MEDIANE ET QUANTILES

Hussein HAMMOUD, IRES de Toulouse
hussein.hammoud@ac-toulouse.fr

Résumé
Il n'existe pas de manière consensuelle de parler des quantiles. Voici un passage en revue succinct de la littérature, qui s'achève par une proposition supplémentaire de formalisation. Une référence incontournable dans le domaine des statistiques descriptives est évoquée : il s'agit de J.W.Tukey et de son diagramme en boite.

Mots-clés
Médiane, quartile, décile, quantile, valeur aberrante, élagage, diagramme en boîte, Tukey, statistiques descriptives, analyse exploratoire.

Introduction par un exemple

Environ 10000 candidats se sont présentés au Concours de Recrutement de Professeurs des Ecoles (CRPE), leurs copies de mathématiques ont été corrigées par plus de 130 correcteurs. On voudrait comparer ces différents correcteurs. Par exemple, pour chacun d'eux, on cherche la note dépassée (on non) par 10% des candidats. Il s'agit dans ce cas du $9^{ème}$ décile. Ou bien, on cherche pour chaque correcteur, la note α telle que la moitié des candidats ont une note inférieure ou égale à α (il s'agit de la médiane). Le problème de la recherche des quantiles, aboutit à la construction d'intervalles contenant un certain pourcentage des observations. On va par exemple chercher l'intervalle contenant 60% des notes les plus basses, ou bien l'intervalle contenant 80% des notes centrales, etc.

Point de vue fréquentiel

Définition
Les quantiles sont des valeurs du caractères qui partagent la distribution en k parties comprenant le même effectif égal à un $k^{ième}$ de l'effectif total.

Valeur de k	Quantile obtenu
4	quartile
10	décile
100	centile

Cas des quartiles
Ce sont les trois valeurs du caractère qui séparent la distribution en 4 parties comprenant le même effectif égal au quart de l'effectif total. Les quartiles sont désignés respectivement par :

- Q_1 le premier quartile (ou quartile inférieur) : 25% des données sont inférieures ou égales à Q_1.
- Q_2 le quartile central : 50% des données sont inférieures ou égales à Q_2. En fait, Q_2 est la médiane.
- Q_3 le troisième quartile (ou quartile supérieur) : 75% des données sont inférieures ou égales à Q_3.

On peut interpréter Q_1 et Q_3 comme les médianes respectives de la série allant de la valeur minimale de la série mère ordonnée jusqu'à Q_2, et celle allant de Q_2 à la valeur maximale.

Diagramme des fréquences cumulées ou diagramme intégral

Les déciles sont les abscisses des points dont les ordonnées sont respectivement $\frac{1}{10}, \frac{2}{10}, ..., \frac{9}{10}$.

Les quartiles sont les abscisses des points dont les ordonnées sont respectivement 0.25, 0.50 et 0.75

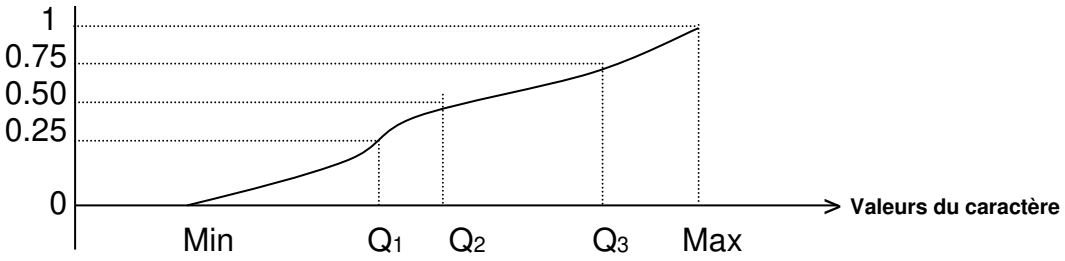
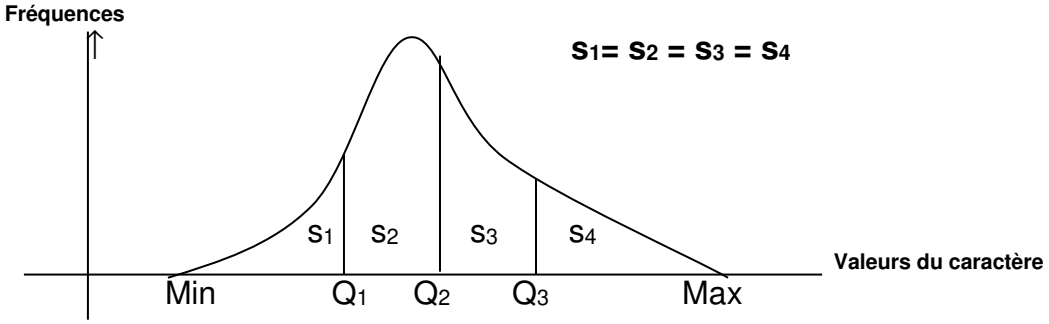


