Une étude sur les quartiles d'une série statistique univariée

Valérie Henry(*)

Résumé : Intuitivement et dans une première approche, les quartiles d'une série statistique univariée devraient diviser l'ensemble des observations en quatre sous-séries d'effectifs égaux.
Dans cette note, nous nous intéressons d'abord à l'analyse comparative de différentes présentations des quartiles proposées dans la littérature et mettons en évidence certaines divergences de vues selon les auteurs. Nous nous penchons également sur les méthodes de calcul utilisées par différents logiciels statistiques ou plus généraux pour fournir les quartiles. Sur la base de nos constatations, nous faisons une proposition de présentation du premier quartile pour une série de données d'un caractère quantitatif qui permette notamment de rester en cohérence avec l'approche graphique usuelle et qui puisse s'étendre aux différents quantiles.

Mots-clés : Quartiles, quantiles, médiane, définitions, registre graphique, registre numérique, fréquence cumulée, fonction de répartition empirique, courbe cumulative, rang d'une observation.

Motivation : Les définitions des quartiles divergent sensiblement d'un auteur à l'autre et d'un logiciel à l'autre, comme le montre la première partie de l'article. Nous pensons qu'il est important pour l'enseignant d'en être conscient car l'élève risque de se trouver confronté à ces différents points de vue ne serait-ce qu'en consultant plusieurs manuels ou en utilisant divers logiciels (simples tableurs ou plus spécifiques).

Introduction

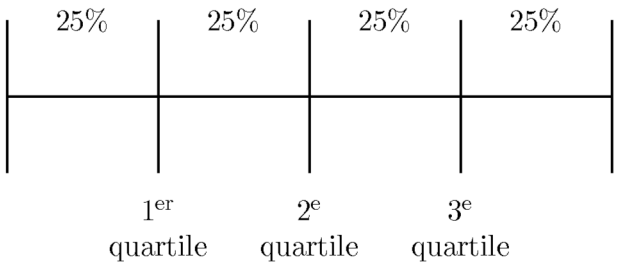
L'homme moderne est constamment assailli d'informations en tous genres. Parmi celles-ci, nombreuses sont celles qui se présentent sous la forme de listes d'observations numériques(1). Pour remédier à l'incapacité de l'esprit humain d'intégrer instantanément un tableau important de nombres, le statisticien propose différentes solutions. Une première catégorie de solutions est constituée de toutes les représentations possibles de ces données : diagrammes (en bâtons, circulaires), histogrammes, polygones de fréquences, boîtes à moustaches, … 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 ceux qui caractérisent la tendance centrale, la dispersion, la dissymétrie ou la forme.

(*) Université de Liège. 7 Boulevard du Rectorat, Bât B31. 4000 Liège. Belgique. E-mail : V.Henry@ulg.ac.be

(1) Nous ne traiterons dans cet article que les séries statistiques d'un seul caractère quantitatif, laissant ainsi de côté les séries multivariées dont l'étude fait appel à des techniques mathématiques plus sophistiquées ; un aperçu de telles méthodes peut être trouvé, par exemple, dans [9].

Dans cette note, nous nous attachons plus particulièrement à trois de ces nombres que sont les *quartiles* d'une série de données statistiques d'un caractère quantitatif. Intuitivement et dans une première approche, les quartiles

> « divisent un ensemble d'observations en quatre parties égales… Voici, schématiquement, une distribution partagée en quartiles. Entre chaque quartile se trouvent 25 % des observations :



> Notons que le deuxième quartile est égal à la médiane. » (Dodge [6], p. 87-88)

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 » (Verdier [16], 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(2).

Après quelques réflexions générales, nous proposons une analyse comparative de présentations des quartiles proposées dans la littérature et mettons en évidence des 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 la base de nos constatations, nous étudions plus particulièrement une présentation du premier quartile qui peut être donnée aussi bien graphiquement que numériquement.

1. Contexte

1.1. Notations et conventions

Nous allons fixer notre attention sur une série statistique d'effectif n et notée $S = \{x_1, x_2, \ldots, x_n\}$, composée de n valeurs observées d'un caractère quantitatif, certains éléments x_i pouvant être égaux. Nous supposerons que la série est ordonnée par valeurs croissantes et que nous avons donc $x_i \leq x_j$ pour tous entiers i, j compris entre 1 et n, tels que $i < j$. Nous appellerons i le rang de x_i ; ultérieurement, nous considérerons des *rangs fictifs* : ce sont des nombres réels qui peuvent être non entiers mais compris entre 1 et n.

(2) Les quartiles et la médiane sont utilisés pour construire la boîte de dispersion en classe de première.

