# Annexe commune aux séries ES, L et S : boîtes et quantiles

## Quantiles

En statistique, pour toute série numérique de données à valeurs dans un intervalle I, on définit la fonction quantile Q, de [0,1] dans I, par :

$$Q(u) = \inf\{x, F(x) \geqslant u\},$$

où $F(x)$ désigne la fréquence des éléments de la série inférieurs ou égaux à $x$.

Soient $a_1,\ldots, a_r$ les valeurs prises par une série de taille $n$, ordonnées par ordre croissant ; la fonction F est discontinue et par ailleurs constante sur les intervalles $[a_i,a_{i+1}[$ ; sa représentation graphique est composée de segments horizontaux.
En pratique, en consultant la liste des nombres $\{F(a_1),\ldots, F(a_r)\}$, il est aisé de déterminer un quantile. Cependant, pour programmer le calcul de Q, on utilise la propriété suivante :
*Soit n la taille de la série ; si on ordonne la série par ordre croissant, Q(u) est la valeur du terme de cette série dont l'indice est le plus petit entier supérieur ou égal à nu.*

Dans le cadre de cette définition, les 3 quartiles sont Q (0,25), Q (0,50), Q (0,75). Les 9 déciles sont les valeurs de $Q(i/10)$, $i = 1\ldots9$ ; les 99 centiles sont les valeurs de $Q(i/100)$, $i = 1\ldots99$. On définit assez souvent la médiane $m$ par $m = Q(0,5)$ : la médiane est alors le second quartile, le cinquième décile, le cinquantième centile, etc.
Mais de nombreux statisticiens, de nombreux logiciels (de qualité) et de nombreux médias utilisent la définition suivante de la médiane d'une série : on ordonne la série des observations par ordre croissant ; si la série est de taille $2n + 1$, la médiane est la valeur du terme de rang $n + 1$ dans cette série ordonnée ; si la série est de taille $2n$, la médiane est la demi-somme des valeurs des termes de rang $n$ et $n + 1$ dans cette série ordonnée.
C'est la définition adoptée dans le programme de seconde. Les deux définitions, Q (0,5) et celle-ci donnent en pratique, pour des séries à valeurs continues de grande taille, des résultats le plus souvent très proches.
La procédure qui consiste à tracer une courbe dite de fréquences cumulées croissantes, continue, obtenue par interpolation linéaire à partir des valeurs $F(a_i)$ définies ci-dessus et à définir la médiane comme l'intersection de cette courbe avec la droite d'équation $y = 0,5$ ou avec une courbe analogue dite des fréquences cumulées décroissantes n'est pas une pratique usuelle en statistique et ne sera pas proposée au lycée.
Si des données sont regroupées en classe, on parle de classe médiane.

### Dans l'enseignement secondaire

Pour les quartiles, la définition proposée sera liée à la fonction quantile :
**Premier quartile :** c'est le plus petit élément $q$ des valeurs des termes de la série tel qu'au moins 25 % des données soient inférieures ou égales à $q$.
**Troisième quartile :** c'est le plus petit élément $q'$ des valeurs des termes de la série tel qu'au moins 75 % des données soient inférieures ou égales à $q'$.

Certains logiciels prennent pour le premier quartile une définition analogue à la médiane : par exemple si $n = 4r$, le premier quartile est la demi-somme des valeurs prises par le terme de rang $r$ et le terme de rang $r + 1$. Nous n'adopterons pas cette définition un peu marginale.

