

Quantiles et diagrammes en boîtes

Document du GEPS de mathématiques – 22/12/ 2000

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 :

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

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

Soient a_1, \dots, a_r les valeurs prises par une série de taille n , ordonnées par ordre croissant ; la fonction F est discontinue et 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 $\{F(a_1), \dots, F(a_r)\}$, 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 $Q(i/10)$, $i = 1 \dots 9$, les 99 centiles sont les valeurs de $Q(i/100)$, $i = 1 \dots 99$. 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 media utilisent la définition suivante de la médiane d'une série :

Médiane : 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 croissante, 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 $y = 0,5$, où 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, nous proposons de garder la définition liée à la fonction quantile :

Premier quartile : c'est le plus petit élément q des valeurs des termes de la série, ordonnées par ordre croissant, 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, ordonnées par ordre croissant, 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.

Nous suggérons de ne pas définir le second quartile mais de manipuler {premier quartile, médiane, troisième quartile} ; il n'y a pas de raisons de signaler qu'avec la définition adoptée, la médiane n'est pas le second quartile, sauf si un élève pose précisément la question. Dans ce cas, on pourra lui expliquer individuellement que certaines séries comportant des æquo (par exemple 1 2 2 2 2 3 5) ne permettent pas une définition agréable de

la médiane comme « le » nombre m tel qu'exactly 50% des termes de la série sont inférieurs à m et exactly 50% supérieurs à m ; à partir de là, plusieurs choix étaient possibles, mais l'idée reste que la médiane coupe la série en deux dans la cas où il n'y a pas d'ex-aequo.

On ne définira que le premier le neuvième décile :

Premier décile : c'est le plus petit élément d des valeurs des termes de la série, ordonnées par ordre croissant, 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, ordonnées par ordre croissant, 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 quartile

Intervalle interdécile : intervalle dont les extrémités sont le premier et le neuvième quartile

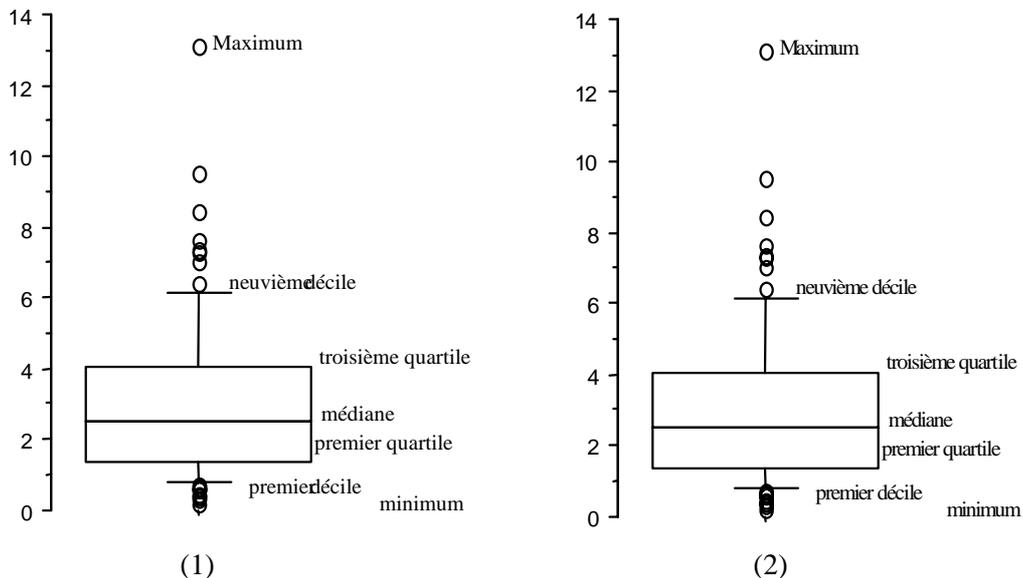
Écart interquartile : longueur de l'intervalle interquartile, i.e. différence entre le troisième et le premier quartile.