Une partie de notre travail va reposer sur la notion de fréquence cumulée qui donne la proportion d'individus, parmi la population considérée, pour lesquels le caractère étudié prend une valeur inférieure ou égale à l'une des valeurs observée. Plus précisément, nous exploiterons la notion de *fonction de répartition empirique*. Or celle-ci n'est pas présentée univoquement dans la littérature ; c'est pourquoi, il nous semble opportun de préciser les définitions et notations que nous emploierons dans la suite.

Nous noterons F la fonction définie, pour tout réel x, par la proportion des valeurs observées qui sont inférieures ou égales à x, c'est-à-dire symboliquement

$$\mathrm{F}(x) = \frac{\mathrm{Card}\left\{x_i \in \mathrm{S} \mid x_i \leq x\right\}}{n}.$$

Nous considérerons également la fonction F_1 définie, pour tout réel x, par

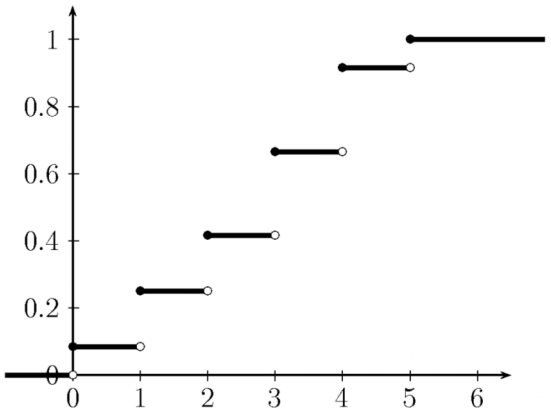
$$\mathrm{F}_1(x) = \frac{\mathrm{Card}\left\{x_i \in \mathrm{S} \mid x_i < x\right\}}{n}.$$

Remarques : a) Plusieurs auteurs notent F la fonction que nous désignons par F_1 ; mentionnons encore que d'aucuns distinguent les valeurs de ces deux fonctions en écrivant $\mathrm{F}\left(x_i^+\right)$ et $\mathrm{F}\left(x_i^-\right)$ à la place de $\mathrm{F}\left(x_i\right)$ et $\mathrm{F}_1\left(x_i\right)$ respectivement (Delmas [5] p. 93).

b) Les deux fonctions F et F_1 sont visiblement proches l'une de l'autre. De manière plus précise, pour tout nombre x ne coïncidant avec aucun élément de S, $\mathrm{F}(x) = \mathrm{F}_1(x)$; par contre, pour toute observation x_j de S,

$$\mathrm{F}\left(x_j\right) = \mathrm{F}_1\left(x_j\right) + \frac{n_j}{n},$$

où n_j désigne le nombre de fois que x_j apparaît dans S. Ces deux fonctions ont une représentation graphique ayant la forme d'un escalier (ascendant), dont les marches horizontales coïncident sauf en leurs extrémités (où F et F_1 sont toutes deux discontinues, F étant continue à droite et F_1 à gauche).



En guise d'exemple, voici ci-dessus la représentation graphique de la fonction d'équation $y = F(x)$ pour la série $S = \{0, 1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 5\}$; ce graphe est souvent appelé la *courbe cumulative des fréquences*.

Par ailleurs, pour tout nombre réel x non négatif, nous noterons $\text{Ent}(x)$ la partie entière de x, c'est-à-dire le plus grand entier inférieur ou égal à x ; dès lors, la partie décimale de x vaut $x - \text{Ent}(x)$; de plus, $\text{Ent}(x) = x$ si et seulement si x est lui-même un entier.

1.2. La médiane

Rappelons que, dans le cas qui nous occupe, la *médiane* 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 $\frac{1}{2}(x_k + x_{k+1})$ lorsque n est le nombre pair $2k$ [11]. Dans la suite, la médiane de S sera notée Q_2 car elle coïncide pour nous avec le deuxième quartile(3).

2. Quelques présentations du premier quartile dans la littérature

Intuitivement, il semble relativement facile d'introduire les deux quartiles, autres que la médiane, d'une série statistique : l'idée première est de suivre le même type de procédé que pour la médiane. Néanmoins, cette première impression se heurte rapidement à plusieurs difficultés notamment liées à l'effectif de la série. Dans ce paragraphe, nous examinons différentes acceptions 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, classé les « définitions » répertoriées selon trois grandes catégories : la première fait appel aux fonctions F et F_1 évoquées ci-dessus, la deuxième utilise le rang (éventuellement fictif) des observations et la dernière est basée sur la notion de médiane.

