Les statistiques descriptives

Première S

L'équipe des professeurs de mathématiques
Lycée Stendhal

J'aimais et j'aime encore les mathématiques pour elles-mêmes comme n'admettant pas l'hypocrisie et le vague, mes deux bêtes d'aversion.

Stendhal

Année 2011-2012

Liste des savoirs et savoir-faire du chapitre :

CODE	INTITULE	Bilan		
		A	EA	NA
S0101	Déterminer moyenne, médiane et quartiles d'une série statistique			
S0102	Déterminer variance et écart-type d'une série statistique			
S0103	Dresser le diagramme en boîte d'une série statistique			
S0104	Décrire le diagramme en boîte d'une série statistique			
S0105	Comparer deux séries statistiques avec les outils appropriés			

Statistiques Descriptives

(En première S)

Dernière mise à jour : Dimanche 31 Décembre 2011

Vincent OBATON, Enseignant au lycée Stendhal de Grenoble (Année 2011-2012)

J'aimais et j'aime encore les mathématiques pour elles-mêmes comme n'admettant pas l'hypocrisie et le vague, mes deux bêtes d'aversion.

Stendhal

Table des matières

1 Notation

La somme de n nombres numérotés de 1 à n peut s'écrire :

$$x_1 + x_2 + x_3 + x_4 + \cdots + x_{n-1} + x_n$$

mais cette écriture est longue et les pointillés ne sont pas satisfaisants.
On écrira, pour faire moins long et éviter les pointillés, cette somme à l'aide du symbole Sigma :

$$\sum_{i=1}^{n} x_i$$

Exemples :

1. $\displaystyle\sum_{i=0}^{n} x_i = x_0 + x_1 + x_2 + \cdots + x_{n-1} + x_n$
2. $\displaystyle\sum_{i=1}^{n-1} i = 1 + 2 + 3 + 4 + \cdots + (n-2) + (n-1)$
3. $\displaystyle\sum_{i=0}^{n} i^2 = 0^2 + 1^2 + 2^2 + 3^2 + \cdots + (n-1)^2 + n^2$

2 Quelques généralités et rappels

2.1 Tri à plat

On note $(x_i; n_i)_{i\in\mathbb{N}}$ la série statistique ci-dessous :
Rappels : L'effectif n_i est le nombre de fois où apparaît la valeur x_i dans la série.

valeurs x_i	x_1	x_2	x_3	x_4	...	...	...	x_{k-2}	x_{k-1}	x_k
Effectifs n_i	n_1	n_2	n_3	n_4	...	...	...	n_{k-2}	n_{k-1}	n_k

2.2 Effectif total d'une série

Définition :

L'effectif total N de la série statistique est la somme de tous les effectifs ou le nombre de valeurs total dans cette série :

$$N = \sum_{i=1}^{k} n_i = n_1 + n_2 + n_3 + \ldots + n_{k-1} + n_k$$

2.3 Fréquence d'apparition d'une valeur

Définition :

La fréquence d'apparition d'une valeur x_i est la proportion de cette valeur par rapport à l'effectif total.

Fréquence par rapport à 1 : $\boxed{f_i = \dfrac{\text{Effectif de la valeur}}{\text{Effectif total}} = \dfrac{n_i}{N}}$

Fréquence par rapport à 100 : $\boxed{F_i = \dfrac{100 n_i}{N}}$

Propriétés :

$$S_f = \sum_{i=1}^{k} f_i = f_1 + f_2 + f_3 + \ldots + f_{k-1} + f_k = 1$$

$$S_F = \sum_{i=1}^{k} F_i = F_1 + F_2 + F_3 + \ldots + F_{k-1} + F_k = 100$$

2.4 Tableau Standard statistique

Le tri à plat d'une série statistique est un tableau contenant les valeurs de la série, les effectifs, les effectifs cumulés croissants, les fréquences, les fréquences cumulées croissantes, les pourcentages et les pourcentages cumulés croissants.

