Chapitre 16 – Statistiques descriptives

I Généralités

Les premières statistiques correctement élaborées ont été celles des recensements démographiques. Ainsi le vocabulaire statistique est essentiellement celui de la démographie.

Définition

L'ensemble étudié est appelé **population**. Les éléments de la population sont appelés **individus** ou unités statistiques et on appelle **effectif total** le nombre de ces individus.
La population est étudiée selon un ou plusieurs **caractères**.

Exemple

- On étudie l'âge des employés d'une entreprise. Ici les individus sont les employés et le caractère étudié est leur âge.
- On s'intéresse à la ville d'origine des élèves d'une classe de BCPST. La population est constituée des élèves de la classe considérée, sur laquelle on regarde le caractère « ville d'origine ».
- On compte le nombre de petits pois dans chaque boîte d'un lot de conserves produites par un industriel de l'agro-alimentaire. La population est ici constituée des boîtes de conserve et le caractère est le nombre de petits pois dans chaque boîte.

Comme on peut le constater sur les exemples, le caractère (grandeur souvent notée $x, y, \ldots$) que l'on observe peut être de nature **quantitative** (taille, âge, poids…) ou **qualitative** (couleur des yeux, ville de naissance…)

Définition

On appelle **modalités** les **valeurs prises par un caractère**.

Remarque.
- Formellement un caractère est une *application* ayant pour ensemble de départ la population, et pour ensemble d'arrivée les modalités.
- Un caractère est parfois appelé une variable statistique.
- Si l'effectif de la population est de N (on dit que la population est de taille N) l'observation se traduit par un N-uplet : $(u_1, u_2, \ldots, u_n)$, où chaque u_i est la valeur du caractère pour l'individu i.

Lorsque ces modalités sont trop nombreuses, on procède à un **regroupement en classes** : les modalités sont rassemblées en intervalles disjoints et complémentaires (d'amplitudes égales ou non), appelées les **classes**. On dit alors qu'on est dans le *cas continu*. Dans le cas contraire, on dit qu'on est dans le *cas discret*.

Exemple

Le nombre d'habitants des villes françaises est un caractère quantitatif dont le nombre de modalités est très grand (de quelques unités pour un hameau à plusieurs centaines de milliers d'habitants dans les grandes agglomérations).
Il convient donc, pour un traitement statistique, de regrouper les villes en classes. Par exemple, on peut distinguer les hameaux (moins de 100 habitants), les villages (entre 101 et 2000 habitants), les villes (entre 2001 et 100000 habitants) et les agglomérations (plus de 100001 habitants).

Dans la suite de ce cours, on s'intéresse uniquement aux **caractères quantitatifs** et on envisage considère :

1. le cas où l'on étudie un seul caractère (statistique univariée)
2. le cas où l'on prend en compte deux caractères sur une même population (statistique bivariée)

II Statistique univariée

Définition (*Notations générales*)

Dans la suite, $x_1, x_2, \ldots, x_p$ ($p \in \mathbb{N}^*$) désignent les modalités. On suppose que les valeurs sont classées dans l'ordre croissant : $x_1 < x_2 < \ldots < x_p$.
Pour chaque modalité x_k, on note $n_k \in \mathbb{N}$ **l'effectif** correspondant, c'est-à-dire le nombre d'individus pour lesquels le caractère prend cette valeur x_k.
L'effectif total de la population est alors : $N = n_1 + n_2 + ... + n_p$.

On obtient la **fréquence** f_k d'une modalité x_k par : $f_k = \dfrac{n_k}{N}$.
On a : $\forall k \in [\![1, p]\!]$, $f_i \in [0, 1]$. De plus, $f_1 + f_2 + ... + f_p = 1$.

