# Une étude didactique sur les quartiles d'une série statistique univariée

par Valérie HENRY

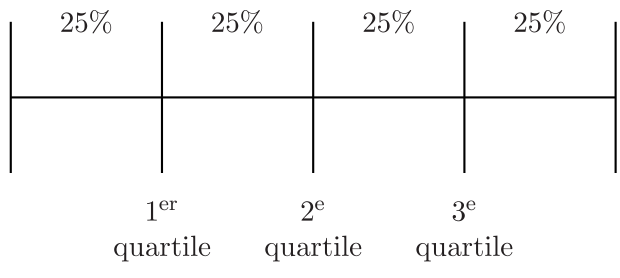
*Université de Liège*
*E-mail : V.Henry@ulg.ac.be*

**Mots-clé :** Série statistique, quartiles, médiane.

## Introduction

L'homme moderne est constamment assailli d'informations en tous genres. Parmi celles-ci, nombreuses sont celles qui se présentent sous la forme de listes d'observations numériques[1]. Pour remédier à l'incapacité de l'esprit humain d'intégrer instantanément un tableau important de nombres, le statisticien propose différentes solutions. Une première catégorie de solutions est constituée de toutes les représentations possibles de ces données : diagramme, histogramme, polygône, ogive, ... sont autant de résumés graphiques d'un ensemble de données. Dans la deuxième catégorie, on retrouve l'ensemble des résumés numériques dont chaque élément cherche à représenter une caractéristique donnée de la série considérée ; parmi ces éléments, on distingue principalement les paramètres de tendance centrale, de dispersion, de dissymétrie et de forme. Au sein de chacune de ces "classes", les paramètres possibles possèdent chacun des propriétés propres qui les rendent plus ou moins bien adaptés à décrire telle ou telle série.

Dans cette note, nous nous attachons plus particulièrement à trois de ces paramètres que sont les *quartiles*. Intuitivement et dans une première approche, les quartiles sont les nombres qui divisent une série (préalablement ordonnée) en quatre sous-séries d'effectifs égaux ([6], p. 88) :




[1]Nous ne traiterons dans cet article que les séries univariées, laissant ainsi de côté les séries multivariées dont l'étude fait appel à des techniques mathématiques plus sophistiquées ; un aperçu de telles méthodes peuvent être trouvées dans [11].

L'idée fondamentale est donc la suivante : 25% de la population se situe en dessous du premier quartile $Q_1$, 25% par-dessus le troisième quartile $Q_3$, et 50% entre les deux ([18], p. 456). Ainsi, le deuxième quartile, qui n'est autre que la médiane, fournit une valeur centrale de la série étudiée, tandis que les deux autres quartiles rendent compte de la dispersion et de la symétrie des valeurs situées au centre de la série observée.

Après quelques réflexions générales, nous proposons une analyse comparative des définitions des quartiles proposées dans la littérature et mettons en évidence les divergences de vues des différents auteurs. Nous nous intéressons également aux méthodes de calcul utilisées par différents logiciels statistiques ou plus généraux pour fournir les quartiles.

Dans la troisième partie, sur base de nos constatations, nous proposons une définition du premier quartile qui permette de rester en cohérence avec la définition de celui-ci pour une variable quantitative continue.

# 1 Contexte

## 1.1 Notation et conventions

Nous allons fixer notre attention sur une série statistique

$$S = \{x_1, x_2, \dots, x_n\}$$

dont les éléments $x_i$ ne sont pas nécessairement distincts.

Lorsque nous ordonnerons les éléments de $S$ par valeurs croissantes, nous travaillerons sur la série ordonnée

$$\bar{S} = \{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$$

où $x_{(i)} \leq x_{(j)}$ pour tous $i, j$ tels que $i < j$. Ainsi, chaque élément $x_i$ de $S$ se voit associer un rang $r_i$ qui désigne sa position dans $\bar{S}$ et est donc défini par $x_i = x_{(r_i)}$.

La *fréquence cumulée* d'un élément $x_i$ est la proportion de valeurs de $S$ qui sont inférieures ou égales à $x_i$ ; plus généralement, pour un réel $x$ arbitraire, nous noterons $F(x)$ la proportion des éléments de $S$ qui sont inférieurs ou égaux à $x$. De même, nous désignerons par $F^*(x)$ le pourcentage des éléments de $S$ qui sont supérieurs ou égaux à $x$. Dès lors, le pourcentage des observations qui sont inférieures (resp. supérieures) à $x$ est égal à $1 - F^*(x)$ (resp. $1 - F(x)$). Le graphe de la fonction *de répartition* d'équation $y = F(x)$ est appelé la *courbe cumulative des fréquences* dans le cas discret et l'*ogive des fréquences cumulées* dans le cas continu.

