La boîte à moustaches pour sensibiliser à la statistique

Monique Le Guen

LA BOITE A MOUSTACHES POUR

SENSIBILISER A LA STATISTIQUE

Monique Le Guen
CNRS- MATISSE[1]

Résumé

La boîte à moustaches une traduction de *Box & Whiskers Plot*, est une invention de TUKEY (1977) pour représenter schématiquement une distribution. Cette représentation graphique peut être un moyen pour approcher les concepts abstraits de la statistique. Nous abordons dans cet article la nécessité de repenser l'initiation à la Statique. En nous appuyant sur les nouvelles connaissances en neuro-sciences nous proposons de placer l'apprenant en situation de **découverte,** en utilisant de vraies données, par l'intermédiaire de logiciels orientés Analyse Exploratoire des Données. Nous détaillons dans une seconde partie, comment lire et interpréter des boîtes à moustaches. Nous montrons comment les élèves peuvent découvrir, en explorant des données, certaines propriétés de la médiane et de la moyenne. En références nous donnons des adresses Internet pour réaliser informatiquement des boîtes à moustaches.

Cet article est destiné aux enseignants et aux praticiens de la Statistique Appliquée.

Mots clés : Sensibilisation à la Statistique, Interactivité, Visualisation, Analyse Exploratoire des Données, AED, J. W. TUKEY, Boîte à moustaches.

Keys Words : Statistics Education, Interactivity, Visualization, Exploratory Data Analysis, EDA, J. W. TUKEY, Box and Whiskers Plot.

Sommaire

[1] MATISSE-CNRS UMR8595, Maison des Sciences Economiques, 106-112 Boulevard de l'Hôpital, 75013 Paris.

1. De la nécessité de repenser l'initiation à la Statistique

L'usage élémentaire de la Statistique vue comme une aide au traitement et au résumé de l'information a envahi notre vie quotidienne. De la lecture d'un journal quotidien, aux travaux plus complexes de la Recherche il n'existe pas de rubriques ou de disciplines, qui ne fassent appel à des notions de base de la Statistique.
Cet élargissement dans les connaissances conduit à ce que tous les élèves sortant de l'enseignement secondaire aient une approche pragmatique des notions de base de la Statistique. Ces notions enseignées dans le secondaire leur permettront d'acquérir une plus grande autonomie dans leurs jugements, ne serait-ce que dans leur vie citoyenne.
Son pré-enseignement peut débuter avant l'entrée à l'Université, si nous en modifions les contenus.

L'approche doit être du domaine de la **découverte,** se faire en situation réelle donc pratique (ROSSMAN A. J. 1995). Apprendre à explorer, à représenter sous des formes multiples, à manipuler les pourcentages, les fréquences, les moyennes, la médiane, les quartiles, le mode, la variabilité, conduit plus *naturellement* au concept abstrait de position centrale, d'écart-type, de variance et de *distribution*.

L'enseignement de la Statistique que nous, enseignants et chercheurs avons supporté, était largement influencé par la théorisation mathématique, donc affaire de matheux, de livres de maths et de formules mathématiques. Depuis les années 1980, se substitue un enseignement *autrement* qui favorise l'émergence des concepts abstraits (LE GUEN 1999, *Voir, Apprendre, Comprendre Autrement)*.

Ces changements reposent sur la micro informatique. Ses nouveaux concepts, les fenêtres, les manipulations via la souris, les visualisations, l'interactivité homme machine, et l'arrivée des tableurs ont favorisé la diffusion dans presque tous les milieux : familial, scolaire, universitaire et professionnel. Les jeunes n'ont plus aucune réticence, contrairement aux adultes novices, à utiliser un clavier. Découvrir les « Maths » et les « Stats » via ce média devient une activité ludique, et non plus une source d'angoisse pour la plupart, voir l'encadré *A propos d'Horace*. L'élève devient actif dans ces choix, il découvre par lui même tout en étant guidé par son enseignant (PAPERT 1980).

Les outils sont maintenant disponibles. Oui, mais pour enseigner la Statistique autrement il faut d'autres ingrédients : Quoi et comment enseigner ? En l'état actuel des connaissances, une unanimité se fait jour au niveau international : Enseigner à partir des outils de l'**A**nalyse **E**xploratoire des **D**onnées initiée par J. W. TUKEY (***E**xploratory **D**ata **A**nalysis*, **EDA** 1977).

Les idées de TUKEY reprises et prolongées par ses nombreux doctorants, collègues devenus à leur tour enseignants et/ou développeurs (BEHRENS J.T., FRIENDLY, FOREST Y., HOAGLIN, HUBER P.J. , MOSTELLER, VELLEMAN, etc.) ont gagné le monde anglo-saxon. Dans le monde francophone l'A.E.D. reste encore peu répandue. En Europe les Sciences sociales ont été les pionnières. L'Allemagne, la Suisse, l'Espagne (BATANERO et al. 1991), par exemple ont des enseignements d'Analyse Exploratoire des Données. La France serait plutôt à la traîne (DESTANDAU S., LADIRAY D., LE GUEN M., 1999 *Analyse Exploratoire des Données*).
La langue est sans conteste le premier handicap.

A cela il faut ajouter les changements de mentalité et de conception importants que nécessite cet enseignement (LE GUEN 1999, *De l'importance de l'image)*. Il est donc nécessaire pour concevoir ces nouveaux programmes de développer une collaboration et des échanges entre toutes les bonnes volontés. La démocratisation, l'accès à l'information, et la diffusion que permet Internet peut être le support d'une telle entreprise.

Lançons une boutade et un espoir. On n'a jamais été aussi près d'une amélioration des enseignements. Les jeunes le réclament et sont même prêts à collaborer par leurs capacités à développer et à voguer sur Internet, sans parler de leur créativité et de leur volonté de changement[2].

A Propos d'HORACE (extrait)

J'étais alors en proie à la *mathématique*
Temps sombre! enfant ému du frisson poétique,
Pauvre oiseau qui heurtait du crâne mes barreaux,
On me livrait tout vif aux *chiffres*, noirs bourreaux ;
On me faisait de force ingurgiter *l'algèbre* ;
On me liait au fond d'un Boisbertrand funèbre ;
On me tordait, depuis les ailes jusqu'au bec,
Sur l'affreux chevalet des *X* et des *Y* ;
Hélas, on me fourrait sous les os des maxillaires
Le *théorème* orné de tous ses *corollaires* ;
Et je me débattais, lugubre patient
Du *diviseur* prêtant main-forte au *quotient*.
De là mes cris.

