Statistiques

Table des matières

I Les indicateurs statistiques

I.1 Effectif

Définition

L'effectif de la valeur d'un caractère est le nombre d'individus ayant cette valeur de caractère.

I.2 Fréquence

Définition

La fréquence f d'une valeur d'un caractère est la proportion d'individus ayant cette valeur de caractère :
$f = \frac{n}{N}$, où n est l'effectif de la valeur du caractère et N l'effectif total.

I.3 Étendue

Définition

L'étendue d'une série statistique est la différence entre les valeurs extrêmes du caractère.

I.4 Mode

Définition

Le mode d'une série statistique est la valeur du caractère ayant l'effectif le plus grand.

I.5 Moyenne (pondérée)

Définition

Soit une série statistique dont les valeurs du caractère sont $x_1, x_2, \cdots, x_k$ et $n_1, n_2, \cdots, c_k$ effectifs associés.
La moyenne de la série statistique, notée $\bar{x}$, a pour valeur :

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_k x_k}{n_1 + n_2 + \cdots + n_k}$$

Conséquence : Lorsqu'on présente la série statistique en ne donnant que la liste des valeurs, alors la moyenne est

$$\frac{x_1 + x_2 + \cdots + x_k}{k}$$

Propriété

Si on appelle f_i la fréquence de la valeur x_i , alors :

$$\bar{x} = f_1 x_1 + f_2 x_2 + \ldots + f_k x_k.$$

La linéarité de la moyenne :

Propriété

Soit k un nombre réel. Soit $x_1, x_2, \cdots, x_n$ les valeurs du caractère d'une série statistique et $\bar{x}$ leur moyenne. Alors :
- la moyenne de la série $kx_1, kx_2, \cdots, kx_n$ est $k\bar{x}$;
- la moyenne de la série $x_1 + k, x_2 + k, \cdots, x_i + k, \cdots, x_n + k$ est $\bar{x} + k$.

Exemples :

- Si la moyenne au contrôle de biologie dans une classe est de 8 sur 20 et que le professeur décide d'augmenter toutes les notes de 10%, alors la nouvelle moyenne est de 8,8.
- Si la moyenne au contrôle d'histoire-géographie est de 8,7 sur 20 et que le professeur décide d'ajouter 1 point à tous les élèves, alors la nouvelle moyenne est de 9,7.

I.6 Médiane

Définition

La médiane d'une série statistique est le nombre tel que :
50 % au moins des individus ont une valeur du caractère inférieure ou égale à ce nombre et 50 % au moins des individus ont une valeur supérieure ou égale à ce nombre.

Médiane d'un caractère quantitatif discret

On considère une série statistique dont les valeurs du caractère sont rangées par ordre croissant, chacune de ces valeurs figurant un nombre de fois égal à son effectif.
- Si le nombre de données est impair, donc de la forme $2n+1$, la médiane est le terme du milieu, c'est-à-dire le rang de terme $n+1$.
- Si le nombre de données est pair, donc de la forme $2n$, la médiane est la demi-somme des termes de rangs n et $n+1$.

I.7 Quartiles

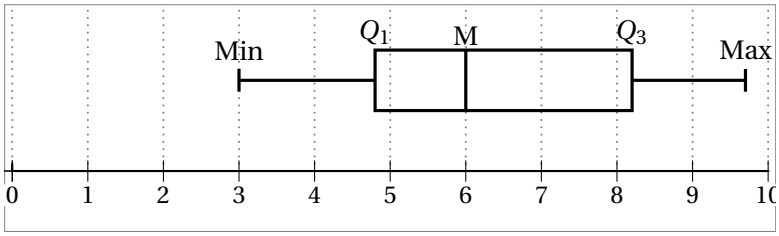
Définition

Le premier quartile d'une série statistique, noté Q_1 est la première valeur de la série, rangée par ordre croissant, tel que 25 % des valeurs de la série soient inférieures ou égales à Q_1.
Le troisième quartile d'une série statistique, noté Q_3 est la première valeur de la série, rangée par ordre croissant, tel que 75 % des valeurs de la série soient inférieures ou égales à Q_3.
Remarque : Q_1 est la valeur x_i de la série dont l'indice i est le premier entier supérieur ou égal à $\frac{n}{4}$ (si n est l'effectif de la série).
Q_3 est la valeur x_i de la série dont l'indice i est le premier entier supérieur ou égal à $\frac{3n}{4}$ (si n est l'effectif de la série).