Écart interdécile : longueur de l'intervalle interdécile, i.e. différence entre le neuvième et le premier décile.

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.

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 :



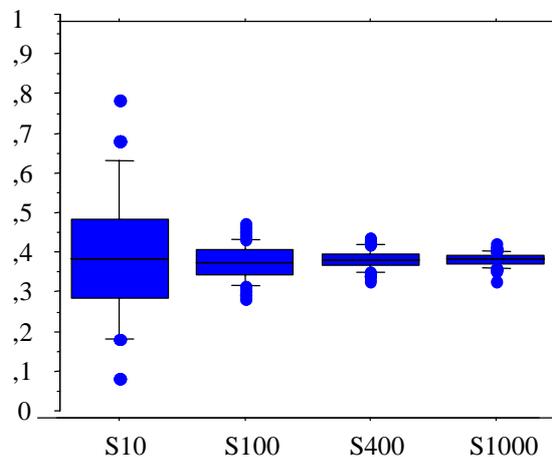
Nous conviendrons de choisir « par défaut » la définition représentée graphiquement en (1), où figurent les premiers et neuvièmes déciles. Si une ou plusieurs valeurs extrêmes sortent résolument des limites du dessin, on indique dessous leurs valeurs sans les représenter.

Néanmoins, les enseignants pourront utiliser des boîtes dont les extrémités sont les 1^{er} et 99^{ème} 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 inter-quartile ; 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^{ème} centile : 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 longueurs 1,5 fois l'écart interquartile ne se justifie plus.

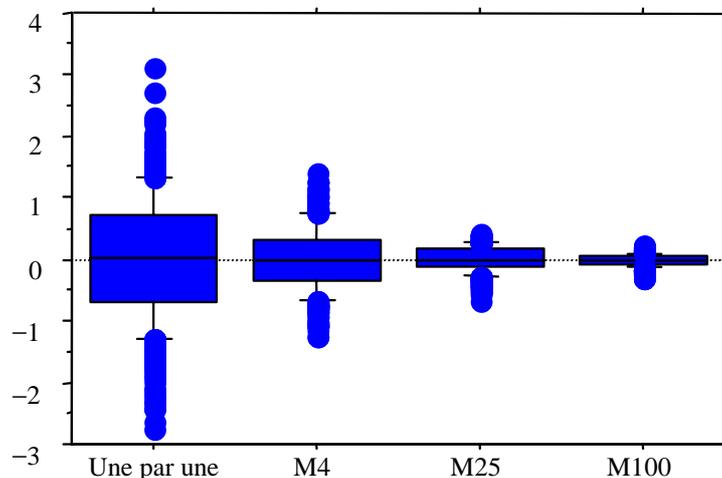
Les diagrammes en boîtes, comme les histogrammes, résument graphiquement une série ; l'idée de base est la suivante : au lieu de partager l'ensemble des valeurs possibles en segments égaux, on les partage en segments (quartile, déciles, centiles) qui contiennent une proportion prédéterminée des valeurs 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.

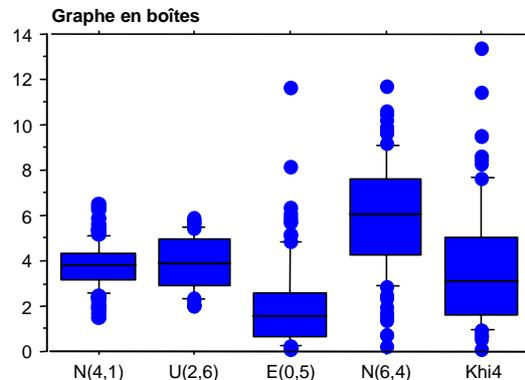
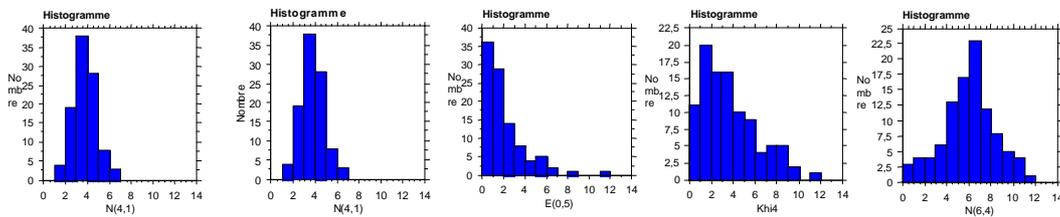


