REPERES - IREM . N° 33 - octobre 1998

A BAS LA MOYENNE !

ou

A propos des paramètres de tendance centrale et de dispersion d'une série statistique

Jean Claude GIRARD,
IREM de Lyon

1. Introduction

Les concepts de tendance centrale et de dispersion sont utilisés pour décrire ou résumer une série statistique. Ces deux concepts sont mal définis ou, plus précisément, on peut les définir de différentes façons. Les mesures du centre ou du milieu [1] d'une série statistique et de la variabilité des valeurs autour de ce milieu ne se sont d'ailleurs pas imposées d'une manière évidente. Le sens de ces paramètres n'est pas intuitif et leur interprétation quelquefois délicate. Si moyenne et écart-type sont employés de façon intensive, d'autres mesures sont possibles et sont utilisées car elles présentent différents avantages.

2. Conceptions à propos de la moyenne

Tout le monde pense savoir ce qu'est la moyenne (arithmétique) d'une série statistique. Peu d'élèves savent cependant donner un sens concret au résultat trouvé et expliquer le pourquoi du calcul effectué.

A la question : "Un étudiant a réussi un examen avec 13,8 de moyenne. Que représente, pour vous, cette valeur ?", 87 étudiants de deuxième année d'IUT ayant

(1) Pour reprendre une expression utilisée pendant longtemps. Voir par exemple l'article "MILIEU à prendre entre les observations" de l'*Encyclopédie Méthodique* (Tome 2, Mathématiques, page 404), Paris, 1785.

reçu un enseignement de statistique en première année (en plus de celui de lycée et collège) ont répondu de la façon suivante :

Paraphrases sur le terme de valeur moyenne
(moyenne pondérée, arithmétique et même algébrique !) : 23

Formule $\frac{1}{n}\sum_{i=1}^{n} x_i$ (ou total divisé par le nombre de coefficients) : 39

Sans réponse, phrases incorrectes ou incompréhensibles : 22

Valeur qui pourrait remplacer toutes les autres (ou représentative de l'ensemble) : 3

(la note que l'étudiant aurait eu s'il avait eu la même note dans toutes les matières,

la note la plus "proche" de toutes les notes obtenues...)

Quelques perles : Il y a autant de notes supérieures à 13,8 que de notes inférieures à 13,8

Il y a des notes supérieures à 10 et d'autres inférieures à 10

Toutes les notes sont proches de 13,8

On constate qu'il est bien difficile pour les étudiants de sortir de la formule qui donne la moyenne et de donner un sens au résultat, même après de nombreuses années de fréquentation de cette valeur caractéristique d'une série statistique.

Trois étudiants seulement citent le fait que l'ensemble des notes obtenues donnent le même résultat (au total, à l'examen) que si chacune avait été de 13,8.

Rares sont ceux qui évoquent (pas toujours avec bonheur) la dispersion des notes autour de la valeur 13,8.

3. Définitions mathématiques

a) Le problème consiste à remplacer la série $x_1, x_2, \ldots, x_n$ par une série constante a,a, ... ,a qui soit la plus "proche" possible de la série de départ.

(i) Si on choisit la métrique euclidienne pour mesurer la distance entre les deux séries c'est-à-dire si on calcule la somme des carrés des écarts $\sum_{i=1}^{n}\left(x_i - a\right)^2$, on peut démontrer (voir §4) que la meilleure valeur de a, c'est-à-dire celle qui minimise cette somme, est la moyenne arithmétique $\bar{x}$ donnée par la formule $\frac{1}{n}\sum_{i=1}^{n} x_i$.

(ii) Si on mesure la distance entre les deux séries en utilisant la norme L_1 c'est-à-dire en calculant la somme des valeurs absolues

des écarts $\sum_{i=1}^{n} |x_i - a|$, on peut démontrer que la meilleure valeur de a, c'est-à-dire celle qui minimise cette somme, est alors la médiane M_e de la série (2).

b) Interprétation géométrique de la moyenne

(i) On représente la série des valeurs x_1, x_2, ... , x_n par le vecteur X de R^n de coordonnées (x_1, x_2, ... , x_n) et on projette orthogonalement ce vecteur sur la n-sectrice des axes c'est-à-dire sur la droite Δ de vecteur directeur 1(1, 1, 1, ... ,1). On obtient le vecteur $\overline{X}$.

Pour le produit scalaire défini par

$$\langle x,y \rangle = \frac{1}{n} \sum_{i=1}^{n} x_i \, . \, y_i,$$

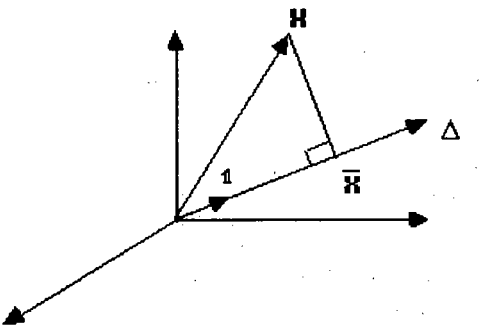
on a :

$$\langle X,1 \rangle = \frac{1}{n} \sum_{i=1}^{n} x_i \, . \, 1 = \bar{x}.$$

Or $\overline{X}$ est le projeté de X sur la droite de vecteur directeur 1, donc $\overline{X} = \langle X,1 \rangle .1$ soit $\overline{X} = \bar{x} \, . \, 1$. Les coordonnées de $\overline{X}$ sont donc toutes égales à $\bar{x}$. Le vecteur $\overline{X}$ est alors le vecteur constant le plus près de X pour la métrique associée au produit scalaire choisi, ce qui signifie que

$$\| X - \overline{X} \| = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

est minimum.



