Quartiles, déciles et centiles

Cette note est suivie des articles : « Quartiles et dispersion », « Incertitude liée à une mesure » et « Diagrammes en boîtes ».

Classes concernées : Médiane, quartiles : à partir de la Troisième ([1] p.59, [2] p.8, [3] p.175 pour la Première S), Déciles : à partir de la Première ([4] pp.85-88).

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* ou sa *longueur*.

Les quartiles dont il est question dans cette note sont quelquefois appelés *empiriques*[2].

On distinguera les valeurs de $(\mathcal{S})$ des termes ou éléments de $(\mathcal{S})$. Les valeurs de $(\mathcal{S})$ sont les valeurs des termes de $(\mathcal{S})$, chacune d'elles étant comptée une seule fois. Dans la série statistique $(1, 3, 2, 5, 2, 3)$, il y a 6 termes - sa taille est 6 - mais seulement 4 valeurs.

Définition 1.1 [Premier quartile] *C'est la plus petite des valeurs q de $(\mathcal{S})$ telles qu'au moins 25 % des termes de $(\mathcal{S})$ soient inférieurs ou égaux 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 nécessairement le cas de la médiane.

Définition 1.2 [Troisième quartile] *C'est la plus petite des valeurs q de $(\mathcal{S})$ telles qu'au moins 75 % des termes de $(\mathcal{S})$ soient inférieurs ou égaux 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] *C'est la plus petite des valeurs q de $(\mathcal{S})$ telles qu'au moins 50 % des termes de $(\mathcal{S})$ soient inférieurs ou égaux à q.* □

Bien sûr, il faudrait démontrer que q_1, q_2 et q_3 existent : c'est facile.

Remarques :

- ✓ 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 autrement, on perd le plaisir de dire que la médiane est le second quartile. Dans la suite, nous ne parlerons plus de q_2.
- ✓ On aurait pu, de la même façon, faire apparaître le minimum de $(\mathcal{S})$ comme le $0^{\text{ème}}$ quartile et son maximum comme le $4^{\text{ème}}$. Ce serait compliquer inutilement les choses.

1. Le vocabulaire utilisé dans cette note est conforme au lexique de [4], pp. 85-86
2. par opposition aux quartiles des lois de probabilité. Dans le cas particulier où $(\mathcal{S})$ est un *échantillon* d'une loi de probabilité, les quartiles empiriques sont utilisés pour estimer les quartiles de la loi. Ceci est tout à fait hors sujet.

Nous laissons au lecteur le soin de vérifier que les définitions de q_1 et q_3 ci-dessus sont respectivement équivalentes aux définitions suivantes :

Définition 1.4 [Premier quartile] q_1 *est l'unique valeur* v *de* $(\mathcal{S})$ *telle que*

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

Commentaires :

✓ Avec « Card », on voit que l'on compte des termes de $(\mathcal{S})$, non des valeurs.

✓ Autrement dit, q_1 est une valeur de $(\mathcal{S})$ qui est caractérisée par la propriété suivante : il y a au moins un quart des éléments de la série $(\mathcal{S})$ qui sont $\leq q_1$ et plus de trois quarts de ces éléments qui sont $\geq q_1$.

Définition 1.5 [Troisième quartile] q_1 *est l'unique valeur* w *de* $(\mathcal{S})$ *telle que*

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

On peut faire pour q_3 les mêmes commentaires que pour q_1. On retiendra que :

> Il y a en gros un quart des termes 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 $\geq q_3$.

et que

> q_1, m et q_3 partagent les éléments de $(\mathcal{S})$ en 4 paquets d'effectifs à peu près égaux.

Notation :

> Jusqu'à la fin de cette note, q et r désigneront le quotient et le reste de la division euclidienne de n par 4, si bien que
> $$n = 4\cdot q + r, \quad 0 \leq r < 4$$

Remarques (on peut sauter ces remarques)

✓ L'égalité parfaite des effectifs n'est réalisée que dans quelques cas. Par exemple, si tous les termes de la série statistique $(\mathcal{S})$ sont *différents* et si sa taille n est *impaire*, les effectifs des termes de $(\mathcal{C})$ appartenant aux intervalles $[min, q_1]$, $[q_1, m]$, $[m, q_3]$ et $[q_3, max]$ sont donnés par le tableau suivant :

Effectifs	$[min\,,\,q_1]$	$[q_1\,,\,m]$	$[m\,,\,q_3]$	$[q_3\,,\,max]$
cas $r = 1$	$q+1$	$q+1$	$q+1$	$q+1$
cas $r = 3$	$q+1$	$q+2$	$q+2$	$q+1$

Dans les deux cas, la somme des effectifs est $n + 3$ parce que q_1, m et q_3 sont comptés deux fois. Seul le premier cas est exactement conforme à nos attentes.

