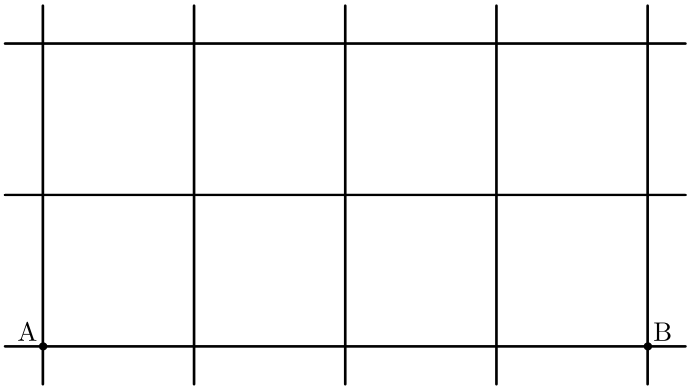
Quartiles : incertitude liée à une mesure

Cette note se rattache à la notion de quartile[1].

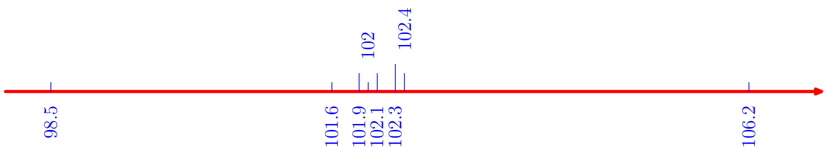
À titre d'exemple, il est impossible de mesurer exactement la distance entre deux points A et B. Cela est dû au fait que les instruments et les opérateurs sont imprécis. De plus, il peut y avoir des impondérables. Ainsi, on a demandé à des élèves de mesurer en cm la distance entre deux points A et B du carrelage de la classe à l'aide d'une réglette graduée de 0 à 40 cm.



Au jugé, ces points sont distants d'à peu près 1 m. Les élèves travaillent par groupes de deux. On obtient la série statistique $(\mathcal{S})$ suivante :

102,3	101,6	102,4	98,5	102,4	102,3	101,9	102	106,2	102,1	101,9	102,1	102,3

que l'on peut représenter sur un axe gradué :



Le problème est maintenant de choisir *la valeur que l'on considérera comme la distance de A à B*. Cette réflexion s'impose puisque les réponses des 13 paires d'élèves ne sont pas tous égales. Ordonnons $(\mathcal{S})$[2]. On obtient

98,5	101,6	101,9	101,9	102	102,1	102,1	102,3	102,3	102,3	102,4	102,4	106,2

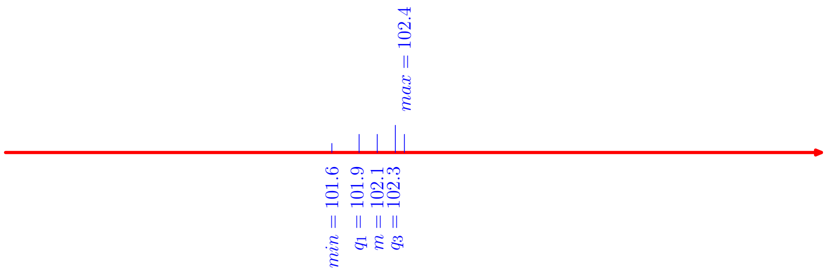
Il saute aux yeux qu'il y a deux anomalies, à savoir les valeurs extrêmes, parce qu'elles sont éloignées des autres valeurs[3]. Cela signifie que deux paires d'élèves ont mal travaillé ; leurs mesures n'ont pas grand'chose à voir avec le problème posé. En statistique, ces anomalies s'appellent des valeurs aberrantes. Il est normal de les effacer. Pour définir la longueur AB, il ne nous reste plus maintenant que la série statistique ordonnée suivante :

1. La notion d'incertitude liée à une mesure est citée dans les *Programmes de l'enseignement des Mathématiques, des SVT, de Physique-Chimie du Collège*, BO n˚6 du 19 avril 2007 Hors-Série, p. 59.
2. Ordonner $(\mathcal{S})$ doit être un réflexe !
3. On s'attend évidemment à ce que les réponses des élèves soient très proches l'une de l'autre

101,6	101,9	101,9	102	102,1	102,1	102,3

102,3	102,3	102,4	102,4

que l'on peut appeler série statistique ordonnée tronquée. Elle est ici de taille 11. On peut la résumer ainsi (q_1 et q_3 désignent ses premier et troisième quartile, m sa médiane, min son minimum, max son maximum) :



ou à l'aide du tableau suivant :

min	q_1	m	q_2	max
101.6	101.9	102.1	102.3	102.4

À noter que la hauteur des traits qui repèrent les valeurs de la série statistique est choisie proportionnellement aux effectifs. Sachant que 101.6 apparaît une seule fois dans la série, on lit sur le dessin ci-dessus que 101.9, 102.1 et 102.4 apparaissent deux fois et que 102.3 apparaît trois fois.

On pourrait choisir comme valeur de la distance de AB la valeur qui apparaît le plus fréquemment dans la série statistique[4], ici 102,3. Cette idée n'est pas bonne car il suffirait d'une seule nouvelle mesure à 101.9 par exemple pour nous obliger à choisir entre 101.9 et 102.4 qui sont des valeurs assez différentes : ce procédé fournirait donc des résultats instables (variant beaucoup d'un opérateur à un autre).

Comme valeur de la distance de AB, on pourrait choisir la médiane, qui est un candidat très valable, mais on préfère la moyenne (arithmétique) parce que, dans la théorie des erreurs initiée par K. Gauss, on considère que chaque mesure est la somme de la valeur exacte de la distance AB et d'une erreur. Cette erreur est envisagée dans un modèle probabiliste qui permet de démontrer que la somme des erreurs est voisine de 0 quand le nombre de mesures est grand. En additionnant les mesures pour en faire la moyenne, on fait donc plus ou moins disparaître l'erreur. L'argument est un peu spécieux car dans les mêmes conditions, ma médiane convergera aussi vers la vraie valeur de la distance de AB.

Pour cette raison, on posera AB=102,2 cm (arrondi de la moyenne de la série tronquée).

Remarque sur les valeurs aberrantes :

Il était facile ici de repérer les valeurs aberrantes et de les éliminer. Mais quand on a à exploiter des séries statistiques de plusieurs milliers de termes, on ne peut pas même pas se rendre compte s'il y a de telles valeurs. On peut essayer d'automatiser le traitement de ces grandes séries statistiques, par exemple en éliminant systématiquement les 10% de valeurs les plus petites et les 10% de valeurs les plus grandes[5]. Ce faisant, on élimine certainement les valeurs aberrantes, mais aussi des valeurs qui ne le sont pas, modifiant ainsi la moyenne indûment. Il ne peut y avoir de traitement miracle qui s'appliquerait à tous les cas. Seul un statisticien ayant de l'expérience saura (?) ce qu'il convient de faire, selon la nature des données.

4. qu'on appelle le mode de la série
5. ce qui introduit les premier et neuvième déciles

Références

1 - Programme de l'enseignement des mathématiques, des SVT, de physique-chimie du collège, B.O. N˚6 19 AVRIL 2007, Hors-série, Annexe 2 Mathématiques

ftp://trf.education.gouv.fr/pub/edutel/bo/2007/hs6/MENE0750668A_annexe2.pdf

2 - Mathématiques Collège, Projet de document d'accompagnement – Probabilités – 17 mars 2008

http://eduscol.education.fr/D0015/doc_acc_clg_probabilites.pdf