(ii) La distance entre les deux séries est mesurée dans le cas de la norme euclidienne par la quantité

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

qui correspond à l'écart type de la série de départ (ou $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ qui correspond à la variance) et dans le cas de la norme L_1 par la quantité $EAM = \frac{1}{n} \sum_{i=1}^{n} |x_i - Me|$, qui correspond à l'écart absolu moyen de la série de départ.

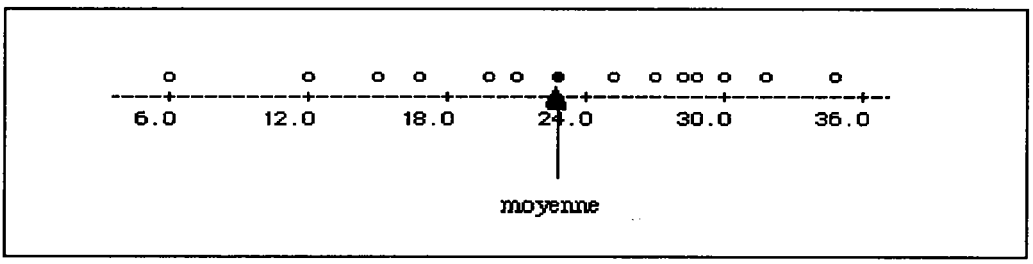
c) Interprétation barycentrique de la moyenne

(i) Propriété : la somme des écarts par rapport à la moyenne est toujours nulle car $\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = 0$. Elle ne fournit

(2) Rappel : une définition simple de la médiane :

Median =	The median is defined to be the single middle value the mean of the two middle values

John W. Tukey, *Exploratory Data Analysis*, 1977. C'est-à-dire : la médiane est définie comme la valeur centrale ou la moyenne des deux valeurs centrales.



donc pas un moyen de mesurer la dispersion. Par contre, c'est peut-être cette idée assez intuitive de compensation des écarts au-dessus et au-dessous de la moyenne qui conduit parfois à une confusion avec la médiane (pour laquelle le *nombre* de valeurs au-dessus et au-dessous est le même).

(ii) La propriété énoncée au paragraphe précédent permet de donner une illustration assez parlante de la moyenne. On peut en effet l'imaginer comme le point d'équilibre (le centre de gravité) d'une barre graduée sur laquelle on représenterait chaque valeur de la série par une masse proportionnelle à son effectif *(cf. figure ci-dessus)*.

4. Retour historique

En dehors de sa simplicité de calcul (à la main ou programmé), la principale caractéristique de la moyenne est qu'elle donne la même importance à chacune des valeurs de la série. Outre le fait que cela peut être un inconvénient si certaines valeurs sont douteuses ou très éloignées des autres valeurs de la série (voir plus loin la définition d'autres paramètres de tendance centrale qui prennent en compte cet état de fait), il est à remarquer que la moyenne ne s'est pas imposée d'emblée aux scientifiques confrontés à une série de valeurs observées.

En 1713, Jacques Bernoulli [3] propose la méthode suivante pour calculer la hauteur moyenne dans un baromètre :

Rassemble toutes les hauteurs que tu as observées, qu'elles soient différentes ou identiques, en une somme que tu divises par le nombre d'observations, ou, ce qui est plus avantageux, si les mêmes hauteurs ont été observées plusieurs fois, les différentes hauteurs sont multipliées par le nombre d'observations qui ont été faites de chacune d'entre elles, la somme de tous ces produits divisée par le nombre de ces observations donne la hauteur "moyenne" (en latin mediam) que les Allemands appellent "eine in die andere gerechnet", les français "l'une portant l'autre" ; si le mercure a été surpris deux fois à la hauteur de 28 pouces, trois fois à la hauteur de 29 et quatre fois à la hauteur de 30, la hauteur moyenne sera :

$$\frac{2.28+3.29+4.30}{9} = 29\ 2/9.$$

Aussi est-il évident qu'ils se trompent ceux qui, pour rechercher la quantité moyenne de mercure, font la moyenne arithmétique entre les extrêmes, et qui avec ce procédé obtiennent une hauteur

(3) *Ars Conjectandi* , traduction N. Meusnier, IREM de Rouen, 1987

moyenne de 29 pouces et non de 29 2/9 pouces.

On reconnaît dans la première méthode le calcul de la moyenne dite pondérée que Bernoulli dit préférable à ce que l'on appelle de nos jours l'**étendue moyenne** (**midrange**) soit : (minimum + maximum)/2. Il est à remarquer que cette dernière méthode est encore employée en météorologie pour déterminer la température d'une journée donnée, alors que la moyenne d'un jour quelconque, par exemple le 14 Juillet, sera la moyenne arithmétique des 14 Juillet des 30 dernières années.

Ce n'est qu'en 1755 que T. Simpson propose l'usage généralisé de la moyenne dans les mesures astronomiques (4) dans le cas de n mesures d'une même grandeur.

En 1805, Legendre (5) établit la liaison entre cette pratique et la méthode des moindres carrés.

" La règle par laquelle on prend le milieu entre les résultats de diverses observations (pour un seul élément), n'est que la conséquence très simple de notre méthode générale, que nous appellerons méthode des moindres carrés. En effet, si l'expérience a donné diverses valeurs a', a", a"' etc. pour une certaine quantité x, la somme des carrés des erreurs sera $(a'-x)^2 + (a''-x)^2 + (a'''-x)^2 + etc.$ *et en égalant cette somme à un minimum, on*

$$a : 0 = (a'-x) + (a''-x) + (a'''-x) + etc$$

d'où résulte $x = \frac{a + a' + a'' + a''' + \dots etc}{n}$, *n étant le nombre des observations".*

5. Problème de sens

On peut toujours calculer mathématiquement un paramètre, pour peu que les fonctions qui interviennent dans le calcul soient définies. Le sens que l'on peut donner au résultat dépend de la situation. Si l'âge moyen des élèves d'une classe est de 15 ans et 3 mois, cela n'a pas le même sens concret (6) que la note moyenne de 13,8 du premier paragraphe car le total des âges ne représente rien. Mathématiquement, cela représente la valeur qui minimise la somme des carrés des écarts comme définie ci-dessus (§3).