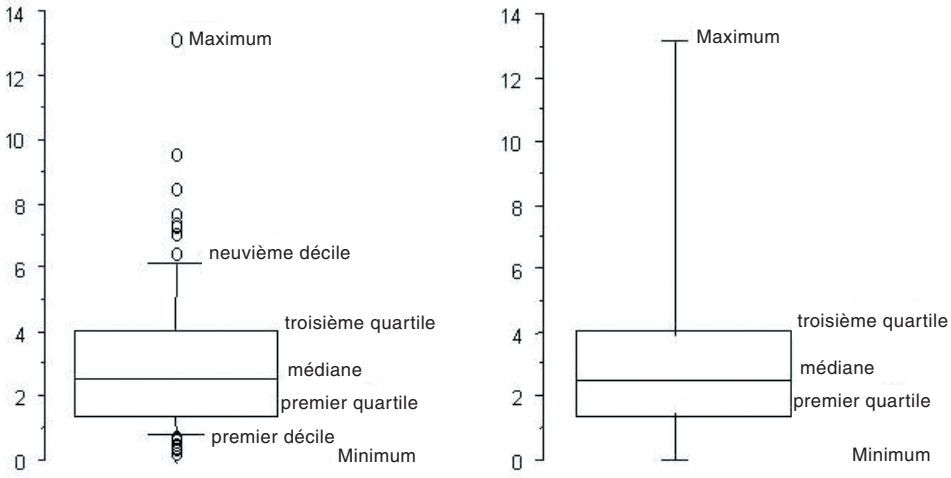
On ne définira que le premier et le neuvième décile :
**Premier décile :** c'est le plus petit élément $d$ des valeurs des termes de la série tel qu'au moins 10 % des données soient inférieures ou égales à $d$
**Neuvième décile :** c'est le plus petit élément $d'$ des valeurs des termes de la série, tel qu'au moins 90 % des données soient inférieures ou égales à $d'$.
On pourra introduire les termes suivants :
**Intervalle interquartile :** intervalle dont les extrémités sont le premier et le troisième quartiles.
**Intervalle interdécile :** intervalle dont les extrémités sont le premier et le neuvième déciles.
**Écart interquartile :** longueur de l'intervalle interquartile, *i.e.* différence entre le troisième et le premier quartiles.
**Écart interdécile :** longueur de l'intervalle interdécile, *i.e.* différence entre le neuvième et le premier déciles.

*Un abus de langage assez fréquent fait qu'on parle aussi d'intervalle interquartile au lieu d'écart interquartile : on évitera cet abus de langage au lycée.*

*Remarque :* si les termes de la série subissent une transformation affine, $x \mapsto ax + b$, où $a$ est positif (ce qui est le cas lors d'un changement d'unité), les quantiles subissent la même tranformation.

## Diagrammes en boîtes

Ces diagrammes sont aussi appelés diagrammes de Tuckey, diagrammes à pattes ou à moustaches (*whiskers plot*). Il n'y a pas que le nom qui varie d'un logiciel à l'autre. Les deux situations les plus classiques sont représentées ci-dessous :




Nous conviendrons de choisir « par défaut » la définition représentée graphiquement en (1), où figurent les premiers et neuvièmes déciles. Si une ou plusieurs valeurs extrêmes sortent résolument des limites du dessin, on indique dessous leurs valeurs sans les représenter. D'autres représentations peuvent être considérées, en spécifiant alors comment sont définies les « moustaches » de la boîte.

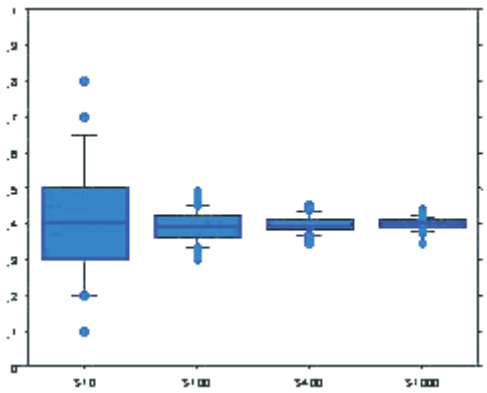
Néanmoins, les enseignants pourront utiliser des boîtes dont les extrémités sont les premiers et 99$^{e}$ centiles, les valeurs extrêmes, etc. L'essentiel est d'avoir compris le principe : un jour d'examen, on demandera simplement à l'élève de spécifier en légende les éléments représentés.

Les premiers diagrammes en boîtes sont les diagrammes de Tuckey où la longueur des « moustaches » est 1,5 fois l'écart interquartile ; les diagrammes de Tuckey étaient utilisés dans des secteurs où les données peuvent le plus souvent être modélisées en utilisant une loi de Gauss ; dans ce cas, au niveau théorique, les extrémités des « moustaches » sont voisines du premier et 99$^{e}$ centiles : ces diagrammes étaient surtout utilisés pour détecter la présence de données exceptionnelles. On utilise aujourd'hui les diagrammes en boîtes pour représenter des distributions empiriques de données quelconques, non nécessairement symétriques autour de la moyenne, et le choix de moustaches de longueur 1,5 fois l'écart interquartile ne se justifie plus.

