MEDIANE ou MOYENNE

Michel MYARA
IRES de Toulouse-Groupe lycée

Lorsque nous enseignons la statistique descriptive au collège ou au lycée, nous commençons très tôt, en classe de quatrième, par le calcul de la moyenne d'une série statistique. Le calcul et l'interprétation de la médiane n'interviennent qu'un an plus tard, en classe de troisième, et uniquement pour les séries de valeurs discrètes.
A aucun moment dans le cursus scolaire, le problème de la pertinence du choix de l'un ou l'autre des paramètres n'est abordé. Dès lors, il n'est guère étonnant que nos élèves, et par suite, les consommateurs de statistiques qu'ils deviendront utilisent sans discernement la moyenne ou la médiane.
Il est inconcevable, qu'à une époque où les moyens de calculs ne sont plus un obstacle, nous puissions encore commettre des erreurs dans les choix de modèles statistiques comme celles décrites par Bernard PY dans son ouvrage "Statistiques sans formule mathématique". Mais examinons plutôt quelques situations mettant en défaut l'un des deux paramètres : médiane ou moyenne.

UN CAS D'ECOLE

Que penser de la distribution suivante d'observations discrètes ? Quel est son "centre"? Quel nombre est significatif comme "valeur centrale"?
(2, 2, 2, 4, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 8, 8, 9, 1000)
Le calcul de la moyenne donne : $\frac{1104}{19} = 58{,}1$.
Est-il raisonnable de penser que le "centre" de cette série de valeurs se situe significativement aux alentours de 58 ? Bien évidemment, non !
Visiblement, il y a un intrus : la valeur 1000 n'est pas du même ordre de grandeur que les autres.
Erreur de saisie ou réalité "hors normes" ?
Sans réponse à cette question, l'élagage de la série statistique permettrait d'aboutir à des résultats convenables, mais qui perdraient toute validité scientifique.

Par conséquent, dans le cas présent, le choix de la moyenne comme paramètre central s'avère être un mauvais choix.

Que donne la médiane ?
La médiane est la 10^{ième} valeur de la série soit 6.

Laissant 9 observations "avant", et 9 observations "après", la médiane se trouve être beaucoup plus significative.

On a utilisé ici toutes les observations de la série. Une médiane de 6 est effectivement plus réaliste qu'une moyenne de 58,1.
Le fait que la médiane ne soit pas calculée directement sur les valeurs observées permet de gommer l'effet de possibles valeurs aberrantes aux bornes de la série statistique.
La moyenne obtenue dans ce cas ne peut en aucune manière être qualifiée de fausse : elle est mal adaptée, c'est tout.

NOTES ET MOYENNES

Les appréciations de qualité de certaines observations se font souvent par des notes : notes d'élèves à un examen, notes du jury aux participants d'une épreuve artistique, notes de qualité de produits testés, etc. Dans la vie courante, on a tendance à comparer telle ou telle note à la moyenne arithmétique de l'ensemble des notes. Implicitement, donc, l'habitude fait en sorte que l'on fait, une fois encore, appel à l'idée de moyenne pour repérer ce fameux centre, qui ne serait, ni fort, ni faible, tout juste "au milieu".
Voyons sur un exemple ! Pour une classe de 35 élèves, les notes obtenues à un devoir sont présentées dans le tableau ci-dessous :

Notes	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Effectifs	1	0	0	1	2	4	3	4	4	2	0	1	0	2	3	2	1	0	1	2	2
ECC	1	1	1	2	4	8	11	15	19	21	21	22	22	24	27	29	30	30	31	33	35

La moyenne de la classe est 9,9 alors qu'il n'y a que 14 notes supérieures à 10.
La note médiane est 8 ($18^{ième}$ valeur sur 35) permet de mieux situer le « centre » de cette série de notes.
L'élagage de la série par suppression des notes 0 et 20 donne les résultats suivants :
Moyenne = 9,6 et Médiane = 8
Dans ce cas, l'élagage de la série statistique ne change que peu de choses.

Là encore, la moyenne n'est pas forcément le bon indicateur car la notion "d'appréciation" qui nous intéresse ici, induit directement une logique d'ordre, de rang, de classement, qui est à la base de la définition de la médiane.
La note de l'étudiant médian à un examen permet de mieux situer un individu par rapport à l'ensemble que classique la note moyenne de 10/20.