De même, la moyenne arithmétique des valeurs 0,05 ; 0,06 ; 0,08 ; 0,07 ; 0,09 n'a pas de sens concret si ces valeurs représentent les taux d'intérêts auxquels on a placé une somme d'argent pendant 5 années consécutives. On montre que dans ce cas la valeur de i telle que

$1 + i = \sqrt[5]{(1+0,05)(1+0,06)(1+0,08)(1+0,07)(1+0,09)}$,

obtenue en effectuant la moyenne géométrique des $1 + x_i$ est préférable puisqu'elle représente le "taux moyen" c'est-à-dire celui qui aurait conduit au même gain si le taux avait été constant pendant les cinq années.

Deux paramètres peuvent être pertinents pour une même série statistique. Ainsi, Christiaan Huygens et son frère Louis, utilisant la première table de mortalité établie par l'anglais John Graunt

(4) A letter to the Right Honorable George Earl of Macclesfield, President of the Royal Society, on the Advantage of taking the mean of a number of observations in practical astronomy. *Phil. Trans.*, 49.

(5) *Nouvelles méthodes pour la détermination des orbites des comètes*, Courcier éditeur, Paris, 1805.

(6) On est dans ce cas pour toutes les variables repérables et non mesurables.

(1662), essayent de déterminer le nombre d'années qui reste à vivre à un individu d'âge donné.

De 100 personnes conçues il en meurt :
36 au bout de 6 ans
24 entre 6 et 16 ans
15 entre 16 et 26 ans
9 entre 26 et 36 ans
6 entre 36 et 46 ans
4 entre 46 et 56 ans
3 entre 56 et 66 ans
2 entre 66 et 76 ans
1 au delà de 76 ans

Louis propose de faire la moyenne pour connaître l'espérance de vie à la naissance, il trouve 18 ans et 2 mois 1/2. Son frère Christiaan (7), construit un contre-exemple (de 100 personnes, 90 meurent avant 6 ans, les autres vivent jusqu'à 155 ans et 2 mois !) qui donne la même espérance de vie de 18 ans et 2 mois 1/2 ce qui montre que ce calcul n'est pas pertinent car *"qui gagerait qu'un enfant conçu parviendrait alors à l'âge de 6 ans seulement aurait grand désavantage puisque de 10 il n'y en a qu'un qui y parvient"*.

Il fait ensuite la distinction entre "moyenne" et "valeur qui a autant de chance d'être dépassée ou non" (la médiane) :

Le calcul que je vous ai envoyé vous aura embarrassé sans doute, auquel ayant songé depuis, et aussi au vôtre, je trouve que nous avons tous deux raison en prenant la chose de différents sens. Vous donnez à un enfant conçu 18 ans et 2 mois 1/2 de vie, et il est vrai que son espérance vaut autant que cela. Cependant il n'est pas apparent qu'il vivra si longtemps, car il est beaucoup plus apparent qu'il mourra devant ce terme. De sorte que si l'on voulait gager qu'il y parviendrait, la partie serait désavantageuse , car on peut seulement gager avec égal avantage qu'il vivra jusqu'à 11 ans environ, ainsi que je le trouve par ma manière...
Ce sont donc deux choses différentes que l'espérance ou la valeur de l'âge futur d'une personne, et l'âge auquel il y a égale apparence qu'il parviendra ou ne parviendra pas. Le premier est pour régler les rentes à vie, et l'autre pour les gageures.

De même, dans une distribution de revenus, les très gros salaires "faussent le calcul de la moyenne". Plus exactement la moyenne n'est pas le paramètre pertinent pour mesurer le centre de la distribution. En effet, dans certaines distributions de ce type, presque toutes les valeurs se trouvent au-dessous de la moyenne, alors que par définition on a toujours 50% des valeurs en dessous de la médiane et 50% en dessus. C'est donc ce dernier paramètre qui est le mieux adapté pour donner une idée du centre d'une distribution de revenus. L'INSEE le complète généralement par le calcul des autres déciles (8).

6. Comparaison de différents paramètres classiques

On constate intuitivement ou graphiquement que certaines distributions

(7) Œuvres complètes de Christiaan Huygens, citées dans *Actes du séminaire d'Histoire des Sciences du lycée Malherbe de Caen*, Juin 1995.

(8) Les n valeurs de la série étant classées dans l'ordre croissant de $x_{(1)}$ à $x_{(n)}$, les déciles sont les valeurs des observations dont les rangs sont $k(n + 1)/10$, pour $k = 1$ à 9. Si ces rangs ne sont pas entiers, on effectue une interpolation.

Nb	Moyenne	Médiane	Ecart type	Min	Max	Q1	Q3	Q3-Q1
19	12	12	2,29	9	15	10	14	4
16	12	12	1,58	9	15	11	13	2

statistiques sont largement étirées et d'autres plus resserrées. Il est logique de penser à comparer deux distributions par des mesures de cette dispersion. Malheureusement ce concept est "flou" (9) en ce sens qu'il y a de nombreuses façons de mesurer la dispersion.

De même, pour une distribution symétrique, la moyenne (ou la médiane puisque ces deux valeurs sont alors confondues) donne une bonne idée de la position centrale de cette distribution, dans le sens où elle correspond géométriquement à la valeur par rapport à laquelle le graphique est symétrique, mais pour une série très dissymétrique, aucune mesure ne s'impose, en particulier la moyenne n'est plus du tout appropriée. Là encore, on est en face du concept flou de position et en particulier de position centrale qui peut être mesuré de différentes façons plus indiquées l'une que l'autre suivant la dispersion de la série.