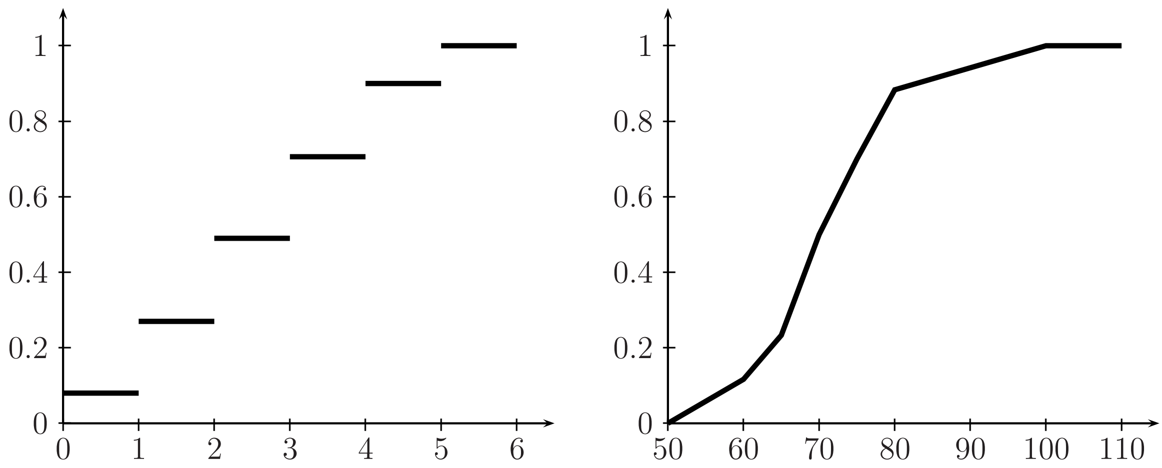



Figure 1: Courbe cumulative des fréquences et ogive des fréquences cumulées

Par ailleurs, pour tout nombre réel $r$ non négatif, nous noterons $\lfloor r \rfloor$ la partie entière de $r$, c'est-à-dire le plus grand entier inférieur ou égal à $r$, tandis que $\lceil r \rceil$ désignera le plus petit entier supérieur ou égal à $r$. La partie décimale de $r$ vaut donc $r - \lfloor r \rfloor$ ; de plus, $\lfloor r \rfloor = \lceil r \rceil = r$ si et seulement si $r$ est lui-même un entier.

## 1.2 La médiane

Rappelons que, dans le cas d'une variable quantitative discrète, la *médiane* $\tilde{x}$ de la série $S$ coïncide avec l'observation $x_{(k+1)}$ lorsque l'effectif $n$ est le nombre impair égal à $2k+1$ ; par contre, il s'agit du nombre $\frac{x_{(k)}+x_{(k+1)}}{2}$ lorsque $n$ est le nombre pair $2k$ ([13]). De manière équivalente, la médiane peut être, dans tous les cas, définie par cette unique formule :

$$\tilde{x} = \frac{x_{(r_1)} + x_{(r_2)}}{2}$$

où les rangs $r_1$ et $r_2$ sont définis par

$$r_1 = \lfloor \frac{n+1}{2} \rfloor = \lceil \frac{n}{2} \rceil \text{ et } r_2 = \lfloor \frac{n+2}{2} \rfloor = \lceil \frac{n+1}{2} \rceil$$

Dans le cas où la variable considérée est de type quantitatif continu, la définition de la médiane est très simple puisqu'elle s'obtient immédiatement à partir de la fonction de répartition $F$ : la médiane est la valeur $\tilde{x}$ telle que $F(\tilde{x}) = 0,5$.

# 2 Définitions du premier quartile dans la littérature

Intuitivement, il semble relativement facile de définir les quartiles, l'idée première étant de suivre le même type de procédé que pour la médiane. Néanmoins, dans le cas discret, cette première impression se heurte rapidement à plusieurs difficultés notamment liées à l'effectif de la série. Dans ce paragraphe, nous examinons les différentes définitions relevées dans la littérature sur le sujet. Nous nous contenterons de nous attacher au premier quartile $Q_1$, laissant aux lecteurs le soin de traiter $Q_3$ par symétrie.

Nous avons, dans un souci de clarté, classés les défintions répertoriées selon trois grandes catégories : la première où les définitions font appel aux fréquences cumulées ou à des notions apparentées, la seconde utilisant le rang des observations et la dernière basée sur la médiane.

Signalons que les définitions qui vont être données se retrouvent parfois dans la littérature sous des formes légèrement différentes ; par exemple, certains auteurs utilisent les effectifs cumulés au lieu des fréquences cumulées.