valeurs x_i	x_1	x_2	x_3	$\ldots$	$\ldots$	$\ldots$	x_{k-1}	x_k
Effectifs n_i	n_1	n_2	n_3	$\ldots$	$\ldots$	$\ldots$	n_{k-1}	n_k
Effectifs Cum Croi N_i	n_1	$N_1 + n_2$	$N_2 + n_3$	$\ldots$	$\ldots$	$\ldots$	$N_{k-2} + n_{k-1}$	N
Fréquences f_i	f_1	f_2	f_3	$\ldots$	$\ldots$	$\ldots$	f_{k-1}	f_k
Fréquences Cum Croi F_i	f_1	$F_1 + f_2$	$F_2 + f_3$	$\ldots$	$\ldots$	$\ldots$	$F_{k-2} + f_{k-1}$	1
Pourcentages p_i	p_1	p_2	p_3	$\ldots$	$\ldots$	$\ldots$	p_{k-1}	p_k
Pourcentages Cum Croi P_i	p_1	$P_1 + p_2$	$P_2 + p_3$	$\ldots$	$\ldots$	$\ldots$	$P_{k-2} + p_{k-1}$	100

3 Critère de position

3.1 Médiane

Définition :

La médiane d'une série statistique est la valeur qui partage cette série en deux séries de même effectif.

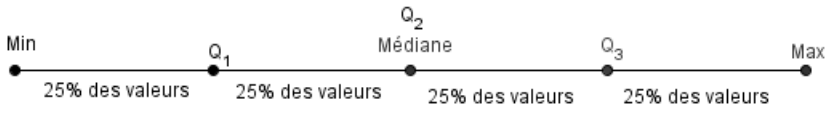
Si M_e est la médiane de la série statistique, alors :

$$50\ \%\text{ des valeurs de la série sont inférieures ou égales à } M_e$$
$$50\ \%\text{ des valeurs de la série sont supérieures ou égales à } M_e$$

Méthode pour trouver la médiane :
Il faut commencer par classer la série dans l'orde croissant.
▷ Si $\frac{N}{2} + 0,5 = d$ est entier alors la médiane est la d ième valeur de la série.
▷ Si $\frac{N}{2} + 0,5 = d,5$ est décimale alors la médiane est entre la d ième et la $d+1$ ième valeur de la série.

3.2 Quartiles

Définition :
Les quartiles d'une série statistique sont les valeurs qui partagent cette série en quatre séries de même effectif.



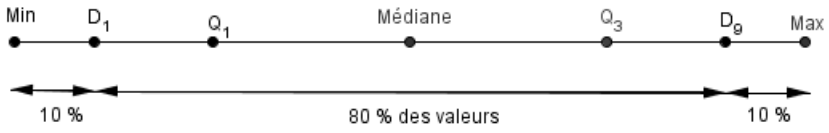
Si Q_1 est le premier quartile et Q_3 le troisième de la série statistique, alors :

$$25\ \%\text{ des valeurs de la série sont dans } [Min, Q_1]$$
$$50\ \%\text{ des valeurs de la série sont dans } [Q_1, Q_3]$$
$$25\ \%\text{ des valeurs de la série sont dans } [Q_3, Max]$$

Méthode pour trouver les quartiles :
Il faut commencer par classer la série dans l'orde croissant.
On utilisera une méthode approximative mais qui donnera des résultats significatifs pour des séries à grands effectifs. (Autement il suffit de couper en deux les deux séries $[Min, M_e]$ et $[M_e, Max]$)
Calculer $\frac{N}{4}$ et on note a l'entier supérieur à $\frac{N}{4}$.
Calculer $\frac{3N}{4}$ et on note b l'entier supérieur à $\frac{3N}{4}$.
▷ Q_1 est la a ième valeur de la série statistique.
▷ Q_3 est la b ième valeur de la série statistique.

3.3 Déciles

Définition :
Les déciles d'une série statistique sont les valeurs qui partagent cette série en dix séries de même effectif.