Diagramme de la distribution ou diagramme différentiel

Les quartiles sont les abscisses des droites (verticales) qui partagent la surface sous la courbe des fréquences, en quatre aires égales.

Les déciles sont les abscisses des droites qui partagent la surface sous la courbe des fréquences, en dix aires égales.

Les centiles sont les abscisses des droites qui partagent la surface sous la courbe des fréquences, en cent aires égales.



Point de vue analytique

Définition

extrait du texte d'accompagnement (2001) de la classe de première.

Soit une série numérique de données à valeurs dans un intervalle.
On définit la fonction Quantile par :

$$Q:[0\,;1] \to I$$

$$u \to \inf\{x \in I,\ F(x) \geq u\}$$

où F(x) est la fréquence des valeurs observées inférieures ou égales à x.

La fonction F est une fonction en escalier dans le cas d'une série discrète finie : F est constante sur les intervalles $[x_i\ ;\ x_{i+1}[$ où x_i sont les valeurs ordonnées de la série.
Si n est la taille de la série, Q(u) est le terme dont le rang est le plus petit entier supérieur ou égal à $n \times u$.

Quantiles particuliers

- quartiles : $Q_i = Q(\frac{1}{4} i)$ pour i = 1, 2, 3 (remarque : $Q(\frac{1}{2})$ est la médiane)
- déciles : $D_i = Q(\frac{1}{10} i)$ pour i = 1 à 9
- centiles : $C_i = Q(\frac{1}{100} i)$ pour i = 1 à 99

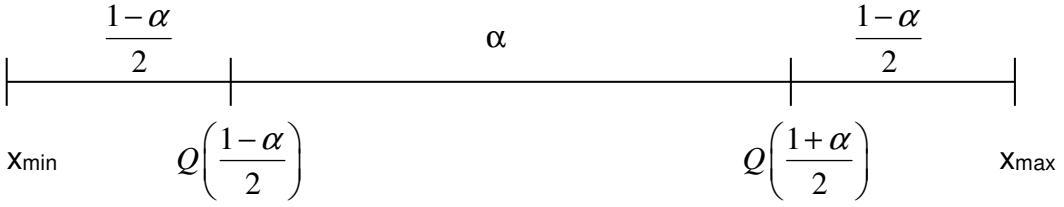
Sens de variation de la fonction Q

La fonction Q est une fonction croissante de u. Pour une série ordonnée, pour tous $u_1, u_2 \in [0\ ;\ 1]$ tels que $u_1 < u_2$ alors $Q(u_1) \leq Q(u_2)$

Démonstration

Pour $u_1 < u_2$ alors $\{x \in I,\ F(x) \geq u_1\} \supseteq \{x \in I,\ F(x) \geq u_2\}$, donc :
$\inf\{x \in I,\ F(x) \geq u_2\} \in \{x \in I,\ F(x) \geq u_1\}$
qui s'écrit plutôt : $Q(u_2) \in \{x \in I,\ F(x) \geq u_1\}$
c'est-à-dire que : $Q(u_2) \geq \inf\{x \in I,\ F(x) \geq u_1\}$
ou encore : $Q(u_2) \geq Q(u_1)$.

Détermination de l'intervalle contenant une proportion α des valeurs centrales



L'intervalle contenant une proportion α des valeurs ″centrales″ est :

$$\left[Q\left(\frac{1-\alpha}{2}\right);Q\left(\frac{1+\alpha}{2}\right)\right]$$

Par exemple, l'intervalle interquartile contenant 50% ($\alpha=\frac{1}{2}$) des observations centrales est donné par l'intervalle [Q_1 ; Q_3] où :

$Q_1= Q(\frac{1-\alpha}{2}) = Q(\frac{1-\frac{1}{2}}{2}) = Q(\frac{1}{4})$ et $Q_3= Q(\frac{1+\alpha}{2}) = Q(\frac{1+\frac{1}{2}}{2}) = Q(\frac{3}{4})$.