SALAIRE MOYEN ET SALAIRE MEDIAN

Plages de salaires		Fréquences en %
B.inf	B.sup	
200	400	0,5
400	600	1
600	800	4
800	1000	7
1000	1200	10
1200	1400	13
1400	1600	12
1600	1800	10
1800	2000	8
2000	2200	6
2200	2400	4
2400	2600	3
2600	2800	3
2800	3000	2
3000	3200	1,5
3200	3400	1,5
3400	3600	1
3600	3800	1
3800	4000	1
4000	4200	0,5
4200	4400	0,5
4400	4600	0,5
4600	10000	9
		100

Dans l'exemple suivant les données sont extraites du site *www.salairemoyen.com* , données mises à disposition par l'INSEE.

Répartition des salaires nets mensuels des salariés en France métropolitaine fin 2012

D'après ce tableau, le salaire net mensuel moyen des français fin 2014 est de 2239 €.

Certains trouveront ce résultat trop élevé, d'autres trop faible. Le fait est que cette valeur n'est qu'une moyenne. Il ne signifie pas que la majorité des Français touchent tous les mois cette somme mais que, si tous les salariés Français touchaient le même salaire mensuel, ce salaire serait de 2239 €.

Plus significatif est le salaire médian net mensuel de 1650 €. Ce dernier sépare la moitié des Français qui gagnent moins que cette somme de l'autre moitié qui gagne plus.

Si l'on recalcule les paramètres précédents en excluant la plage des salaires les plus élevés les résultats obtenus sont alors :
Salaire mensuel net moyen = 1730 €
Salaire médian = 1580 €

L'influence des hauts salaires est donc considérable sur la moyenne alors qu'elle est minime sur la médiane. Le salaire moyen cache de grandes disparités et ne peut donc pas être utilisé pour des comparaisons.

On ne peut pas dire que la moyenne soit fausse, elle est simplement inadaptée à l'utilisation que l'on en fait, à savoir : situer un salaire donné par rapport à l'ensemble.

LE CAS D'UNE MOYENNE PARTIELLE

Ici, il ne s'agit, non plus de détails de calcul, mais simplement de la prise de conscience de la "puissance psychologique" que nous impose traditionnellement cette notion de moyenne dans nos perceptions des événements de la vie courante.
Prenons, par exemple, le cas d'un professeur de philosophie qui, pour répondre à la demande insistante des 800 étudiants qu'il a devant lui dans l'amphithéâtre, dirait : "Je ne peux pas vous donner vos notes, car je n'ai pas fini de corriger vos copies, mais la moyenne de celles que j'ai corrigées jusqu'à présent se situe aux alentours de 11/20."
Il percevra sans doute un grand soupir de soulagement dans l'auditoire. Soupir complètement injustifié de la part de chacun des étudiants puisque, d'une part, aucun d'eux n'est capable de se situer par rapport aux autres en la matière et que, d'autre part, ils ne connaissent pas la variabilité de l'échelle de notation du professeur.

UNE ETUDE DE LA CEPEC-S.A.

D'après le CEPEC (Centre d'Etude de Projet Economique, organisme suisse de benchmarking économique)
La médiane et la moyenne sont des outils statistiques couramment utilisés pour les comparaisons de salaires. Ces deux outils, qui ont chacun leurs avantages et leurs inconvénients, devraient être employés de manière complémentaire selon le cadre de l'étude.
La médiane est généralement plus représentative pour l'analyse des petits groupes, car la présence de quelques cas extrêmes peut tirer fortement la moyenne vers le haut ou vers le bas et donner ainsi une image qui n'est pas représentative de la grande majorité des données.
Prenons un exemple simple pour illustrer le phénomène : un groupe de 10 salaires, dont 8 de 1000 € et 2 de 5000 €. La moyenne est 1800 et la médiane de 1000. La moyenne est tirée fortement vers le haut, dans une zone où il n'y a justement pas de salaires. La médiane est par contre représentative de la majorité des salaires analysés.
Dans les grands groupes, de plusieurs dizaines ou centaines de données analysées, la médiane et la moyenne tendent à se confondre. Les dispersions vers le haut et vers le bas tendent à se compenser et d'autre part les cas extrêmes isolés ont peu d'influence sur la moyenne qui est fonction de l'effectif de la population.
Dans les grands groupes, c'est toutefois la médiane qui peut donner une image peu représentative de la réalité. C'est le cas lorsque les groupes analysés sont formés de sous-groupes eux-mêmes assez homogènes, mais bien distincts les uns des autres.
Pour illustrer ce phénomène, prenons à nouveau un cas simple avec deux groupes de 100 salaires. Dans chacun des deux groupes, il y a un sous-groupe de salaires de 2000 € et un sous-groupe de salaires de 3000 € :

