Chapitre 5

Statistiques descriptives

I. Vocabulaire

Population : Une population est un ensemble de personnes ou d'objets, appelés **individus**, définis par une propriété commune.
Exemple : les habitants d'un pays, les automobiles fabriquées en 2010,…

Caractère : Pour une population choisie, on peut étudier un caractère de ses individus.
Exemple : on peut étudier le caractère « taille » des élèves d'un lycée.

Caractère quantitatif : Un caractère est dit quantitatif lorsqu'il est possible de le mesurer en associant un nombre à chaque individu. Un caractère quantitatif est aussi appelé **variable**.
Exemple : l'âge, la taille, le nombre de frères et sœurs, ...

- Un caractère quantitatif est dit **continu** lorsque les nombres qui le mesurent peuvent prendre, à priori, toutes les valeurs d'un intervalle.
  Exemple : le poids, la taille, la durée de vie d'un moteur, ...

- Il est **discret** dans le cas contraire.
  Exemple : l'année de naissance, le nombre d'enfants par famille, ...

Caractère qualitatif : On appelle ainsi tout caractère non quantitatif.
Exemple : la couleur des yeux.

Exemple :
Voici les notes obtenues à un Bac blanc par 10 élèves de Première S :

12 ; 16 ; 10 ; 19 ; 5 ; 20 ; 11 ; 10 ; 15 ; 8

- La population étudiée correspond aux élèves de la classe.
- Il s'agit donc ici d'une série statistique dont le caractère étudié est quantitatif discret.

Note	5	8	10	11	12	15	16	19	20
Effectif	1	1	2	1	1	1	1	1	1

Une représentation graphique possible de la série est :

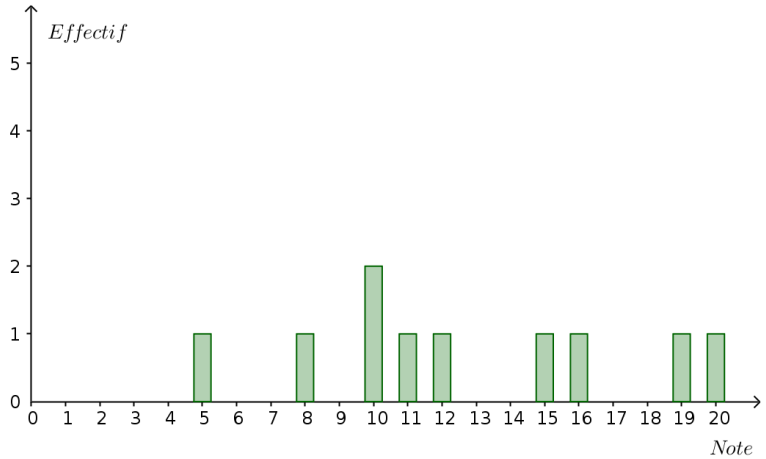


Diagramme en barres

II. Moyenne et écart type

1) Moyenne

Définition :

La **moyenne** d'une série statistique, dont les valeurs du caractère sont x_1, x_2, …, x_k et les effectifs correspondants n_1, n_2, …, n_k est notée $\overline{x}$ et vaut :

$$\overline{x}=\frac{1}{N}\sum_{i=1}^{k} n_i x_i \text{ où } N=\sum_{i=1}^{k} n_i=n_1+n_2+\ldots+n_k \text{ est l'effectif total.}$$

Exemple :

La moyenne de la série statistique est :

$$\overline{x}=\frac{5+8+2\times10+11+12+15+16+19+20}{1+1+2+1+1+1+1+1+1}=12{,}6$$

La note moyenne des élèves est donc 12,6.

Propriété :

Soit x_1, x_2, …, x_k les valeurs du caractère d'une série statistique et f_1, f_2, …, f_k les fréquences associées.

