\leq q_1$, trois quarts qui sont $\geq q_1$; de même, il y en a en gros trois quarts qui sont $\leq q_3$, un quart qui sont $\geq q_3$.

Il est facile de vérifier que les définitions qui suivent sont équivalentes aux définitions 1.1 et 1.3 :

Définition 1.4 [Premier quartile (empirique)] q_1 *est l'unique terme de la série tel que*

$$\operatorname{Card}(i\,;\, 1 \leq i \leq n\,,\, x_i \leq q_1) \geq \frac{n}{4} \quad \text{et} \quad \operatorname{Card}(i\,;\, 1 \leq i \leq n\,,\, x_i \geq q_1) \geq \frac{3n}{4}\cdot \qquad \square$$

Définition 1.5 [Troisième quartile (empirique)] q_1 *est l'unique terme de la série tel que*

$$\operatorname{Card}(i\,;\, 1 \leq i \leq n\,,\, x_i \leq q_3) \geq \frac{3n}{4} \quad \text{et} \quad \operatorname{Card}(i\,;\, 1 \leq i \leq n\,,\, x_i \geq q_3) \geq \frac{n}{4}\cdot \qquad \square$$

Les définitions 1.1, 1.2, 1.4 et 1.5 s'appliquent à toutes les séries, même les plus artificielles, comme le montre le tableau suivant :

Série $\mathcal{S}$	Taille	q_1	q_3
(0)	1	0	0
$(1, 4, 3, 2)$	4	1	3
$(3, 2, 5, 1, 4)$	5	2	4
$(3, 6, 2, 4, 5, 1)$	6	2	5
$(1, 7, 3, 5, 4, 6, 2)$	7	2	6
$(3, 3, 2, 3, 1, 2)$	6	2	3
$(1, 3, 2, 2, 2)$	5	2	2

2 Détermination pratique de q_1 et q_3

En statistique, on travaille habituellement avec des séries statistiques longues. De ce point de vue, le tableau précédent est particulièrement caricatural. On utilise alors un tableur ou une calculette qui donneront des quartiles q_1 et q_3 faux, sauf cas fortuit ou si vous utilisez une calculatrice qui respecte nos programmes [5]. Cela vient du fait que les définitions utilisées par ces machines ou logiciels sont différentes des

[5] la TI-Collège Plus de Texas Instruments qui, de surcroît, est plus facile à utiliser

définitions 1.1 et 1.2 [6]. Cette différence de définitions n'apparaît pas dans le calcul de la médiane pour lequel on peut utiliser la fonction MEDIANE des tableurs.

On ne peut donc pas calculer les quartiles d'une série statistique à l'aide de ces machines [7]. Dans tous les cas, nous recommandons la procédure suivante, parce qu'elle montre bien la signification des quartiles :

- ordonner $(\mathcal{S})$: $\mathcal{S}' = (y_1, \ldots, y_n)$ désignera la suite ordonnée obtenue ;
- lire q_1 et q_3 au bon endroit dans la série ordonnée.

Remarque Une bonne manière de montrer aux élèves le rôle de q_1 et de q_3 est de représenter la valeur de la série statistique considérée sur un axe gradué. Une telle représentation graphique rend inutile le rangement dans l'ordre croissant de la série initiale. On ne l'utilise pas en général parce que c'est long alors que les tris avec un tableur sont instantanés et parce que cela ne marcherait que pour de petites séries statistiques.

Voici quelques exemples dans lesquels Q_1 et Q_3 désignent les valeurs des premier et troisième quartiles calculés à l'aide de la fonction QUARTILE du tableur Calc de OOo :

$(\mathcal{S})$	$(\mathcal{S}')$	q_1	q_3	Q_1	Q_3
$(-2, 1.87, -3.17, 4.5, -6)$	$(-6, -3.17, -2, 1.87, 4.5)$	-3.17	1.87	-3.17	1.87
$(-1, -4, 0, -4, 7, 5, -4)$	$(-4, -4, -4, -1, 0, 5, 7)$	-4	5	-4	2.5

Bien sûr, $(\mathcal{S})$ et $(\mathcal{S}')$ ont la même taille. On constate que les quartiles q_3 et Q_3 de la deuxième série statistique sont différents. Q_3 est faux [8]. L'exercice 2 indique quels termes de $(\mathcal{S}')$ il faut lire pour obtenir q_1 et q_3.

