

# Médiane et moyenne

Stéphane Bortzmeyer

<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 14 Novembre 2006. Dernière mise à jour le 12  
Septembre 2008

<http://www.bortzmeyer.org/mediane-et-moyenne.html>

---

Beaucoup de logiciels de test des réseaux, comme le célèbre ping, permettent de répéter le test et affichent la moyenne des résultats obtenus. C'est presque toujours une mauvaise idée, la médiane devrait être utilisée à la place. La médiane est la valeur telle qu'une moitié des mesures est située en dessous.

Un excellent article <<http://shlang.com/writing/mean-delay-considered-harmful.html>> de Stanislav Shalunov (un expert de la mesure des réseaux et un des piliers du groupe de travail IPPM <<http://www.ietf.org/html.charters/ippm-charter.html>> de l'IETF) explique fort bien pourquoi : la moyenne a plusieurs défauts et, notamment, elle n'est pas **robuste**. Une seule mesure franchement hors des limites suffit à la changer beaucoup. Or, de telles mesures sont fréquentes sur Internet. Par exemple, si je mesure le RTT d'un serveur, je peux obtenir un tableau comme (3, 3, 4, 3, 198, 3, 4) (RTT en millisecondes). Sa médiane sera de 3 ms (reflétant le cas de loin le plus courant) mais sa moyenne sera complètement « faussée » par la valeur exceptionnelle de 198 ms (la moyenne vaudra 31 ms).

Si on tient à utiliser la moyenne, il faut au moins l'accompagner de son indispensable auxiliaire, l'écart-type (merci à Stéphane Bunel pour ce rappel et pour son excellente page sur la question <<http://stephane.bpf.st/div/math/moyenne-et-ecart-type>>). Une autre technique est d'éliminer du calcul de la moyenne les "outliers", les valeurs qui s'éloignent « trop » de la moyenne, par exemple de plus de deux écarts-types.

Si je peux me permettre un peu de publicité, le programme echoping <<http://echoping.sourceforge.net>> est un des rares programmes de mesure réseaux à afficher la médiane des valeurs mesurées (et à afficher systématiquement l'écart-type à côté de la moyenne ; quant à l'élimination des "outliers", elle a fait l'objet d'un "patch" d'Andy Juniper, intégré depuis la version 6). Voici un exemple des résultats affichés par echoping (notez la différence entre la médiane et la moyenne, pour un site dont le temps de réponse varie considérablement) :

```
Minimum time: 0.993819 seconds (258 bytes per sec.)
Maximum time: 7.264942 seconds (35 bytes per sec.)
Average time: 2.677578 seconds (96 bytes per sec.)
Standard deviation: 2.059113
Median time: 1.860476 seconds (138 bytes per sec.)
```

Avec la dernière version d'echoping (qui inclus le "*patch*" "*outlier*") et l'option `-N 1`, cela donnerait :

```
Minimum time: 0.858710 seconds (298 bytes per sec.)
Maximum time: 24.057933 seconds (11 bytes per sec.)
Average time: 4.521010 seconds (57 bytes per sec.)
Standard deviation: 6.765369
Median time: 1.488127 seconds (172 bytes per sec.)
Average of values within 1 standard deviations: 2.350241
```

D'autres exemples d'un tel programme sont [thrulay](http://www.internet2.edu/~shalunov/thrulay/) <<http://www.internet2.edu/~shalunov/thrulay/>> de Stanislav Shalunov ou bien le [owamp](http://e2epi.internet2.edu/owamp/) d'Internet2 <<http://e2epi.internet2.edu/owamp/>>.

Malheureusement, les experts en réseaux ont en général une connaissance très limitée des statistiques. Pour arranger cela, une excellente source est "*Basic Statistics*" <<http://www.statsoft.com/textbook/stbasic.html>>, un texte de statistiques expliquées aux ingénieurs non-mathématicien. Pour plus de réflexions sur le manque gênant de culture statistique chez les programmeurs, j'invite à lire l'excellent article de Zed Shaw <[http://www.zedshaw.com/essays/programmer\\_stats.html](http://www.zedshaw.com/essays/programmer_stats.html)>, avec plein d'exemples écrits en R.

Et pour implémenter un calcul de médiane ? echoping, écrit en C, appelle simplement `qsort` avant de choisir l'élément du milieu du tableau ainsi trié. Cet algorithme est en  $O(n \cdot \log(n))$  mais d'autres algorithmes sont en  $O(n)$  (par exemple le "*median of medians*").

Pour PostgreSQL, Sébastien Dinot a écrit une jolie implémentation <<http://archives.postgresql.org/pgsql-fr-generale/2008-06/msg00003.php>>, qui est utilisée pour mon système de statistiques <[/auto/summary.html](http://auto/summary.html)>.

L'utilisation abusive de la moyenne à la place de la médiane ne se rencontre pas uniquement dans les réseaux informatiques. On la trouve aussi souvent en politique. Par exemple, lorsqu'il faut gommer un peu les inégalités salariales, on annonce un salaire moyen car il est nettement plus élevé que le salaire médian qui est pourtant bien plus représentatif du vécu de la majorité. L'écart entre les deux nombres est d'autant plus élevé que les inégalités sont accentuées.

C'est ainsi que l'INSEE nous apprend <[http://www.insee.fr/fr/ffc/docs\\_ffc/IP1067.pdf](http://www.insee.fr/fr/ffc/docs_ffc/IP1067.pdf)> qu'en France en 2004, le salaire moyen annuel net était de 22 193 € alors que le salaire médian (celui qui sépare les salariés en deux parties égales, ceux qui gagnent plus et ceux qui gagnent moins) était de seulement 17 802 €. Paradoxe intéressant, la grande majorité des français vivent donc avec moins de la moyenne... C'est logique du point de vue mathématique, mais contraire à l'intuition habituelle de ce que signifie « moyen ».