Groupe 1	
Salaire	Effectif
2000	60
3000	40
Moyenne : 2400	
Médiane : 2000	

Groupe 2	
Salaire	Effectif
2000	40
3000	60
Moyenne : 2600	
Médiane : 3000	

L'écart entre les moyennes est de 200 soit 4% de la moyenne globale des deux groupes qui est de 2500.
L'écart entre les médianes est de 1000 soit 40% de la médiane globale des deux groupes qui est de 2500.
L'image donnée par les médianes est donc, dans ce cas, fortement biaisé.

L'exemple ci-dessus est évidemment simpliste. Dans son principe, il correspond toutefois bien à certaines analyses statistiques basées sur les médianes, qui sont publiées par des organismes officiels. Ainsi par exemple, l'Office Fédéral de la Statistique (Suisse) publie une statistique des salaires selon le niveau des qualifications requises pour un poste de travail, en comparant les salaires des femmes et des hommes. Cette statistique distingue 4 niveaux de qualifications, des plus exigeantes aux plus simples. Les deux niveaux les plus élevés sont présentés de manière regroupée. Dans la mesure où les proportions de femmes et d'hommes sont probablement inversées entre les deux niveaux de qualifications, il est probable que l'écart calculé sur la base de médianes biaise l'image de la réalité, alors que les moyennes seraient plus représentatives. Cette même critique s'applique encore plus nettement aux comparaisons des salaires médians pour l'ensemble des niveaux de qualifications.

PRIX MEDIAN OU PRIX MOYEN ?

Dans leurs communiqués de presse, la Fédération des chambres immobilières du commentent désormais l'évolution du prix des propriétés en se basant sur le prix médian.
Aux États-Unis, la National Association of Realtors utilise cette mesure depuis longtemps pour rapporter le prix des propriétés.

Extrait du communiqué de presse de la FCIQ :
Voyons pourquoi, en immobilier, la médiane est généralement un meilleur indicateur pour traiter du prix des propriétés que la moyenne.
La médiane est la valeur qui permet de partager une série en deux parties égales.
Dans le cas qui nous intéresse, le prix médian est celui qui indique que la moitié des transactions ont eu cours à un prix inférieur et l'autre moitié à un prix supérieur. Par exemple, un prix médian de 150 000 $ indique que 50 % des propriétés se sont vendues en deçà de 150 000 $ et l'autre 50 % à un prix supérieur.

L'avantage de la médiane comme mesure de tendance centrale est qu'elle n'est pas influencée par les valeurs extrêmes.
À l'inverse, l'inconvénient du prix moyen, comme toute moyenne d'ailleurs, est justement qu'il est influencé par ces valeurs extrêmes, ce qui peut créer des distorsions majeures susceptibles de fausser l'interprétation des données. On n'a qu'à penser à un secteur géographique où le prix des propriétés tourne généralement autour de 150 000 $ à 200 000 $ et où, pour un mois donné, une propriété qui n'est pas représentative du secteur est vendue à 2 000 000 $. Cette transaction vient tirer la moyenne vers le haut et du même coup, la croissance des prix dans ce secteur sera surestimée. Le prix médian, lui, n'est pas influencé par cette transaction à 2 000 000 $. Il donne donc une meilleure lecture du marché, tant en ce qui a trait aux prix qu'au taux de croissance entre deux périodes.
En conclusion, qu'on utilise le prix médian ou le prix moyen, plus le nombre de transactions à partir desquelles ils sont calculés est faible, plus il faut les interpréter avec prudence. La norme à cet effet est d'avoir un minimum de 30 ventes. Sous ce seuil, le risque est très important.

Ce qui précède met en cause la validité de l'utilisation d'une moyenne dans le cas où l'effectif de la population est faible. En effet, plus l'effectif total est faible, plus la moyenne est influencée par les valeurs extrêmes.

ALORS, MEDIANE OU MOYENNE ?