II Diagramme en boîte (ou diagramme de Tukey ou boîte à moustaches)

Les deux quartiles Q_1, Q_3, la médiane M d'une série statistique, associés aux valeurs extrêmes (minimum et maximum) permettent d'appréhender certaines caractéristiques de la répartition des valeurs.

Exemple :
Voici la série des températures (en degré Celcius) relevées sous abri à différents moments de la journée. Elles sont classées par ordre croissant.
3 ; 3,8 ; 4,5 ; 4,8 ; 5 ; 5,5 ; 5,7 ; 5,8 ; 6,2 ; 7 ; 7,3 ; 8,2 ; 9 ; 9,2 ; 9,5 ; 9,7
Les valeurs extrêmes sont 3 et 9,7.
La médiane vaut 6 (moyenne entre 5,8 et 6,2).
Le premier quartile est $Q_1 = 4,8$; le troisième quartile est $Q_3 = 8,2$. Le diagramme en boîte est alors :



Les diagrammes en boîte servent à faire des comparaisons de deux séries statistiques.

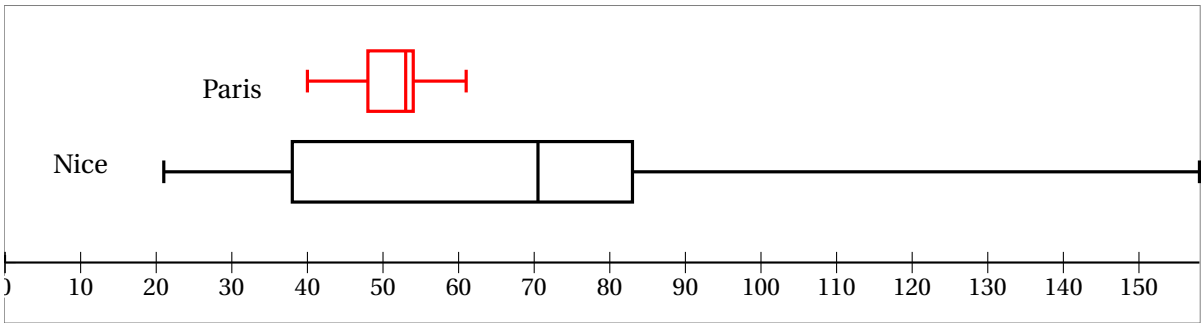
Exemple :
Les séries suivantes donnent les précipitations moyennes mensuelles en millimètres à Nice et à Paris :

	J	F	M	A	M	J	J	A	S	O	N	D
Nice	67	83	71	70	39	37	21	38	83	109	158	92
Paris	53	48	40	45	53	57	54	61	54	50	58	51

Pour effectuer la comparaison, on va ranger chaque série par ordre croissant :
Nice : 21 ; 37 ; 38 ; 39 ; 67 ; 70 ; 71 ; 83 ; 83 ; 92 ; 109 ; 158
Paris 40 ; 45 ; 48 ; 50 ; 51 ; 53 ; 53 ; 54 ; 54 ; 57 ; 58 ; 61

Pour Nice, on a : Min =21 ; Max = 158 ; $Q_1 = 38$; $M_1 = 70,5$ et $Q_3 = 83$
Pour Paris, on a : Min =40 ; Max = 61 ; $Q_1 = 48$; $M_1 = 53$ et $Q_3 = 54$

Diagrammes en boîtes :



Les précipitations sont plus régulières tout au long de l'année à Paris (série moins dispersée). LA totalité des valeurs de la série des précipitations à Paris est comprise entre le premier quartile et la médiane de la série des précipitations à Nice. Pour la ville de Nice, plus de la moitié des mois ont des précipitations supérieures au maximum de Paris.