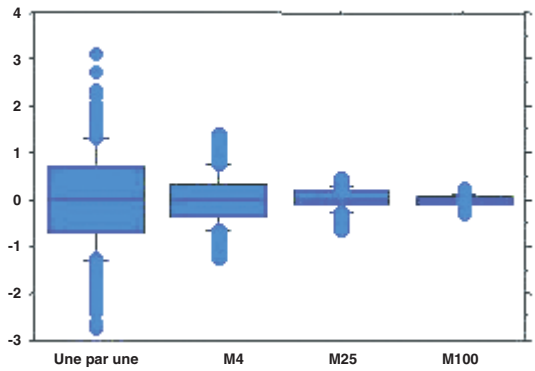
Les diagrammes en boîtes, comme les histogrammes, résument graphiquement une série ; l'idée de base est la suivante : au lieu de partager l'ensemble des valeurs possibles en segments égaux, on les partage en segments (quartiles, déciles, centiles) qui contiennent une proportion prédéterminée des valeurs de la série. Les diagrammes en boîtes permettent de visualiser certains phénomènes et notamment de comparer plusieurs répartitions de valeurs. Ainsi, dans la figure ci-dessous, on a représenté les diagrammes en boîtes de :

– 100 simulations d'un sondage de taille 10 dans une population dont les individus sont codés 0 ou 1, la proportion de 1 étant ce qu'on cherche à déterminer (un sondage

de taille $n$ est ici le tirage au hasard – et avec remise – de $n$ individus dans une population de taille N) ;
– 100 simulations d'un sondage de taille 100 dans la même population ;
– 100 simulations d'un sondage de taille 400 dans la même population ;
– 100 simulations d'un sondage de taille 1000 dans la même population.



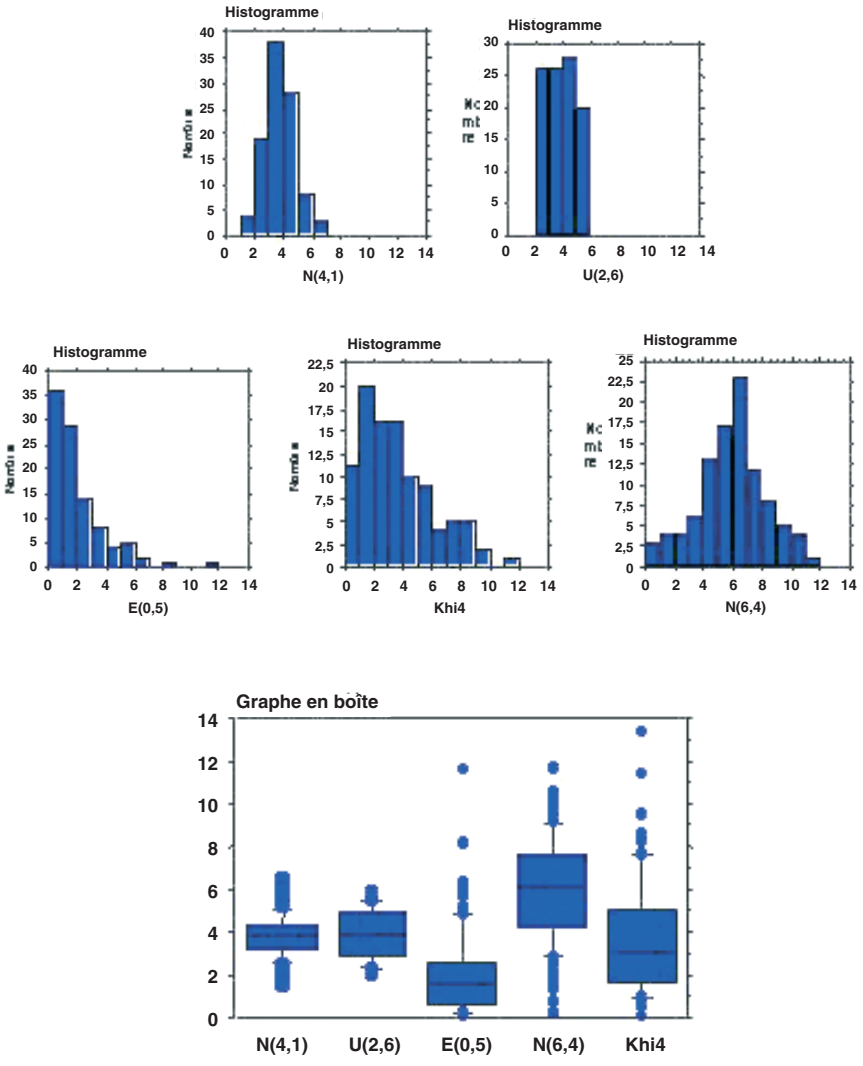
La deuxième figure représente des mesures de hauteur d'eau dans un barrage par rapport à un niveau fixé : on met 100 appareils qui mesurent cette hauteur d'eau. On a essayé quatre sortes d'appareils de mesure :
– ceux qui font une seule mesure ;
– ceux qui font 4 mesures et donnent leur moyenne ;
– ceux qui font 25 mesures et donnent leur moyenne ;
– ceux qui font 100 mesures et donnent leur moyenne.