Définition (*Effectifs et fréquences cumulés croissants*)

Pour tout $k \in [\![1, p]\!]$, on appelle

- **effectif cumulé (croissant)** de la modalité x_k le nombre $n_1 + n_2 + \ldots + n_k$ d'individus pour lesquels le caractère est inférieur ou égal à x_k.
- **fréquence cumulée (croissante)** de la modalité x_k la fréquence $f_1 + f_2 + \ldots + f_k \in [0, 1]$ associée à l'effectif cumulé croissant.

1. Représentations graphiques

On peut représenter visuellement une série statistique à l'aide de différentes représentations graphiques :

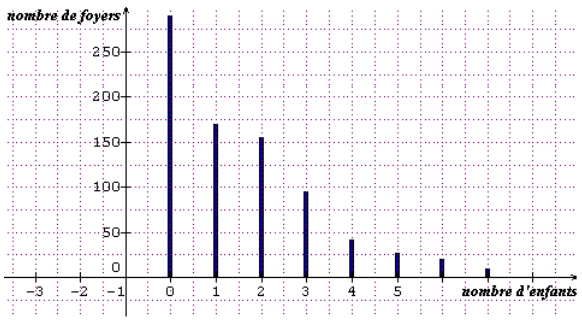
- Un **diagramme en bâtons** est une suite de rectangles de largeurs très fines et de hauteurs égales aux différentes fréquences (ou effectifs). C'est le mode de représentation privilégié pour le cas discret.
- Un **histogramme** est au contraire adapté à la représentation des séries quantitatives dans le cas continu. Cette fois, on dessine des rectangles dont les largeurs sont égales aux amplitudes des différentes classes, en faisant en sorte que **la fréquence associée à chaque classe est proportionnelle à l'aire du rectangle**. Ainsi, les hauteurs sont proportionnelles aux fréquences (ou aux effectifs) divisées par les amplitudes.

Exemple

Dans un village, on a recensé le nombre d'enfants par foyer.

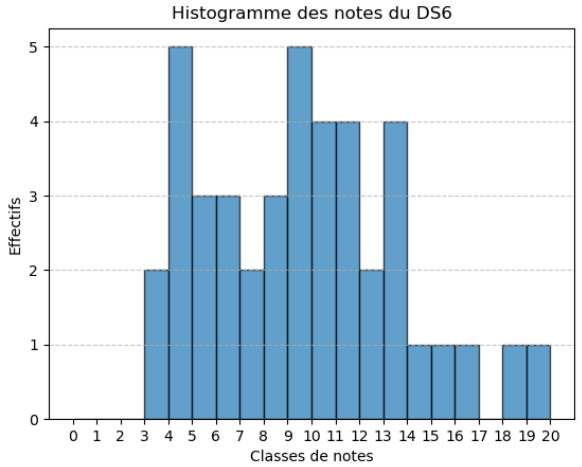
Nombre d'enfants	0	1	2	3	4	5	6	7
Nombre de foyers	302	158	145	95	53	27	22	11
Effectifs cumulés croissants								

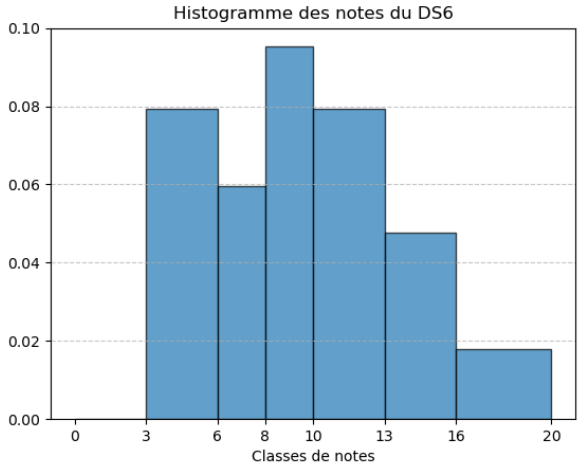
Population étudiée :
Effectif total : $N =$
Caractère étudié :