Dans les définitions qui suivent, les noms cités entre crochets sont donnés à titre indicatif, certains auteurs introduisant d'ailleurs les quartiles de plusieurs manières ; les références précises se trouvent en fin de travail.

## A. Définitions par les fréquences cumulées ou notions apparentées

- **Définition A$_1$.** $Q_1$ est défini implicitement par l'égalité $F(Q_1) = \frac{1}{4}$ ; en d'autres termes, il y a exactement 25% des observations inférieures ou égales à $Q_1$. La façon la plus courante de présenter cette définition consiste à dire que les trois quartiles divisent la série ordonnée en quatre sous-séries de même effectif [Lethielleux ; Chauvat-Réau].

- **Définition A$_2$.** $Q_1$ est défini implicitement par $F^*(Q_1) = \frac{3}{4}$ ; en d'autres termes, il y a exactement 75% des observations supérieures ou égales à $Q_1$, ou encore 25% des observations inférieures à $Q_1$ [Chareille-Pinaut].

- **Définition A$_3$.** $Q_1$ est le plus petit élément $q$ de $S$ tel que $F(q) \geq \frac{1}{4}$ ; c'est la plus petite valeur telle qu'au moins 25% des observations lui sont inférieures ou égales [Groupe d'Experts pour les Programmes Scolaires].

- **Définition A$_4$.** $Q_1$ est le plus petit élément $q$ de $S$ tel que $F^*(q) \leq \frac{3}{4}$, ce qui est équivalent à $1 - F^*(q) \geq \frac{1}{4}$ : c'est la plus petite valeur en-dessous de laquelle on trouve au moins 25% des observations [Lambert].

- **Définition A$_5$.** $Q_1$ est défini implicitement par les inégalités $F(Q_1) \geq \frac{1}{4}$ et $F^*(Q_1) \geq \frac{3}{4}$ ; en d'autres termes, il y a au moins 25% des observations supérieures ou égales à $Q_1$ et au moins 75% des observations qui lui sont supérieures ou égales (ou au plus 25% qui lui sont inférieures) [Lessard-Monga] ; dans le cas où plusieurs nombres vérifient cette définition, on peut prendre pour $Q_1$ leur moyenne arithmétique.

- **Définition A$_6$.** S'il existe un indice $k$ tel que $F\left(x_{(k)}\right) < \frac{1}{4} < F\left(x_{(k+1)}\right)$, alors $Q_1 = x_{(k+1)}$ ; sinon, c'est-à-dire s'il existe un indice $k$ tel que $F\left(x_{(k)}\right) = \frac{1}{4}$, alors $Q_1 = \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right)$ [Droesbeke].

## B. Définitions à partir des rangs

- **Définition B$_1$.** Le rang de $Q_1$ est égal à $\frac{1}{2}(1 + \lfloor r_m \rfloor)$, où $r_m$ désigne le rang de la médiane et est égal à $\frac{n+1}{2}$, avec la convention que si le rang n'est pas entier, on prend la moyenne arithmétique des deux valeurs dont les rangs sont les plus proches [Dodge].

- **Définition B$_2$.** Le rang de $Q_1$ vaut $\frac{r_m+1}{2}$ et est donc égal à $\frac{n+3}{4}$, avec la convention que si le rang n'est pas entier, on effectue une interpolation linéaire entre les valeurs dont les rangs sont les plus proches [Verdier].

- **Définition B$_3$.** $Q_1 = x_{(e)} + f\left(x_{(e+1)} - x_{(e)}\right)$, où $e = \lfloor \frac{n+1}{4} \rfloor$ et $f = \frac{n+1}{4} - \lfloor \frac{n+1}{4} \rfloor$ ; de façon plus intuitive, le rang de $Q_1$ vaut $\frac{n+1}{4}$, avec la convention que si ce rang n'est pas entier, on effectue une interpolation linéaire entre les valeurs dont les rangs sont les plus proches [Dagnelie].

- **Définition B$_4$.** Le rang de $Q_1$ est égal à $\frac{n}{4}$, avec la convention que si ce rang n'est pas entier, on effectue une interpolation linéaire entre les valeurs dont les rangs sont les plus proches.

- **Définition B$_5$.** $Q_1$ est la moyenne arithmétique des valeurs de rangs $r_1 = \lceil \frac{n}{4} \rceil$ et $r_2 = \lceil \frac{n+1}{4} \rceil$ : en formule, $Q_1 = \frac{1}{2}\left(x_{(r_1)} + x_{(r_2)}\right)$.