Signalons que les présentations qui vont être données se retrouvent dans la littérature : après les avoir exposées, nous citerons, toujours entre guillemets, des extraits d'ouvrages (dont les références précises se retrouvent dans la bibliographie), puis d'éventuels commentaires ; mentionnons encore que les citations que nous avons choisies, parmi d'autres, sont généralement commentées par leurs auteurs et parfois accompagnées d'autres présentations que celles reprises dans notre texte.

A. Présentations par les fréquences cumulées ou notions apparentées

- **Présentation 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.

(3) Notons que, pour l'actuel programme de la classe de première, la médiane ne coïncide pas nécessairement avec le deuxième quartile.

Ce paramètre Q_1 est en fait un cas particulier de α-quantile, à savoir celui pour lequel $\alpha = \frac{1}{4}$. D'une manière générale, on peut définir, pour tout α compris entre 0 et 1, un α-quantile, encore noté dans la suite x_α, comme étant « tel qu'une proportion α des individus ait une valeur du caractère inférieure ou égale à x_α » (Goldfarb-Pardoux [8], p. 21).

Les trois quartiles Q_1, Q_2 et Q_3 sont respectivement des quantiles d'ordre $\frac{1}{4}$, $\frac{1}{2}$ et $\frac{3}{4}$. On peut alors écrire que « empiriquement, ces trois nombres partagent l'ensemble des observations en quatre parties " de même effectif " » (Droesbeke [7], p. 97).
Notons que la précaution prise par l'auteur de cette citation en plaçant entre guillemets les mots « de même effectif » se justifie par le fait que cette présentation n'est pas toujours applicable telle quelle, parce que l'équation implicite en question ne possède pas toujours une solution unique. Cette présentation ne peut donc pas être retenue rigoureusement en toute généralité.

- **Présentation A_2**. Q_1 est défini implicitement par $F_1(Q_1) = \frac{1}{4}$. De façon équivalente, on peut affirmer que « le quart des valeurs observées sont inférieures à Q_1 (et donc les trois autres quarts supérieures ou égales à Q_1 » (Chareille-Pinaut [1], p. 90).
Les mêmes remarques et objections que celles émises pour A_1 peuvent être formulées à propos de cette présentation.

- **Présentation A_3**. Q_1 est le plus petit élément q de S tel que $F(q) \geq \frac{1}{4}$. Cette version est celle mentionnée notamment dans le document d'accompagnement du programme de Première (p. 85) : « Premier quartile : c'est le plus petit élément q des valeurs des termes de la série tel qu'au moins 25 % des données soient inférieures ou égales à q ».

- **Présentation A_4**. Q_1 est défini implicitement par les inégalités $F(Q_1) \geq \frac{1}{4}$ et $F_1(Q_1) \leq \frac{1}{4}$. En d'autres termes, il y a au moins 25 % des observations inférieures ou égales à Q_1 et au plus 25 % lui sont inférieures (et donc au moins 75 % des observations lui sont supérieures ou égales) ; en formule :

  « (pourcentage des données $< C_{25}$) $\leq \frac{1}{4} \leq$ (pourcentage des données $\leq C_{25}$) »

  (Lessard-Monga [13], p. 79)(4).

(4) Q_1 est appelé par ces auteurs le 25e centile et noté C_{25}.

Il convient de signaler que cette version ne garantit pas l'unicité de la solution ; toutefois, « si deux valeurs consécutives satisfont à cette double inégalité, on prend (par convention) pour quartile leur moyenne arithmétique » (Droesbeke [7], p. 96).

B. Présentations à partir des « rangs » (éventuellement fictifs)

- **Présentation B_1**. Le « rang », éventuellement fictif, de Q_1 est égal à $\frac{1}{2}\left(1+\text{Ent}\left(\frac{n+1}{2}\right)\right)$; il est donc introduit à partir de celui de la médiane : « rang médiane = $\frac{n+1}{2}$, rang quartile = $\frac{[\text{rang médiane}]+1}{2}$ où $[x]$ est la valeur de x tronquée à l'entier inférieur. La médiane et les quartiles seront les données correspondant aux rangs calculés (…) Des rangs non-entiers signifient que l'on calculera la moyenne entre les deux valeurs les plus proches » (Dodge [6], p. 112).

- **Présentation B_2**. Le « rang », éventuellement fictif, de Q_1 vaut $\frac{n+3}{4}$. En d'autres termes, sachant que le « rang » du premier élément est 1 et celui de la médiane est $\frac{n+1}{2}$, le « rang » de Q_1 est « la demi-somme de ces rangs. [Lorsque cette valeur n'est pas un entier, le quartile sera] le barycentre des deux valeurs les plus proches affectées des coefficients égaux à ceux qui interviennent pour exprimer que le rang fictif est le barycentre de deux rangs successifs » (Verdier [16], p. 457) ; en formule, on peut écrire : $Q_1 = x_i + f(x_{i+1} - x_i)$, avec $i = \text{Ent}\left(\frac{n+3}{4}\right)$ et $f = \frac{n+3}{4} - i$.