VICTOR HUGO, 1831
Les contemplations, Aurore
GF Flammarion p57

« La mathophobie endémique de la culture contemporaine empêche quantités de personnes d'assimiler toute notion reconnue pour "mathématique", alors que d'autres notions mathématiques sont acquises sans difficultés, dès lors qu'elles ne sont pas perçues comme telles ».

PAPERT S. (1980)
« *Jaillissement de l'esprit*
Ordinateurs et apprentissage »

« We Believe that data should be at the heart of all statistics education and that students should be introduced to statistics through data-centered courses ».

THOMAS MOORE & ROSEMAY ROBERTS (1989).

« Automate calculation and graphics as much as possible ».

DAVID MOORE (1992)

[2] Voir le site et la lettre d'information mensuelle du Mouvement de étudiants pour la réforme de l'enseignement de l'économie. http://www.autisme-economie.org/

1.1 John Wilder TUKEY (1915-2000).

Sur le Web plusieurs sites retracent la biographie de TUKEY (cf. Références), nous présentons ici quelques points de repère.

J. W. TUKEY[3] est né dans le Massachussets. Il suit d'abord un enseignement de chimie à l'Université de Brown, concrétisé par un PHD, puis s'oriente vers les mathématiques à l'Université de Princeton et obtient deux PHD en Mathématiques en 2 ans. Entre 1939 et 1945 il découvre la Statistique en travaillant avec l'armée. À partir de 1945 et tout au long de sa carrière, TUKEY se partagera entre l'enseignement de la statistique, à l'Université de Princeton, et la Recherche & Développement, au sein de la direction technique des laboratoires AT&T Bell Company à Murray Hill.

Son œuvre est considérable. On lui doit, mais la liste n'est pas exhaustive, la technique de la *Median Polish*, le lissage par médianes mobiles, l'algorithme de la transformée de Fourier rapide (FFT), quelques lois de probabilités, le *Jackknife* (qu'il a lui-même baptisé ainsi, du nom du couteau multi-usages du boy-scout), les graphiques *Stem and Leaf* (tige et feuille) , *Box Plot*, *Box & Whiskers Plot*, sans oublier, bien sûr, la *Tukey's Line*, le *Tukey's Quick Test*, le *Tukey's Test for Non-Additivity*, le test de *Siegel-Tukey* et le critère de *Tukey-Kramer* etc.

Son influence majeure est d'avoir apporté une distinction entre l'Analyse Exploratoire des données et l'Analyse confirmatoire des données, dans un esprit analogue à J. P. BENZECRI.

En avance sur son temps, il a également proposé une révision de l'enseignement de la Statistique. Le développement des techniques informatiques, *hardware* et *software*, ont permis récemment les réalisations et la diffusion de ses idées.

1.2 Tukey et l'image - Des mots nouveaux , Des expressions nouvelles

Trimming, *Winsorized Mean*, *Software*, *Brainware* et *Bit* (Binary digIT), sont autant de mots, d'expressions que TUKEY a inventés.

L'accès aux articles et ouvrages de TUKEY sera plus facile si l'on commence par lire les écrits de ses élèves et collègues. Son style d'écriture est en effet particulier, et parfois très imagé. Sous sa plume, les quartiles peuvent devenir des « hindges » (littéralement «pivots, gonds ou charnières »), les valeurs extrêmes des « ones », la transformation d'une variable une « re-expression ». Lorsqu'il compare l'aplatissement d'une distribution observée à la loi normale, il parle, de « sharpness » ou de « spikyness » plutôt que de Kurtosis, ce qui est plus compréhensible par le novice anglophone.

Et les exemples de même nature sont foison ! Pour les francophones, traduire l'esprit TUKEY n'est donc pas toujours évidente.

Depuis quelques années JACQUES VANPOUCKE de l'Université Paul Sabatier de Toulouse, co-fondateur et animateur de l'Association MIRAGE[4], nous propose des traductions originales et pertinentes dans l'esprit TUKEY.

[3] TUKEY s'orthographie **T U K E Y** et non avec un C comme dans TUCKEY.

[4] Association MIRAGE (**M**ouvement **I**nternational pour le Développement de la **R**echerche en **A**nalyse **G**raphique et **E**xploratoire) organise chaque année en Septembre une école d'été à Carcassonne, sur l'Analyse Exploratoire des données**.** http://www.unige.ch/ses/sococ/mirage/

Ainsi le *Box et Whiskers Plot*[5] sera traduit par boîte à pattes (BàP) ou boîte à moustaches. Autre exemple, le *stem & leaf* devient le branchage, et l'étude d'une distribution par les quantiles (fractiles) devient une fractilogénèse.

Arrêtons nous sur les variétés de Box Plots.

Le **terme générique** *Box Plot* et le **terme spécifique** *Box & Whiskers Plot* recouvrent une grande variété de diagrammes en forme de boîtes qui se différencient par leur construction, leurs interprétations, et leurs usages. E. HORBER qui a effectué des recherches bibliographiques sur ce thème a repéré une soixantaine de formes et de constructions différentes. Le lecteur pourra se faire une opinion en lisant sa note disponible sur Internet[6] . La conclusion est que le vocabulaire anglo-saxon n'est pas unifié, les termes sont souvent employés les uns pour les autres. Pour les francophones se rajoute la (ou une) traduction. Ainsi la traduction de *Box & Whiskers Plot* par boîte à moustaches n'est pas unique. Nos amis Québécois disent boîte à moustaches. Nos collègues de l'Association MIRAGE utilisent plus volontiers le terme Boîte à Pattes. Il fallait choisir.

Nous avons choisi dans cet article, la traduction boîte à moustaches et nous allons décrire la boîte à moustaches la plus couramment utilisée par les explorateurs de données. C'est aussi celle que l'on trouve dans la plupart des logiciels statistiques.

2. A l'utilité de la boîte à moustaches de TUKEY

La boîte à moustaches est une représentation schématique de la distribution d'une variable. Cette représentation graphique peut être un moyen pour approcher les concepts abstraits de la statistique, si l'on pratique son usage sur différents jeux de données.
Tout d'abord nous montrons une représentation[7] d'une boîte à moustaches, construite sur un jeu de données. L'interprétation d'une boîte à moustaches nécessite un apprentissage aussi nous détaillons comment lire et interpréter ce graphique. Nous montrons comment les élèves peuvent découvrir, en explorant des données, certaines propriétés de la médiane et de la moyenne.
En références nous donnons des adresses Internet pour réaliser informatiquement différentes formes de boîtes à moustaches et de Box Plots.

1. Les données

Pour chaque élève d'une classe mixte, d'effectif 59, sont collectés son **poids** en kilogrammes, sa **taille** exprimée en centimètres et son **sexe** (code 1 pour masculin, code 2 pour féminin), cf. Annexe.
Le fichier des données comporte 3 variables POIDS, TAILLE et SEXE, et 59 observations (élèves) réparties selon le sexe (23 garçons et 36 filles).