Les deux exemples situés ci-dessus sont spectaculaires et aisés à interpréter. Pour des séries de données quelconques, interpréter un diagramme en boîte demande un peu d'expérience et d'honnêteté pour ne pas transformer en affirmation théorique une observation lue sur un diagramme, que ce soit un histogramme ou un diagramme en boîte.
Ci-dessous, nous présentons pour des séries de taille 100 simulées à partir de modèles classiquement utilisés divers résumés numériques et graphiques qu'on pourra s'exercer à lire :
– deux séries simulées à partir de lois de Gauss de moyennes 4 et 6 et de variances 1 et 4 et une série simulée à partir de la loi uniforme sur 1,6. Ces lois sont symétriques autour de leur moyenne : l'espérance et la médiane théoriques coïncident et les graphiques théoriques sont symétriques.
– une série simulée à partir de la loi exponentielle d'espérance 2 et une série simulée à partir d'une loi du $\chi^2$ à 4 degrés de libertés : ces lois de probabilité n'admettent pas de symétrie.

**Statistiques descriptives**

| | Moy. | Dév. Std | Nombre | Minimum | Maximum | Médiane | Interquartile | 10% Moy. élaguée |
|---|---|---|---|---|---|---|---|---|
| N(4,1) | 3,782 | ,999 | 100 | 1,556 | 6,597 | 3,799 | 1,186 | 3,741 |
| U(2,6) | 3,927 | 1,153 | 100 | 2,006 | 5,927 | 3,881 | 1,999 | 3,932 |
| E(0,5) | 2,029 | 1,945 | 100 | ,055 | 11,633 | 1,562 | 1,926 | 1,706 |
| N(6,4) | 5,995 | 2,366 | 100 | ,185 | 11,762 | 6,056 | 3,376 | 6,021 |
| Khi4 | 3,745 | 2,630 | 100 | ,095 | 13,430 | 3,085 | 3,428 | 3,423 |







Enfin, à ce propos et pour sa propre formation, l'enseignant pourra utiliser le logiciel *SEL* (présent sur le cédérom joint). Plus précisément, il pourra :

– dans l'applet située à la page « Diagrammes en boîte » du lexique, voir comment fluctuent ces diagrammes lorsqu'on tire des échantillons au hasard dans un série de données réelles (tailles d'enfants de six ans) ;

– dans l'applet de simulations « diagrammes en bâtons, histogrammes et quantiles », superposer les histogrammes, fonctions de répartitions, fonctions quantiles et diagrammes en boîtes de différentes lois classiques avec celles d'échantillons simulés ;

– dans l'applet « ajustement par quantiles », visualiser une technique classique d'ajustement de lois à des données.