En effet, une mesure de tendance centrale représente d'autant mieux l'ensemble des valeurs de la série que celles-ci sont moins dispersées. Ceci illustre le fait que ces deux concepts, bien que mal définis, sont liés entre eux.

Prenons par exemple les deux séries statistiques suivantes données par leur graphique en tiges et feuilles (stem and leaf (10)) et qui pourraient représenter les notes de deux classes.

9	0000
10	000
11	00
12	0
13	00
14	000
15	0000

9	0
10	00
11	000
12	0000
13	000
14	00
15	0

Explications (11) : pour la première série, on a 4 fois la note "9", 3 fois la note 10, etc.

Le calcul de différents paramètres classiques (12) pour chacune des séries donne les résultats suivants *(ci-dessus)*.

Elles ont la même moyenne et la même médiane et pourtant personne ne dira que le profil des deux classes est semblable. L'une a des résultats plus homogènes que

(9) Pour reprendre le terme de F. Mosteller et J.W. Tukey dans *Data Analysis and Regression. A second course in statistics*, Addison Wesley, Reading, Massachusetts, 1977.

(10) John W. Tukey-*Exploratory Data*, Addison Wesley, Reading, Massachusetts, 1977.

(11) Pour une présentation des graphiques en tiges et feuilles, voir par exemple, J.-C. Girard, "Des diagrammes à l'histogramme", *Enseigner la Statistique du CM à la Seconde. Pourquoi ? Comment ?*, Groupe Probabilités et Statistique, IREM de Lyon, 1998.

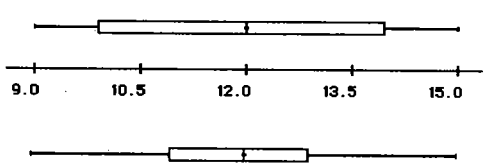
(12) Q_1 est le premier quartile et Q_3 le troisième quartile. Les n valeurs de la série étant classées dans l'ordre croissant de $x_{(1)}$ à $x_{(n)}$, Q_1 représente la valeur de l'observation de rang $(n + 1)/4$ et Q_3 l'observation de rang $3(n + 1)/4$. Si ces rangs ne sont pas entiers, on effectue une interpolation.

Nb	Moyenne	Médiane	Ecart type	Min	Max	Q1	Q3	Q3-Q1
20	12,4	12,5	2,838	9	20	10	14,75	4,75
17	12,471	12	2,428	9	20	11	13,5	2,5

l'autre. Les écarts types qui mesurent la dispersion des séries le montrent (2,29 > 1,58) mais sans que l'on sache réellement apprécier la différence 2,29 – 1,58 (augmentation absolue) ou le rapport 2,29/1,58 (augmentation relative).

Un autre moyen de comparer les dispersions est l'**intervalle** ou **écart interquartile** c'est-à-dire $Q_3 - Q_1$. Le sens en est plus facile à appréhender que celui de l'écart type. C'est la longueur de l'intervalle qui contient la moitié centrale de la distribution. Il vaut 4 pour la première série et 2 pour la seconde. En quelque sorte la première série est deux fois plus dispersée que la deuxième pour cette mesure.

De plus, on peut représenter graphiquement cette dispersion dans un graphique en boîte, ou boîte à moustaches (boxplot [13]).



Explications : Les extrémités de la boîte sont Q1 et Q3. Les moustaches s'étendent jusqu'au minimum et au maximum. Le trait central correspond à la médiane [14].

L'inconvénient de la moyenne et, partant, des paramètres de dispersion calculés à partir de celle-ci est d'être sensible à une valeur anormalement grande ou petite. Par exemple, voyons quelles sont les conséquences d'une note supplémentaire égale à 20.

Le calcul des mêmes paramètres pour chacune des deux séries modifiées donne les résultats suivants *(ci-dessus)*.

La moyenne est généralement affectée par une valeur aussi éloignée du centre de la distribution. En effet la moyenne augmente de 0,4 dans la première série et de 0,471 dans la deuxième. Un des intérêts de la médiane est d'être moins sensible, en général, à ces valeurs. En effet, dans la deuxième série la médiane reste inchangée (12). Par contre dans la première série, elle est plus affectée que la moyenne puisqu'elle augmente de 0,5. L'explication est que cette série présente une faible densité dans la partie centrale alors que la première série présente au contraire une forte densité puisque 12 est la valeur la

(13) John W. Tukey-*Exploratory Data Analysis*, Addison Wesley, Reading, Massachusetts, 1977.

(14) Pour la construction des graphiques en boîte et un exemple détaillé d'utilisation, voir J.-C. Girard, "La médiane, pour quoi faire", *Enseigner la Statistique du CM à la Seconde. Pourquoi ? Comment ?*, Groupe Probabilités et Statistique, IREM de Lyon, 1998.

plus fréquente (ce qu'on appelle le **mode** de la série) et que les valeurs voisines (11 et 13) sont aussi très fréquentes. La médiane est également sensible aux discontinuités dans la partie centrale de la distribution.

Pour ce qui concerne les paramètres de dispersion, l'écart type passe de 2,29 à 2,8 (respectivement de 1,58 à 2,4) alors que l'écart interquartile qui était de 4 (respectivement 2) passe à 4,75 (respectivement 2,5) c'est-à-dire que l'on conserve à peu près le même rapport entre les deux.

7. D'autres paramètres de tendance centrale

Différents statisticiens ont proposé des moyens autres que la moyenne et la médiane pour mesurer le "milieu" d'une distribution statistique en particulier dans le cas d'estimation de la tendance centrale à partir d'un échantillon. Ils ont en commun de ne pas prendre en compte de la même façon toutes les observations, soit en éliminant les valeurs aberrantes, soit en donnant moins de poids aux valeurs éloignées de la partie centrale de la distribution. On qualifie ces mesures de "robustes".

