Statistiques, cours, première S

F.Gaudon

19 novembre 2009

Table des matières

1 Quantiles d'une série statistique

Définition :

- Le *premier quartile* noté Q_1 de la série statistique est la plus petite valeur telle qu'au moins 25% des valeurs lui soient inférieures ou égales ;
- Le *troisième quartile* noté Q_3 de la série est la plus petite valeur telle qu'au moins 75% des valeurs lui soient inférieures ou égales.

Détermination pratique :

On suppose la série ordonnée dans l'ordre croissant des valeurs du caractère. Soit N l'effectif total.

- Si $\frac{N}{4}$ est une entier alors Q_1 est la valeur de rang $\frac{N}{4}$ et Q_3 est la valeur de rang $\frac{3N}{4}$;
- si $\frac{N}{4}$ n'est pas entier, alors Q_1 est la valeur dont le rang suit $\frac{N}{4}$ et Q_3 la valeur dont le rang suit $\frac{3N}{4}$.

Remarque :

Il existe d'autres définitions des quartiles. En particulier, on peut définir le premier quartile comme la médiane de la sous série formée des valeurs inférieures ou égales à la médiane et le troisième quartile comme la médiane de la sous série formée des valeurs supérieures ou égales à la médiane.

Définition :

Le n^e *décile* est la plus petite valeur telle que $n \times 10$ % des effectifs au moins lui soient inférieurs ou égales.

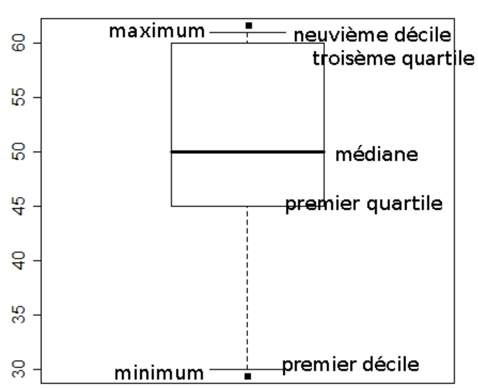
Définition :

Un *diagramme* en boîte représente les caractéristiques statistiques suivantes :

- minimum et maximum ;
- médiane ;
- le premier et le troisième quartile ;
- le premier et le neuvième décile.

Remarque :

On fait parfois figurer le minimum et le maximum de la série au lieu des premier et neuvième déciles.



2 Caractéristiques de dispersion

On considère une série statistique $(n_i; x_i)_{i=1}^{p}$ d'effectif total $N = \sum_{i=1}^{p} n_i$.

Définition :

La *variance* de la série est le nombre noté V défini par :

$$V = \frac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \ldots + n_p(x_p - \bar{x})^2}{N}$$

Propriété :

$$V = \frac{n_1 x_1^2 + n_2 x_2^2 + \ldots + n_p x_p^2}{N} - \bar{x}^2$$

Preuve :

$$V = \frac{n_1(x_1 - \bar{x})^2 + \ldots + n_p(x_p - \bar{x})^2}{N}$$
$$V = \frac{1}{N}(n_1(x_1^2 - 2 \times x_1 \times \bar{x} + \bar{x}^2) + \ldots + n_p(x_p^2 - 2 \times x_p \times \bar{x} + \bar{x}^2))$$
$$V = \frac{1}{N}(n_1 x_1^2 + \ldots + n_p x_p^2 - 2 \times \bar{x}(x_1 + \ldots + x_p) + \bar{x}^2 + \ldots + \bar{x}^2)$$
$$V = \frac{n_1 x_1^2 + \ldots + n_p x_p^2}{N} - 2 \times \bar{x}^2 \times \frac{x_1 + \ldots + x_p}{N} + \frac{\bar{x}^2 + \ldots + \bar{x}^2}{N}$$
$$V = \frac{n_1 x_1^2 + \ldots + n_p x_p^2}{N} - 2 \times \bar{x}^2 + \frac{N \times \bar{x}^2}{N}$$
$$V = \frac{n_1 x_1^2 + \ldots + n_p x_p^2}{N} - 2 \times \bar{x}^2 + \bar{x}^2$$
$$V = \frac{n_1 x_1^2 + n_2 x_2^2 + \ldots + n_p x_p^2}{N} - \bar{x}^2$$