<u>Une proposition de formalisation pour la médiane et les quartiles : la notion de profondeur</u>

Voici maintenant une manière supplémentaire de déterminer des quartiles. On considère une variable quantitative et un ensemble P d'observations constituant le paquet de données à explorer. Les données sont classées dans l'ordre croissant et on définit alors, pour tout $x\in P$, $r_c(x)$ le rang de x selon cet ordre croissant. Puis les données sont classées dans l'ordre décroissant et on définit pour tout $x\in P$, $r_d(x)$ le rang de x selon cet ordre décroissant.
Enfin, on définit la profondeur de $x\in P$ par la plus petite distance entre x et l'une des deux extrémités du paquet : *prof(x) = inf($r_c(x)$, $r_d(x)$)*
Pour tout $x\in P$, on a évidemment *: $r_c(x)+ r_d(x) = 1+$ Nombre de données*
La médiane étant située au milieu, *prof(Me)* = $\frac{1+Nombre\ de\ données}{2}$
Chacun des quartiles étant situé à égale distance de la médiane et l'une des extrémités, ils ont la même profondeur, soit : $\frac{1+E(prof(Me))}{2}$
E(x) désignant comme d'habitude, la fonction partie entière.

<u>Exemple</u>
On considère les paquets : P_1 = {7, 60, 61, 62, 63, 65, 69, 73, 76}
et P_2 = {60, 61, 62, 63, 65, 69, 73, 76}.

$x\in P_1$	7	60	61	62	63	65	69	73	76
Rang croissant	1	2	3	4	5	6	7	8	9
Rang décroissant	9	8	7	6	5	4	3	2	1
Profondeur	1	2	3	4	5	4	3	2	1

$x\in P_2$	60	61	62	63	65	69	73	76
Rang croissant	1	2	3	4	5	6	7	8
Rang décroissant	8	7	6	5	4	3	2	1
Profondeur	1	2	3	4	4	3	2	1

On obtient les résultats suivants :

	prof(Me)	Me	prof(Q_1)	Q_1	prof(Q_2)	Q_2
P_1	5	63	3	61	3	69
P_2	4.5	64	2.5	61.5	2.5	71

Le diagramme en boîte de TUKEY

Définition

Extrait du document d'accompagnement des programmes de Premières - 2001

"On divise l'intervalle des valeurs prises par une série statistique en intervalles contenant certains pourcentages des données".

Il y a 4 subdivisions contenant chacune 25 % des données rangées dans l'ordre croissant.

Description

" Box and whiskers plot " nous vient de Tukey (1915-2000) dans son ouvrage historique publié en 1977, Exploratory Data Analysis.

Le diagramme en boîte ou boîte à pattes est constitué d'une boîte rectangulaire avec un trait vertical à l'intérieur. Les deux bords de la boîte correspondent aux deux quartiles. Leur différence est l'écart interquartile. A l'extérieur, un trait de chaque côté, prolonge la boîte jusqu'à la valeur minimum et la valeur maximum. Ce sont les "moustaches".

Le diagramme en boîte permet de détecter les valeurs aberrantes.

Valeur aberrante : qui se démarque des autres *(Tukey).*

Sur le diagramme, la valeur aberrante se trouve à plus d'un saut de l'un ou de l'autre des deux bords de la boîte.

Un saut est égal à une fois et demie la longueur de l'écart interquartile *(Tukey)*

Le diagramme en boîte permet, visuellement, de :

- évaluer la médiane
- évaluer les quartiles
- évaluer l'étendue interquartile
- détecter les valeurs aberrantes
- évaluer le degré de symétrie entre les deux quartiles (il y a symétrie lorsque la médiane se trouve au milieu des deux quartiles)

En bref, l'utilité du diagramme en boîte peut se résumer à :

- décrire un paquet de données
- comparer plusieurs paquets de données en ce qui concerne la médiane, les quartiles, l'écart interquartile, les valeurs aberrantes.