- **Présentation B_3**. Le « rang », éventuellement fictif, de Q_1 vaut $\frac{n+1}{4}$. Il est donc égal au quart de la somme des rangs extrêmes. On pourra encore effectuer « éventuellement une interpolation entre les valeurs situées à proximité de ces rangs, lorsque ceux-ci ne sont pas entiers » (Dagnelie [4], p. 86).
On peut alors calculer le j^e quartile selon la formule « $Q_j = x_i + (k(x_{i+1} - x_i))$, où i est la partie entière de $\frac{j(n+1)}{4}$ et k la partie fractionnelle de $\frac{j(n+1)}{4}$. » (Dodge [6], p. 88).

- **Présentation B_4**. Le « rang », éventuellement fictif, de Q_1 est égal à $\frac{n}{4}$; il est donc égal au quart de l'effectif global. Lorsque ce « rang » n'est pas entier, on effectue à nouveau une interpolation linéaire entre les valeurs dont les rangs sont les plus proches.

C. Présentations à partir de la notion de médiane

- **Présentation C_1**. Q_1 est la médiane de la sous-série composée des « valeurs inférieure ou égales à la médiane » (Comte-Gaden [3], p. 86).
  Cette version simple ne peut toutefois être acceptée en toute généralité, comme en témoigne l'exemple de la série S = {1 , 2 , 3 , 3 , 3 , 3} : dans ce cas, la sous-série en question coïncide avec la série entière, alors qu'il semble clair que la médiane et le premier quartile ne peuvent coïncider.
- **Présentation C_2**. Q_1 est la médiane de la sous-série composée des valeurs inférieures à la médiane. Cette variante de C_1 ne peut pas non plus être toujours retenue, ainsi qu'en atteste l'exemple de la série S = {1, 1, 1, 1, 2} pour lequel la sous-série considérée n'existe pas.
- **Présentation C_3**. Q_1 est la médiane de la sous-série composée des i premières valeurs, où i désigne la partie entière de $\frac{n}{2}$. Intuitivement, cette version « consiste à prendre pour premier quartile la médiane de la première moitié de la série » (Verdier [16], p. 457).

Comparaison de ces présentations

Éliminons d'emblée les présentations A_1, A_2, C_1 et C_2 qui ne peuvent être retenues en toute généralité comme il a été expliqué ci-dessus.
Toutes les versions restantes fournissent évidemment des résultats proches, mais il existe des différences qui sont liées au reste de la division de l'effectif par 4.
Lorsque n est de la forme $n = 4k + r$, avec k entier et r égal à 0 , 1 , 2 ou 3, presque toutes ces présentations choisissent comme premier quartile un nombre compris dans l'intervalle $I_k = [x_k ; x_{k+1}]$, celui-ci pouvant être réduit à un singleton lorsque x_k et x_{k+1} coïncident ; seules les présentations B_1 et B_2 optent, lorsque $r = 3$, pour un point pouvant ne pas appartenir à I_k puisqu'il s'agit alors du milieu de l'intervalle $I_{k+1} = [x_{k+1} ; x_{k+2}]$. L'extrémité gauche x_k de I_k est choisie comme premier quartile par A_3 et A_4 lorsque $r = 0$. Par ailleurs, l'extrémité droite x_{k+1} de I_k est retenue comme premier quartile par A_3 et A_4 pour $r \neq 0$, par B_1 pour r égal à 1 ou à 2, par B_2 pour $r = 1$, par B_3 pour $r = 3$, par C_3 pour r égal à 2 ou à 3 . Dans tous les autres cas, Q_1 est choisi soit au milieu de l'intervalle I_k, soit au quart, c'est-à-dire en $\frac{3}{4}x_k + \frac{1}{4}x_{k+1}$, soit encore aux trois-quarts de I_k, c'est-à-dire en $\frac{1}{4}x_k + \frac{3}{4}x_{k+1}$.
De façon plus précise, le tableau ci-dessous reprend tous les cas possibles.

La lecture de ce tableau met en évidence la variété des résultats obtenus et la possibilité d'introduire le premier quartile autrement que par ces présentations.