Si D_1 est le premier décile et D_9 le neuvième de la série statistique, alors :

10 % des valeurs de la série sont dans $[Min, D_1]$
80 % des valeurs de la série sont dans $[D_1, D_9]$
10 % des valeurs de la série sont dans $[D_9, Max]$

Méthode pour trouver les déciles :
Il faut commencer par classer la série dans l'orde croissant.
On utilisera une méthode approximative mais qui donnera des résultats significatifs pour des séries à grands effectifs.
Calculer $\frac{N}{10}$ et on note a l'entier supérieur à $\frac{N}{10}$.
Calculer $\frac{9N}{10}$ et on note b l'entier supérieur à $\frac{9N}{10}$.
▷ D_1 est la a ième valeur de la série statistique.
▷ D_9 est la b ième valeur de la série statistique.

3.4 Moyenne

Définition :

La moyenne arithmétique de la série statistique est le nombre :

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{k} x_i \times ni = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3 + \ldots + x_{k-1} n_{k-1} + x_k n_k}{N}$$

ou

$$\overline{x} = \sum_{i=1}^{k} x_i \times f_i = x_1 f_1 + x_2 f_2 + x_3 f_3 + \ldots + x_k f_k$$

Propriétés de la moyenne :

1. Si $\overline{x}$ est la moyenne d'un groupe d'effectif N_1 et $\overline{y}$ la moyenne d'un groupe d'effectif N_2 alors la moyenne $\overline{z}$ de la série constituée de l'ensemble des deux groupes est :

$$\overline{z} = \frac{N_1\overline{x} + N_2\overline{y}}{N_1 + N_2}$$

2. Si $\overline{x}$ est la moyenne d'une série (x_i, n_i) alors la moyenne de la série $(ax_i + b, n_i)$ est :

$$\overline{y} = a\overline{x} + b$$

3. **Moyenne élaguée** :
Quand une valeur aberrante, correspondant à une erreur de mesure ou à une situation exceptionnelle, est présente dans une série, elle influence considérablement la valeur moyenne. Une moyenne calculée après avoir enlevé certaines valeurs est appelée **Moyenne elaguée**.

3.5 Modes

Définition :

Les modes d'une série sont les valeurs ayant le plus grand effectif.

4 Critère de dispersion

4.1 Etendue

Définition :

L'étendue d'une série statistique est la différence entre la plus grande valeur et la plus petite, de la série.

$$\boxed{Et = Max - Min}$$

4.2 Ecart Inter-Quartiles

Définition :

L'écart inter-quartiles est la différence entre Q_3 et Q_1

$$\boxed{E_Q = Q_3 - Q_1}$$

L'intervalle inter-quartiles est l'intervalle entre Q_1 et Q_3

$$\boxed{I_Q = [Q_1, Q_3]}$$

4.3 Ecart Inter-Déciles

Définition :

L'écart inter-déciles est la différence entre D_9 et D_1

$$\boxed{E_D = D_9 - D_1}$$

L'intervalle inter-déciles est l'intervalle entre D_1 et D_9

$$\boxed{I_D = [D_1, D_9]}$$

4.4 Variance

Certaines séries statistiques peuvent avoir les mêmes critères de position comme la médiane et la moyenne.
Pour les différencier on va utiliser un nouvel outil qui va mesurer la dispersion de la série autour de la moyenne. On souhaite trouver une mesure de l'écart entre les valeurs de la série et sa moyenne. Si cet écart est grand alors la série est très hétérogène et les valeurs sont éloignées de la moyenne sinon si cet écart est petit la série est homogène et les valeurs rapprochées autour de la moyenne.
On pourrait calculer la moyenne des écarts à la moyenne mais celle-ci donne toujours 0 à cause des écarts qui sont opposés.
Démonstration :

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

$$= \overline{x} \times \frac{1}{N}\sum_{i=1}^{k} ni - \overline{x} = \overline{x} \times \frac{N}{N} - \overline{x} = \overline{x} - \overline{x} = 0$$