Exemple

Les notes au DS sont usuellement mises en ligne sous forme d'un histogramme, où les classes sont régulières : $[0,2[, [2,4[, [4,6[, \ldots, [16,18[, [18,20]$ (pour une amplitude de 2), ou parfois avec une amplitude de 1. De la sorte, un histogramme est strictement équivalent à un diagramme en barres. On peut cependant faire des histogrammes avec des classes d'amplitude variées, où c'est bien l'aire qui apporte une information. Par exemple, voilà deux regroupement en classes différents pour les mêmes notes :





Représentation des effectifs ou fréquences cumulé(e)s croissant(e)s

On a souvent intérêt à représenter les effectifs ou fréquences cumulés croissants, notamment pour déterminer efficacement la médiane ou les quartiles par exemple.

Pour ce faire, on place dans un repère les points (x_k, ECC_k) où ECC_k est l'effectif cumulé associé à la modalité x_k.

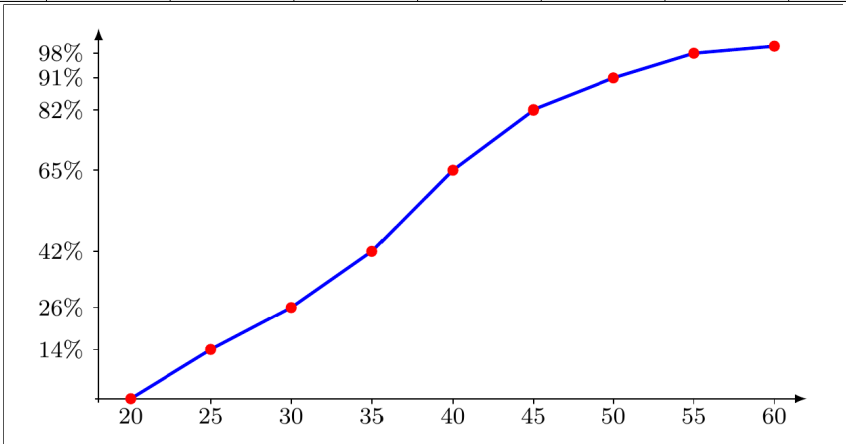
Dans le cas continu : si les classes sont les intervalles $([a_{k-1}, a_k[)_{1\leq k\leq p}$, avec $a_0 < a_1 < \ldots < a_p$, on place dans un repère les points (a_k, ECC_k), et on les relie, en commençant par le point $(a_0, 0)$.

On obtient alors une courbe *affine par morceaux* (et continue), appelée le **polygone des effectifs (resp. fréquences) cumulés croissants**.

Exemple

On recense l'âge de 100 employés d'une entreprise. On obtient :

Classes d'âge	$[20,25[$	$[25,30[$	$[30,35[$	$[35,40[$	$[40,45[$	$[45,50[$	$[50,55[$	$[50,55[$
Effectifs	14	12	16	23	17	9	7	2
Effectifs cumulés croissants								



2. Paramètres de position

• Moyenne

Définition

La **moyenne**, notée $\overline{x}$, est définie par :

$$\overline{x} \quad = \quad \frac{n_1\,x_1 + n_2\,x_2 + \ ... \ + n_p\,x_p}{N} \quad = \quad \frac{1}{N}\sum_{k=1}^{p} n_k x_k \quad = \quad \sum_{k=1}^{p} f_k x_k$$

Dans le cas continu, on prend pour valeurs x_i les **centres de classes** , c'est-à-dire les **milieux des intervalles.**

Théorème

Linéarité de la moyenne

La moyenne est linéaire. Cela signifie que pour tous réels a et b, on a :
- la moyenne de la nouvelle série ax_1, ax_2, ...ax_p est égale à $a \times \overline{x}$.
- la moyenne de la nouvelle série $x_1 + b$, $x_2 + b$, ...$x_p + b$ est égale à $\overline{x} + b$.