Présentations	$n = 4k$	$n = 4k + 1$	$n = 4k + 2$	$n = 4k + 3$
A_3	x_k	x_{k+1}	x_{k+1}	x_{k+1}
A_4	$\frac{1}{2}(x_k + x_{k+1})$	x_{k+1}	x_{k+1}	x_{k+1}
B_1	$\frac{1}{2}(x_k + x_{k+1})$	x_{k+1}	x_{k+1}	$\frac{1}{2}(x_{k+1} + x_{k+2})$
B_2	$\frac{1}{4}x_k + \frac{3}{4}x_{k+1}$	x_{k+1}	$\frac{3}{4}x_{k+1} + \frac{1}{4}x_{k+2}$	$\frac{1}{2}(x_{k+1} + x_{k+2})$
B_3	$\frac{3}{4}x_k + \frac{1}{4}x_{k+1}$	$\frac{1}{2}(x_k + x_{k+1})$	$\frac{1}{4}x_k + \frac{3}{4}x_{k+1}$	x_{k+1}
B_4	x_k	$\frac{3}{4}x_k + \frac{1}{4}x_{k+1}$	$\frac{1}{2}(x_k + x_{k+1})$	$\frac{1}{4}x_k + \frac{3}{4}x_{k+1}$
C_3	$\frac{1}{2}(x_k + x_{k+1})$	$\frac{1}{2}(x_k + x_{k+1})$	x_{k+1}	x_{k+1}

Premier quartile et logiciels

Une analyse sommaire des méthodes de calcul utilisées par différents logiciels montre également une grande diversité dans les résultats obtenus.
On constate en effet que *Statistica* opte pour la définition A_4, *R* pour B_1, *Mathematica* pour A_3, *Excel* pour B_2, *Maple* pour B_4 et les calculatrices *TI* pour C_3.

En guise d'exemples simples et suggestifs, voici les résultats fournis par ces différents calculateurs pour les séries S_j composées des j premiers entiers positifs (j = 4, 5, 6, 7) ; le lecteur pourra traiter d'autres exemples dont l'effectif est plus imposant et comprenant, éventuellement, des observations qui peuvent être égales.

Séries	S4	S5	S6	S7
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

Ces exemples très simples montrent déjà clairement la disparité des réponses fournies par les différents logiciels.

3. Proposition de présentation

Toutes les présentations traitées dans les tableaux ci-dessus finissent par être assez proches les unes des autres lorsque l'effectif de la série grandit (sauf cas aberrant, bien sûr). L'on pourrait donc penser a priori que le choix de l'une de ces versions est purement conventionnel.

Néanmoins, nous nous proposons d'approfondir plus particulièrement la présentation A_4, car elle nous semble posséder diverses particularités intéressantes qui vont être analysées ci-dessous.
Un premier avantage que possède A_4 réside dans le fait que cette version peut être présentée dans divers cadres et registres.

3.1. Présentation graphique

Partons de la représentation graphique de la fonction F pour introduire le premier quartile : l'idée de base retenue consiste à s'intéresser à l'intersection de ce graphe avec une droite horizontale d'ordonnée $\frac{1}{4}$. Mais cette intersection est soit vide, en raison des points de discontinuité de F, soit composée de plusieurs points.
Pour éviter qu'une droite horizontale ne rencontre pas la courbe cumulative des fréquences, il suffit de compléter ce graphe 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 $(x_n, 1)$. Formellement, on trace ainsi une ligne brisée composée successivement de

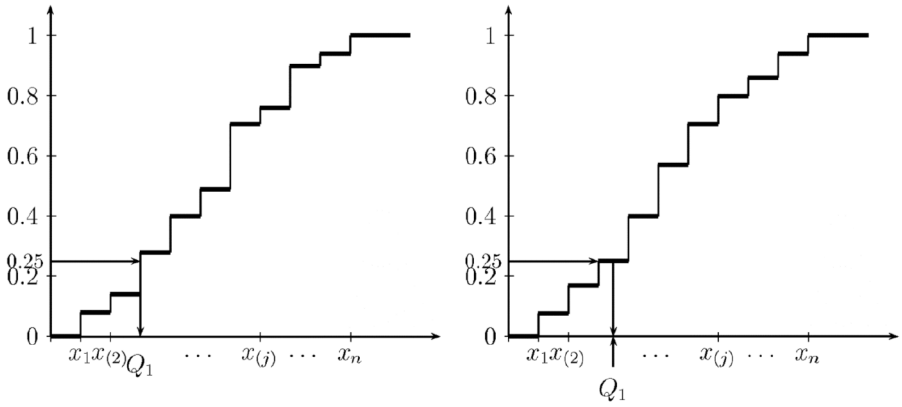
- la portion de l'axe des abscisses pour tout nombre inférieur ou égal à x_1 ;
- le segment vertical reliant les points $(x_1, 0)$ et $\left(x_1, \frac{n_1}{n}\right)$, où n_1 désigne le nombre de fois que x_1 intervient dans la série S ;
- le segment horizontal joignant $\left(x_1, \frac{n_1}{n}\right)$ et $\left(x_{(2)}, \frac{n_1 + n_{(2)}}{n}\right)$, où $x_{(2)}$ est la plus petite valeur de la série strictement supérieure à x_1 et $n(2)$ est le nombre de fois que $x_{(2)}$ apparaît au sein de S ;
- et ainsi de suite jusqu'à ce que le point $(x_n, 1)$ soit atteint, auquel cas la ligne brisée se prolonge par la portion de droite horizontale d'ordonnée unitaire pour les points d'abscisse supérieure ou égale à x_n.