Cet exemple est inspiré des données de BATANERO, ESTEPA & GODINO (1991) disponibles également sur Internet[8].

Pour de jeunes élèves, en collège et lycée, les ouvrages de ROSSMAN A. J. (1995, 2001) rassemblent de nombreux jeux de données et exemples d'activités pour découvrir la Statistique.

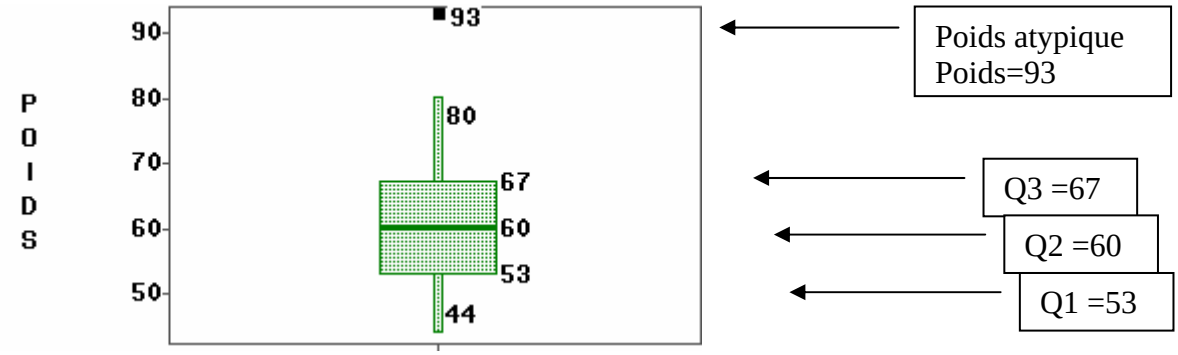
---

[5] *Whiskers* en anglais signifie moustaches et favoris (pattes). Sans doute un jeu de mots de TUKEY pour « imager » l'asymétrie souvent rencontrée dans les distributions observées.

[6] Site Internet : http://www.unige.ch/ses/sococ/mirage/ dans la rubrique Nouvelles Juin 2001.
[7] Les graphiques ont été réalisés avec le logiciel SAS®, par la Procédure BoxPlot ou par le module SAS/INSIGHT.
[8] Site Internet : http://www.ugr.es/~batanero/ListadoEstadistica.htm

2. La boîte à moustaches

La représentation graphique de la boîte à moustaches est mystérieuse lorsqu'on la découvre pour la première fois, cf. Graphique 1: *Boîte à moustaches de la variable POIDS*. Pour lire et interpréter, il est nécessaire de connaître sa construction.
La boîte à moustaches utilise 5 valeurs qui résument des données : le minimum, les 3 quartiles Q1, Q2 (médiane), Q3, et le maximum.



Graphique 1 *: Boîte à moustaches de la variable POIDS*

Les quartiles Q1, Q2, Q3 sont les éléments essentiels de ce graphique. Après une présentation des quartiles sur un exemple simple, nous détaillerons les étapes de la construction des quartiles et de l'écart interquartile qui s'en déduit.

2.1 Les quartiles et l'écart interquartile

Pour illustrer notre propos, nous montrons sur un cas très simple[9] comment sont calculer les quartiles.
Soit la série des 9 valeurs ordonnées : 1 , 3 **, 4 ,**5 , **6** ,7 , **9** ,10, 15

La médiane Q2 partage la série en deux groupes d'effectif égaux, ce qui donne :
Q2=6.
Le Quartile Q1 repartage le groupe du bas (5 valeurs inférieures) en deux groupes d'effectif égaux, ce qui donne : Q1=4.
Le Quartile Q3 repartage le groupe du haut (5 valeurs supérieures) en deux groupes d'effectif égaux, ce qui donne : Q3=9.

Selon que l'effectif n des valeurs est pair ou impair, on procédera différemment pour évaluer les quartiles.

Procédure:

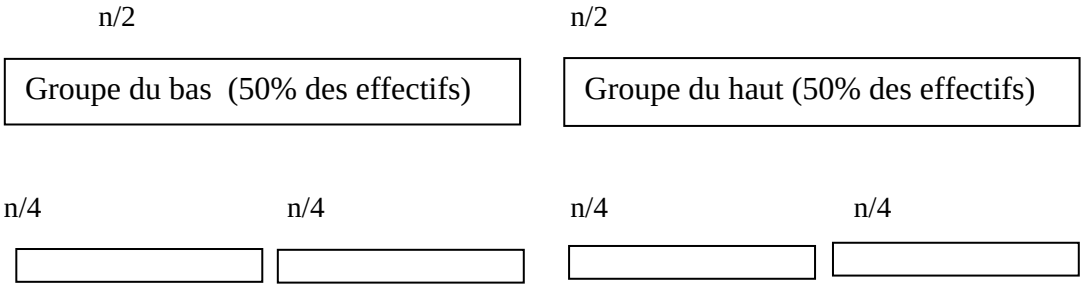
1- Classer les n données par ordre croissant.
2- Diviser les données en 2 groupes de tailles égales.
On obtient le groupe du bas et le groupe du haut, chacun contenant 50% des observations.

Si n est pair ➔ la médiane est la moyenne des 2 points milieu.
Si n est impair ➔ la médiane est le point milieu.

[9] En pratique le calcul des quartiles s'effectue lorsque le nombre d'observations est plus important.

Dans ce cas il faut, pour permettre les calculs qui vont suivre, reproduire la valeur de ce point dans les 2 groupes.

3- Calculer à nouveau la médiane du *groupe du bas*.
On obtient le quartile Q1, qui correspond à 25 % des observations.
4- Calculer à nouveau la médiane du *groupe du haut*.
On obtient le quartile Q3, qui correspond à 75 % des observations.



L'écart interquartile (*InterQuartile Range*) est utilisé comme indicateur de dispersion. Il correspond à 50% des effectifs situés dans la partie centrale de la distribution. Pour la variable POIDS l'écart interquartile vaut 14, cf. Graphique 1.

$$Ecart Interquartile = Q3 - Q1 = 67 - 53 = 14$$

2.2 Lecture d'une boîte à moustaches

On repère sur la boîte à moustaches d'une variable:

- l'échelle des valeurs de la variable, située sur l'axe vertical.
- la valeur du 1er quartile Q1 (25% des effectifs), correspondant au trait inférieur de la boîte,
- la valeur du 2ème quartile Q2 (50% des effectifs), représentée par un trait horizontal à l'intérieur de la boîte,
- la valeur du 3ème quartile Q3 (75% des effectifs), correspondant au trait supérieur de la boîte,
- les 2 « moustaches» inférieure et supérieure, représentées ici par les petits rectangles verticaux de part et d'autre de la boîte. Ces 2 moustaches, délimitent les valeurs dites *adjacentes* qui sont déterminées à partir de l'écart interquartile (Q3-Q1).
- les valeurs dites extrêmes, atypiques, exceptionnelles, *(outliers*) situées au-delà des valeurs adjacentes sont individualisées. Elles sont représentées par des marqueurs (carré, ou étoile, etc.).