Alors la moyenne de la série statistique est :

$$\overline{x}=\sum_{i=1}^{k} f_i x_i$$

Exemple :

Dans l'exemple précédent on a :

Note	5	8	10	11	12	15	16	19	20
Fréquence	0,1	0,1	0,2	0,1	0,1	0,1	0,1	0,1	0,1

On a donc la moyenne :

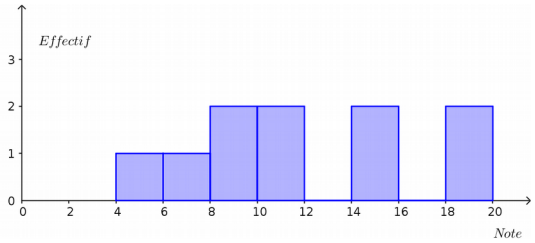
$\overline{x}=5\times0{,}1+8\times0{,}1+10\times0{,}2+11\times0{,}1+12\times0{,}1+15\times0{,}1+16\times0{,}1+19\times0{,}1+20\times0{,}1=12{,}6$

Remarque :

Lorsque le caractère est quantitatif et regroupé par classe, les x_i représentent les centres des classes.

Dans le cas de la série étudiée on peut effectuer des regroupements par classe :

Note	$4<x\leqslant6$	$6<x\leqslant8$	$8<x\leqslant10$	$10<x\leqslant12$	$12<x\leqslant14$	$14<x\leqslant16$	$16<x\leqslant18$	$18<x\leqslant20$
Effectif	1	1	2	2	0	2	0	2



Histogramme

On a donc la moyenne :

$$\frac{5+7+2\times9+2\times11+2\times15+2\times19}{1+1+2+2+2+2}=12$$

2) *Écart-type*

Définitions :

- La **variance** d'une série statistique dont les valeurs du caractère sont x_1, x_2, …, x_k, les effectifs correspondants n_1, n_2, …, n_k et la moyenne $\overline{x}$, est égale à :

$$V=\frac{1}{N}\sum_{i=1}^{k} n_i(x_i-\overline{x})^2 \text{ où } N=\sum_{i=1}^{k} n_i \text{ est l'effectif total.}$$

- L'**écart-type** d'une série statistique, noté σ, est égal à la racine carrée de la variance :

$$\sigma=\sqrt{V}$$

Remarques :

- La variance est la moyenne des carrés des « écarts à la moyenne ».
- L'intérêt de l'écart-type par rapport à la variance est son unité : c'est la même que celle du caractère étudié.
- La moyenne donne la tendance centrale (c'est un paramètre de position).
- L'écart-type est un paramètre de dispersion indiquant l'éloignement des valeurs de la série autour de la moyenne.

Exemple :

$$V=\frac{1}{N}\sum_{i=1}^{k} n_i(x_i-\overline{x})^2=\frac{1}{10}[1\times(5-12{,}6)^2+1\times(8-12{,}6)^2+2\times(10-12{,}6)^2+\ldots+1\times(20-12{,}6)^2]=20{,}84$$

$$\sigma=\sqrt{V}=\sqrt{20{,}84}\simeq 4{,}57$$

Propriété :

Avec les notations précédentes, la variance d'une série statistique est égale à :

$$V=\frac{1}{N}\sum_{i=1}^{k} n_i x_i^2-\overline{x}^2$$

Démonstration :

$$V=\frac{1}{N}\sum_{i=1}^{k} n_i(x_i-\overline{x})^2$$

$$V=\frac{1}{N}\sum_{i=1}^{k} n_i(x_i^2-2x_i\overline{x}+\overline{x}^2)$$

$$V=\frac{1}{N}\sum_{i=1}^{k} (n_i x_i^2-2n_i x_i\overline{x}+n_i\overline{x}^2)$$

$$V=\frac{1}{N}\sum_{i=1}^{k} n_i x_i^2-\frac{1}{N}\sum_{i=1}^{k} 2n_i x_i\overline{x}+\frac{1}{N}\sum_{i=1}^{k} n_i\overline{x}^2$$

$$V=\frac{1}{N}\sum_{i=1}^{k} n_i x_i^2-2\overline{x}\frac{1}{N}\sum_{i=1}^{k} n_i x_i+\overline{x}^2\frac{1}{N}\sum_{i=1}^{k} n_i$$

$$V=\frac{1}{N}\sum_{i=1}^{k} n_i x_i^2-2\overline{x}\,\overline{x}+\overline{x}^2$$

