



Médiane d'une série statistique

1 Médiane d'une série statistique

Soit (\mathcal{S}) une *série statistique*, c'est à dire une *suite* ou *liste* de nombres réels¹, notée

$$(\mathcal{S}) = (x_1, \dots, x_n).$$

n s'appelle la *taille* ou *longueur* de (\mathcal{S}) . Nous n'imposons aucune contrainte à ces nombres. En particulier, nous ne supposons pas que ce sont des valeurs prises par des variables aléatoires^{2, 3}. Notons aussi que des égalités sont possibles et que x_1, \dots, x_n ne sont pas forcément rangés dans l'ordre croissant⁴. Par exemple,

$$(\mathcal{S}_1) = (3, 2, 2, 3, 3) \quad (\mathcal{S}_2) = (2, 3, 3, 2) \quad (\mathcal{S}_3) = (2, 2, 2, 2) \quad \text{et} \quad (\mathcal{S}_4) = (4, 7, 1, 3, -2)$$

sont des séries statistiques. Disons tout de suite que ce sont des séries statistiques sans intérêt. La statistique s'est développée en tant que science pour extraire *le plus d'information possible de grandes listes de nombres*. Ceci dit, les définitions doivent s'appliquer même aux cas les plus simples.

1.1 Définition de la médiane

On voudrait que la médiane⁵ soit le « milieu » de la série statistique (\mathcal{S}) , c'est à dire un nombre m tel qu'il y ait dans (\mathcal{S}) autant de valeurs inférieures ou égales à m que de valeurs supérieures ou égales à m . Ainsi présentée, ce serait une des notions les plus simples de la statistique descriptive (avec le minimum, le maximum et la moyenne de (\mathcal{S})).

Mais cette idée simple réserve quelques surprises :

- (\mathcal{S}_1) n'aurait pas de médiane
- (\mathcal{S}_2) en aurait une infinité, à savoir tout nombre de l'intervalle $]2, 3[$
- (\mathcal{S}_3) en aurait une seule, le nombre 2
- (\mathcal{S}_4) en aurait également une seule, le nombre 3.

De plus, le terme « milieu », issu de la géométrie, est trompeur et devrait être évité. En effet, si l'on représente les séries statistiques précédentes sur un axe gradué, (\mathcal{S}_1) et (\mathcal{S}_2) seront représentées par 2 points seulement (éventuellement accompagnés de leurs effectifs), (\mathcal{S}_3) par un seul point : cette représentation graphique complique les choses au lieu de les simplifier.

Ordonner (\mathcal{S}) ⁶ : On fait un pas décisif si l'on pense à *ordonner* la série statistique donnée,

1. que l'on a quelque bonne raison de vouloir étudier
2. Ranger, classer, regrouper des nombres sont des activités antérieures à l'apparition du Calcul des probabilités.
3. Quand c'est le cas, on dit plutôt que (\mathcal{S}) est un échantillon.
4. au sens large
5. Cela concerne toutes les classes à partir de la Troisième, voir [3], p. 34. Nous respectons la terminologie du lexique [2], pp. 85-86.
6. Les machines (tableurs, calculatrices, logiciels de calcul standard) savent ordonner, pratiquement instantanément, des listes très longues de nombres. Leur fonction de tri s'appelle habituellement « sort ». Les élèves

c'est à dire à la transformer en une série statistique de même longueur notée

$$(\mathcal{S}') = (y_1, \dots, y_n)$$

formée des nombres x_1, \dots, x_n rangés dans l'ordre croissant. Autrement dit, (\mathcal{S}') vérifie les inégalités

$$y_1 \leq \dots \leq y_n.$$

Cette définition est claire et sans ambiguïté. Si un nombre apparaît k fois dans (\mathcal{S}) , il apparaît également k fois dans (\mathcal{S}') . Par exemple,

$$(\mathcal{S}'_1) = (2, 2, 3, 3, 3), \quad (\mathcal{S}'_2) = (2, 2, 3, 3), \quad (\mathcal{S}'_3) = (2, 2, 2, 2), \quad (\mathcal{S}'_4) = (-2, 1, 3, 4, 7)$$

On a maintenant envie de dire que les médianes des séries statistiques ci-dessus devraient être successivement 3, tout nombre de l'intervalle $]2, 3[$, 2 et 3. Cette bonne idée appelle deux remarques :

1 - (\mathcal{S}'_1) aurait maintenant une médiane, notons $m = 3$, mais le nombre de valeurs de (\mathcal{S}) inférieures ou égales à m (à savoir 5) serait différent du nombre de valeurs de (\mathcal{S}) supérieures ou égales à m (à savoir 3) ;

2 - La médiane de (\mathcal{S}'_2) ne serait toujours pas unique. On remédie facilement à ce problème en convenant que dans ce cas, la médiane sera le milieu de l'intervalle $]2, 3[$, soit 2.5.

Ces remarques justifient la définition suivante adoptée partout dans le monde et qui, on l'a vu, fait intervenir une convention, autrement dit, est une définition artificielle.

Définition 1.1 Soit $(\mathcal{S}) = (x_1, \dots, x_n)$ une série statistique de taille n , $(\mathcal{S}') = (y_1, \dots, y_n)$ la série statistique ordonnée associée à (\mathcal{S}) .

Si n est impair, noté $n = 2k + 1$, on appelle médiane de (\mathcal{S}) le nombre y_{k+1} ; sinon, et si n est noté $n = 2k$, on appelle médiane de m le nombre $\frac{y_k + y_{k+1}}{2}$. \square

1.2 À quoi sert la médiane ?