2.3 Délimitation des longueurs des moustaches (valeurs adjacentes)

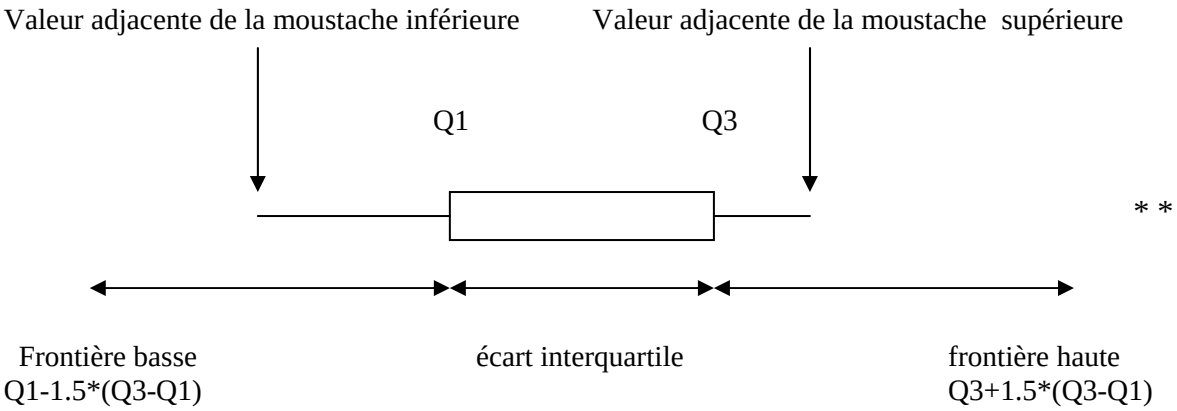
L'extrémité de la moustache inférieure est la valeur minimum dans les données qui est supérieure à la valeur **frontière basse** :

Q1 -**1,5***(Q3-Q1) soit 32 pour la variable POIDS

L'extrémité de la moustache supérieure est la valeur maximum dans les données qui est inférieure à la valeur **frontière haute** :

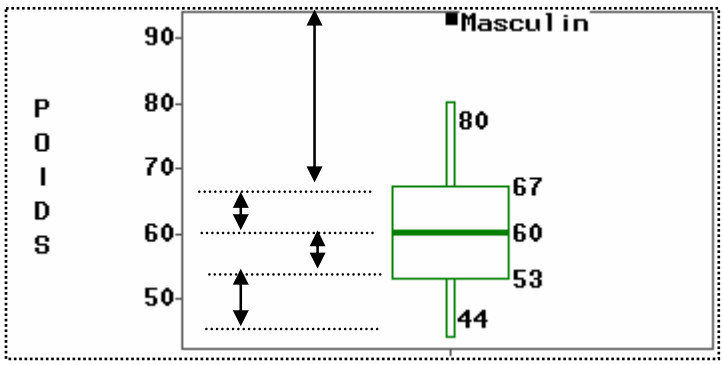
Q3 +**1,5***(Q3-Q1) soit 88 pour la variable POIDS

Dans le schéma suivant deux valeurs sont atypiques car situées au delà de la frontière haute.



2.4 Lecture de la boîte à moustaches de la variable POIDS

Sur le Graphique 1 : *Boîte à moustaches de la variable POIDS*, la médiane des élèves est à 60 kilos, le quart des élèves de poids faible se situe entre 44 et 53 kilos. La moitié des élèves de poids moyen se situe entre 53 et 67 kilos et le dernier quart des élèves se situe entre 67 et 93 kilos. Un élève a un poids de 93 kgs, atypique par rapport à ses camarades.
Une seule valeur est atypique (93) car elle est située au delà de la frontière haute (88). Aucune valeur atypique ne se trouve au delà de la frontière basse (32).



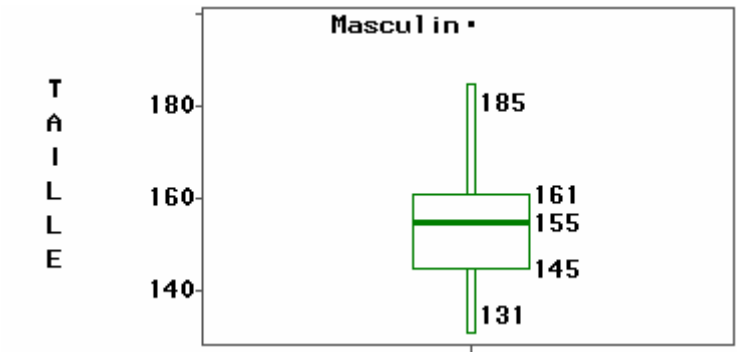
La distribution est décomposée en 4 zones de même effectif (25%) .

Graphique 2 *: Le point atypique correspond au poids d'un garçon.*

Bien que la distribution soit découpée en 4 zones (quartiles) de même effectif (25%) les plages de valeurs du poids ne sont pas égales (Graphique 2). La distribution est plus allongée vers les valeurs élevées du poids.
C'est une première lecture de la boîte à moustaches : allure générale de la distribution avec individualisation des points atypiques.

Selon les logiciels il est possible de cliquer sur les points extrêmes pour les identifier par une étiquette, voir Graphique 1 (➔ repérer 93 le poids le plus élevé) et Graphique 2 (➔ repérer le sexe du plus lourd, c'est un garçon).

Si le fichier des données contenait le nom des élèves, on pourrait afficher le nom de l'élève qui a un poids atypique. Après le **diagnostic,** les informations supplémentaires facilitent le début d'une explication du « pourquoi » ce point est atypique.



Graphique 3 *: Boîte à moustaches de la variable TAILLE*

En changeant de variable, cf. le Graphique 3 : *Boîte à moustaches de la variable TAILLE*, l'élève peut faire les remarques suivantes :

- la médiane de la distribution des points n'est plus centrée dans la boîte,
- les moustaches ne sont pas toujours symétriques,
- dans les hautes valeurs, une seule observation est atypique

Pour le praticien qui analyse une distribution observée, la boîte à moustaches permet de répondre à certaines questions :

- Existe-t-il des observations atypiques ? ➔ en les repérant et les identifiant
- La distribution est-elle symétrique? ➔ en repérant la position de la médiane dans la boîte, et la dissymétrie des moustaches.
- Quelle est l'allure des queues de distribution ?
- La partie centrale (50% des effectifs) est-elle plus ou moins concentrée ou étalée par rapport au reste de la distribution?