Les deux exemples situés ci-dessus sont spectaculaires et aisés à interpréter. Pour des séries de données quelconques, interpréter un diagramme en boîte demande un peu d'expérience et d'honnêteté pour ne pas transformer en affirmation théorique une observation lue sur un diagramme, que ce soit un histogramme 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 moyenne 4 et 6 et de variance 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 théorique 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 khi-deux à 4 degrés de libertés : ces lois 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 (voir annexe). 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 une série de données réelles (tailles d'enfants de 6 ans).
- dans l'applet de simulations « diagrammes en bâtons, histogrammes et quantiles » superposer les histogrammes, fonctions de répartition, 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.

Annexe : les logiciels SEL et SMEL

Ces logiciels ont pour origine la question suivante :

Comment se servir des outils proposés par le Web pour l'auto formation ou l'enseignement ?

Une réponse possible réside dans la mise à disposition de logiciels interactifs et dont la structure permet à la fois une mise à jour aisée et la transposition dans d'autres domaines.

Le domaine choisi est la statistique ; un premier logiciel SEL (statistique en ligne) est destiné à l'auto formation des enseignants de sciences des lycées et collèges et concerne la statistique descriptive :

<http://www.inrialpes.fr/sel>

Un second logiciel, SMEL, qui englobe le premier, contient aussi des éléments de base de statistique inférentielle :

<http://www.math-info.univ-paris5.fr/smel>

Ces deux logiciels sont structurés en quatre couches : une couche *articles* (26 articles dans SMEL), une couche *lexique*, une couche *cours*, une couche *simulation*. Dans les articles, les mots du lexique (environ 200 dans SMEL : loi normale, variance, loi de probabilité, etc.) sont cliquables : apparaît alors la page correspondante du lexique (page html), où le mot est brièvement défini et où figure dans certains cas une applet interactive (applet en java le plus souvent construite sur un jeu de données réelles accessibles à partir de la page d'accueil du logiciel ; les graphiques comportent des curseurs mobiles et des fenêtres à contenus modifiables) ; dans cette page il y a de plus des fenêtres à menu déroulant :

- une fenêtre *Voir aussi* permet d'accéder directement à des mots en lien avec celui de la page ouverte,
- une fenêtre *Plus de détails* renvoie à un paragraphe du cours (où les mots du lexique sont aussi cliquables),
- une fenêtre *Lecture* renvoie aux articles ayant trait au contenu de la page.

Enfin, un renvoi à la couche simulations propose des applets de simulations (comparaison d'estimateurs, loi des grands nombres, analyse de variance, tests, diagrammes, etc.).

Les pages de logiciel sont écrites automatiquement à partir d'un programme contenant tous les éléments à mettre dans cette page, sauf l'applet qui est spécifique du mot concerné dans la plupart des cas.

Chaque utilisateur se définira peu à peu des parcours de formation à l'intérieur du logiciel et grâce à l'interactivité pourra expérimenter soit à partir d'un jeu de données soit avec des données simulées. Ainsi, l'article « courbes de croissance » ouvre le chemin à un travail sur la loi de Gauss Le néophyte pourra se familiariser avec les éléments de base de la statistique descriptive (quantiles, écart-type, divers diagrammes, etc.), voir ce que signifie le loi des grands nombres (et le contre-exemple de la loi de Student de paramètre 1), avoir une idée de ce qu'est l'analyse de la variance, et le moins néophyte pourra s'étonner devant les résultats proposés dans l'applet « variance biaisée et non biaisée »

Les logiciels SEL et SMEL sont téléchargeables et...gratuits !