III Variance ; Écart type

III.1 Variance

Considérons deux groupes d'élèves, l'un de dix élèves et l'autre de huit élèves ; leurs notes de mathématiques à un contrôle sont :

Première série :

note x_i	1	2	3	17	20
effectif n_i	3	1	1	1	4

Deuxième série :

note x_i	8	10	11	12
effectif n_i	1	2	4	1

La moyenne de la première série est : $\dfrac{n_1x_1+\cdots+n_5x_5}{n_1+\cdots+n_5}=\dfrac{105}{10}=10,5$.

La moyenne de la deuxième série est : $\dfrac{84}{8}=10,5$.

Les deux moyennes sont égales ; pourtant, la répartition des notes n'est pas du tout la même.

Il faut donc trouver un moyen de mesurer la dispersion des nombres autour de la moyenne.

Un premier moyen est l'étendue, mais ce n'est pas très fiable.

Nous allons voir un deuxième moyen, qui est l'écart type.

Définition

Soit une série statistique donnée par le tableau :

Valeur du caractère	x_1	x_2	$\cdots$	x_p	Total
Effectif	n_1	n_2	$\cdots$	n_p	N

La moyenne de cette série est : $\overline{x}=\dfrac{n_1x_1+n_2x_2+\cdots+n_Px_p}{N}$.

La **variance** est le nombre V défini par :

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

V est donc la **moyenne des carrés des écarts entre chaque valeur x_i et la moyenne**.

Autre formulation de la variance :

Pour chaque indice i, on a : $(x_i-\overline{x})^2=x_i^2-2x_i\overline{x}+\overline{x}^2$.

En remplaçant dans le calcul de la variance chaque $(x_i-\overline{x})^2$ par ce que l'on vient de trouver, on obtient :

$$V=\frac{1}{N}\left[n_1\left(x_1^2-2x_1\overline{x}+\overline{x}^2\right)+n_2\left(x_2^2-2x_2\overline{x}+\overline{x}^2\right)+\cdots+n_p\left(x_p^2-2x_p\overline{x}+\overline{x}^2\right)\right]$$
$$=\frac{1}{N}\left[n_1x_1^2+n_2x_2^2+\cdots+n_px_p^2-2n_1x_1\overline{x}-2n_2x_2\overline{x}-\cdots-n_px_p\overline{x}+n_1\overline{x}^2+n_2\overline{x}^2+\cdots+n_p\overline{x}^2\right]$$
$$=\frac{1}{N}\left[n_1x_1^2+n_2x_2^2+\cdots+n_px_p^2-2\overline{x}(n_1x_1+n_2x_2+\cdots+n_px_p)+\overline{x}^2\left(n_1+n_2+\cdots+n_p\right)\right]$$
$$=\frac{1}{N}\left[n_1x_1^2+n_2x_2^2+\cdots+n_px_p^2-2\overline{x}\times N\overline{x}+\overline{x}^2N\right]$$
$$=\frac{1}{N}\left[n_1x_1^2+n_2x_2^2+\cdots+n_px_p^2-N\overline{x}^2\right]$$
$$=\frac{n_1x_1^2+n_2x_2^2+\cdots+n_px_p^2}{N}-\overline{x}^2.$$

donc :

$$\boxed{\mathbf{V=\frac{n_1x_1^2+n_2x_2^2+\cdots+n_px_p^2}{N}-\overline{x}^2}}$$

Exemple : pour la deuxième série de notes :

note x_i	8	10	11	12
x_i^2	64	100	121	144
effectif n_i	1	2	4	1

$$V=\frac{(1\times64)+(2\times100)+(4\times121)+(1\times144)}{8}-10,5^2=\frac{892}{8}-10,5^2=111,5-110,25=1,25.$$

III.2 Écart type

Définition

La variance est homogène aux carrés des valeurs de la série. Pour avoir une grandeur homogène aux valeurs de la série, on définit **l'écart type** de la série par : $\sigma=\sqrt{\mathbf{V}}$.
L'écart type est la racine carrée de la variance.