2.5 Pourquoi la valeur 1.5 pour déterminer les moustaches?

Dans la boîte à moustaches définie par TUKEY, la boîte a pour hauteur la distance interquartile (Q3-Q1), et les moustaches sont basées généralement sur **1,5** fois la hauteur de la boîte. Dans ce cas, une valeur est atypique si elle dépasse de 1.5 fois l'écart interquartile au dessous du 1er quartile ou au dessus du 3ème quartile.
En se basant sur les quartiles, c'est à dire des statistiques d'ordre, la médiane et l'écart interquartile ne sont jamais influencés par les valeurs extrêmes.

La valeur 1.5 est selon TUKEY une valeur pragmatique (*rule of thumb*), qui a une raison probabiliste.

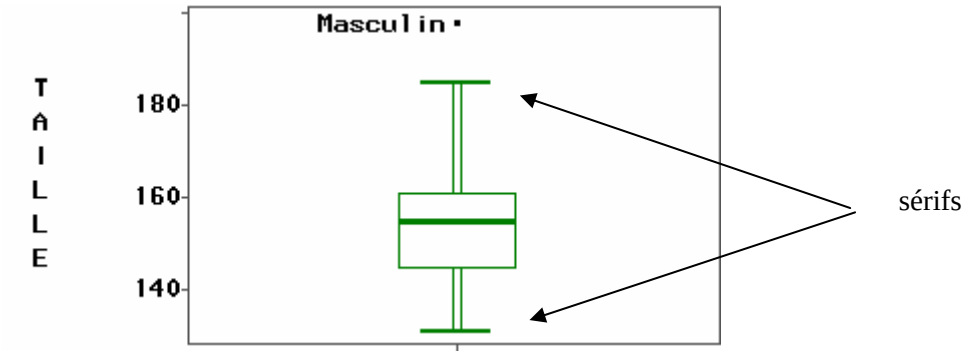
Si une variable suit une distribution normale, alors la zone délimitée par la boîte et les moustaches devrait contenir **99,3 %** des observations. On ne devrait donc trouver que **0.7%** d'observations atypiques (*outliers*). Si le coefficient vaut 1, la probabilité serait de **0.957**, et elle vaudrait **0.999** si le coefficient est égal à 2.

Pour TUKEY la valeur 1.5 est donc un compromis pour retenir comme atypiques assez d'observations mais pas trop d'observations.

Selon les logiciels le coefficient **1,5** est imposé ou paramétrable.

2.6 Représentations variées des boîtes à moustaches

La largeur de la boîte n'a aucune signification. Il existe des variantes dans la forme de la boîte. Certains logiciels représentent la boîte avec un simple trait. De même pour les moustaches elles peuvent être délimitées par des crochets ou des *sérifs* (empattements), traits horizontaux délimiteurs qui aident l'œil à mieux repérer les valeurs adjacentes cf. Graphique 4 : *Boîte à moustaches avec sérif*, etc.

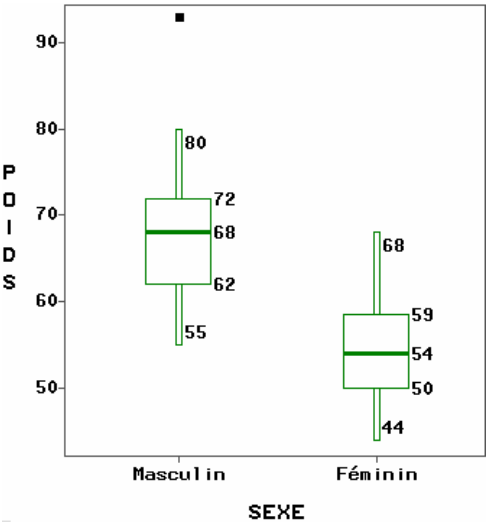


Graphique 4: *Boîte à moustaches avec sérif*

3. Les boîtes à moustaches juxtaposées

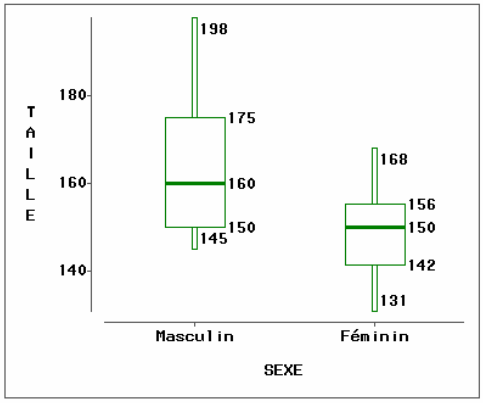
3.1 Comparaisons de distributions selon des groupes

Pour comparer les distributions de la variable POIDS selon les 2 groupes Masculin/Féminin, on juxtapose sur le même graphique les 2 boîtes à moustaches définies respectivement pour le groupe Masculin et le groupe Féminin, en utilisant la même échelle.



Graphique 5 : *Comparaison des distributions de la variable POIDS selon le sexe.*

Sur le Graphique 5 : *Comparaison des distributions des POIDS des élèves selon le sexe*, est visualisée une différence de poids entre filles et garçons (médiane à 68 pour le groupe Masculin et 54 pour le groupe Féminin, 1er quartile à 62 pour le groupe Masculin et 50 pour le groupe Féminin etc.). Il n'y a pas de poids atypique pour le groupe Féminin.