Dans ces conditions, toute droite horizontale d'ordonnée comprise entre 0 et 1 rencontre effectivement la ligne brisée en question.

Pour déterminer le premier quartile Q_1, on distingue alors deux cas selon que la droite horizontale d'ordonnée $\frac{1}{4}$ rencontre la ligne brisée selon un segment vertical ou un segment horizontal

a) Lorsque la droite horizontale d'ordonnée $\frac{1}{4}$ rencontre la ligne brisée selon un segment vertical (en dehors de ses extrémités) d'abscisse x_i, on a $F(x_i) > \frac{1}{4}$

et $F_1(x_i) < \frac{1}{4}$. Dans ce cas, on choisit[5] $Q_1 = x_i$. Remarquons que cette situation se présente nécessairement lorsque n n'est pas un multiple de 4, c'est-à-dire $n = 4k + r$ avec $r \neq 0$: dans ce cas i est égal à $k + 1$; cela peut également se produire lorsque n est un multiple de 4, donc de la forme $n = 4k$ pour autant que x_k et x_{k+1} coïncident : alors, i vaut indifféremment k ou $k + 1$.



b) Lorsque la droite horizontale d'ordonnée $\frac{1}{4}$ rencontre la ligne brisée selon un segment horizontal, il existe un indice i tel que $F(x_i) = \frac{1}{4}$. Cette éventualité ne peut se produire que lorsque $n = 4k$, avec $i = k$ et $x_{k+1} > x_k$. Dans ce cas, on choisit $Q_1 = \frac{x_k + x_{k+1}}{2}$, à savoir l'abscisse du milieu du segment horizontal en question.

Remarquons que cette présentation 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 présentations. Ainsi, le troisième quartile Q_3 peut être obtenu comme ci-dessus mais en travaillant sur la série ordonnée par valeurs décroissantes ; c'est également l'opposé du premier quartile de la série des opposés des x_i.

3.2. Présentation à l'aide de la notion de médiane

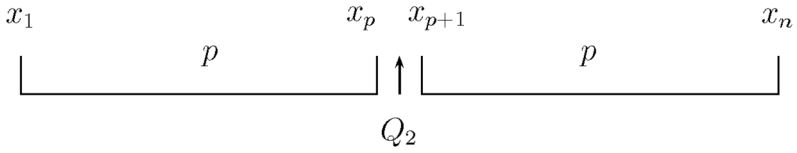
En accord avec l'approche graphique et le critère de cohérence énoncé ci-dessus, nous allons définir la notion de quartile en partant de l'idée intuitive sous-jacente, à

(5) La norme AFNOR pose quant à elle $Q_1 = \frac{x_k + x_{k+1}}{2}$.

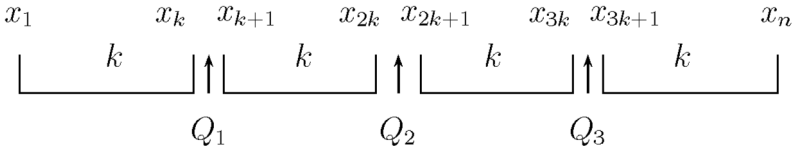
savoir que les trois quartiles divisent la série initiale en quatre sous-séries ordonnées comprenant chacune au moins 25 % d'observations consécutives dans la série initiale ; cette approche naïve ne peut, bien entendu, suffire puisque, comme l'effectif de la série S n'est pas toujours divisible par 4, certaines observations pourraient être omises dans cette subdivision.
Il est préconisé ici 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 Q_2 de S. 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 (Verdier [16]).
Il reste à décider si la médiane Q_2 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.