Remarque : l'écart-type et la variance se calculent à la calculatrice

Exemple : pour la première série de notes, on a : $V=\frac{1905}{10}-10,5^2=80,25$.
L'écart type de la première série est $\sigma=\sqrt{V}=\sqrt{80,25}\approx8,96$.
Celui de la deuxième série est $\sigma=\sqrt{1,25}\approx1,118$.

L'écart type de la première série est plus grand que celui de la deuxième série : les notes sont **plus dispersées** dans le premier cas que dans le second.

IV Nuage de points, point moyen

On considère une série statistique à deux variables $(x_i\ ;\ y_i)$ telle que $\sigma_x\neq0$ et $\sigma_y\neq0$. ($\sigma_x=0$ correspond au cas où toutes les valeurs x_i seraient égales).

IV.1 Nuage de points

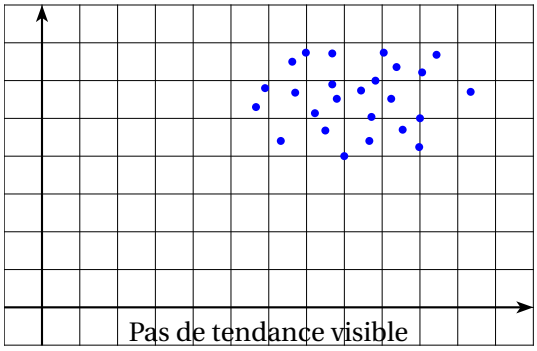
Définition

Dans un repère orthogonal, l'ensemble des points A_i de coordonnées $(x_I\ ;\ y_i)$, $1\leqslant i\leqslant n$ est appelé **nuage de points** associé à cette série statistique.

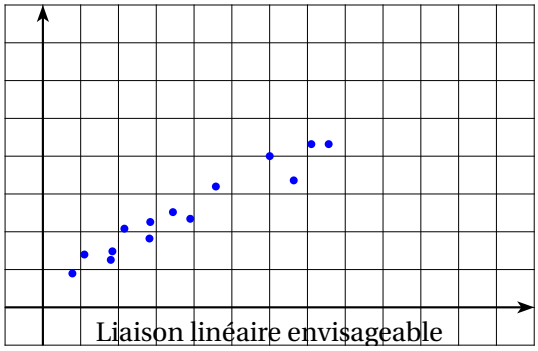
La forme du nuage de points peut sembler indiquer une liaison ou corrélation entre les deux variables x et y.
On peut chercher une corrélation linéaire (les points sont à peu près alignés) ou une corrélation avec une courbe liée à

une fonction connue (fonction inverse, logarithme, puissance...)

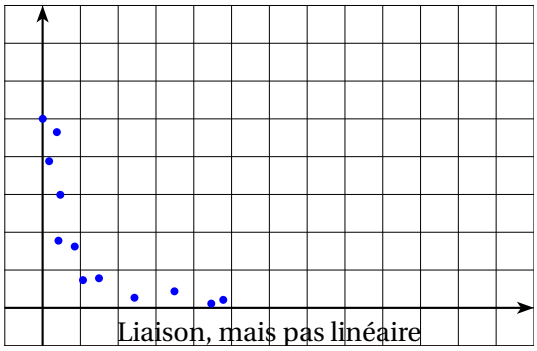
Exemples de nuages :



Pas de tendance visible



Liaison linéaire envisageable



Liaison, mais pas linéaire

Dans le cas où les points sont sensiblement alignés, on cherche une droite d'équation $y = ax+b$, qui passe « proche »des points du nuage.
On dit qu'on a réalisé un ajustement linéaire.