Un solution pour éviter la trop grande influence des valeurs éloignées est de faire la **moyenne tronquée** (trimmed mean). Pour cela on enlève, aux deux extrémités de la série ordonnée, un nombre d'observations (arrondi à l'entier le plus proche) correspondant au même pourcentage (par exemple 5%), puis on fait la moyenne des valeurs restantes. On trouve 12 dans les deux séries de départ, 12,167 et 12,2 dans les séries modifiées. Elle est donc moins affectée que la moyenne et la médiane dans le deuxième cas.

Si on effectue une moyenne tronquée à 25%, on obtient alors la moyenne des valeurs de la moitié centrale de la distribution. On l'appelle **midmean** [15]. Elle vaut 12 dans les deux séries de départ, 12,2 et 12,11 dans les deux séries modifiées. C'est un bon compromis entre la moyenne et la médiane.

Le premier à préconiser la suppression d'une partie des observations semble avoir été l'astronome Boscovich [16] qui calculait la moyenne d'une série d'observations après avoir enlevé la plus petite et la plus grande des valeurs.

Cette méthode sera reprise par les fermiers généraux de l'Ancien Régime pour calculer l'impôt dû par un paysan à partir de la récolte moyenne des 5 dernières années après avoir enlevé la meilleure et la moins bonne.

La moyenne bipondérée

Faire une moyenne tronquée revient à affecter certaine valeur d'un poids 0 et d'autres d'un poids 1. Une alternative est d'affecter à chacune des observations un poids qui diminue en fonction de son éloignement au centre de la distribution. F. Mosteller et J.W. Tukey [17] proposent de calculer la médiane Me de la série puis de remplacer chaque observation X par $Z = \frac{X - Me}{31}$, où I est l'écart interquartile,

(15) Voir par exemple : *Understanding Robust and Exploratory Data Analysis*, Hoaglin, Mosteller & Tukey, J. Wiley, New York, 1983.

(16) Voir *Histoire de la Statistique*- J.-J. Droesbeke et P. Tassi, P. U. F., Paris, 1990.

(17) *Data Analysis and Regresion : A Second Course in Statistics*, Addison Wesley, Reading, Massachusetts, 1977.

puis de faire la moyenne des valeurs de la série, pondérées par :

$w = 0$ si $|Z| > 1$
$w = (1 - Z^2)^2$ si $|Z| < 1$.

Cela revient à considérer aberrantes les observations à plus de 3I de la médiane.

On peut procéder ensuite par itération c'est-à-dire recommencer le calcul en remplaçant la médiane Me par la moyenne bipondérée trouvée jusqu'à ce qu'il n'y ait plus de variation significative (voir un exemple à la fin).

Trimean [18] : Si n est le nombre d'observations, m = (n + 1)/2 est la position de la médiane. On définit les quarts supérieur et inférieur F_U et F_L (upper and lower fourths) ou H_U et H_L (upper and lower hinges) comme les médianes de la première moitié et de la deuxième moitié de la série (médiane comprise) c'est-à-dire les observations de rang ([m] + 1)/2 en partant du maximum et du minimum [19].

L'intervalle $[F_L ; F_U]$ contient environ la moitié centrale des observations. Compte tenu des définitions il peut être différent de l'intervalle interquartile $[Q_1 ; Q_3]$

La moyenne "Trimean" est alors calculée par $1/4(F_L + 2Me + F_U)$.

F–Mean : C'est la moyenne des valeurs de la moitié centrale des observations, c'est-à-dire de l'intervalle $[F_L ; F_U]$, F_L et F_U compris. Compte tenu des définitions il peut être différent de la "midmean".

(18) Tukey J. W. - *Exploratoty Data Analysis*, Addison Wesley, Reading, Massachusetts, 1977.

(19) Où [m] représente la partie entière de m.

Formule de Gastwirth [20] : Si $Q_{1/3}$ et $Q_{2/3}$ sont les valeurs qui partagent la série en trois [21], on calcule $0,3\,Q_{1/3} + 0,4\,Me + 0,3\,Q_{2/3}$.

Broadened Median [22] : Pour conserver la stabilité de la médiane par rapport aux valeurs aberrantes et diminuer sa sensibilité aux discontinuités dans le milieu de la série, J.W. Tukey a proposé le calcul suivant : si n est impair, la médiane élargie est la moyenne des trois valeurs centrales si $5 \leq n \leq 12$, des 5 valeurs centrales si $n \geq 13$. Si n est pair, la médiane élargie est une moyenne pondérée des 4 valeurs centrales si $5 \leq n \leq 12$ avec comme poids 1/6,1/3, 1/3 et 1/6, et si $n \geq 13$, la moyenne pondérée des 6 valeurs centrales avec comme poids 1/5 pour les quatre valeurs centrales et 1/10 pour les deux autres.

8. D'autres paramètres de dispersion

On peut définir différentes mesures de la variabilité d'une distribution statistique. Un paramètre, assez frustre, est l'**étendue** de la série c'est-à-dire l'écart entre le minimum et le maximum. Pour l'exemple du paragraphe 6, il est le même dans les deux séries de départ (15 – 9 = 6) mais il évident qu'il est très affecté par des valeurs éloignées : il passe ainsi à 11 dans les séries modifiées.

(20) Voir par exemple : *Robust Statistics : A Review*, P. Huber, The Annals of Mathematical Statistics, 1972, Vol. 43.

(21) $Q_{1/3}$ et $Q_{2/3}$ sont deux cas particuliers de quantiles. Ils partagent la série dans les proportions indiquées. Ils correspondent aux observations (classées dans l'ordre croissant) de rangs (n + 1)/3 et 2(n + 1)/3. On les calcule par interpolation.

(22) *Understanding Robust and Exploratory Data Analysis*, Hoaglin, Mosteller & Tukey, J. Wiley, New York, 1983.