## C. Définitions à partir de la notion de médiane

- **Définition C$_1$.** $Q_1$ est la médiane de la première moitié de la série, c'est-à-dire la médiane de la sous série $S = \left\{x_{(1)}, x_{(2)}, \ldots, x_{\left(\lfloor \frac{n}{2} \rfloor\right)}\right\}$ [Verdier].

- **Définition C$_2$.** $Q_1$ est la médiane de la série des valeurs inférieures ou égales à la médiane de $S$ [Comte - Goden].

## Comparaison de ces définitions

Toutes ces définitions donnent évidemment des résultats relativement proches mais il existe des différences qui sont, comme nous l'avons dit plus haut, liées au reste de la division de l'effectif par 4.

Le tableau ci-dessous reprend les résultats fournis par les différentes définitions lorsque ce reste vaut respectivement 0, 1, 2 et 3. Un trait horizontal $-$ indique que la définition en question ne peut pas être appliquée au cas considéré.

| Définitions | $n = 4k$ | $n = 4k+1$ | $n = 4k+2$ | $n = 4k+3$ |
|---|---|---|---|---|
| $A_1$ | $x_{(k)}$ | $-$ | $-$ | $-$ |
| $A_2$ | $x_{(k+1)}$ | $-$ | $-$ | $-$ |
| $A_3$ | $x_{(k)}$ | $x_{(k+1)}$ | $x_{(k+1)}$ | $x_{(k+1)}$ |
| $A_4$ | $x_{(k+1)}$ | $x_{(k+2)}$ | $x_{(k+2)}$ | $x_{(k+2)}$ |
| $A_5, A_6, B_5$ | $\frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right)$ | $x_{(k+1)}$ | $x_{(k+1)}$ | $x_{(k+1)}$ |
| $B_1, C_2$ | $\frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right)$ | $x_{(k+1)}$ | $x_{(k+1)}$ | $\frac{1}{2}\left(x_{(k+1)} + x_{(k+2)}\right)$ |
| $B_2$ | $\frac{1}{4}x_{(k)} + \frac{3}{4}x_{(k+1)}$ | $x_{(k+1)}$ | $\frac{3}{4}x_{(k+1)} + \frac{1}{4}x_{(k+2)}$ | $\frac{1}{2}\left(x_{(k+1)} + x_{(k+2)}\right)$ |
| $B_3$ | $\frac{3}{4}x_{(k)} + \frac{1}{4}x_{(k+1)}$ | $\frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right)$ | $\frac{1}{4}x_{(k)} + \frac{3}{4}x_{(k+1)}$ | $x_{(k+1)}$ |
| $B_4$ | $x_{(k)}$ | $\frac{3}{4}x_{(k)} + \frac{1}{4}x_{(k+1)}$ | $\frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right)$ | $\frac{1}{4}x_{(k)} + \frac{3}{4}x_{(k+1)}$ |
| $C_1$ | $\frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right)$ | $\frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right)$ | $x_{(k+1)}$ | $x_{(k+1)}$ |

Remarquons la variété des résultats, ainsi que l'équivalence de certaines de ces définitions.

### Premier quartile et logiciels

Une analyse sommaire des méthodes de calcul utilisées par différents logiciels montrent également une grande variété dans les résultats obtenus. On constate en effet que *Statistica* opte pour la définition $A_5$, *R* pour $B_1$, *Mathematica* pour $A_3$, *Excel* pour $B_2$, *Maple* pour $B_4$ et les calculatrices *TI* pour $C_1$.

En guise d'exemples, voici les résultats fournis par ces différents calculateurs pour les séries simples $S_j$ composées des $j$ premiers entiers positifs ($j = 4, 5, 6, 7$).

| Séries | $S_4$ | $S_5$ | $S_6$ | $S_7$ |
|---|---|---|---|---|
| *Statistica* | 1.5 | 2 | 2 | 2 |
| *R* | 1.5 | 2 | 2 | 2.5 |
| *Mathematica* | 1 | 2 | 2 | 2 |
| *Excel* | 1.75 | 2 | 2.25 | 2.5 |
| *Maple* | 1 | 1.25 | 1.5 | 1.75 |
| *TI* | 1.5 | 1.5 | 2 | 2 |

## 3 Proposition de définition

Au premier abord, toutes ces définitions se valent et le choix de l'une ou de l'autre n'est finalement affaire que de convention. D'ailleurs, toutes ces définitions finissent par être assez proches les unes des autres lorsque l'effectif de la série grandit.

Néanmoins, nous allons tenter, dans le paragraphe suivant, d'objectiver certains critères qui nous permettront de proposer une définition en concordance avec ceux-ci.