Exemple

1. Un professeur considère les notes des ses élèves au dernier partiel. Le caractère, noté x, a pour moyenne $\overline{x} = 8,1$. Jugeant qu'il a été trop sévère, il décide d'augmenter les notes de 50% avant de leur retrancher 2 points. Autrement dit, il définit un nouveau caractère $y = 1,5x - 2$.
   Sans avoir à étudier la série statistique de y, on sait que la nouvelle moyenne est $\overline{y} =$ .....
2. Considérons deux séries statistiques : celle d'un caractère x_1 sur une population d'effectif total n_1 et celle d'un caractère x_2 sur une population d'effectif total n_2 . Si on réunit les deux populations (ce qui n'a de sens que si elles sont de même nature) et que l'on considère, sur cette nouvelle population, le caractère x obtenu en réunissant les modalités de x_1 et x_2, alors :

   $$\overline{x} = \frac{n_1\overline{x_1} + n_2\overline{x_2}}{n_1 + n_2}$$

   Si par exemple, à un devoir commun, la moyenne des notes d'une classe de 46 étudiants est $9,3$ et celle de l'autre classe, de 38 étudiants, est $10,6$.
   Alors, la moyenne de la promotion est de : .....

• Médiane

Définition

La **médiane** d'une série, notée M, est la valeur centrale de la série si N est impair, ou la moyenne des deux valeurs centrales si N est pair.

- Si N est impair, il s'agit de la valeur associée à l'individu médian, de rang $\dfrac{N+1}{2}$.
- Si N est pair, il s'agit de la moyenne des deux valeurs associés aux individus médians, de rang respectif : $\dfrac{N}{2}$ et $\dfrac{N}{2} + 1$.

Au moins 50% des valeurs de la série sont inférieures ou égales à M, et au moins 50% des valeurs sont supérieures ou égales à M.

Dans le cas continu, la médiane se détermine à l'aide de la courbe des fréquences cumulées croissantes : il suffit pour cela de déterminer **l'abscisse du point dont l'ordonnée est égale à** $1/2$.
Cela revient à faire l'hypothèse que les valeurs sont réparties uniformément dans les classes (hypothèse déjà émise par exemple pour le calcul de la moyenne).

Exemple

Déterminer la médiane dans chacun des exemples précédents :

Remarque. La moyenne est sensible aux valeurs extrêmes de la série, mais ce n'est pas le cas de la médiane ! Par exemple : on préfère étudier le salaire médian en France, plutôt que le salaire moyen.

3. Paramètres de dispersion

Les paramètres de position donnent une idée de l'ordre de grandeur des valeurs de la série, mais n'indiquent pas comment se répartissent les valeurs autour de ce paramètre de position. Ça, c'est le rôle des paramètres de dispersion.

- Étendue

Définition

L'*étendue* d'une série statistique est la **différence entre la plus grande valeur et la plus petite** pour lesquelles l'effectif n'est pas nul. C'est donc l'écart maximal entre deux modalités de la série.

- **Variance et écart-type**

Définition

On appelle **variance** de la série statistique le nombre, noté V_x ou V, défini par :

$$V = \frac{1}{N}\sum_{k=1}^{p} n_k(x_k - \overline{x})^2$$

La variance est *la moyenne des carrés des écarts à la moyenne !* On appelle **écart-type** de cette série le réel, noté s_x ou s, défini par $s = \sqrt{V}$.

Remarque.

- L'écart-type mesure un écart moyen par rapport à la moyenne $\overline{x}$.
  Plus il est élevé, plus les valeurs sont dispersées. Plus il est faible, plus les valeurs sont regroupées autour de la moyenne de la série.
- $s_x = 0$ si et seulement si la série statistique est constante !
- L'écart-type est exprimé dans la **même unité** que le caractère étudié (si x est une masse en kg alors s_x est aussi exprimée en kg) : il est **homogène** au caractère. Ce n'est pas le cas de la variance !
- Pour comparer deux écart-types, on rapporte ces écart-types à leurs moyennes en considérant le coefficient de variation $\frac{s_x}{\overline{x}}$ (dont la valeur, sans unité, est souvent donnée en pourcentage).

