Diagrammes en boîtes

Ce sujet apparaît dans les programmes de Première (S, ES et L, cf. [2], pp.85-88).

Les diagrammes en boîtes [1] sont aussi appelés boîtes à moustaches, boîtes à pattes, diagrammes de Tukey, du nom du mathématicien qui les a introduits. Nous allons les disposer selon un axe de symétrie vertical comme dans les 2 exemples ci-dessous (extrait de [2], p. 86).

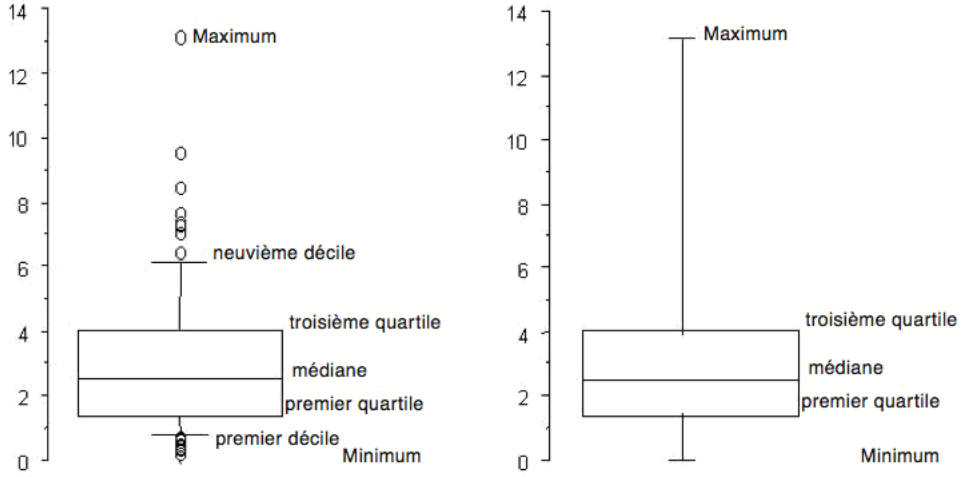


Diagramme 1

Il se trouve que *la largeur des boîtes n'a aucune signification* [2]. Par conséquent, *seules importent les ordonnées des différents traits horizontaux du dessin.* Ce qui veut dire que l'on aurait pu se contenter de les repérer sur un axe. Voilà pourquoi nous pensons que les diagrammes en boîtes sont essentiellement des gadgets et un phénomène de mode.

Avec beaucoup de bonne volonté, on peut dire que c'est *une manière astucieuse de présenter le résumé d'une série statistique.* Mais *résumer une seule série statistique* par un diagramme en boîtes est sans grand intérêt. Il faut limiter cet usage aux comparaisons de *plusieurs* séries statistiques *se rapportant à un même sujet* parce que c'est surtout dans ce cas que les propriétés graphiques des diagrammes en boîtes peuvent apporter quelque chose.

Du bon usage de la statistique descriptive : Bien sûr, vos élèves ne verront pas l'intérêt de faire le résumé d'une série statistique courte [3]. Au contraire, si sa taille est de l'ordre de grandeur du millier ou plus [4], vous aurez besoin, pour vous faire une idée de la localisation et de la dispersion de cette série, de ses minimum, maximum, étendue, moyenne, écart-type [5], médiane, premier et troisième quartiles, premier et quatre-vingt-dix-neuvième centiles (à la rigueur) de la série statistique.

Vous ne résumerez pas une série statistique par le même type de diagramme en boîtes suivant qu'elle est courte ou longue, c'est ce que montre les diagrammes que nous avons reproduits ci-

1. selon la terminologie des programmes de mathématiques des classes de Première, séries ES, L et S.
2. Les diagrammes en boîtes introduits par Tukey échappent à cette critique parce qu'il se plaçait dans un cadre bien balisé, qui lui permettait de donner une signification à la largeur des diagrammes. En voulant faire des diagrammes en boîtes un outil universel, on allait évidemment au-devant de sérieuses difficultés.
3. série statistique dont la taille n est de l'ordre de grandeur de la dizaine
4. Des séries de plusieurs millions de termes sont courantes.
5. Pour simplifier, nous omettrons l'écart-type dans ce qui suit.

dessus (par simple capture d'écran du document cité). Le choix d'un diagramme en boîtes dépendra aussi de la *nature* des données.

Le diagramme de gauche de (1) fait apparaître toutes les valeurs de la série statistique considérée, qui est inconnue, entre le minimum et le premier décile d'une part, entre le neuvième décile et le maximum d'autre part. On en déduit qu'il s'agit d'une série d'à peu près 80 termes, que les 10% de valeurs les plus petites sont peu dispersés tandis que les 10% de valeurs les plus grandes le sont beaucoup. En fait, les valeurs sont peu dispersées entre le minimum et le neuvième décile. En particulier, l'*écart interquartile* $q_3 - q_1$ est faible. On notera que commenter des nombres dont on ne connaît pas l'origine ressemble assez à un jeu idiot. Remarquons aussi que, dans le cas d'une série statistique de 3000 termes, par exemple, il y aurait à peu près 300 termes entre le neuvième décile et le maximum ou entre le minimum et le premier décile. Il ne serait donc pas possible de représenter toutes les valeurs concernées, simplement parce que cela ferait deux grosses taches !

