Programmer le calcul de la médiane et des quartiles

2e

Fiche professeur

Auteurs : Pierre Lapôtre, Emmanuel Ostenne et Martijn van Brugghe

But de l'activité : Les tableurs, calculatrices et logiciels de calcul employés au lycée contiennent des fonctions pré-programmées qui calculent les premier et troisième quartiles d'une série statistique. Ces fonctions sont basées sur des définitions de Q_1 et Q_3 différentes des définitions des programmes français d'enseignement. Elles nous donnent donc des résultats faux. Nous allons établir des algorithmes corrects exécutables soit sur le tableur « Calc », soit sur une calculatrice programmable, soit sur « scilab pour les lycées », « Xcas » ou « Javascript ».

Compétences engagées :

- ✓ Écrire un algorithme en langage naturel.
- ✓ Écrire un algorithme exécutable simple.
- ✓ Comprendre un algorithme exécutable donné.
- ✓ Tester un algorithme.
- ✓ Engendrer des nombres au hasard sur un intervalle.
- ✓ Comprendre et appliquer les notions de médiane et de quartiles.

Pré-requis :

- ✓ Algorithmique : notions d'affectation et de boucle conditionnelle simple.
- ✓ Algorithmes en langage naturel, algorithme exécutable.
- ✓ Notions de médiane et de quartiles.
- ✓ Secondairement, simuler des nombres au hasard sur l'intervalle [0,1] .

Matériels utilisés :

- ✓ Calculatrice programmable ou ordinateur équipé de « Calc » (tableur d'OOo) ou de « scilab pour les lycées », de « Xcas » ou de « Javascript ».

Durée indicative : 2 heures

Documents utiles à télécharger :

- ✓ « Fiche Élève » et le fichier de calcul correspondant au système utilisé.

Déroulement de la séance :

Suivre la « Fiche Élève ».

Avertissement :

À notre connaissance, seule la « TI Collège Plus » calcule les quartiles correctement, c'est à dire en utilisant les définitions de nos programmes.

Les définitions des quartiles utilisées par « scilab », « Xcas », « Calc » ou les calculatrices programmables d'une part, les définitions de nos programmes d'enseignement d'autre part, traduisent correctement la même idée. Mais pour éviter de recourir à la bonne définition, qui, se référant à la *fonction quantile*, est très technique et serait incompréhensible pour les élèves, on fait des choix simplificateurs, autrement dit, on adopte des conventions. Malheureusement, ces conventions varient d'un pays à l'autre. Voilà pourquoi les systèmes de calcul utilisés en classe donnent des résultats faux (pour nous).

Par exemple, si A désigne la série statistique $(-1,-4{,}0,-4{,}7{,}5,-4)$, « scilab » donne au troisième quartile Q_3 de A la valeur 3.75, tandis que « Xcas » lui donne la valeur 0. « Calc », le tableur d'OOo, lui donne la valeur 2.5. En réalité, $Q_3 = 5$.

Ces logiciels calculent également Q_1 de manière erronée.

Sauf erreur, le calcul de la médiane m est toujours correct, quelque soit le logiciel de calcul utilisé.

Comment détermine-t-on m, Q_1 et Q_3 conformément aux définitions du programme ?

- ✓ **Médiane :** On ordonne la série statistique observée dans l'ordre croissant. Si elle est de longueur 2p+1, la médiane est la valeur du terme de rang p+1 de la série ordonnée (terme du milieu) ; si elle est de taille 2p, la médiane est la demi-somme des termes de rang p et p+1 (la demi-somme des deux termes du milieu) de la série ordonnée.
- ✓ **Premier quartile :** Le premier quartile Q_1 est le plus petit élément x de la série statistique tel qu'*au moins 25% des données soient inférieures ou égales à* x. Cela signifie, en clair, que si la longueur de la série statistique observée est 4p (respectivement 4p+1, 4p+2, 4p+3), c'est la valeur du terme de rang p

(respectivement p+1, p+1, p+1) de la série ordonnée (en gros, un quart des termes de la série ordonnée sont à gauche du premier quartile, trois quarts à droite).

✓ **Troisième quartile :** Le troisième quartile Q_3 est le plus petit élément y de la série statistique tel qu'*au moins 75% des données soient inférieures ou égales à* y. Cela signifie, en clair, que si la taille de la série statistique observée est 4p (respectivement 4p+1, 4p+2, 4p+3), c'est la valeur du terme de rang 3p (respectivement 3p+1, 3p+2, 3p+3) de la série ordonnée.