Proposition (*Formule de König-Huygens*)

$$V = \overline{x^2} - \overline{x}^2$$

La variance est donc la différence de la moyenne du carré et du carré de la moyenne.

Démonstration.

□

- **Quartiles et écart interquartile**

Définition

Le *premier quartile* Q_1 est la plus petite valeur de la série telle que au moins 25% des valeurs lui soient inférieures ou égales.
Le *deuxième quartile* Q_2 est égal à la médiane de la série.
Le *troisième quartile* Q_3 est la plus petite valeur de la série telle que au moins 75% des valeurs lui soient inférieures ou égales.

L'*écart interquartile* est égal à $Q_3 - Q_1$.

Remarque.

- Q_1 est donc la valeur de rang $\left\lceil \frac{N}{4} \right\rceil$, et Q_3 la valeur de rang $\left\lceil \frac{3N}{4} \right\rceil$, où $\lceil x \rceil$ désigne la *partie entière supérieure* de x, ie le plus petit entier $k \in \mathbb{N}$ tel que $x \leq k$.
- **Avec le polygone des fréquences cumulées croissantes :** Q_1 correspond à l'abscisse du point dont l'ordonnée est égale à $1/4$, et Q_3 correspond à l'abscisse du point dont l'ordonnée est égale à $3/4$.
- L'intervalle interquartile $[Q_1, Q_3]$ contient, à peu près, 50% des individus de la population. L'écart interquartile est donc un indicateur de la **dispersion des valeurs « centrales »** de la série.

Exemple

Déterminer les quartiles et l'écart interquartile dans chacun des exemples précédents.

III Statistique bivariée

Lorsqu'il semble qu'il y ait un lien entre deux caractères d'une population, il est intéressant d'étudier ces deux caractères simultanément. On étudie alors une série à deux variables.

1. Généralités et nuages de points

On considère une population d'effectif total N sur laquelle est observé un couple (x, y) de caractères quantitatifs, x prenant ses valeurs dans $\{x_1, x_2, \ldots, x_p\}$ et y dans $\{y_1, y_2, \ldots, y_q\}$, qui sont ordonnés par ordre croissant.
Les couples de valeurs (x_k, y_l) sont appelées les modalités (conjointes) du couple (x, y).

Pour tout $(k, l) \in [\![1, p]\!] \times [\![1, q]\!]$, on appelle **effectif conjoint** de la modalité (x_k, y_l) le nombre d'individus $n_{k,l}$ pour lesquels le caractère x prend la valeur x_k et simultanément, le caractère y est égal à y_l.
La **fréquence conjointe** est égale à $f_{k,l} = \dfrac{n_{k,l}}{N}$.

On peut alors regrouper les effectifs conjoints (ou les fréquences conjointes) dans un **tableau à double entrée** (avec p lignes et q colonnes). On rajoute à ce tableau une ligne et une colonne (les marges) dans lesquelles on indique "sous-totaux", c'est-à-dire les effectifs ou fréquences correspondant à chacun des deux caractères indépendamment de l'autre. On les appelle les **effectifs marginaux** (resp. **fréquences marginales**).

Exemple

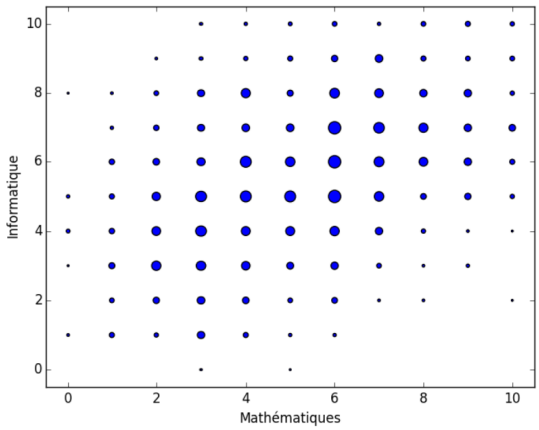
Le tableau suivant indique les résultats des candidats à une épreuve de concours comptant 10 points d'informatique et 10 points de mathématiques. Les deux caractères étudiés sont donc la note en maths et la note en info de chaque candidat.

