

Médiane

31 août 2017

Étant donné un ensemble E de valeurs, une médiane est une valeur m (pas forcément dans E) tel qu'il y a autant de valeurs au dessus que en dessous dans E :

$$\#\{v \in E \mid v < m\} = \#\{v \in E \mid v > m\} .$$

Il se peut qu'il y ai plusieurs médiane (comme dans l'ensemble $E = \{2, 3, 5, 8\}$ par exemple); mais il y a généralement une seul médiane :

1. Montrez que si E est de taille impaire, alors il n'y a qu'un médiane,

En fait, si l'échantillon est représentatif (si la *distribution* des valeurs est une *Gaussienne* ou uniforme), assez grand, et si les valeurs sont entières, alors il n'y aura qu'une seule médiane. En fait, Dans la grande majorité des études statistiques, ces caractéristiques sont réunies.

Nous allons maintenant faire une petite expérience :

2. Vous allez choisir deux classes (ou plus) de l'école de niveaux différents.
3. Pour chacune de ces classes, vous allez récupérer l'âge de tous les élèves ainsi que d'un prof (celui que faisait cours à ce moment là).
4. Calculez l'âge moyen dans chaque classe.
5. Pouvez-vous dire à coup sure le niveau de la classe à l'aide de l'âge moyen ?
6. Maintenant, calculez l'âge médian dans chaque classe.
7. Que pouvez-vous en dire ?

J'espère vous avoir convaincu de l'importance de la médiane. Nous allons maintenant essayer de la calculer efficacement.

8. Décrivez un algorithme pour calculer votre médiane.
9. Implémentez votre algorithme.
10. Écrire un programme de test qui va tester votre implémentation sur un échantillon aléatoire 10 000 valeurs (demandez de l'aide au professeur si besoin).
11. Le premier objectif est la médiane d'un tel échantillon aléatoire en moins de 10 secondes. Faites tourner le test plusieurs fois, si il est trop long, réfléchissez à améliorer votre algorithme.
12. Prouvez que votre programme fait au plus 100 000 000 opérations (attention, une comparaison est une opération).

L'objectif est maintenant d'améliorer votre programme pour qu'il tourne plus vite et puisse traiter de plus grands ensembles de valeurs.

13. Supposons que l'on l'on sache que les valeurs représentent l'âge de personnes. On peut alors supposé qu'elles varient entre 0 et 120. Trouvez un algorithme qui ne face que 1 000 200 opérations sur un ensemble de 1 000 000 valeurs.
14. implémentez-le et faites le tourner sur un échantillon aléatoire 1 000 000 valeurs.
15. Essayez d'autres tailles d'échantillons aléatoires et comparez avec votre programme précédent. Que constatez-vous ?
16. (difficile) Trouvez un algorithme qui ne face que 50 000 000 opérations sur un ensemble de 1 000 000 valeurs quelconques (pas forcément des âges).
17. implémentez-le et faites le tourner sur un échantillon aléatoire 1 000 000 valeurs.