Outre les deux paramètres déjà cités dans le paragraphe 3, à savoir, la **variance** $\sigma^2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$, c'est-à-dire la moyenne des carrés des écarts entre les observations et leur moyenne, doublement sensible aux valeurs aberrantes, et l'**écart absolu moyen** par rapport à la médiane, $EAM = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - Me\right|$, qui donne l'écart absolu minimum, on définit encore l'**écart absolu moyen** par rapport à la moyenne $EAM = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$.

On peut enfin citer la médiane des valeurs absolues des écarts entre les observations et leur médiane (**Median Absolute Deviation**) : MAD = Médiane ($|x_i - Me|$) qui bénéficie doublement des avantages de la médiane, ou encore l'étendue entre les quarts, **F-spread** ou $d_F = F_U - F_L$, pratiquement égale à l'**écart interquartile.**

9. Conclusion

Comme on l'a vu, la valeur d'un paramètre de tendance centrale n'a pas beaucoup de signification sans une mesure de la dispersion, et réciproquement.

Que ce soit pour caractériser une série entièrement connue ou pour donner des informations à partir d'un échantillon, une alternative au couple $(\bar{X}, \sigma)$, qui cumule les défauts déjà cités des deux paramètres, pourrait être le couple (Me, MAD) ou encore le couple (Me, d_F) ou (Me, I) qui conduisent à des estimations moins sensibles aux valeurs aberrantes et à des intervalles de confiance plus robustes (c'est-à-dire qui ne font pas appel à l'hypothèse d'une distribution supposée connue, par exemple la distribution normale).

D'autre part, on ne peut raisonnablement chercher à comparer deux séries statistiques par leurs moyennes que si elles ont à peu près la même forme de distribution et la même dispersion.

De plus, il ne faut pas oublier que l'écart type a une unité, qui est la même que celle des valeurs de la série et de leur moyenne. Si on multiplie ces valeurs par 10, par un changement d'unité par exemple, l'écart type sera aussi multiplié par 10 sans que la distribution soit évidemment changée.

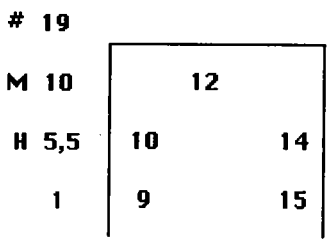
Il n'est donc légitime de comparer les écarts types de deux séries que si les moyennes sont du même ordre de grandeur.

Dans le cas contraire, un autre paramètre, le **coefficient de variation** défini par $CV = \frac{\bar{x}}{\sigma}$ et qui est un pourcentage, donc sans unité, permet d'éviter ces phénomènes d'échelle.

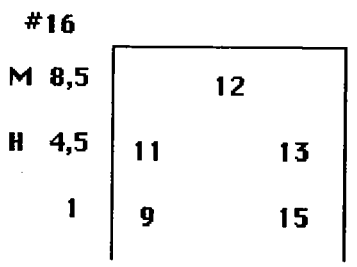
Enfin, il est peut-être illusoire de vouloir décrire une distribution statistique par deux nombres, quels qu'ils soient. Tukey, encore lui, propose de résumer une série par 5 nombres : le minimum, le maximum, la médiane, les deux quartiles (ou les deux quarts) et de présenter les résultats dans un tableau [23] comme illus-

(23) Letter-Values Display, J.W. Tukey, *Exploratory Data Analysis*, Addison Wesley, Reading, Massachusetts, 1977.

tré dans l'exemple ci-dessous (qui reprend les séries de départ du §6) :



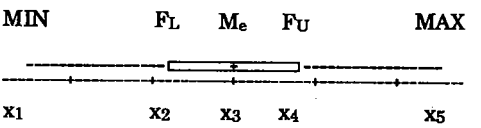
Pour la première série, le nombre d'observations est 19, la médiane (notée M ; c'est la 10^e valeur) est 12, les quarts (égaux ici aux quartiles et notés H ; ce sont les valeurs de rang 5,5 à partir des deux extrémités) sont 10 et 14 donc $d_F = I = 4$, le minimum est 9 et le maximum 15 donc l'étendue est égale à 6.



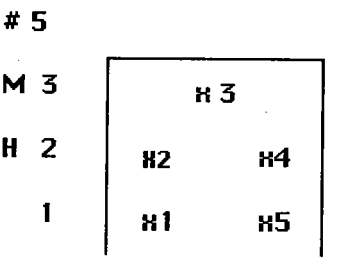
Pour la deuxième série, il y a 16 observations, les quarts sont 11 et 13 et l'écart interquartile est égal à 2, le reste est inchangé.

Ces schémas et ces paramètres robustes trouvent une application intéressante dans le contrôle de qualité où l'on prélève couramment 5 objets pour estimer, par exemple, la moyenne des poids ou des diamètres des articles d'une production [24].

Si les 5 observations sont rangées dans l'ordre croissant et notées : x_1, x_2, x_3, x_4 et x_5, le graphique en boîte devient :



Le premier schéma devient alors :



L'écart interquartile est $x_4 - x_2$ et la F-mean qui est égale à la médiane élargie et à la moyenne tronquée à 20% : $(x_2 + x_3 + x_4)/3$.

(24) Voir, par exemple, White E.M. et Schroeder R. : "A simultaneous control chart", *Journal of Quality Technology* Vol. 19, N° 1, Janvier 1987 ; ou Iglewicz B. et Hoaglin D.C. : " Use of Boxplots for Process Evaluation" *Journal of Quality Technology* Vol. 19, N° 4, Octobre 1987.

UN EXEMPLE : LES DÉPARTEMENTS FRANÇAIS

1. Objectifs

On se propose de faire l'analyse d'une distribution statistique du point de vue des paramètres de tendance centrale. L'exemple choisi porte sur la population des départements français métropolitains (chiffres du recensement de 1990, source : *Quid* (25)).