Solution :

Algorithme 1 : Tri d'une série statistique avec une fonction Trier décroissante

Entrées :
A : suite de nombres réels (série statistique)
Sorties :
B : suite obtenue en rangeant A dans l'ordre croissant
1 - **début**
  $C \leftarrow (-A)$
  Trier C
  $B \leftarrow (-C)$
  Afficher B
fin

2 - Une activité de l'IREM de Lille porte sur les tris : « Algorithmique : le tri à bulles », à l'adresse : `http://irem.univ-lille1.fr/activites/article129.html`

Algorithme 2 : Calcul de la médiane d'une série statistique

Entrées :
A : série statistique ;
Sorties :
médiane de A ;
début
  $n \leftarrow$ longueur de A ;
  $B \leftarrow$ suite A rangée dans l'ordre croissant ;
  si n *est impair* **alors**
    $m \leftarrow b_{(\frac{n-1}{2}+1)}$;
  sinon
    $m \leftarrow \frac{1}{2}(b_{\frac{n}{2}} + b_{(\frac{n}{2}+1)})$;
  fin
  Afficher m
fin

3 - Voir le fichier du système utilisé.

4.a - En fait, « Programmer le calcul de la moyenne et de la médiane » ne devrait avoir aucun rapport avec le Calcul des probabilités. Mais quand on a besoin de longues suites de nombres, *il est commode d'utiliser un générateur de nombres au hasard*. Avec « scilab » ou « Xcas », on peut parfaitement travailler sur des suites de longueur 1000000.

Il est indispensable de traiter des séries statistiques longues. Ce n'est pas la peine de faire tant d'efforts pour calculer des médianes qu'on pourrait trouver à la main. De nos jours, des séries de plusieurs millions de nombres sont courantes.

4.b - À partir du moment où on utilise un générateur de nombres aléatoires, il est difficile de ne pas poser cette question, qui se rattache à la signification de la médiane. Chaque nombre de la suite A a autant de chances de tomber dans l'intervalle $[0,\frac{1}{2}]$ que dans l'intervalle $[\frac{1}{2},1]$. Par conséquent, il devrait y avoir à peu près autant de nombres $\leqslant \frac{1}{2}$ que de nombres $\geqslant \frac{1}{2}$ dans la suite A. Donc m devrait être très voisin de $\frac{1}{2}$.

Il y a des résultats théoriques là-dessus : si on faisait tendre n vers $+\infty$, m - qui dépend de n - tendrait vers $\frac{1}{2}$.

5.a & 5.b - Le terme de rang q a bien la propriété qu'au moins un quart des termes de A lui sont inférieurs ou égaux. C'est bien le plus petit élément de A qui ait cette propriété puisque tout terme de A qui lui serait strictement plus petit serait de rang dans B strictement plus petit que q. Toutes les égalités du tableau qui suit se démontrent de même.

6 -

Algorithme 3 : Calcul de Q_1 et Q_3

Entrées :
A : série statistique ;
Sorties :
Q_1 et Q_3, premier et troisième quartiles de A ;
début
  $n \leftarrow$ longueur de A ;
  $B \leftarrow$ suite A rangée dans l'ordre croissant ;
  $r \leftarrow$ reste de la division euclidienne de n par 4 ;
  $q \leftarrow$ quotient de la division euclidienne de n par 4 ;
  si $r = 0$ **alors**
    $Q_1 \leftarrow b_q$;
    $Q_3 \leftarrow b_{3q}$;
  fin
  si $r = 1$ **alors**
    $Q_1 \leftarrow b_{q+1}$;
    $Q_3 \leftarrow b_{3q+1}$;
  fin
  si $r = 2$ **alors**
    $Q_1 \leftarrow b_{q+1}$;
    $Q_3 \leftarrow b_{3q+2}$;
  fin
  si $r = 4$ **alors**
    $Q_1 \leftarrow b_{q+1}$;
    $Q_3 \leftarrow b_{3q+3}$;
  fin
  Afficher Q_1, Q_3 ;
fin

7 - Le professeur choisira peut-être de demander à ses élèves d'écrire eux-mêmes l'algorithme exécutable. Sinon, ils prendront connaissance du fichier de calcul ad hoc.

8 - Voir le corrigé de la question **4**. Q_1 et Q_3 auront des valeurs voisines de $\frac{1}{4}$ et de $\frac{3}{4}$. Il y a aussi des résultats théoriques là-dessus : si on faisait tendre n vers $+\infty$, Q_1 et Q_3 - qui dépendent de n - tendraient respectivement vers $\frac{1}{4}$ et $\frac{3}{4}$.

Pour aller plus loin :
Calculer les déciles d_1 et d_9 de la série statistique (en vue du tracé de la boîte à moustaches associée). C'est facile et fastidieux.