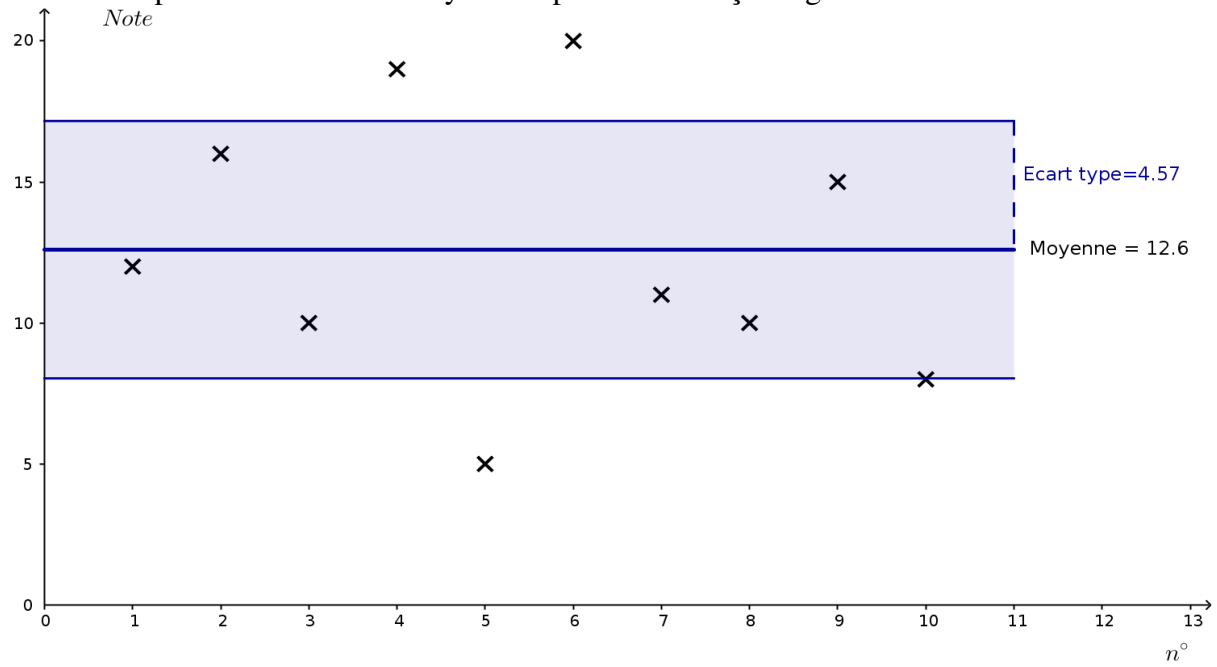
$$V=\frac{1}{N}\sum_{i=1}^{k} n_i x_i^2-\overline{x}^2$$

Exemple :

$$\frac{1}{N}\sum_{i=1}^{k} n_i x_i^2 - \overline{x^2} = \frac{1}{10}(1\times 5^2 + 1\times 8^2 + 2\times 10^2 + \ldots + 1\times 20^2) - 12{,}6^2 = 20{,}84$$

3) *Utilisation du couple moyenne-écart-type*

La moyenne et l'écart-type prennent en compte toutes les valeurs de la série et sont, de ce fait, influencées par les valeurs extrêmes.
L'écart-type mesure la dispersion des valeurs autour de la moyenne : plus il est grand, plus les valeurs sont dispersées et moins la moyenne représente de façon significative la série.



Nuage de points

III. Médiane et écart inter-quartile

1) *Médiane*

Définition :

La **médiane Me** d'une série statistique est telle que :
50% au moins des individus ont une valeur du caractère inférieure ou égale à **Me** et 50% au moins des individus ont une valeur supérieure ou égale à **Me**.

Remarques :

- Pour un caractère quantitatif discret, on peut calculer la médiane de 2 façons :
  - On ordonne les valeurs de la série par ordre croissant ; si la série est de taille $2p+1$, **Me** est la valeur du terme de rang $p+1$, si la série est de taille $2p$, **Me** est la moyenne des valeurs des termes de rang p et $p+1$.
  - La médiane est la plus petite valeur dont la fréquence cumulée croissante dépasse 0,5
- Pour un caractère quantitatif continue, la médiane est l'abscisse du point de la courbe des fréquences cumulées croissantes d'ordonnée 0,5.