Exemple d'utilisation

Les trois paquets de données ci-dessous donnent les performances de trois groupes indépendants de sportifs de haut niveau, respectivement après un an (P_1), deux ans (P_2) et trois ans (P_3) d'entraînement :

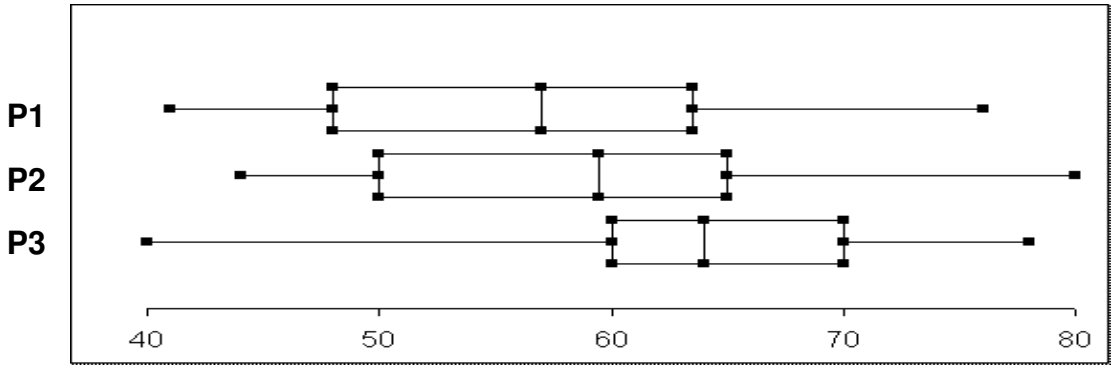
P_1 = {41, 45, 46, 47, 47, 48, 48, 49, 50, 51, 56, 57, 57, 59, 61, 62, 63, 64, 65, 66, 67, 74, 76}

P_2 = {44, 46, 50, 53, 55, 64, 65, 65, 73, 80}

P_3 = {40, 50, 51, 55, 60, 61, 61, 61, 64, 66, 67, 68, 70, 71, 71, 74, 78}

	Min	Q_1	Me	Q_3	Max
P_1	41	48	57	63.5	76
P_2	44	50	59.5	65	80
P_3	40	60	64	70	78

Comparaison des trois groupes par juxtaposition de leurs diagrammes en boîte :



Il devient clair que le nombre d'années d'entrainement induit globalement une performance supérieure.

Pertinence d'un diagramme en boîte

Cependant, en pratique, il ne faut jamais s'en tenir uniquement au diagramme en boîte pour étudier une distribution. On a souvent recours à un autre outil d'analyse exploratoire, par exemple, le diagramme tige-feuilles. Voici deux situations où la représentation par un diagramme en boite montre ses limites.

Le premier exemple concerne deux distributions différentes, peuvent avoir des boîtes identiques.

Considérons les deux paquets suivants :

P_1 = {1,1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8, 8, 8, 9, 9, 9,10,10}

P_2 = {1,1, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 6, 7, 7, 7, 8, 8, 8, 9, 9, 9, 9,10,10,10,10,10,10}

	Min	Q_1	Me	Q_3	Max
P_1	1	3	6	8	10
P_2	1	3	6	8	10

Le second exemple montre qu'il peut arriver que le diagramme en boîte ne reflète pas du tout la situation étudiée, et la médiane devenant franchement un mauvais indicateur. La distribution du parc automobile d'une entreprise selon la puissance fiscale des véhicules est présentée ci-dessous :

Puissance en CV	4	5	6	7	8	9
Nombre de véhicules	52	63	254	70	25	13
Effectifs cumulés	52	115	369	439	464	477

Le résumé de cette distribution est le suivant :

Min	Q_1	Me	Q_3	Max
4	6	6	6	9

Nous laissons au lecteur le soin d'en juger.
Dans ce genre de situations, on doit recourir à d'autres types de résumés, par exemple un diagramme en barres ou circulaire, on peut aussi bien calculer la moyenne des puissances fiscales, l'écart-type, etc.