1	73000	289000	485000	795000	
2	113000	293000	494000	799000	
3	118000	296000	509000	815000	
4	131000	299000	514000	839000	
5	131000	306000	514000	926000	
6	132000	306000	527000	953000	
7	134000	311000	529000	972000	
8	136000	322000	537000	1011000	
9	156000	323000	538000	1016000	
10	159000	342000	548000	1050000	
11	175000	343000	558000	1052000	
12	196000	346000	559000	1078000	
13	200000	348000	568000	1085000	
14	204000	354000	578000	1213000	Gironde
15	207000	358000	580000	1216000	Val-de-Marne
16	225000	364000	585000	1223000	Seine-Maritime
17	230000	380000	598000	1307000	Yvelines
18	233000	386000	618000	1381000	Seine-St-Denis
19	238000	386000	620000	1391000	Hauts-de-Seine
20	238000	396000	671000	1433000	Pas-de-Calais
21	249000	414000	706000	1509000	Rhône
22	270000	467000	712000	1759000	Bouches-du-Rhône
23	278000	471000	726000	2152000	Paris
24	278000	480000	746000	2532000	Nord

Existe-t-il un département moyen ? Cela a-t-il un sens ? Si oui, quel paramètre choisir pour le mesurer ?

(25) La Corse du nord et la Corse du Sud sont comptées séparément, le Finistère Nord et le Finistère Sud ne forment qu'un département. On a donc 96 départements dans ce recensement. Les effectifs sont arrondis au millier le plus proche.

2. Première analyse

Le calcul des paramètres les plus courants donne les résultats suivants (26) :

```
MTB > Describe C2.

                N     MEAN   MEDIAN  TRMEAN    STDEV
C2             96   589698   482500  542558   452138

               MIN      MAX       Q1       Q3
C2           73000  2532000   280750   782750
```

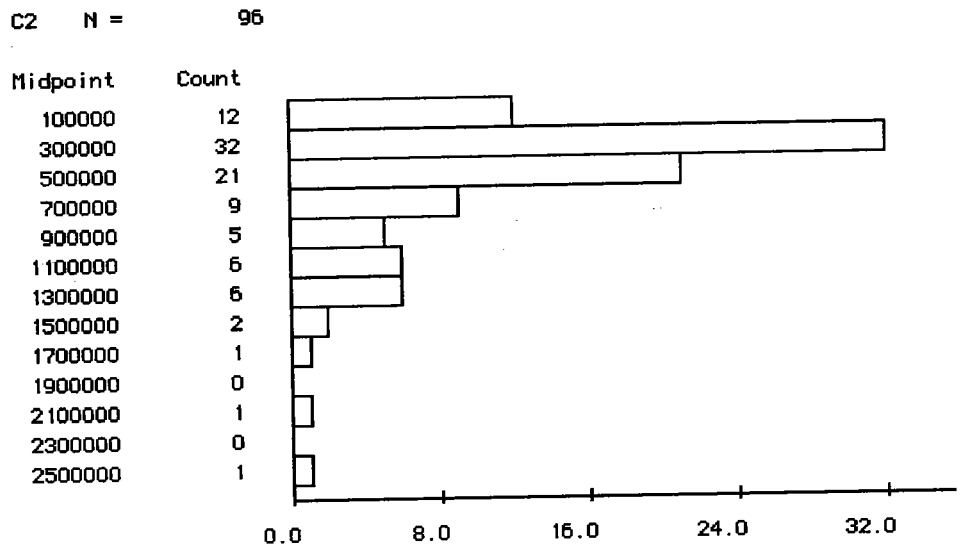
La moyenne est assez éloignée de la médiane (22% de plus). L'écart type est très important par rapport à la moyenne (le coefficient de variation est égal à 452.138 / 589.698 soit 77%). On est donc en présence d'une série très dispersée et non symétrique (distribution étalée à droite). Le rapport entre l'écart interquartile, qui est égal à 502.000, et l'écart type est de 1,11 ce qui est très peu puisque pour la loi normale centrée réduite ce rapport est de 1,35. Les valeurs de la série sont donc assez concentrées entre les deux quartiles Q_1 et Q_3 et assez dispersées au delà du troisième quartile.

3. Analyse graphique

Les graphiques, que ce soit l'histogramme ou le graphique en tige et feuilles, confirment ces premières conclusions.

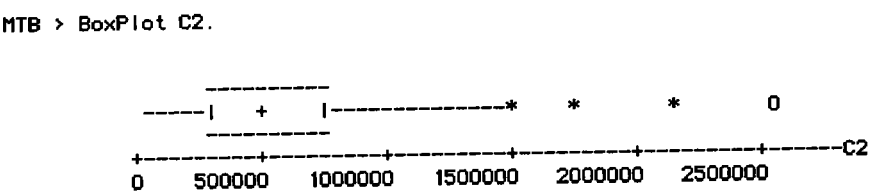
```
MTB > Stem-and-Leaf C2.

Stem-and-leaf of C2        N   = 96
Leaf Unit = 100000

   12     0 011111111111
   44     0 22222222222222223333333333333333
  (21)    0 444444555555555555555
   31     0 666777777
   22     0 88999
   17     1 000000
   11     1 222333
    5     1 45
    3     1 7
    2     1
    2     2 1
    1     2
    1     2 5
```

(26) Tous les graphiques et calculs ont été faits avec le logiciel MINITAB : Mean représente la moyenne, Median la médiane, Trmean la moyenne tronquée à 5% et Stdev l'écart type.