Exemple :

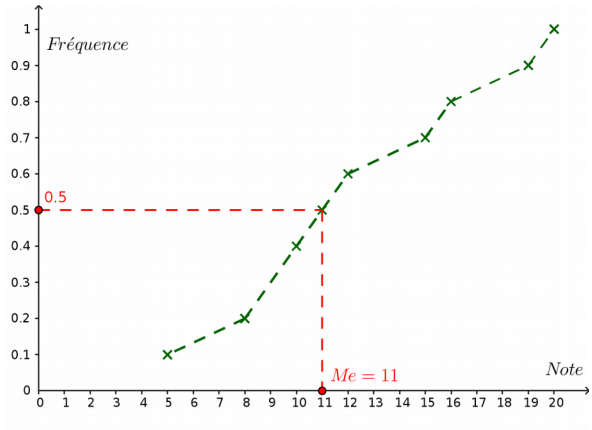
La moitié de l'effectif est $\frac{10}{2}=5$.

Les notes sont déjà rangés par ordre croissant :

5 ; 8 ; 10; 10 ; 11 ; 12 ; 15 ; 16 ; 19 ; 20

L'effectif est pair, donc la médiane est la demi-somme des 5e et 6e prix, c'est-à-dire :

$$\frac{11+12}{2}=11,5$$



Fréquences cumulées croissantes

Remarques :

- La médiane est une mesure de tendance centrale (comme la moyenne).
- Contrairement à la moyenne, la médiane n'est pas influencée par les valeurs extrêmes.

2) *Quartiles*

Définitions :

- Le **premier quartile** d'une série statistique, noté Q_1, correspond au plus petit nombre de la série tel qu'au moins 25% des données soient inférieures ou égales à ce nombre.
- Le **troisième quartile** d'une série statistique, noté Q_3, correspond au plus petit nombre de la série tel qu'au moins 75% des données soient inférieures ou égales à ce nombre.

Remarque :

Les quartiles et la médiane sont des paramètres de position ; quartiles et médiane partagent la série en 4 groupes de (presque) même taille.

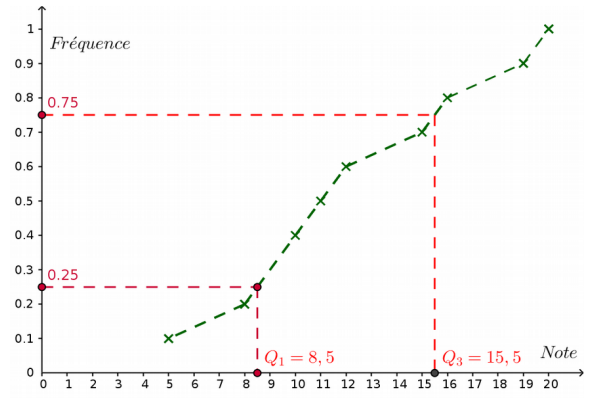
Exemple :

25% de l'effectif donne $\frac{10}{4}=2,5$.

On prend donc la 3e valeur : $Q_1=10$

75% de l'effectif donne $\frac{3}{4}\times 10=7,5$.

On prend donc la 8e valeur : $Q_3=16$



Fréquences cumulées croissantes

Définition :

$[Q_1;Q_3]$ est l'**intervalle interquartile** ; Q_3-Q_1 est l'**écart interquartile**.