1. Envisageons tout d'abord le cas où n est pair, égal à $2p$.

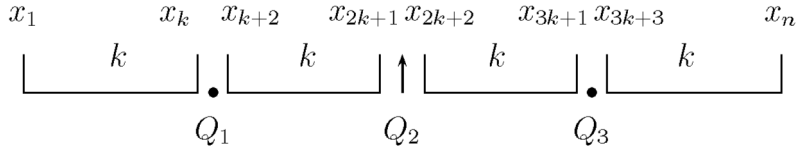
Dans ce cas, la médiane pragmatique est donnée par $Q_2 = \dfrac{x_p + x_{p+1}}{2}$, et la sous-série $S = \{x_1, x_2, \ldots, x_p\}$ contient exactement la moitié des éléments de S.



On définit alors Q_1 comme étant la médiane de S_1. Deux cas sont possibles.
1.1. Ou bien p est pair, d'où n est un multiple de 4 égal à $4k$.



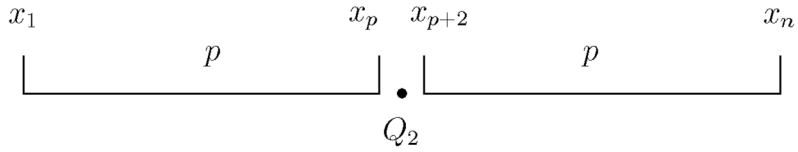
Alors, $Q_1 = \dfrac{x_k + x_{k+1}}{2}$ et la série S comprend alors exactement 25 % de ses valeurs dont le rang est inférieur au rang fictif de Q_1 .
1.2. Ou bien p est impair, d'où n est un multiple de 4 plus 2, soit $n = 4k + 2$.



Dans ce cas, $Q_1 = x_{k+1}$: il y a dans ce cas plus de 25 % des valeurs de S qui sont inférieures ou égales à Q_1.

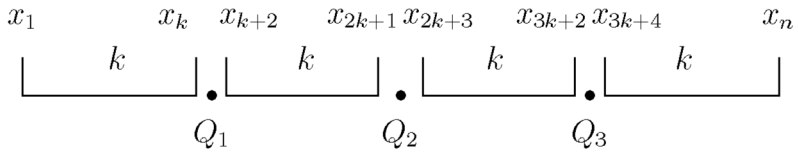
Ainsi, lorsque n est pair, il y a toujours au moins 25 % des valeurs de S qui sont inférieures ou égales à Q_1.

2. Considérons à présent le cas où n est impair, égal à $2p + 1$: la médiane de S vaut $Q_2 = x_{p+1}$.



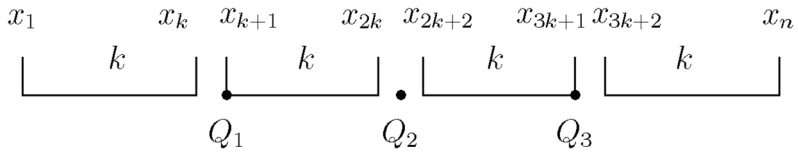
Il est alors évidemment impossible de répartir les éléments de S en deux sous-séries de même effectif et ne comprenant aucun élément de même rang. Considérons encore la sous-série composée des éléments de rang inférieur à celui de cette médiane, soit $S_1 = \{x_1, \ldots, x_p\}$ et distinguons à nouveau deux cas.

2.1. Lorsque p est impair, n est un multiple de 4 plus 3, c'est-à-dire $n = 4k + 3$.



On peut alors choisir pour Q_1 la médiane de S_1, c'est-à-dire $Q_1 = x_{k+1}$. Il y a encore dans ce cas plus de 25 % des valeurs de S qui sont inférieures ou égales à Q_1.

2.2. Le dernier cas, le plus problématique, est rencontré lorsque p est pair ; donc n est un multiple de 4 plus 1, soit $n = 4k + 1$.



Il existe alors moins de 25 % des observations dont le rang est inférieur au rang fictif de la médiane de la sous-série S_1. En raison de la contrainte imposant qu'au moins 25 % des observations soient inférieures ou égales à Q_1, il y a lieu ici de considérer la sous-série comprenant Q_2, à savoir la sous-série $S_2 = \{x_1, \ldots, x_{p+1}\}$; on choisit pour Q_1 la médiane de S_2, soit $Q_1 = x_{k+1}$.

Avec cette présentation, on obtient dans tous les cas un nombre Q_1 tel qu'au moins 25 % des observations lui soient inférieures ou égales. Ainsi, en définitive, cette version est une sorte de « mélange » de C_1 et de C_2, puisqu'il s'agit généralement de considérer des sous-séries incluant ou non la valeur de rang médian (selon que n est un multiple de 4 plus 1 ou non). Cette présentation peut alors être formulée dans un registre littéraire : le premier quartile est obtenu en prenant la médiane de la sous-série contenant les observations dont le rang est strictement inférieur à celui de la médiane pour autant qu'au moins 25 % des observations soient inférieures ou égales à cette valeur (sinon, ce qui se présente dans le cas où $n = 4k + 1$, il faut inclure la médiane dans la sous-série en question).