Le problème de l'élagage

Le problème en réalité est de savoir où faut-il arrêter les extrémités de la boîte. Bien-entendu, si les valeurs minimale et maximale ne s'éloignent pas trop de la "fenêtre" de représentation, la question est sans intérêt sauf si l'on désire encore découper l'intervalle pour le "quantiler" davantage. Mais lorsque les valeurs sont aberrantes, il serait peut-être salutaire de les élaguer, c'est-à-dire les éliminer tout simplement.

Il y a plusieurs méthodes d'élagage

- L'élagage selon Tukey : une valeur située à plus d'un saut d'un quartile à l'une ou à l'autre extrémité, est une valeur aberrante, donc à élaguer.

On rappelle qu'un saut est égal à 1,5 fois l'écart interquartile. Une valeur aberrante est donc inférieure à $\boldsymbol{Q_1 - saut}$ ou bien supérieure à $\boldsymbol{Q_3 + saut}$

Exemple
Soit le paquet de données suivant :
{30, 59, 60, 64, 64, 68, 68, 69, 75, 75, 77, 81, 83}

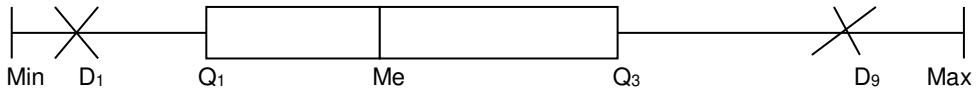
$Q_1=(60+64)/2$ et $Q_3=(75+77)/2$

Min	Q_1	Me	Q_3	Max
30	62	68	76	83

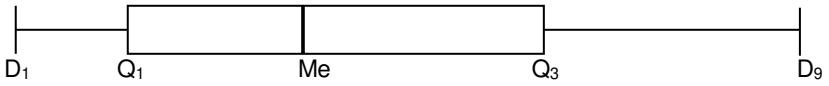
L'écart interquartile est : $\Delta Q = Q_3 - Q_1 = 76 - 62 = 14$
Le saut est : $1{,}5 \times 14 = 21$

Les données se trouvant en-decà de 62–21 = 41 ou au-delà de 76+21 = 97 sont suspectes. On les identifie lorsque cela est possible, sur le diagramme en boite par un symbole spécifique. Ici, seule la valeur 30 sera décrétée aberrante.
Les traits à l'extérieur de la boîte vont jusqu'à la dernière valeur qui ne soit pas aberrante.

- L'élagage à partir des déciles : les valeurs situées avant le premier décile ou après le neuvième décile sont à élaguer. En général, dans ce cas, on représente le minimum, le maximum et les deux déciles. La position des déciles est marquée par une petite croix.



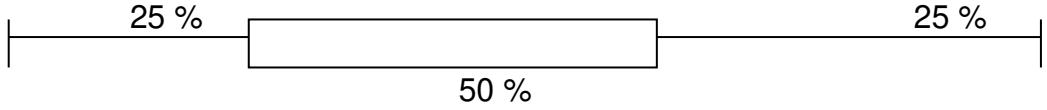
ou bien :



- L'élagage à 2α % , qui consiste à élaguer α% de chaque côté, soit α % des valeurs se trouvant immédiatement après le minimum, et α % des valeurs se trouvant immédiatement avant le maximum.

Que devons-nous enseigner aux élèves

Le diagramme en boîte revient à procéder à un partage des données en 4 parties contenant chacune 25% des données, et pour chacune de ces parties, on peut donner la valeur minimale et la valeur maximale. En particulier, on pourra donner les valeurs prises par les 50 % du centre.



Extrait du document d'accompagnement des programmes de Premières - 2001

- les techniques de détermination de la médiane sont souples, adaptables, et par extension, celles de détermination des quantiles.
- examiner toujours la légitimité d'un élagage, son utilité, son sens

Conclusion

Le diagramme en boîte fait partie des outils de l'analyse exploratoire, il fournit visuellement les principales caractéristiques d'une distribution, en particulier, il met en évidence la dispersion et les valeurs aberrantes. Comme sommaires numériques, on choisit la médiane et l'écart interquartile, plus fiables et plus résistants en général que les habituels moyenne et écart-type.
Nous terminons en conseillons au lecteur intéressé par le sujet, la consultation de l'excellent ouvrage deTukey publié en 1977, Exploratory Data Analysis, aux éditions Addison-Wesley, ouvrage que l'on trouve dans les bibliothèques universitaires.