Propriété :

Pour tout réel x, on a :

$$\sum_{i=1}^{p} n_i(x_i - x)^2 \geq \sum_{i=1}^{p} n_i(x_i - \bar{x})^2$$

La variance minimise donc la distance des valeurs du caractère à la moyenne, c'est donc un indicateur approprié pour mesurer la dispersion des valeurs autour de la moyenne.

Preuve :
Soit f définie par $f(x) = \sum_{i=1}^{p} n_i(x_i - x)^2$ pour tout réel x. f est dérivable sur $\mathbb{R}$.

On a $f'(x) = -\sum_{i=1}^{p} 2n_i(x_i - x) = -2\sum_{i=1}^{p} n_i x_i + 2x\sum_{i=1}^{p} x_i = -2\sum_{i=1}^{p} n_i x_i + 2Nx$.

Donc $f'(x) \leq 0 \Leftrightarrow -\sum_{i=1}^{p} n_i x_i + aN \leq 0 \Leftrightarrow x \leq \frac{1}{N}\sum_{i=1}^{p} n_i x_i \Leftrightarrow x \leq \bar{x}$. Donc f est décroissante sur $]-\infty; \bar{x}]$ et croissante sur $[\bar{x}; +\infty[$ donc minimale pou $x = \bar{x}$.

Définition :

L'écart-type de la série statistique est le nombre noté σ défini par :

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

Définition :

On appelle *intervalle inter quartile* l'intervalle $[Q_1\,;Q_2]$ et *écart inter quartile* le nombre $Q_3 - Q_1$.

3 Transformation affine de valeurs

Propriété :

Soient a et b deux réels et x_i pour i allant de 1 à p le valeurs non nécessairement distinctes d'une série statistique.
On note $\bar{x}$ sa moyenne, s son écart type, V sa variance, M_e, Q_1 et Q_3 sa médiane et ses quartiles.
On définit alors une nouvelle série en posant $X_i = ax_i + b$.
La moyenne de la nouvelle série est alors $a\bar{x} + b$, sa variance a^2V, son écart type $|a|\,s$ et, si $a > 0$, son premier quartile est $aQ_1 + b$, sa médiane $aM_e + b$, son troisième quartile $aQ_2 + b$ et son écart inter quartile est $a(Q_3 - Q_1)$.

Preuve :

$$\begin{aligned}
\bar{X} &= \frac{(ax_1 + b) + (ax_2 + b) + \ldots + (ax_p + b)}{N} \\
&= \frac{a(x_1 + x_2 + \ldots + x_p) + Nb}{N} \\
&= a\frac{x_1 + x_2 + \ldots + x_p}{N} + \frac{Nb}{N} \\
&= a\bar{x} + b
\end{aligned}$$

D'autre part,

$$\begin{aligned}
\frac{\sum_{i=1}^{p}(X_i - \bar{X})^2}{N} &= \frac{\sum_{i=1}^{p}(ax_i + b - a\bar{x} + b)^2}{N} \\
&= \frac{\sum_{i=1}^{p}(a(x_i - \bar{x}))^2}{N} \\
&= a^2\frac{\sum_{i=1}^{p}(x_i - \bar{x})^2}{N} \\
&= a^2V
\end{aligned}$$

D'où l'écart type de la série des X_i est donnée par $\sqrt{a^2V} = |a|\sqrt{V} = |a|\, s$.

Les propriétés de la médiane et des quartiles découlent directement de la définition et de la conservation de l'ordre puisque $a > 0$.