Comme le montrent les exemples précédents, le paramètre central idéal n'existe pas. Tout ce que l'on peut faire c'est essayer de choisir, suivant la situation étudiée, le paramètre central le mieux adapté.
Dans ce but, le statisticien Yule (XIXème siècle) a défini six propriétés souhaitables pour les valeurs centrales et comparé sur ces critères la médiane, la moyenne et le mode. Ces propriétés sont les suivantes :

(1) Etre définie de façon objective
Deux personnes différentes traitant la même information doivent trouver le même résultat en ce qui concerne le calcul des valeurs centrales. Ceci est vrai pour la moyenne et la médiane mais pas pour le mode qui dépend du choix de la partition en classe adoptée.

(2) dépendre de toutes les observations
La modification d'une seule observation doit entraîner une modification de la valeur centrale. Ceci est vrai de la moyenne mais pas du mode et de la médiane.

(3) avoir une signification concrète
Bien que la moyenne paraisse "naturelle" elle est en fait très abstraite alors que le mode peut être défini comme la situation la "plus fréquente" et la médiane comme celle "qui divise en deux la distribution" (un individu sur deux a une valeur inférieure ou supérieure à celle-ci). Le caractère abstrait de la moyenne ressort bien quand on l'applique à des caractères discrets (e.g. que signifie 2.5 enfants par femmes ?)

(4) être simple à calculer
Cette préoccupation du XIXe siècle n'est plus de mise à l'époque des ordinateurs ... Toutes les valeurs centrales sont actuellement simples à calculer.

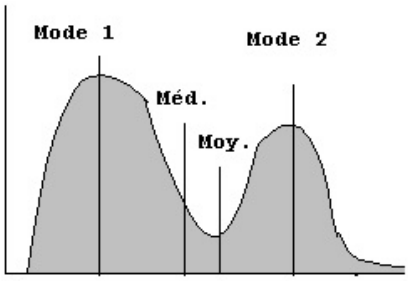
(5) être peu sensible aux fluctuations d'échantillonage
Il s'agit en apparence de l'inverse de la propriété (2). Mais on peut dire que cette propriété définit la robustesse de la mesure face à des erreurs qui peuvent apparaître (données mal codées, valeurs aberrantes). La médiane est la plus trois paramètres.

(6) se prête au calcul algébrique
Lorsque l'on connaît les valeurs centrales de k échantillons *E1...Ek* d'effectifs respectifs *P1...Pk*, peut-on retrouver la valeur centrale de E qui est la réunion de tous ces échantillons ? La réponse est affirmative dans le cas de la moyenne mais négative dans ceux du mode et de la médiane. Ceci est un gros avantage pour le stockage de l'information.
Le tableau ci-dessous permet de résumer les avantages et inconvénients des trois valeurs centrales.

Propriété de Yule	Mode	Mediane	Moyenne
1) est définie de façon objective	-	+	+
2) dépend de toutes les observations	-	-	+
3) a une signification concrète	+	+	-
4) est simple à calculer	+	+	+
5) est peu sensible aux fluctuations d'échantillonnage	-	+	-
6) se prête au calcul algébrique	-	-	+

Ces propriétés n'aident que très peu au choix qui nous intéresse dans cet article.
Après lecture des exemples précédents, il ressort que dans le cas de faibles effectifs la médiane est plus significative que la moyenne car plus l'effectif est faible plus les valeurs extrêmes influent sur la moyenne.

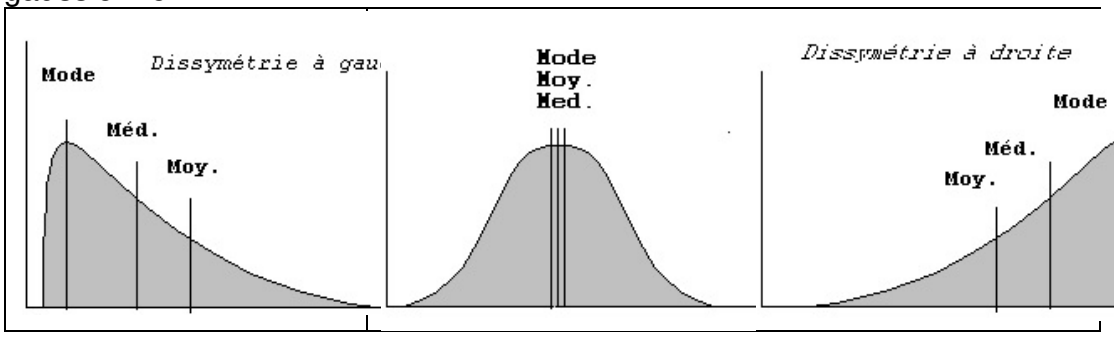
D'autre part, dans le cas d'une distribution multimodale, si l'effectif est faible, nous sommes ramenés au cas précédent sinon, l'exemple 5) (CEPEC) montre que la moyenne est moins biaisée que la médiane et donc plus significative.