Pour éviter ce problème, on va faire la moyenne des carrées des écarts à la moyenne.
On note ce résultat, la **variance** de la série.
Définition :

La variance d'une série statistique est la moyenne des carrées des écarts à la moyenne de chacune des valeurs.

valeurs x_i	x_1	x_2	x_3	$\dots$	x_{k-1}	x_k
Effectifs n_i	n_1	n_2	n_3	$\dots$	n_{k-1}	n_k
$(\overline{x}-x_i)^2$	$(\overline{x}-x_1)^2$	$(\overline{x}-x_2)^2$	$(\overline{x}-x_3)^2$	$\dots$	$(\overline{x}-x_{k-1})^2$	$(\overline{x}-x_k)^2$

La variance de la série est donc la moyenne de la dernière ligne du tableau ci-dessus :

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

ou

$$V(x) = \sum_{i=1}^{k}(x-x_i)^2 f_i$$

Propriété :

$$V(ax) = a^2 V(x)$$

$$V(x+b) = V(x)$$

donc

$$V(ax+b) = a^2 V(x)$$

4.5 Ecart-Type

Définition :

L'écart-type σ est la racine carrée de la variance pour revenir aux même unités que les valeurs de la série statistique.

$$\boxed{\sigma = \sqrt{V(x)}}$$

Propriété :
Si on a deux séries S_1 et S_2 d'écart-type respectifs σ_1 et σ_2
Si $\sigma_1 < \sigma_2$ alors la série S_1 est plus homgène que la série S_2 ou la série S_2 est plus hétérogène que la série S_1.
Propriété :

$$\sigma(ax) = |a|\sigma(x)$$
$$\sigma(x+b) = \sigma(x)$$

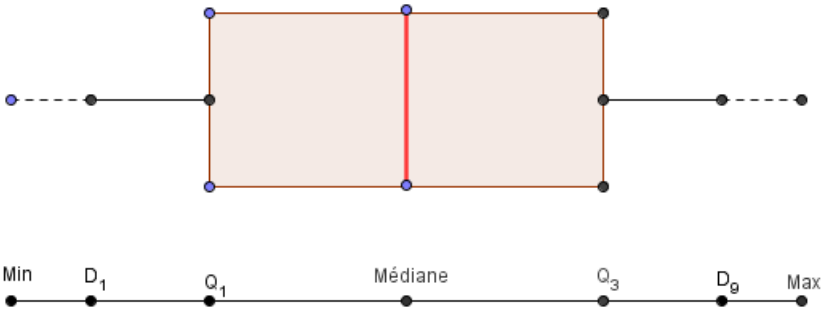
donc

$$\sigma(ax+b) = |a|\sigma(x)$$

5 Représentation statistiques

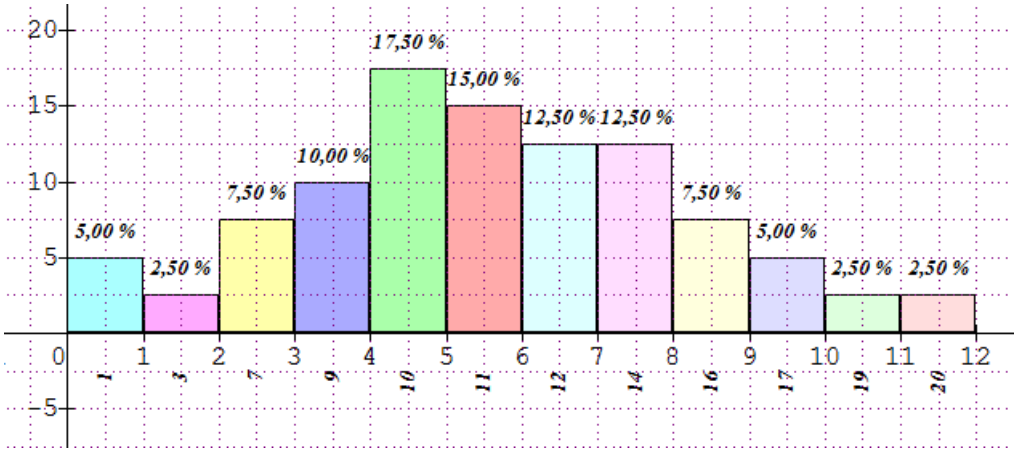
5.1 Diagramme en boîte

Les diagrammes en boîte, ou boîtes à moustaches, sont des diagrammes permettant de comparer rapidement des séries statistiques.