Info / Maths	0	1	2	3	4	5	6	7	8	9	10	Effectifs maths
0	0	2	0	1	4	3	0	0	1	0	0	11
1	0	7	6	10	8	7	8	3	2	0	0	51
2	0	5	11	24	21	19	12	8	6	2	0	108
3	1	13	4	22	28	29	16	12	12	3	2	158
4	0	7	12	20	22	33	33	16	23	5	3	174
5	1	3	6	13	22	33	25	16	10	7	4	140
6	0	3	9	15	24	42	42	42	26	11	6	220
7	0	0	2	6	15	25	27	31	21	16	3	146
8	0	0	2	2	5	9	21	23	15	7	6	90
9	0	0	0	3	2	11	15	15	15	6	7	74
10	0	0	1	0	1	5	7	12	5	6	5	42
Effectifs Info	2	40	63	116	152	216	206	178	136	63	36	1208

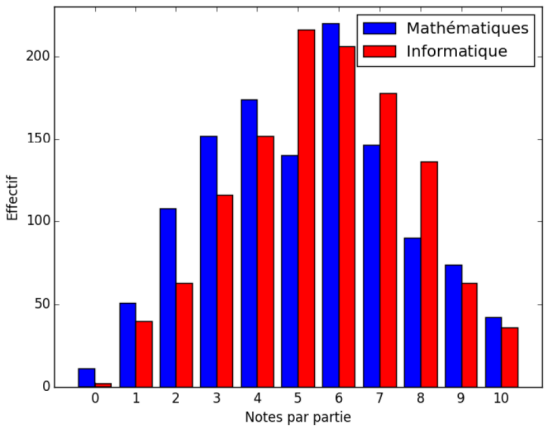
Pour représenter graphiquement une série statistique double, on utilise un nuage de points :

- On représente chaque apparition de modalité conjointe (x_k, y_l) par un point aux coordonnées (x_k, y_l) dans un repère.
- Si des modalités apparaissent plusieurs fois, comme c'est le cas ci-dessus, chaque modalité (x_k, y_l) est représentée par un disque (ou un carré) de centre (x_k, y_l) et d'aire proportionnelle à la fréquence conjointe $f_{k,l}$.

Nuage de points



Histogrammes des marginales



2. Point moyen

En statistique bivariée, on utilise une seule caractéristique de position : le point moyen (qui est une généralisation de la moyenne).

Définition

On définit le point moyen de la série statistique double le point, noté habituellement G, dont les coordonnées sont les moyennes marginales de x et de y. Autrement dit, on a : $\boxed{G\left(\overline{x}, \overline{y}\right)}$.

Remarque. Il s'agit du barycentre du nuage de point !

Exemple

Placer le point moyen dans l'exemple précédent, en sachant que la note moyenne en informatique est de $5,5$ environ, et la note moyenne en maths est d'environ $5,2$.

3. Covariance et coefficient de corrélation

Définition

On appelle **covariance de la série statistique double** le nombre réel, noté $\text{cov}(x, y)$, défini par :

$$\text{cov}(x, y) = \frac{1}{N} \sum_{k=1}^{p} \sum_{l=1}^{q} n_{k,l} \left(x_k - \overline{x}\right) \left(y_l - \overline{y}\right)$$

Ce qui revient à : $\text{cov}(x, y) = \overline{xy} - \overline{x} \cdot \overline{y}$ (la différence de la moyenne du produit et du produit des moyennes).