IV.2 Ajustement affine graphique

1. Ajustement à vue :
   On trace une droite passant approximativement le plus près possible de tous les points.
   C'est une méthode rapide, mais peu précise.
2. Ajustement à l'aide du point moyen.
   On appelle point moyen du nuage le point G dont les coordonnées sont $(\overline{x}\ ;\ \overline{y})$, c'est-à-dire les moyennes des abscisses et des ordonnées.
   On fait alors passer la droite d'ajustement par le point G et proche (le plus possible) de tous les points.

IV.3 Ajustement affine par la méthode des moindres carrés

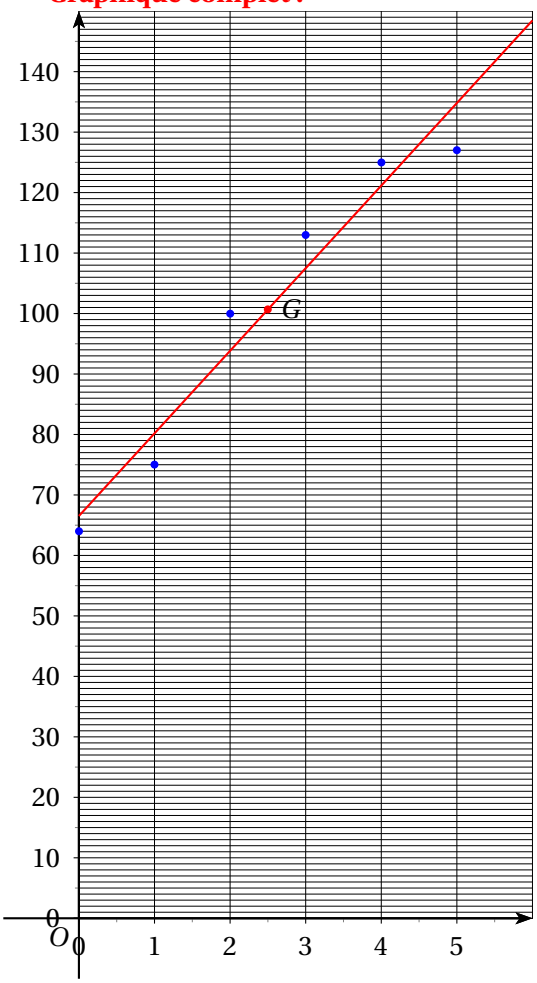
- il s'agit de trouver la droite d'ajustement $\mathscr{D}$,d'équation $y = ax + b$, telle que la somme des carrés des écarts entre les points de même abscisses $M_i(x_i\ ;\ y_i)$ du nuage et $P_i(x_i\ ;\ y_i')$ de $\mathscr{D}$ soit minimale.
- Cette droite passe par le point moyen G
- Les coefficients a et b sont obtenus à la calculatrice.

Exemple :
La société MERCURE vend des machines agricoles. Suite à une restructuration en 1998 elle a pu relancer sa production et ses bénéfices annuels ont évolué comme indiqué dans le tableau suivant :

Année	1999	2000	2001	2002	2003	2004
Rang de l'année : x_i	0	1	2	3	4	5
Bénéfice en k€ : y_i	64	75	100	113	125	127

1. Construire le nuage de points associé à la série statistique $(x_i\ ;\ y_i)$ dans un repère orthogonal.
   Les unités graphiques seront : 2 cm pour une unité sur l'axe des abscisses ; 1 cm pour 10 unités sur l'axe des ordonnées.
2. Donner les coordonnées du point moyen G du nuage (arrondir au dixième). Placer le point G dans le repère.
3. On envisage de représenter le bénéfice y comme une fonction affine du rang x de l'année.
4. Donner une équation de la droite d'ajustement (D) obtenue par la méthode des moindres carrés (arrondir les coefficients au centième).
5. Tracer cette droite (D) dans le repère.
6. Quelle prévision ferait-on pour le bénéfice en 2005 avec cette approximation ?

Graphique complet :



Les coordonnées de G sont $(2,5\ ;\ 100,667)$.
La droite d'ajustement a pour équation $y = ax + b$ avec $a \approx 13,657$ et $b \approx 66,523$
Pour 2005, on peut envisager un bénéfice de 147 k€.