5.2 Autres diagrammes

Diagramme en barres (Histogrammes) :



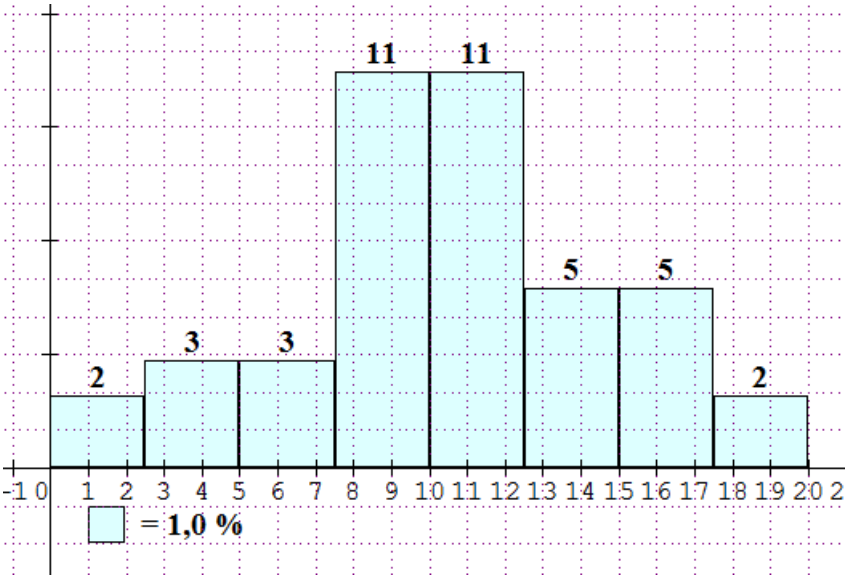


Diagramme en bâtons :

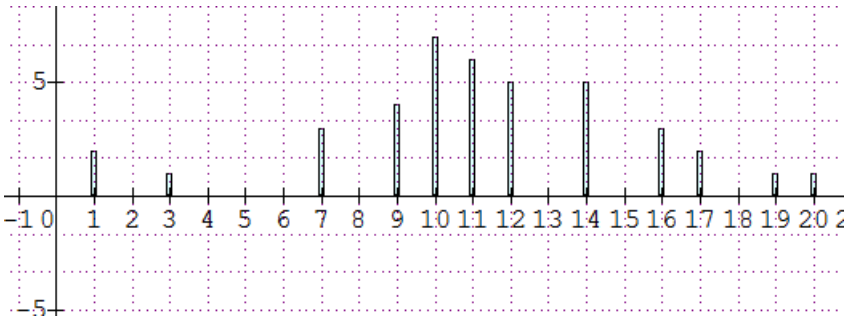
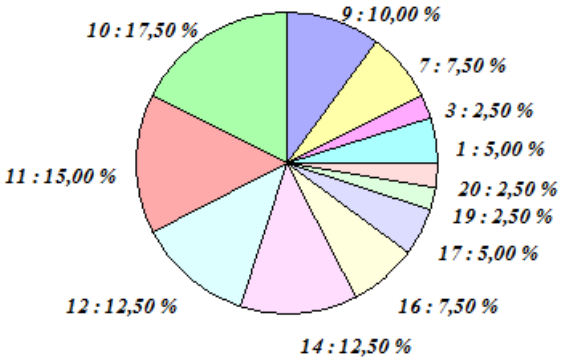
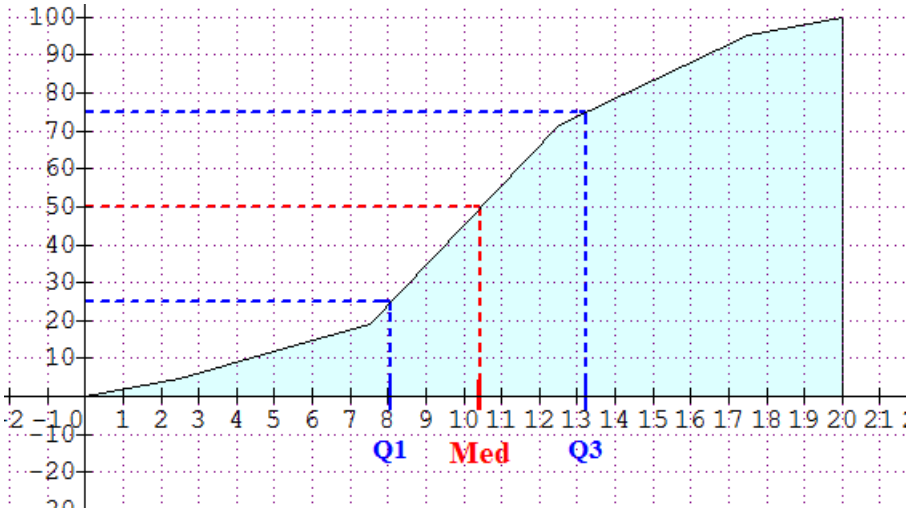