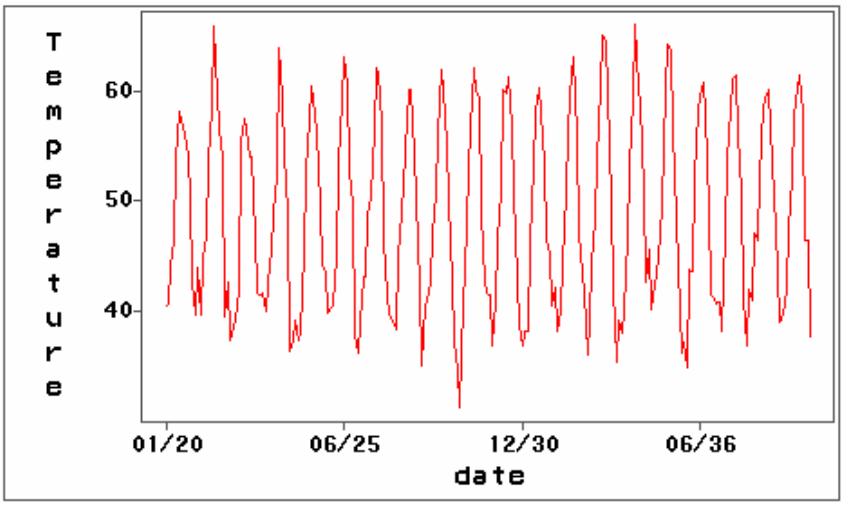
Graphique 6 : *Comparaison des distributions des tailles des élèves selon le sexe.*

Sur le Graphique 6 : *Comparaison des distributions des tailles des élèves selon le sexe*, l'écart interquartile est plus étalé pour le groupe Masculin que pour le groupe Féminin et la distribution est plus dissymétrique. Compte tenu de l'étalement dans la partie centrale de la distribution, il n'y a plus de taille atypique pour le groupe Masculin. Les moustaches s'étendent dans ce cas, jusqu'à la **valeur minimum** et la **valeur maximum**.

C'est précisément la facilité de comparaison qu'offre l'œil qui fait l'intérêt et la force de cette représentation visuelle. Cette visualisation conduit plus facilement à l'Analyse de la Variance (Comparaisons des moyennes compte tenu de leurs variances).

3.2 Utilisation des boîtes à moustaches pour visualiser des séries chronologiques

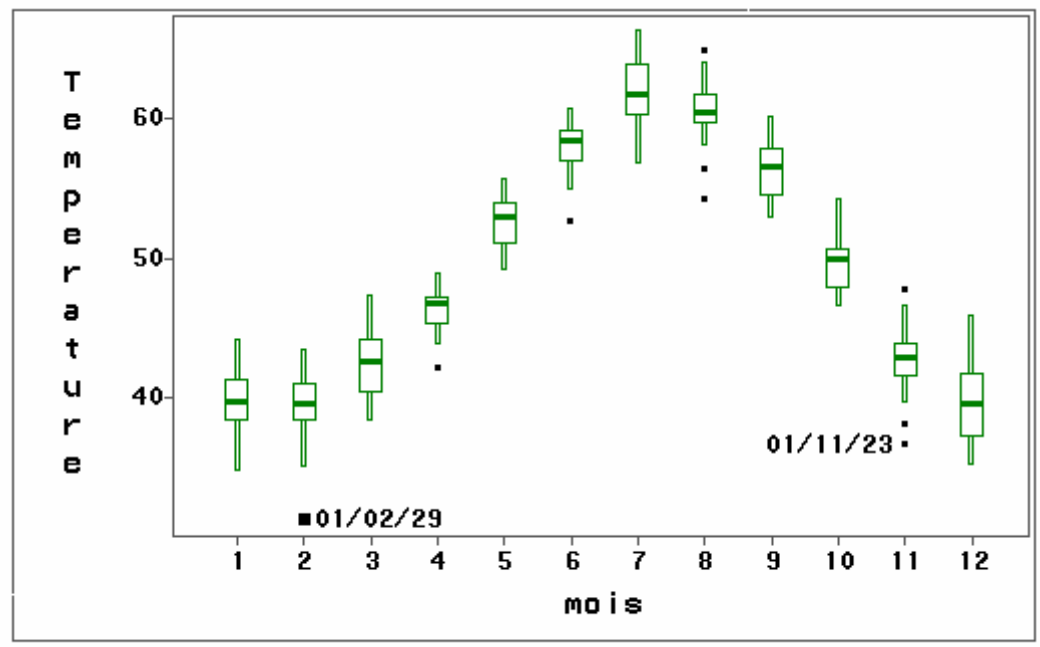
Soit la série[10] des températures mensuelles moyennes à Nottingham de 1920 à 1939. Cette série de 240 valeurs est représentée sous forme chronologique cf. le Graphique 7 : *Série des températures mensuelles moyennes à Nottingham de 1920 à 1939.*



[10] Site Internet des données source
http://wwwpersonal.buseco.monash.edu.au/~hyndman/TSDL/books/anderson.DAT

Graphique 7 : *Série des températures mensuelles moyennes à Nottingham de 1920 à 1939*

Ces mêmes données sont regroupées par mois et représentées sous forme de boîtes à moustaches cf. Graphique 8 **:** *Série des températures mensuelles moyennes à Nottingham regroupées par mois.*

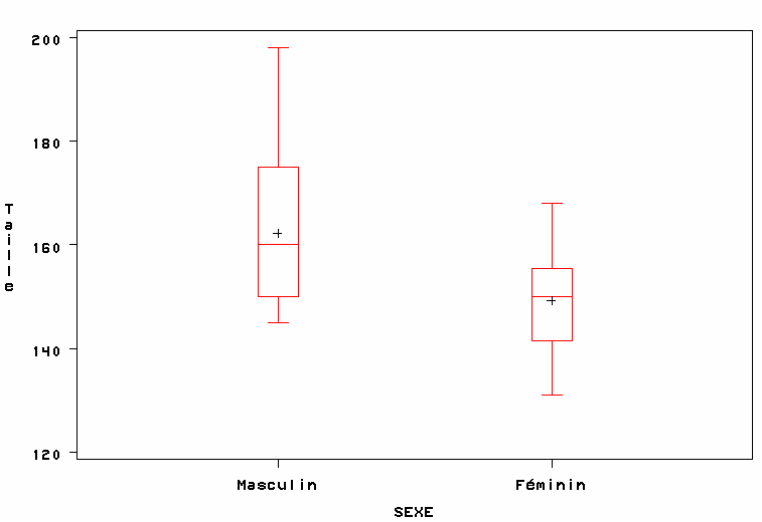


Graphique 8: *Série des températures mensuelles moyennes à Nottingham regroupées par mois.*

Ces deux graphiques donnent une vision différente des données. Les objectifs d'analyse diffèrent dans chacune des représentations.
Les graphiques qui utilisent des boîtes à moustaches permettent d'avoir une vue synthétique, **globale** et en même temps une vue **locale** sur les données (cf. valeurs atypiques).

4. Découvertes par l'élève des propriétés de la médiane et de la moyenne

Avec certains logiciels il est possible de positionner la moyenne et de la comparer visuellement à la médiane. Ainsi dans le Graphique 7 : *Comparaison des médianes et des moyennes*, la médiane (trait horizontal dans la boîte) est inférieure à la moyenne (symbolisée par une croix) pour le groupe Masculin, tandis qu'elle est très légèrement supérieure pour le groupe Féminin.