Le diagramme de droite de (1) est beaucoup plus clair, mais est moins bien renseigné. On ne peut rien en déduire sur la taille de la série correspondante ou sur l'écart entre le minimum et le premier décile ou entre le neuvième décile et le maximum, qui sont des quantités assez importantes.

Notre conclusion est que *suivant la nature et la taille de la série statistique, on choisit de fabriquer tel type de diagramme en boîtes plutôt que tel autre*[6].

Par exemple, refaisons l'intéressant exercice proposé dans [2], pp. 86-87 :

Exercice 1 *On fait 100 simulations d'un sondage*[7] *de taille 10 dans une population dont les individus sont codés 0 ou 1 et à chaque sondage, on calcule la proportion de 1 sortis ; puis on recommence dans les mêmes conditions 100 simulations d'un sondage de taille 100, 100 simulations d'un sondage de taille 400 et 100 simulations d'un sondage de taille 1000. On obtient 4 séries statistiques de taille 100 dont les éléments sont des proportions de 1. Le problème est de tracer les diagrammes en boîtes de ces 4 séries et de les commenter.*

Pour tracer ces diagrammes en boîtes relatifs à 4 séries statistiques de taille 100, nous avons choisi de faire apparaître, dans cet ordre évidemment quand on remonte l'axe des ordonnées, les minimum, C1, D1, Q1, m, Q3, D9, C99, maximum[8]. Voilà ce que nous avons obtenu en exécutant le fichier « Boite » (fichier « scilab », en ligne) :

6. Nous imaginons que les macros ou algorithmes que l'on trouve sur Internet et qui produisent des diagrammes en boîtes se limitent au minimum (du style du diagramme de droite de (1)) et par conséquent manquent une partie du problème.

7. Il s'agit simplement de tirages au hasard avec remise dans la population.

8. tout en ayant fortement l'impression que l'on pourrait se passer des centiles.

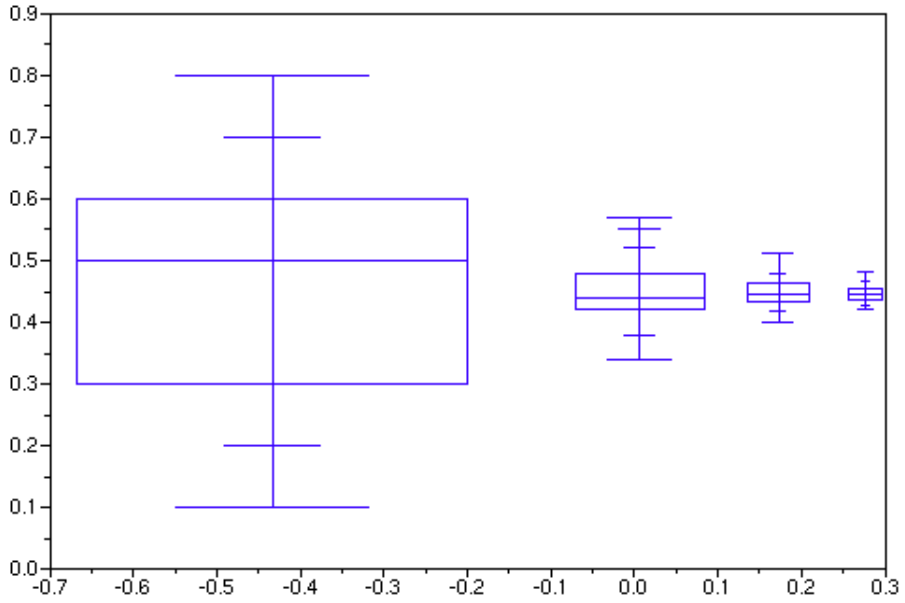


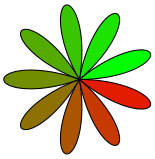
Diagramme 2

Constatant que la dispersion des séries statistiques diminue autour d'une valeur qui semble se rapprocher de 0.45, on pourra dire, en pensant à la loi des grands nombres[9] que les 4 séries statistiques se régularisent (quand la taille des sondages augmente) et se rapprochent de l'état idéal d'une série dont tous les termes vaudraient 0.45, ce qui explique les observations. L'intérêt de cet exercice est donc sa relation avec la loi des grands nombres.

Références

[1] - Mathématiques, Classe de Première, Série scientifique, BO n° 7, 31 août 2000 Hors-série, probabilité et statistique :
ftp://trf.education.gouv.fr/pub/edutel/bo/2000/hs7/v5proba.pdf

[2] - Mathématiques, classes de première des séries générales, collection Lycée – voie générale et technologique, série *Accompagnement des programmes*
http://www.cndp.fr/archivage/valid/86906/86906-13718-17372.pdf



9. D'après la loi des grands nombres, si l'on faisait grandir indéfiniment la taille des sondages à partir desquels on calcule les proportions, ces proportions convergeraient vers 0.45, qui est la proportion de 1 choisie *au moment de la simulation des sondages.*