## 3.1 Critère retenu

Commençons par observer que tous les auteurs consultés s'accordent sur une même présentation pour une variable statistique continue pour laquelle les observations sont groupées en classes au sein desquelles les données sont supposées équitablement réparties.

Chez certains auteurs, le cas des séries groupées est même le seul envisagé [2]. Cette situation est évoquée par tous les auteurs parce qu'elle est fréquemment rencontrée dans la pratique, mais aussi par sa simplicité de traitement provenant essentiellement du fait que la fonction de répartition $F$ est alors continue et injective (voir figure 1.1). En effet, dans ce cas, l'ogive des fréquences cumulées rencontre toute droite horizontale d'ordonnée comprise entre 0 et 1 en un et un seul point. Dans ces conditions, il suffit de rechercher l'abscisse de l'unique point d'intersection de cette courbe avec une droite horizontale d'ordonnée $\frac{1}{4}$ (resp. $\frac{1}{2}$ ; $\frac{3}{4}$) pour obtenir le premier quartile (resp. la médiane ; le troisième quartile) ; formellement, on peut donc définir alors $Q_1$ (resp. $\tilde{x}$ ; $Q_3$) comme l'unique solution de l'équation $F(x) = \frac{1}{4}$ (resp. $F(x) = \frac{1}{2}$ ; $F(x) = \frac{3}{4}$). Dans la pratique, on détermine tout d'abord la classe contenant $Q_1$ (resp. $\tilde{x}$ ; $Q_3$), puis on effectue une interpolation linéaire pour obtenir la valeur souhaitée. On détermine d'abord la classe $c_{Q_1}$ qui contient le quartile $Q_1$ : elle est telle que

$$N_{Q_1-1} \leq \frac{n}{4} < N_{Q_1}$$

où $N_{Q_1}$ et $N_{Q_1-1}$ désignent respectivement les effectifs cumulés de la classe $c_{Q_1}$ et de la classe précédant $c_{Q_1}$. Le quartile $Q_1$ est alors solution de l'équation

$$N\left(Q_1\right) = \frac{n}{4},$$

qui, en notant $e^-_{Q_1}$ (resp. $e^+_{Q_1}$ ; $a_{Q_1}$ ; $n_{Q_1}$) la borne inférieure (resp. la borne supérieure ; l'amplitude ; l'effectif) de la classe $c_{Q_1}$, est donnée par la formule suivante ([10]) :

$$Q_1 = e^-_{Q_1} + a_{Q_1}\frac{\frac{n}{4} - N_{Q_1-1}}{n_{Q_1}}.$$

Pour une série non groupée, la situation est moins simple puisque la courbe cumulative des fréquences n'est pas continue (voir figure 1.1), de sorte que l'équation $F(x) = \alpha$, avec $\alpha$ égal à $\frac{1}{4}$, $\frac{1}{2}$ ou $\frac{3}{4}$ n'admet pas toujours de solution. Néanmoins, il semblerait opportun, par souci de cohérence, d'opter pour une définition relative à une série simple qui admette une même interprétation graphique que dans le cas d'une série groupée.

Pour éviter qu'une droite horizontale ne rencontre pas la courbe cumulative d'une série donnée $S$, il suffit de compléter le graphe de $F$ par des segments de droite verticaux reliant les paliers horizontaux du graphe, de manière à obtenir une ligne, en forme d'escalier, joignant sans interruption les points $(x_{(1)}, 0)$ et $(x_{(n)}, 1)$ : dans ces conditions, toute

[2] malheureusement, parfois sans mention de l'hypothèse fondamentale d'équirépartition dans les classes.

droite horizontale d'ordonnée comprise entre 0 et 1 rencontre effectivement la ligne brisée en question ([1]).

Pour déterminer le premier quartile $Q_1$, on distingue alors deux cas selon que la droite horizontale d'ordonnée $\frac{n}{4}$ rencontre la ligne brisée selon un segment vertical ou un segment horizontal : dans le premier cas, $Q_1$ est l'abscisse de cette droite verticale, dans le second cas $Q_1$ est l'abscisse du milieu du segment horizontal en question.