Dans la suite, nous nous limiterons aux distributions unimodales.
La médiane induit, de par sa définition, une notion d'ordre ce qui peut être intéressant lorsque la valeur centrale doit permettre de situer une valeur donnée par rapport à l'ensemble de la population.
De plus elle n'est pas sensible aux valeurs extrêmes. (*exemples 1) et 2)*)

Dans le cas d'une distribution unimodale, plus la distribution est symétrique plus la moyenne est significative. Ceci nous amène à étudier les distributions dissymétriques.

Enfin, on peut également remarquer que l'écart entre la moyenne et la médiane est fonction du caractère symétrique de la population étudiée. Cela va jusqu'à l'égalité entre moyenne et médiane pour une distribution gaussienne.



Une remarque s'impose : en cas de dissymétrie, la médiane est toujours plus proche du mode que la moyenne et permet donc de mieux situer une valeur par rapport à l'ensemble.
Nous sommes ainsi ramenés à évaluer la dissymétrie d'une distribution.
Nous disposons pour cela de plusieurs outils plus ou moins simples : le coefficient d'asymétrie de Fisher, le coefficient d'asymétrie de Pearson et le coefficient d'asymétrie de Yule et Kendall (ou de Bowley).

Les deux premiers sont de nature algébrique alors que le dernier (Yule et Kendall), beaucoup plus simple ne fait intervenir que les quartiles de la distribution étudiée :

$$U = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$

Nous aurons toujours : $-1 \leq U \leq 1$ et U est d'autant plus proche de 0 que la distribution est symétrique.
Comme il n'existe pas de critère précis de séparation entre symétrie et asymétrie, il reste à définir des valeurs seuils de ce coefficient permettant choisir entre médiane et moyenne.

Essayons d'appliquer le coefficient d'asymétrie de Yule au choix moyenne-médiane sur les exemples précédents en divisant l'intervalle [-1 ; 1] en trois intervalles de même longueur : [-1 ; -1/3], [-1/3 ; 1/3] et [1/3 ; 1]. Pour se faire, considèrera qu'une série est dissymétrique à gauche si U $\in$ [-1 ; -1/3], dissymétrique à droite si U $\in$ [1/3 ; 1] et symétrique si U $\in$ [-1/3 ; 1/3].

Cas de l'exemple 1 : $Q_1 = 6$; $Q_2 = Me = 6$; $Q_3 = 7$ et $U = \frac{(7-6)-(6-6)}{(7-6)+(6-6)} = 1$
D'après l'étude précédente, la série est considérée comme dissymétrique à gauche. Le paramètre central le plus significatif est donc la médiane.

Cas de l'exemple 2 : $Q_1 = 6$; $Q_2 = Me = 8$; $Q_3 = 14$ et $U = \frac{(14-8)-(8-6)}{(14-8)+(8-6)} = 1/2$
Ici, $U \in [1/3 ; 1]$ donc la série est considérée comme dissymétrique à gauche. Le paramètre central le plus significatif est donc, ici encore, la médiane.

CONCLUSION

D'après ce qui précède, nous pouvons à présent donner quelques indications permettant de choisir de façon pertinente entre la moyenne ou la médiane d'une série statistique.

Critère 1 :
Lorsque le caractère étudié est de type ordinal, la médiane est, dans tous les cas, préférable à la moyenne.

Critère 2 :
Lorsque le caractère étudié n'est pas de type ordinal, le coefficient de dissymétrie de Yule-Kendall (ou de Bowley), noté U, peut être utilisé de la façon suivante :
Si $-1 \leq U \leq -\frac{1}{3}$ ou $\frac{1}{3} \leq U \leq 1$ **alors** la médiane est préférable à la moyenne
Sinon la moyenne est préférable à la médiane.