✓ Les définitions 1.1 et 1.2 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

- ✓ Une bonne manière de montrer aux élèves le rôle de q_1 et de q_3 est de représenter les valeurs de la série statistique considérée sur un axe gradué. Une telle représentation graphique revient à ordonner la série sans le dire, puisque leurs yeux balaieront l'axe, probablement de la gauche vers la droite. Évidemment, ceci est impraticable quand la série statistique considérée est longue.
- ✓ D'une manière générale, la statistique est faite pour essayer de tirer des enseignements de séries statistiques très longues, où l'on ne voit rien (fréquemment, plusieurs milliers de termes). Les activités en classe doivent en tenir compte. Les exemples ci-dessus sont caricaturaux et à éviter. Quand on a affaire à peu de nombres, on ouvre les yeux et on regarde : inutile de faire de la statistique !

2 Détermination pratique de q_1 et q_3

En statistique, on travaille habituellement sur des séries statistiques longues qui imposent l'usage d'un tableur, d'une calculatrice ou d'un logiciel de calcul *ad hoc*. Ceux-ci donneront des quartiles q_1 et q_3 faux, sauf cas fortuit ou sauf si vous utilisez une calculatrice qui respecte nos programmes[3]. Cela vient du fait que les définitions utilisées par ces machines ou logiciels sont différentes des définitions 1.1 et 1.2[4]. Cette difficulté n'apparaît pas dans le calcul de la médiane dont la définition est la définition internationale. On peut donc utiliser la fonction MEDIANE des tableurs ou les fonctions analogues des calculatrices ou des logiciels de calcul standard. Mais il faut retenir que

On ne peut pas calculer les quartiles q_1 et q_3 d'une série statistique[5] à l'aide de ces machines[6].

Pour en convaincre le lecteur, 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, q_1 et q_3 étant les valeurs exactes :

($\mathcal{S}$)	($\mathcal{S}'$)	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

On constate que les quartiles q_3 et Q_3 de la deuxième série statistique sont différents. Q_3 est faux[7].

Dans tous les cas, nous recommandons la procédure suivante, qui donne des résultats exacts et

3. la TI-Collège Plus de Texas Instruments qui, de surcroît, parle français
4. sauf la machine citée ci-dessus
5. ni son diagramme en boîte
6. sauf, de nouveau, la machine citée ci-dessus
7. On voit à quel point il est regrettable que les définitions de nos programmes soient différentes des définitions internationales, qui traduisent de toute façon les mêmes idées.

qui a l'avantage de bien montrer la signification des quartiles :

- ordonner $(\mathcal{S})$. On notera $(\mathcal{S}') = (y_1, \ldots, y_n)$ la suite ordonnée obtenue ;
- lire q_1 et q_3 dans $(\mathcal{S}')$ en appliquant les définitions 1.1 et 1.2.

Comment identifier q_1 et q_3 dans $(\mathcal{S}')$ sans avoir à réfléchir ?
Nous laissons au lecteur le soin de vérifier le tableau suivant (cf. [8]) :

r	0	1	2	3
q_1	y_q	y_{q+1}	y_{q+1}	y_{q+1}
q_3	y_{3q}	y_{3q+1}	y_{3q+2}	y_{3q+3}

Tableau 2.1

Par exemple, si $n = 4q + 3$, le troisième quartile de $(\mathcal{S})$ est la valeur du terme de rang $3q + 3$ de $(\mathcal{S}')$. Ce procédé est facilement programmable.

3 Programmer le calcul de q_1 et q_3

- ✓ Comme on l'a déjà dit, en période d'apprentissage, il vaut mieux déterminer à la main la médiane et les quartiles d'une série statistique donnée (le calculateur électronique étant seulement utilisé pour ordonner cette série).
- ✓ Ces notions étant connues, si on désire les appliquer à l'étude d'un tableau statistique intéressant, par exemple, on pourra éviter de gaspiller du temps en utilisant un algorithme qui les calcule directement.

3.1 Programmer le calcul de q_1 et q_3 avec un tableur

Programmer le tableur « Calc » d'OOo, en s'appuyant sur le tableau 2.1 ci-dessus est facile mais pas spécialement agréable. On pourra utiliser « modeleMQE », modèle de fichier ods téléchargeable et commenté, qui contient une macro calculant la médiane m, q_1, q_3 et l'étendue de la série statistique.