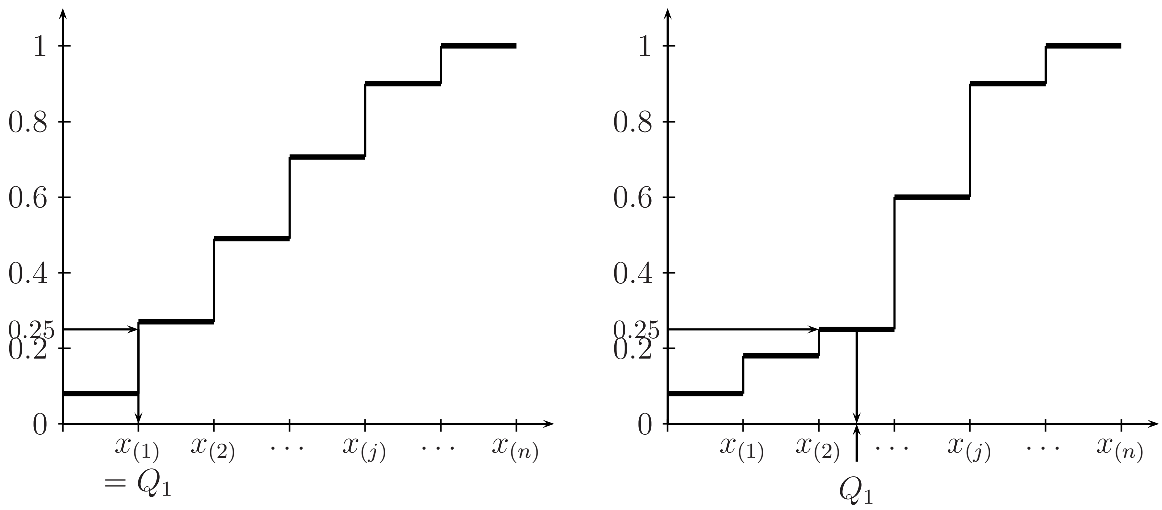



Figure 2: Calcul du premier quartile à partir de la courbe cumulative des fréquences

# 4 Sélection d'une définition

En accord avec l'approche graphique et le critère de cohérence énoncé ci-dessus, nous allons définir la notion de quartile en partant de l'idée intuitive sous-jacente, à savoir que les trois quartiles divisent la série ordonnée en quatre sous-séries comprenant environ chacune 25% des observations.

En réalité, le concept à introduire est une extension naturelle de l'idée de la médiane ([19], p. 147). Comme la médiane a été introduite au préalable, il est préconisé de présenter les deux quartiles extrêmes $Q_1$ et $Q_3$ comme étant les médianes des deux sous-séries construites à partir de la série ordonnée de départ et délimitées par la médiane $\tilde{x}$ connue. C'est d'ailleurs cette idée "naturelle" qui semble venir spontanément à l'esprit des élèves qui connaissent déjà la notion de médiane et à qui l'on demande de partager la série en quatre parties grosso modo de même effectif ([18]).

Il reste à décider si la médiane $\tilde{x}$ doit être comprise ou non dans les deux sous-séries envisagées. Nous allons voir que la réponse à cette question n'est pas aussi simple qu'on le souhaiterait, car elle va dépendre de l'effectif $n$, essentiellement du reste de la division de $n$ par 4.

Envisageons tout d'abord le cas où $n$ est pair, égal à $2p$. Dans ce cas, la médiane pragmatique est donnée par $\tilde{x} = \frac{1}{2}\left(x_{(p)} + x_{(p+1)}\right)$, et la sous-série $S_1 = \left\{x_{(1)}, x_{(2)}, \ldots, x_{(p)}\right\}$ contient exactement la moitié des éléments de $S$. On définit alors $Q_1$ comme étant la médiane de $S_1$. Ou bien $p$ est pair, d'où $n$ est un multiple de 4 égal à $4k$, d'où $Q_1 = \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right)$ et la série $S$ comprend alors exactement 25% de ses valeurs inférieures à $Q_1$ et 75% des valeurs supérieures à $Q_1$. Ou bien $p$ est impair, d'où $n$ est un multiple de 4 plus 2, soit $n = 4k + 2$, auquel cas, $Q_1 = x_{(k+1)}$ : il y a dans ce cas plus de 25% (resp. plus de 75%) des valeurs de $S$ qui sont inférieures ou égales (resp. supérieures ou égales) à $Q_1$. Ainsi, lorsque $n$ est pair, il y a donc toujours au moins 25% (resp. au moins 75%) des valeurs de $S$ qui sont inférieures ou égales (resp. supérieures ou égales) à $Q_1$.

Considérons à présent le cas où $n$ est impair, égal à $2p + 1$ : la médiane de $S$ vaut $\tilde{x} = x_{(p+1)}$ ; il est alors évidemment impossible de répartir les éléments de $S$ en deux sous-séries de même effectif et ne comprenant aucun élément de même rang.