Remarque :

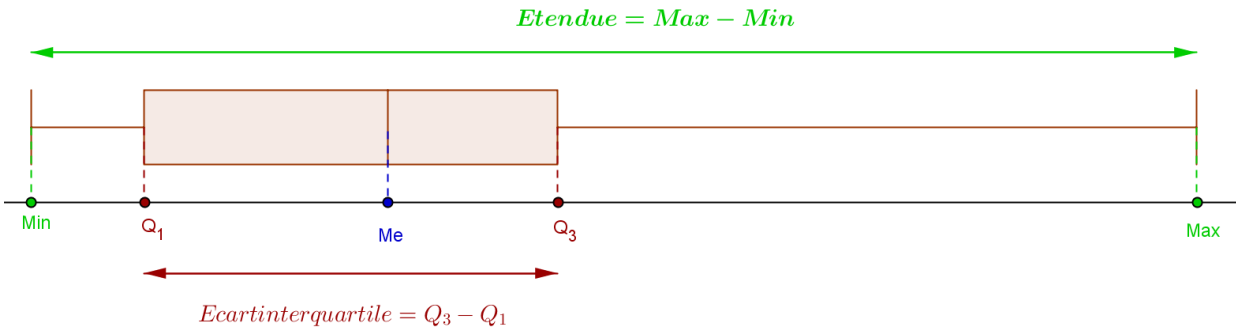
L'écart interquartile est un paramètre de dispersion.

Exemple :

$Q_3-Q_1=16-10=6$

3) Diagramme en boîte

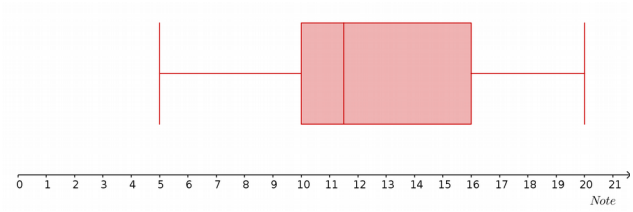
Définition :
On appelle **diagramme en boîte** d'une série statistique la représentation graphique ci-dessous. Elle se compose de deux rectangles et de deux segments dont les longueurs correspondent aux principaux paramètres de position de la série.



Remarques :

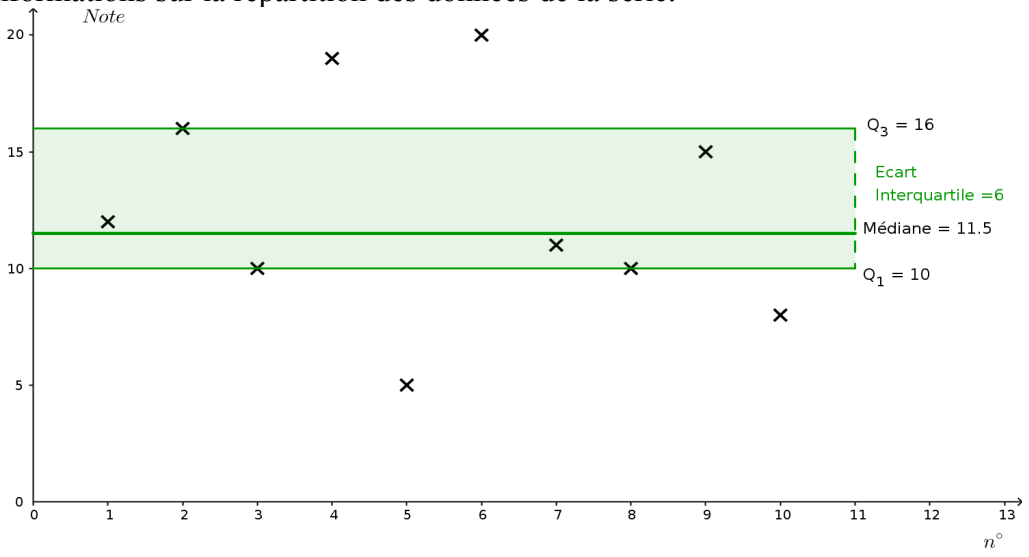
- Le diagramme en boîte est aussi appelé boîte à moustache ou encore diagramme de Tuckey.
- L'épaisseur des rectangles n'a pas de signification.

Exemple :
Le diagramme en boîte correspondant aux notes est donné ci-contre.