3 Références

1 - Programme de l'enseignement des mathématiques, des SVT, de physique-chimie du collège, B.O. N°6 19 AVRIL 2007, Hors-série, Annexe 2 Mathématiques
`ftp://trf.education.gouv.fr/pub/edutel/bo/2007/hs6/MENE0750668A_annexe2.pdf`

2 - Mathématiques Collège, Projet de document d'accompagnement – Probabilités – 17 mars 2008
`http://eduscol.education.fr/D0015/doc_acc_clg_probabilites.pdf`

4 Exercices

Exercice 1 Démontrer que $q_1 \leq q_3$ quelle que soit la série $(\mathcal{S})$.

Exercice 2 Posons

$$n = 4 \cdot q + r \quad \text{où} \quad 0 \leq r < 4 \quad \text{ce qui implique} \quad q = [\frac{n}{4}] \quad \text{et} \quad n \equiv r\ (4) \qquad (1) \quad \triangle$$

[6] sauf la machine citée ci-dessus

[7] idem

[8] Ces différences en fait ne sont pas importantes : nos définitions comme les définitions des machines traduisent de toute façon la même idée. Mais les élèves ont besoin d'avoir une réponse précise, celle des programmes.

L'égalité est l'égalité de la division euclidienne de n par 4 : q est le quotient et r est le reste ; q est la partie entière de $\frac{n}{4}$, n est congru à r modulo 4. Nous utilisons ces notations ci-dessous.

Vérifier le tableau suivant [9] :

r	0	1	2	3
q_1	$y_{[\frac{n}{4}]}$	$y_{[\frac{n}{4}]+1}$	$y_{[\frac{n}{4}]+1}$	$y_{[\frac{n}{4}]+1}$
q_3	$y_{3[\frac{n}{4}]}$	$y_{3[\frac{n}{4}]+1}$	$y_{3[\frac{n}{4}]+2}$	$y_{3[\frac{n}{4}]+3}$

Exercice 3 On suppose que tous les termes de la série statistique $(\mathcal{S})$ sont différents et que sa taille est impaire. Calculer les effectifs des valeurs de $(\mathcal{S})$ dans les intervalles $[min\,,\,q_1]$, $[q_1\,,\,m]$, $[m\,,\,q_3]$ et $[q_3\,,\,max]$ et remplir le tableau suivant :

Effectifs	$[min\,,\,q_1]$	$[q_1\,,\,m]$	$[m\,,\,q_3]$	$[q_3\,,\,max]$
$r \equiv 1\ (4)$	$[\frac{n}{4}]+1$	$[\frac{n}{4}]+1$	$[\frac{n}{4}]+1$	$[\frac{n}{4}]+1$
$r \equiv 3\ (4)$	$[\frac{n}{4}]+1$	$[\frac{n}{4}]+2$	$[\frac{n}{4}]+2$	$[\frac{n}{4}]+1$

La somme des effectifs est-elle égale à n ? Pourquoi ?

Commentaire : Le premier cas est exactement conforme à nos attentes.

[9] *Extrait du manuel d'utilisation de la TI-Collège Plus (livré avec la calculatrice) pp. 35 et 36* : Dans une série de n données, $Q1$ est la valeur classée au rang $n/4$. Si n n'est pas un multiple de 4, $Q1$ est la valeur de la série qui a pour rang l'entier immédiatement supérieur à $n/4$. $Q3$ est la valeur classée au rang $3n/4$. Si n n'est pas un multiple de 4, $Q3$ est la valeur de la série qui a pour rang l'entier immédiatement supérieur à $3n/4$. C'est exactement ce que dit le tableau.