Lorsque $p$ est pair, $n$ est un multiple de 4 plus 3, c'est-à-dire $n = 4k + 3$ : on peut alors former la sous-série ordonnée $S_1$ comprenant les valeurs de $\bar{S}$ qui sont inférieures à $\tilde{x}$, soit $S_1 = \left\{x_{(1)}, x_{(2)}, \ldots, x_{(2k+1)}\right\}$, et choisir alors pour $Q_1$ la médiane de $S_1$, c'est-à-dire $Q_1 = x_{(k+1)}$. I l y a encore dans ce cas plus de 25% (resp. plus de 75%) des valeurs de $S$ qui sont inférieures ou égales (resp. supérieures ou égales) à $Q_1$.

Le dernier cas, le plus problématique, est rencontré lorsque $p$ est pair ; $n$ est donc un multiple de 4 plus 1, soit $n = 4k + 1$ ; il existe alors moins de 25% des observations qui sont inférieures ou égales à la médiane $m$ de la sous-série ordonnée comprenant les valeurs inférieures à $\tilde{x}$. Dès lors, une droite horizontale d'ordonnée $\frac{1}{4}$ ne rencontre pas la ligne brisée construite au départ de la courbe cumulative des fréquences de $S$ en un point d'abscisse $m$ ; c'est pourquoi, il y a lieu ici de considérer la sous-série comprenant $\tilde{x}$, soit à nouveau $S_1 = \left\{x_{(1)}, x_{(2)}, \ldots, x_{(2k+1)}\right\}$ ; on choisit encore pour $Q_1$ la médiane de $S_1$, soit encore $Q_1 = x_{(k+1)}$.

Avec cette présentation, on obtient dans tous les cas un nombre $Q_1$ tel qu'au moins 25% (resp. au moins 75%) des observations lui sont inférieures ou égales (resp. supérieures ou égales).

En définitive, cette version est un “mélange” de $C_1$ et de $C_2$, puisqu'il s'agit de considérer des sous-séries incluant ou non la médiane (selon que $n$ est un multiple de 4 plus 1 ou non). Elle peut se présenter indifféremment sous les formes $A_5$, $A_6$ ou $B_4$.

Formulons quelques remarques à propos de cette définition.

- Elle s'exprime relativement aisément puisqu'il suffit de considérer deux cas : $n = 4k + 1$ ou non. Dans le cas le plus général, on l'obtient en prenant la médiane de la sous-série contenant les observations strictement inférieures à la médiane de la série complète. Dans le premier cas, $Q_1$ est obtenu suivant la même démarche que précédemment mais en incluant la médiane de la série complète dans la sous-série considérée.

- Elle peut être formulée en travaillant aussi bien sur la série ordonnée par valeurs croissantes que sur celle ordonnée par valeurs décroissantes, ce qui n'est pas le cas pour toutes les autres définitions. Ainsi, le troisième quartile $Q_3$ peut être obtenu comme ci-dessus mais en série ordonnée par valeurs décroissantes ; c'est également l'opposé du premier quartile de la série des opposés des $x_i$.

- Elle se traduit techniquement en considérant deux cas, selon que l'effectif $n$ est un multiple de 4 ou non[3]. On a en effet, avec $k = \lfloor \frac{n}{4} \rfloor$ :

  - Si $n$ n'est pas un multiple de 4, alors $Q_1 = x_{(\lfloor \frac{n}{4} \rfloor + 1)}$ et $Q_3 = x_{(n - \lfloor \frac{n}{4} \rfloor)}$

  - Si $n$ est un multiple de 4, alors $Q_1 = \frac{1}{2}\left(x_{(\frac{n}{4})} + x_{(\frac{n}{4}+1)}\right)$ et $Q_3 = \frac{1}{2}\left(x_{(\frac{3n}{4})} + x_{(\frac{3n}{4}+1)}\right)$.

- Les diverses possibilités pour les quartiles sont résumées dans ce tableau :

| $n$ | $n = 4k$ | $n = 4k+1$ | $n = 4k+2$ | $n = 4k+3$ |
|---|---|---|---|---|
| $Q_1$ | $\frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right)$ | $x_{(k+1)}$ | $x_{(k+1)}$ | $x_{(k+1)}$ |
| $\tilde{x}$ | $\frac{1}{2}\left(x_{(2k)} + x_{(2k+1)}\right)$ | $x_{(2k+1)}$ | $\frac{1}{2}\left(x_{(2k+1)} + x_{(2k+2)}\right)$ | $x_{(2k+3)}$ |
| $Q_3$ | $\frac{1}{2}\left(x_{(3k)} + x_{(3k+1)}\right)$ | $x_{(3k+1)}$ | $x_{(3k+2)}$ | $x_{(3k+3)}$ |

- Toutes ces formules peuvent encore être condensées en une seule capable de couvrir tous les cas possibles, quel que soit le reste de la division de $n$ par 4, et pour tous les quartiles d'ordre $j$ pour $j \in \{1, 2, 3\}$ :

$$Q_j = \frac{1}{2}\left(x_{(\lceil \frac{jn}{4} \rceil)} + x_{(\lceil \frac{jn+1}{4} \rceil)}\right).$$

- Elle admet une interprétation graphique qui pourra être reprise pour estimer les quartiles dans le cas de séries groupées, ainsi qu'il a déjà été exposé ci-dessus.