4) Utilisation du couple médiane-écart interquartile

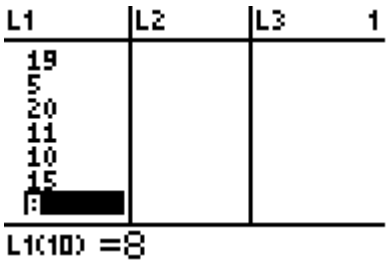
La médiane et l'écart interquartile sont déterminés par le nombre de valeurs de la série et non par la taille de ces valeurs : ils ne sont donc pas influencés par les valeurs extrêmes, qui peuvent parfois être trompeuses.
Le diagramme en boîte, qui met en évidence les informations fournies par ces deux paramètres, est très utile pour comparer un même caractère dans plusieurs séries de tailles différentes. Il donne aussi des informations sur la répartition des données de la série.

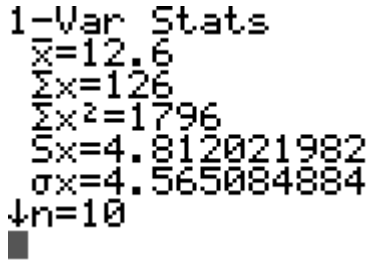


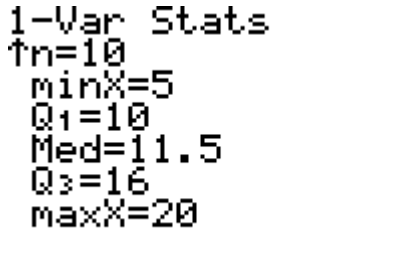
Nuage de points

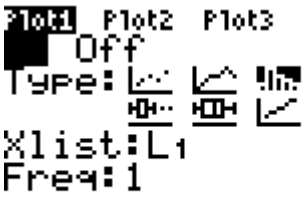
Utilisation de la calculatrice :