Ces deux graphiques font apparaître une classe modale (celle qui contient la valeur la plus fréquente : le mode) au début de la distribution mais après une forme décroissante on observe une remontée vers la fin de la distribution. Ne serait-on pas alors en présence d'une distribution bimodale qui traduit généralement un mélange de deux populations ? Dans ce cas, il est bien évident que chercher une valeur centrale n'a pas beaucoup de sens. La première idée qui vient à l'esprit est que les départements les plus gros pourraient avoir quelque chose en commun et former une sous population. Cela pourrait être le cas, par exemple, si les départements à forte population étaient ceux de la région parisienne. La liste des populations des 96 départements permet d'écarter, en partie, cette hypothèse. On trouve trois départements de province parmi les quatre plus peuplés, cinq parmi les onze plus peuplés.

Le graphique en boîte à moustache (boxplot), en plus de la dissymétrie, fait apparaître 3 valeurs éloignées (à plus de 1,5 I de Q_3 et notées *) et une très éloignée (à plus de 3 I de Q_3 et notée o). Les départements correspondants sont le Nord, Paris, les Bouches-du-Rhône et le Rhône. Il n'y a aucune raison de les écarter de l'analyse et du calcul de la valeur centrale.



4. Calculs de la valeur centrale

La moyenne est très supérieure à la médiane car la distribution est "étalée vers la droite". Les départements à forte population "tirent la moyenne vers le haut". Quelle est dans ce cas là, somme toute très fréquent, la meilleure mesure de la tendance centrale ? Quelle est donc la population "moyenne" c'est-à-dire celle du département le plus représentatif de l'ensemble des départements de France ? On a l'habitude de dire que la médiane est un meilleur paramètre dans ce type de distributions (par exemple dans les distributions de salaires), mais que donnent les autres paramètres présentés précédemment ?

Commençons par les **moyennes tronquées :**

Moyenne tronquée à :	
0% (moyenne ordinaire	589 698
5%	542 558
10%	518 750
20%	484 559
25% (midmean)	470 833
49% (médiane	482 500

Plus le pourcentage de valeurs tronquées augmente, plus l'influence des fortes populations diminuent.

Le calcul de la **moyenne bipondérée** en 10 itérations donne les résultats suivants :

Médiane	482 500
(Ecart interquartile	502 000)
1re itération	493 880
2e itération	496 244
3e itération	496 736
4e itération	496 839
5e itération	496 860
6e itération	496 865
7e itération	496 866
Moyenne bipondérée	496 866

On trouve une valeur qui n'est pas très éloignée des moyennes tronquées.

Le graphique en lettres (letter-values display) fait apparaître non seulement les quarts (H) mais aussi les huitièmes (E), les seizièmes (D), les trente-deuxièmes (C), etc [27].

```
MTB > LVals C1.
      DEPTH       LOWER       UPPER         MID       SPREAD
N=    96
M     48.5       482.500                 482.500
H     24.5       283.500     770.500     527.000      487.000
E     12.5       198.000    1081.500     639.750      883.500
D      6.5       133.000    1386.000     759.500     1253.000
C      3.5       124.500    1634.000     879.250     1509.500
B      2.0       113.000    2152.000    1132.500     2039.000
       1          73.000    2532.000    1302.500     2459.000
```

(Remarque : les populations sont en milliers)

Si la série était symétrique, les centres de tous les intervalles (MID) seraient égaux à la médiane. On retrouve que la série est très dissymétrique (elle est étalée vers la droite).

Les quarts sont : $F_L = 283.500$ $F_U = 770.500$ ($d_F = 487.000$)

On peut donc calculer la moyenne **trimean** :
$1/4 (F_L + Me + F_U) = 1/4 (283.500 + 482.500 + 770.500) = 504.750$

et la **F-mean** :
$(283.500 + 48*470.833 + 770.500)/ 50 = 473.080$

La médiane élargie (**broadened median**) prend en compte ici les 6 valeurs centrales avec les pondérations 0,1, 0,2, 0,2, 0,2, 0,1, 0,1 soit :

```
467000    0.1     46700
471000    0.2     94200
480000    0.2     96000
485000    0.2     97000
494000    0.2     98800
509000    0.1     50900
                    *
                 483600
```

La médiane élargie est donc égale à 483 600.

(27) Depth = profondeur, c'est-à-dire le rang en partant du maximum ou du minimum
Lower = valeur inférieure Upper = valeur supérieure
mid = moyenne entre les deux valeurs précédentes
spread = étendue entre les deux valeurs précédentes

Les valeurs $Q_{1/3}$ et $Q_{2/3}$ qui partagent la série en trois sont :

$Q_{1/3} = x_{32} + 1/3(x_{33} - x_{32}) = 322.000 + 1/3(323.000 - 322.000) = 322.333$
$Q_{2/3} = x_{64} + 1/3(x_{65} - x_{64}) = 585.000 + 2/3(598.000 - 585.000) = 593.667$

d'où la **moyenne de Gastwirth** :

0,3*322.333 + 0,4*482.500 + 0,3* 593.667 = 467.800.

5. Conclusion

Tous les paramètres calculés sont assez éloignés de la moyenne. Par contre leurs valeurs sont assez voisines les unes des autres. La majorité se trouve entre 470 000 et 500 000.

Si la moyenne des populations des départements français a un sens, la valeur qui mesure ce département moyen (en 1990) est plus certainement autour de 500.000 que de 600.000 comme l'indique la moyenne arithmétique (589.698).

Il faut par conséquent se méfier de la moyenne simple et ne pas l'utiliser sans discernement, en particulier si l'on est en face d'une distribution très dissymétrique ce qui peut se constater sur un graphique ou en comparant la moyenne et la médiane.

Il est donc particulièrement indiqué de ne pas priver les élèves de cette dernière notion, finalement assez simple, et de toujours relier la moyenne ou la médiane à une idée de dispersion, ce qui peut se faire par une mesure (l'écart interquartile, par exemple) et être illustré par un graphique en boîte à moustaches.

D'une manière générale, comme on a essayé de le montrer dans l'exemple des départements, l'analyse d'une série statistique doit s'appuyer sur l'étude de plusieurs graphiques et le calcul de plusieurs paramètres.