3.3. Présentation dans un cadre numérique

La présentation retenue A_4 peut se traduire techniquement en considérant deux cas, selon que l'effectif n est un multiple de 4 ou non. En effet, notons $k = \text{Ent}\left(\frac{n}{4}\right)$.

a) Si n n'est pas un multiple de 4, alors $Q_1 = x_{k+1}$; de fait, on a alors $F(x_{k+1}) \geq \frac{k+1}{n} > \frac{1}{4}$ et $F_1(x_{k+1}) \leq \frac{k}{n} < \frac{1}{4}$.

b) Si n est un multiple de 4, alors $Q_1 = \frac{x_k + x_{k+1}}{2}$; effectivement, on peut alors écrire $F(x_k) \geq \frac{k}{n} = \frac{1}{4}$ et $F_1(x_k) < \frac{k}{n} = \frac{1}{4}$, tandis que $F(x_{k+1}) \geq \frac{k+1}{n} > \frac{1}{4}$ et $F_1(x_{k+1}) \leq \frac{k}{n} < \frac{1}{4}$.

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

N	$n = 4k$	$n = 4k + 1$	$n = 4k + 2$	$n = 4k + 3$
Q_1	$\frac{1}{2}(x_k + x_{k+1})$	x_{k+1}	x_{k+1}	x_{k+1}
Q_2	$\frac{1}{2}(x_{2k} + x_{2k+1})$	x_{2k+1}	$\frac{1}{2}(x_{2k+1} + x_{2k+2})$	x_{2k+2}
Q_3	$\frac{1}{2}(x_{3k} + x_{3k+1})$	x_{3k+1}	x_{3k+2}	x_{3k+3}

Un avantage de cette présentation numérique, outre la simplicité manifeste de son application, est la possibilité de la généraliser sans peine au cas de n'importe quel p-quantile (pour tout entier p) ; en particulier, on peut l'utiliser pour introduire la médiane pragmatique ($p = 2$), le premier décile ($p = 10$) ou le premier centile ($p = 100$). En effet, si n n'est pas un multiple de p, donc de la forme $n = pk + r$, avec k entier et r strictement compris entre 0 et p, alors le premier p-quantile vaut x_{k+1}. Si n est un multiple de p, de la forme $n = pk$, alors le premier p-quantile est égal à $\frac{1}{2}(x_k + x_{k+1})$.

Signalons enfin que toutes les formules rencontrées dans le dernier tableau peuvent 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 $j \in \{1 , 2 , 3\}$:

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

où $[x]$ désigne le plus petit entier supérieur ou égal à x.

4. Courte conclusion

Notre étude nous a rappelé que les quartiles peuvent être introduits de différentes manières. Parmi celles-ci, nous avons porté plus particulièrement notre attention sur une présentation qui nous semble posséder un quadruple avantage :

- elle peut être développée dans différents cadres et registres, à savoir graphiquement ou numériquement, en faisant ou non appel à la notion de médiane ;
- sa présentation dans le cadre graphique rejoint ce qui est fait généralement dans le cas des séries classées ;
- le premier quartile peut être calculé très systématiquement et aisément ;
- cette version, tant dans le cadre graphique que numérique, est facilement généralisable à tout fractile.

Bien entendu, d'autres conventions pourraient également être adoptées.

Références

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

[2] CHAUVAT G. - REAU J.P., *Statistiques descriptives*, VUEF/Armand Colin, Paris, 2002.

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

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

[5] DELMAS B., *Statistique descriptive*, Éditions Nathan/HER, Paris, 2000.

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

[7] DROESBEKE J. J., *Éléments de statistique*, Éditions de l'Université de Bruxelles – Éditions Ellipses, Bruxelles - Paris, 1992.

[8] GOLDFARB B. – PARDOUX C., *Introduction à la méthode statistique. Gestion. Économie*. 3e édition, Dunod, Paris, 2000.

[9] HAESBROECK G., Analyse exploratoire des données à l'aide de boîtes à moustaches, *Cahiers de l'IREM de Bruxelles*, nº 1, Éditions F. Ferrer et du Céfal, Bruxelles et Liège, 2004, p. 67-85.

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

[11] 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.

[12] JANVIER M., *Statistique descriptive avec ou sans tableur*, Dunod, Paris, 1999.

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

[14] LETHIELLEUX M., *Statistique descriptive*, Éd. Dunod, Collection « Express », Paris, 1998.

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

[16] 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, p. 456-465.