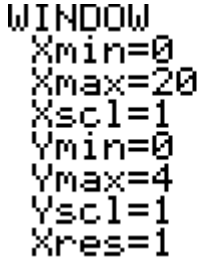
TI

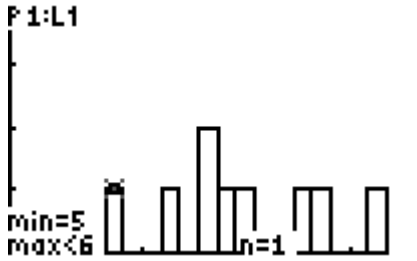


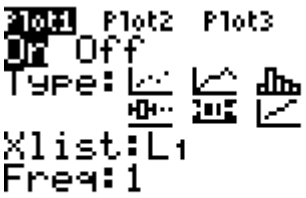


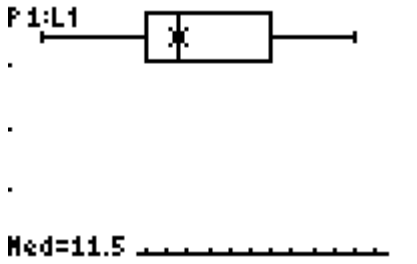




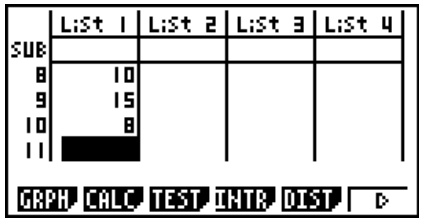


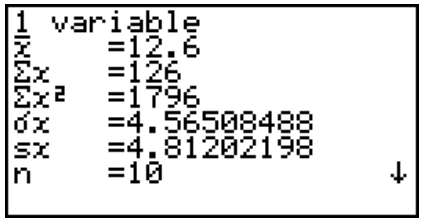


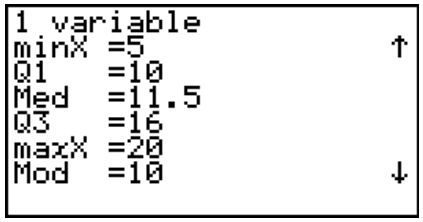


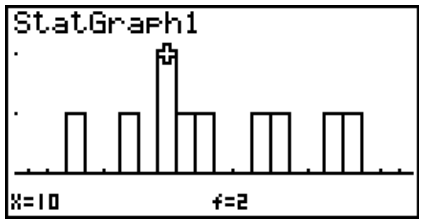


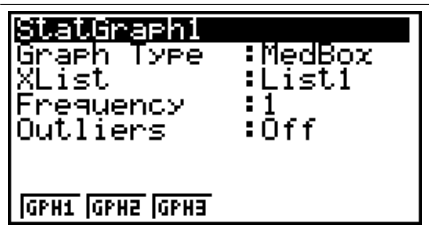
Casio

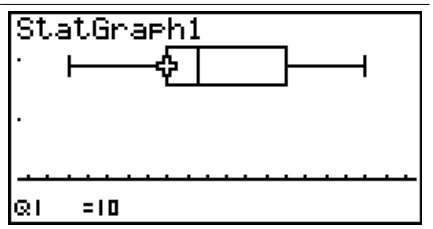






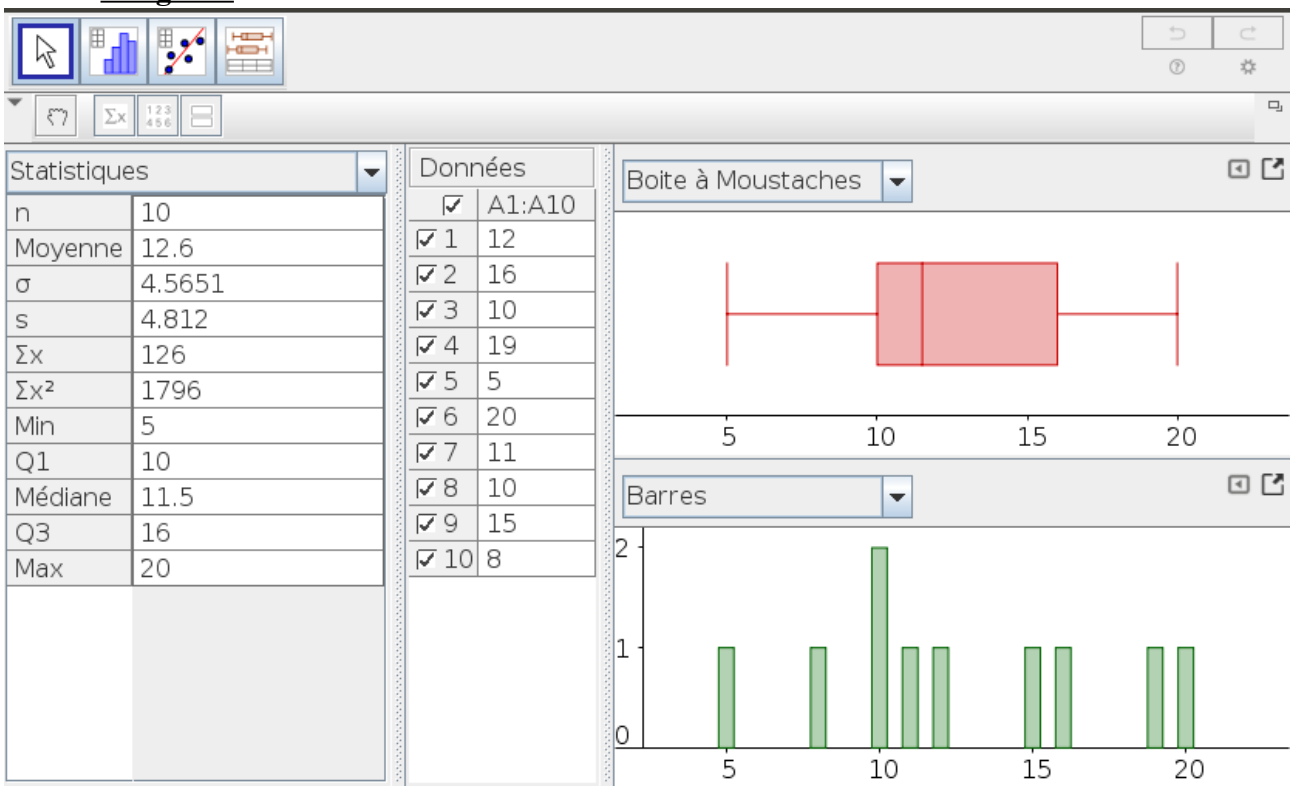






Remarque :
Certains logiciels permettent également d'étudier les séries statistiques.

- **Geogebra**



- **R**