3.2 Programmer le calcul de q_1 et q_3 avec « scilab pour les lycées »

Le tableau 2.1 montre que l'on calcule q_1 et q_3 en tenant compte de la valeur de r. Il y a quatre cas. On peut raccourcir l'algorithme si l'on dispose d'une fonction comme « ceil » [9], [10] de « scilab ». Pour cela, ré-écrivons les définitions 1.1 et 1.2 en utilisant la série ordonnée $(\mathcal{S}')$.

Définition 3.1 *Le premier quartile d'une série statistique $(\mathcal{S})$ ordonnée en une série notée $(\mathcal{S}') = (y_1, \ldots, y_n)$ est égal à y_k si k est le plus petit entier l tel que $\frac{l}{n} \geq \dfrac{1}{4}$.* □

Définition 3.2 *Le troisième quartile d'une série statistique $(\mathcal{S})$ ordonnée en une série notée $(\mathcal{S}') = (y_1, \ldots, y_n)$ est égal à y_k si k est le plus petit entier l tel que $\frac{l}{n} \geq \dfrac{3}{4}$.* □

---

8. *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 $3n/4$.
9. Par définition, $ceil(x)$ est le plus petit entier qui est supérieur ou égal au réel x.
10. À comparer avec $E(x)$ ou $[x]$, partie entière de x, qui est le plus grand entier inférieur ou égal à x.

Ces définitions montrent que

$$\boxed{q_1 = y_k \quad \text{où} \quad k = ceil(\tfrac{n}{4}) \quad \text{et} \quad q_3 = y_l \quad \text{où} \quad l = ceil(\tfrac{3n}{4})}$$

On pourra utiliser l'algorithme « MQDC » téléchargeable et exécutable par « scilab pour les lycées » qui produit la médiane, les quartiles q_1 et q_3, les déciles d_1 et d_9 et enfin les centiles c_1 et c_{99}, voir ci-dessous. Cet algorithme est très compact grâce « ceil ».

4 Déciles

Les déciles reposent sur la même idée que les quartiles : ils divisent les termes de $(\mathcal{S})$ en 10 paquets d'effectifs à peu près égaux, au lieu de 4 pour les quartiles.

Définition 4.1 *Pour $j = 1, \ldots, 9$, on appelle j ème décile d_j d'une série statistique $(\mathcal{S})$ ordonnée en une série $(\mathcal{S}') = (y_1, \ldots, y_n)$ la plus petite valeur d de $(\mathcal{S})$ telle qu'au moins $(10 \times j)\,\%$ des termes de $(\mathcal{S})$ soient inférieurs ou égaux à d. Autrement dit, d_j est égal à y_k si k est le plus petit entier l tel que $\dfrac{l}{n} \geq \dfrac{j}{10}$, autrement dit si $k = ceil(\dfrac{n \cdot j}{10})$.* □

On utilise surtout les premier et neuvième déciles, qui renseignent sur les 10 % de plus petits termes et sur les 10 % de plus grands termes de $(\mathcal{S})$. d_1 et d_9 n'ont aucun intérêt si $n < 10$ (il s'agit dans ce cas du minimum et du maximum de $(\mathcal{S})$) !

5 Centiles

Définition 5.1 *Pour $j = 1, \ldots, 99$, on appelle j ème centile c_j d'une série statistique $(\mathcal{S})$ ordonnée en une série $(\mathcal{S}') = (y_1, \ldots, y_n)$ la plus petite valeur c de $(\mathcal{S})$ telle qu'au moins $j\,\%$ des termes de $(\mathcal{S})$ soient inférieurs ou égaux à c. Autrement dit, c_j est égal à y_k si k est le plus petit entier l tel que $\dfrac{l}{n} \geq \dfrac{j}{100}$, autrement si $l = ceil(\dfrac{n \cdot j}{100})$.* □

Bien sûr, $d_1 = c_{10}$, $d_9 = c_{90}$. c_1 et c_{99} s'utilisent pour avoir des renseignements sur les valeurs extrêmes de séries statistiques longues. Si $n < 99$, on retrouve un cas sans intérêt, comme précédemment.

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] - Programme de mathématiques de la classe de Seconde, Bulletin officiel n° 30 du 23 juillet 2009 :
http://media.education.gouv.fr/file/30/52/3/programme_mathematiques_seconde_65523.pdf

[3] - Mathématiques, Classe de Première, Série scientifique, BO n° 7, 31 août 2000 Hors-série :
http://www.education.gouv.fr/bo/2000/hs7/vol5mathsc.htm

[4] - Mathématiques, classes de première des séries générales, collection Lycée – voie générale et technologique, série *Accompagnement des programmes*
http://www.cndp.fr/archivage/valid/86906/86906-13718-17372.pdf