Graphique 7 : *Comparaison des médianes (trait horizontal) et des moyennes (symbolisées par une croix) de la variable TAILLE.*

En explorant, l'élève peut donner un sens concret à la moyenne et à la médiane et découvrir certaines de leurs propriétés.

- La médiane tout comme la moyenne n'est pas forcément égale à une valeur rencontrée dans les données.
- La médiane et la moyenne sont des représentants d'une position centrale dans les données.
- La médiane et la moyenne ont chacune une valeur comprise entre les valeurs extrêmes de la distribution. Les deux valeurs peuvent être égales ou différentes.
- Elles sont égales si la distribution est symétrique.
- Lorsque la distribution est plus allongée vers les grandes valeurs, la médiane est **inférieure** à la moyenne. Lorsque la distribution est plus allongée vers les petites valeurs, la médiane est **supérieure** à la moyenne.
- Plus la distribution est dissymétrique, plus la médiane s'écarte de la moyenne.
- En supprimant un point atypique dans les données, l'élève peut réaliser que la moyenne est très influencée par les valeurs extrêmes, ce qui n'est pas le cas de la médiane. Il peut ainsi approcher la notion de contribution.

Après avoir visualiser par des boîtes à moustaches différentes variables, les notions de variabilité, de distributions prendront un sens plus concret. L'élève pourra comprendre que si sur un jeu de données, il existe une différence entre la moyenne et la médiane, c'est un **diagnostic de dissymétrie**.

5. Réalisations informatiques des boîtes à moustaches

Pratiquement tous les logiciels actuels de Statistique permettent de réaliser des boîtes à moustaches. Par contre, dans le monde de la bureautique cette fonctionnalité est plus rare. Le tableur EXCEL de MS ne permet pas la réalisation immédiate d'un tel graphique. Il est nécessaire avant de réaliser le graphique, de calculer les différents éléments d'une boîte à moustaches en utilisant les fonctions statistiques de EXCEL.

Les sites Internet[11] de NEVILLE HUNT et CHRISTINE SPENCER donnent des exemples de réalisations de boîtes à moustaches avec EXCEL.

[11] *Site de* NEVILLE HUNT http://www.mis.coventry.ac.uk/~nhunt/boxplot.htm
Site de CHRISTINE SPENCER http://www.elementkjournals.com/ime/0008/ime0081.htm

Conclusion

Voir une distribution à partir d'une boîte à moustaches est un point de vue sur les données. Les données sont vues sous un certain angle qui ne permet pas de tout voir. Par exemple on ne peut repérer sur une boîte à moustaches les pics (modes) d'une distribution. Il existe d'autres représentations comme l'histogramme, classiquement enseigné, mais également le Branchage (*Stem & Leaf)*, le *dot plot* qui font partie comme la boîte à moustaches de l'arsenal de l'explorateur de données et qui viendront compléter son point de vue. Toutes ces formes graphiques font partie d'une nouvelle culture, et elles nécessitent un apprentissage.

La visualisation des données sous des formes graphiques variées, l'interactivité que permet la micro-informatique, l'accès généralisé à Internet comme source de documentation, et le développement d'Applets sont des moyens qui concourent à l'amélioration de l'enseignement et de la pratique de la Statistique. Mais tous ces changements ont un coût en formation pour les enseignants.

Remerciements

Nous remercions nos collègues de l'Ecole d'été EEDA 2001 à Carcassonne, et tout particulièrement E. HORBER, R. LAFOSSE, D. LADIRAY et J. VANPOUCKE pour leur apport et leurs conseils quant à la réalisation de ce document.

Nous remercions également Annie Morin Directrice de l'IREM de Rennes et Responsable de la Revue « Statistiquement Votre » accessible sur le Web, qui nous a permis de publier la première version de cet article.

Annexe : Les données

SEXE	POIDS	TAILLE
1	65	185
1	68	180
1	72	178
1	55	148
1	64	145
1	70	161
1	66	159
1	74	165
1	75	155
1	70	165
1	93	198
1	64	150
1	60	175
1	62	148
1	70	160
1	80	155
1	61	153
1	60	145
1	62	170
1	68	175
1	70	145
1	72	157
1	71	161
2	60	150
2	45	148
2	46	149
2	50	138
2	47	131
2	55	146
2	49	136
2	52	145
2	50	150
2	46	132
2	50	155
2	52	150
2	52	141
2	48	139
2	52	152
2	63	131
2	53	160
2	54	158
2	54	155
2	54	135
2	53	155
2	55	168
2	57	162
2	44	155
2	56	142
2	56	145
2	56	150
2	53	150
2	60	156
2	65	168
2	67	165
2	61	155
2	68	133
2	55	160
2	64	150
2	60	160

Données au format Excel

disponibles sur le site de la SFDS :
http://www.sfds.asso.fr/groupes/statvotre/SxPoiTai.xls

Séries Chronologiques accessibles sur Internet

Températures à Nottingham.
http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/books/anderson.DAT

Références

Articles et Ouvrages

BATANERO C., ESTEPA A., GODINO J.D. (1991), « *Analysis Exploratorio de Datos : Sus posibilidades en la enseñanza secundaria* », Suma, n°9, 1991, pp25-31.
disponible sur le site Web :
http://www.ugr.es/~batanero/ListadoEstadistica.htm

BATANERO C., GODINO J. D., GREEN D. R., HOLMES P., VALLECILLOS A., (1991), « *Errores y difficultades en la comprension de los conceptos estadisticos elementales* », International Journal of Mathematics Education in Science and Technology, 25(4), 527-547.
disponible sur le site Web :
http://www.ugr.es/~batanero/.htm

CLEVELAND W.S., (1993), « *Visualizing Data* », Hobart Press, Summit, New Jersey, USA.

CLEVELAND W.S., (1994), « *The Elements of Graphing » Data*, Hobart Press, Summit, New Jersey, USA.

CHAMBERS J.M., CLEVELAND W.S., KLEINER B.,TUKEY P.A., (1983) « *Graphical Methods For Data Analysis* », Wadsworth International Group, Monterey, Californie.

DEHAENE S., (1997), « *La bosse des maths* », Éditions O. Jacob.

DEHAENE S., & POSTEL-VINAY O., (2001), Entretien « *Stanilas Dehaene : Qu'est-ce qu'un nombre* », La Recherche, Octobre 2001, n°346, pp46-48.

DESTANDAU S., & LE GUEN M., (1998), « *l'Analyse Exploratoire des données avec SAS/INSIGHT* », Insee-Guide n°7-8.

DESTANDAU S., LADIRAY D., LE GUEN M., (1999), « *l'Analyse Exploratoire des données et SAS/INSIGHT* », Courrier des Statistiques, n°90, juin 1999, INSEE, pp3-44.