# Références


[1] BRENY H., *Introduction élémentaire aux principes et méthodes de la théorie des probabilités, y compris l'analyse statistique*, Presses Universitaires de Bruxelles, 1969.

[2] CHAREILLE P. - PINAUT Y., *Statistique descriptive, DEUG : méthodes, cours, exercices corrigés*, Ed. Montchrestien, Paris, 2000.

[3] CHAUVAT G. - REAU J.P., *Statistique descriptive*, Ed. Hachette, Paris, 1995.


[3]La démarche devient alors fort semblable à celle formulée pour introduire la médiane.

[4] COMTE M. - GODEN J., *Statistiques et probabilités pour les sciences économiques et sociales*, Presses Universitaires de France, Paris, 2000.

[5] DAGNELIE P., *Statistique théorique et appliquée : tome 1 : statistique descriptive et vbases de l'inférence statistique*, Ed. De Boeck Université, Bruxelles, 1998.

[6] DODGE Y., *Premiers pas en statistique*, Springer-Verlag France, Paris, 1999.

[7] DROESBEKE J.J., *Eléments de statistique*, Editions de l'Université de Bruxelles - Editions Ellipses, Beuxelles - Paris, 1992.

[8] DUPERRET J.C., Mode, moyenne, médiane au collège : pourquoi ?, *Repères Irem*, numéro 6, 1992, pp. 5-19.

[9] GEPS, *Quantiles et diagrammes en boîtes*, document du groupe chargé de rédiger les programmes et les documents d'accompagnement en France, 22/12/2000.

[10] HAESBROECK G., *Statistique et exercices pratiques : statistique descriptive*, notes provisoires de cours, Université de Liège, 2003.

[11] HAESBROECK G., Analyse exploratoire des données à l'aide de boîtes à moustaches, *Cahiers de l'IREM de Bruxelles*, n°1, Editions F.Ferrer et Céfal, Bruxelles et Liège, 2004, pp. 67-85.

[12] HAESBROECK G. - HENRY V., *Pratique de la statistique descriptive : résolution et interprétations de problèmes*, Editions F.Ferrer et Céfal, Bruxelles et Liège, 2004.

[13] IREM de Liège - Luxembourg, *La médiane d'une série statistique univariée*, dossier préparé par le Groupe de Recherches "Statistique et Probabilités" sous la direction de G. HAESBROECK, Liège, année académique 2000-2001.

[14] LAMBERT P., *Statistique descriptive*, notes de cours, Université de Liège, 1998.

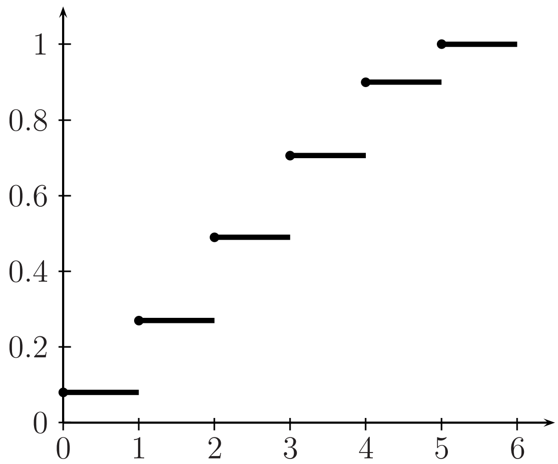
[15] LESSARD S. - MONGA, *Statistique : concepts et méthodes*, Presses Universitaires de Montréal - Ed. Masson, Montréal - Paris, 1993.

[16] LETHIELLEUX M., *Statistique descriptive*, Ed. Dunod, Collection "Express", Paris, 1998.

[17] VERDIER J., Deux ou trois petites choses que je sais de la médiane, *Revue de l'A.P.M.E.P.*, 430, 2000, pp. 557-568.

[18] VERDIER J., Deux ou trois choses que je sais des quartiles et des boîtes à moustaches, *Revue de l'A.P.M.E.P.*, 435, 2001, pp. 456-465.

[19] YULE G.U. - KENDALL M.G., *An Introduction to the Theory of Statistics*, Ed. Charles Griffin & Company, London, 1946.


1
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
6