Diagramme en camembert :



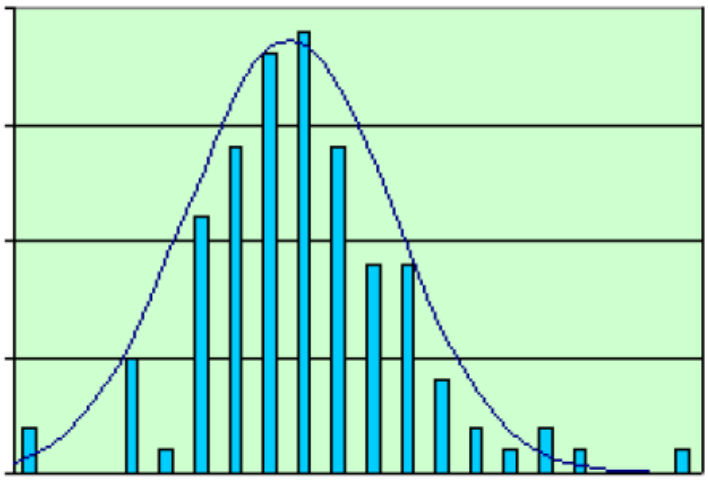
5.3 Polygône des fréquences cumulées

Le polygône des fréquences cumulées (en pourcentages) permet de lire rapidement la médiane et les quartiles d'une séries statistique :



6 Plages de normalité des distributions normales Gaussiennes

Lorsque l'on fait des statistiques sur une grande quantitié de valeurs, il arrivent souvent que l'on obtienne des diagramme ayant sensiblement la même forme dite EN CLOCHE ou COURBE DE GAUSS, comme ci-dessous :



Lorsque la série statistique donne une représentation graphique de la forme d'une courbe de Gauss, les données sont qualifiées de données Gaussiennes.

Proriétés (Plages de normalité) :

On note $\overline{x}$ la moyenne de la série et σ l'écart-type de la série.

1. Environ 68 % des données se trouvent dans l'intervalle $[\overline{x}-\sigma, \overline{x}+\sigma]$
   On nomme cet intervalle la plage de normalité pour le niveau de confiance 0.68
2. Environ 95 % des données se trouvent dans l'intervalle $[\overline{x}-2\sigma, \overline{x}+2\sigma]$
   On nomme cet intervalle la plage de normalité pour le niveau de confiance 0.95
3. Environ 99 % des données se trouvent dans l'intervalle $[\overline{x}-3\sigma, \overline{x}+3\sigma]$
   On nomme cet intervalle la plage de normalité pour le niveau de confiance 0.99