ERICKSON, B. H., & NOSANCHUK,T. A., (1992), « *Understanding Data : An introduction to exploratory and confirmatory data analysis for students in the Social Sciences* », Milton Keynes, Open University Press, 1977, 2e édition 1992.

FOX J. & LONG J.S., (1990), « *Modern Methods Of Data Analysis*», Sage Publications.

GARFIELD J., (1995) « *How Students Learn Statistics*», International Statistics Review, 63, 1, pp. 25-34.

GONICK L. et SMITH W., (1993), « *The Cartoon Guide to Statistics* », HarperPerennial.

LE GUEN M., (1995), « *Statistique, imagerie et sciences cognitives* », Bulletin de méthodologie sociologique, n° 49, pp. 90-100.

LE GUEN M., (1999), « *L'Analyse exploratoire des données est au cerveau droit, ce que l'analyse confirmatoire est au cerveau gauche, les deux doivent communiquer pour traiter l'information* ».
Document de Travail MATISSE - LES n°99-05.

LE GUEN M., (1999) « *De l'importance de l'image*», Courrier des Statistiques, n°90, juin 1999, INSEE, pp7-9.
disponible sur le site de la SFDS :
http://www.sfds.asso.fr/groupes/statvotre/Importance-Image.pdf

LE GUEN M., (1999), « *Enseignement de la Statistique, Voir, Apprendre, Comprendre Autrement* », Courrier des Statistiques, n°90, juin 1999, INSEE, pp37-38.
disponible sur le site de la SFDS :
http://www.sfds.asso.fr/groupes/statvotre/Voir-Apprendre-Comprendre.pdf

LE GUEN M., (2001), « *Repenser l'Initiation à la Statistique* », Revue Statistiquement Votre, n°4, 2001.
disponible sur le site de la SFDS :
http://www.sfds.asso.fr/groupes/statvotre/repenser.pdf

LE GUEN M., (2001), « *La Boîte à moustaches de TUKEY, Un outil pour initier à la Statistique*», Revue Statistiquement Votre, n°4, 2001.
disponible sur le site de la SFDS :
http://www.sfds.asso.fr/groupes/statvotre/Boite-a-moustaches.pdf

MARASINGHE M.G., MEEKER W.Q., COOK D. & SHIN T., (1996), « *Using Graphics and Simulation to Teach Statical Concepts* », The American Statistician, Vol 50, n° 4, pp. 342-351.

MOORE D.S., COBB G. W., GARFIELD J. & MEEKER W.Q., (1995), « *Statistics Education Fin de Siècle* », The American Statistician, Vol 49, n° 3, pp. 250-260, 1995.

PAPERT S. (1980), « *Jaillissement de l'esprit Ordinateurs et Apprentissage* », Flammarion.
titre original « *Mindstorms Children, Computers and Powerful ideas* ».

ROSSMAN A. J., (1995), « *Workshop Statistics, Discovery with Data* », Springer.

ROSSMAN A. J., (2001), « *Workshop Statistics, Discovery with Data and FATHOM* », Springer.

STUART M. , (1995), *« Changing the Teaching of Statistics »*, The Statistician, 44, n° 1, pp. 45-54.

TUKEY, J. W. (1977), *« Exploratory Data Analysis »*, EDA, Reading, MA, (Addison-Wesley).
Cet ouvrage serait en réécriture.

WILLIAMS L.V., (1997), « *Deux cerveaux pour apprendre, le gauche et le droit* », traduit par Trocme-Fabre H., Éditions Organisations.

Liens vers des sites Internet

Articles ou documents sur les *Box Plots* et les *Box & Whiskers Plots*

http://www.itl.nist.gov/div898/handbook/eda/section3/boxplot.htm
http://www.ruf.rice.edu/~lane/hyperstat/A37797.html
http://research.ed.asu.edu/siip/briefs/boxplots.computing.html
http://www.cmh.edu/stats/fund/boxplot.htm
http://www.math.sfu.ca/stats/Courses/Stat-301/Handouts/node32.html

Applet **d'un site français** permettant le calcul des différents éléments d'une boîte à moustaches

http://www.math-info.univ-paris5.fr/~ycart/mst99/demiguel/demiguel.html

Pour réaliser des Box Plots avec EXCEL

Site de NEVILLE HUNT
http://www.mis.coventry.ac.uk/~nhunt/boxplot.htm
Site de CHRISTINE SPENCER
http://www.elementkjournals.com/ime/0008/ime0081.htm

Logiciels d'enseignement de la Statistique orienté Analyse Exploratoire des Données.

EDA : Logiciel de E. HORBER
http://www.unige.ch/ses/sococ/eda/edasoft.html

DATADESK : Logiciel de VELLEMAN
http://www.datadesk.com/
ACTIVSTATS : environnement d'enseignement de la Statistique s'appuyant sur le logiciel Datadesk.
http://www.datadesk.com/ActivStats/

VISTA : Logiciel de FOREST W. Y.
The **Vi**sual **Sta**tistics System avec le slogan « *Vista help you see what your data seem to say* »
http://forrest.psych.unc.edu/research/index.html

MS Excel et Enseignement de la Statistique

Les Cercles d'EXCEL'ense
http://www.cisia.com/cisia/Actualite/Excelense/Excelense.htm
Spreadsheets in Education
http://sunsite.univie.ac.at/Spreadsite/ contient de nombreuses références en Excel

Projets d'environnement et d'enseignement de la Statistique **en français**

SEL Statistique en ligne de l'INRIA
http://www.inrialpes.fr/sel/
avec des réalisations de boîtes à déciles http://www.inrialpes.fr/sel/lexique/diag_boite/diag_boite.html
SMEL Simulations en ligne de l'INRIA
http://www.inrialpes.fr/sel/simulations/cadre_simulations.html
St@tNet du CNAM
http://www.cnam.agropolis.fr/statnet/

Bibliographie de JOHN W. TUKEY

http://www-groups.dcs.st-and.ac.uk/~history/Mathematicians/Tukey.html
http://stat.bell-labs.com/who/tukey/index.html

Association **MIRAGE** (**M**ouvement **I**nternational pour le Développement de la **R**echerche en **A**nalyse **G**raphique et **E**xploratoire).

http://www.unige.ch/ses/sococ/mirage/ nombreuses références sur la Visualisation et l'Exploration.
Voir la rubrique Nouvelles Juin 2001 pour une discussion sur la terminologie française des *Box Plots :* boîte à moustaches, boîte à pattes et boîte à déciles.

Association **Pénombre** propose un espace public de réflexion et d'échange sur **l'usage du nombre** dans les débats de société: justice, sociologie, médias, statistiques.

http://www.unil.ch/penombre