Il est clair que c'est la présence de valeurs multiples dans la série statistique (\mathcal{S}) qui a compliqué la définition de la médiane. Un peu de réflexion montre que

1 - si n est impair (noté $n = 2k + 1$) et si $y_k < m = y_{k+1} < y_{k+2}$, il y a $k + 1$ éléments de (\mathcal{S}) inférieurs ou égaux à m et $k + 1$ éléments de (\mathcal{S}) supérieurs ou égaux à m : l'égalité des effectifs est réalisée. C'est le cas de (\mathcal{S}'_4) , ce n'est pas le cas de (\mathcal{S}'_1) comme on l'a déjà remarqué ;

2 - Le cas n pair (noté $n = 2k$) et $y_k < y_{k+1}$ est un autre cas d'égalité des effectifs.

On peut donc dire que la définition (1.1) fournit souvent le milieu de la série statistique au sens naïf, c'est à dire au sens qu'on voulait lui donner au départ. On comprend alors pourquoi on considère la médiane comme un paramètre de position ou de localisation.

La médiane est un paramètre très significatif :

- ✓ Si un employé constate que son salaire est supérieur au *salaire médian* de sa société, il en déduira qu'il appartient à la moitié des employés les mieux payés. La comparaison de son salaire et du *salaire moyen* ne lui permettrait pas d'en tirer cette conclusion.

les utilisent sans problème et bien entendu sans les avoir programmées eux-mêmes, de la Troisième aux classes terminales.

- ✓ Dans l'exemple grand public suivant : « En 2007, le *revenu fiscal médian* par ménage d'Anzin était de 11.153 euros, ce qui plaçait cette ville au 30.453^{ème} rang parmi les 30.714 communes de plus de 50 ménages en métropole » (cf. [1]), la médiane a été préférée à la moyenne.

1.3 Quelques propriétés de la médiane

Si les valeurs d'une série statistique (\mathcal{S}) subissent une transformation affine notée $x \mapsto a \cdot x + b$, les valeurs x_k deviennent $a \cdot x_k + b$; évidemment les valeurs y_k deviennent $a \cdot y_k + b$ (puisque'il s'agit en fait des nouvelles valeurs des x_k écrites dans un ordre différent) si bien que m devient $a \cdot m + b$, d'après la définition (1.1) : la médiane subit la même transformation affine. On sait que la moyenne a également cette propriété.

Le lecteur constatera à l'usage qu'il n'y a pas de relation simple entre la médiane et la moyenne d'une série statistique

- ✓ Par exemple, les moyennes de (\mathcal{S}_1), (\mathcal{S}_2), (\mathcal{S}_3) et (\mathcal{S}_4) sont respectivement 2.6, 2.5, 2 et 2.6. Seules les moyenne et médiane de (\mathcal{S}_2) (respectivement (\mathcal{S}_3)) sont égales.
- ✓ Si l'on modifiait (\mathcal{S}_2) en remplaçant l'un des 2 par -10, la moyenne deviendrait - 0.5 alors que la médiane ne changerait pas.
- ✓ Plus généralement, si on modifie les valeurs des éléments d'une série statistique, il est clair que sa médiane ne change pas *tant que les inégalités larges qui existent entre ces éléments ne sont pas affectées*. Par exemple, si le maximum grandit ou si le minimum diminue, la médiane ne change pas alors que la moyenne change. On dit que *la médiane est insensible aux variations des valeurs extrêmes*.

2 Programmer le calcul de la médiane

2.1 Avec un tableur

Au moins en Troisième, il semble préférable d'utiliser un tableur. Dans ce cas, il est facile, après avoir saisi la série statistique, de l'ordonner et de lire le terme médian de cette série si sa longueur n est impaire ou de calculer la demi-somme des termes médians si elle est paire (autrement dit, d'appliquer la définition (1.1)).

2.2 Avec un logiciel de calcul

Le calcul de la médiane est facile. Comme avec un tableur, le plus long est en général de saisir les données. Ensuite, grâce à sa fonction de tri, on les ordonne en un instant. Le calcul comportera nécessairement une instruction conditionnelle (if ... then ... else ... end) puisqu'il faut tester si n est pair ou impair.

Voici un algorithme exécutable par « scilab »⁷ : voir le fichier téléchargeable « Mediane » (fichier sce)⁸ :

7. scilab pour les lycées

8. On aurait pu extraire la valeur de n de S à l'aide d'une commande « `n=size(S,"c");` », ce qui aurait permis de supprimer la commande « `n=input("n=");` », mais dans le cas considéré où la série statistique S est saisie, sa longueur est évidemment connue.

Listing 1 – source scilab

```
// Médiane d'une série statistique S
// Entrées
S=input("S=");
// Traitement des données
n=size(S,"c");// n est la longueur de S.
Y=trier(S);
r=reste(n,2);
q=quotient(n,2);
  if r==0 then
    med=(Y(q)+Y(q+1))/2;
  else
    med=Y(q+1);
  end
// Sorties
disp('La médiane de S est égale à' +string(med))
```

Références

- [1] - Anzin, article de l'encyclopédie en ligne Wikipedia
<http://fr.wikipedia.org/wiki/Anzin>
- [2] - 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>
- [3] - Programmes du collège, programmes de l'enseignement de mathématiques, Classe de troisième (BO spécial n° 6 du 28 août 2008)
http://media.education.gouv.fr/file/special_6/52/5/Programme_math_33525.pdf
- [4] - Ressources pour faire la classe, Ressources pour le Collège - Probabilités au Collège - Éduscol, mars 2008
http://media.education.gouv.fr/file/Programmes/17/6/doc_acc_clg_probabilites_109176.pdf
- [5] - Scilab pour les lycées
<http://www.scilab.org/fr/education/lycee>