Remarque.
- On peut considérer la covariance d'une variable avec elle-même. On obtient $\text{cov}(x, x) = V_x$.
- ⚠ La covariance peut être positive ou négative !
  Elle est positive lorsque x et y ont tendance à varier dans le même sens et négative s'ils ont tendance à varier en sens contraire.

Définition

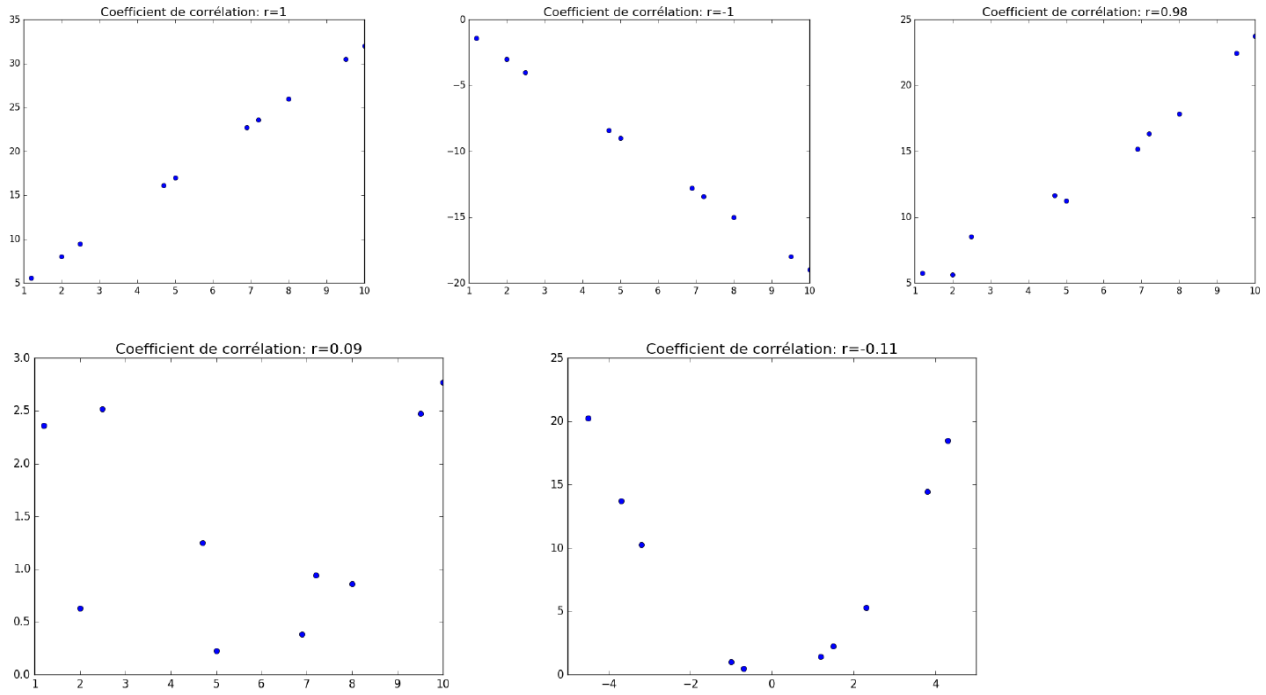
On suppose que les variables x et y ne sont pas constantes de sorte que $s_x \neq 0$ et $s_y \neq 0$.
On appelle **coefficient de corrélation de la série statistique double** le nombre réel, noté $r_{x,y}$, défini par :

$$r_{x,y} = \frac{\text{cov}(x, y)}{s_x s_y}$$

Propriété

- $r_{x,y} \in [-1, 1]$
- $r_{x,y} = \pm 1 \iff \exists a, b \in \mathbb{R}, y = ax + b.$

Remarque.
- Le coefficient de corrélation est sans unité
- Lorsque le coefficient de corrélation est proche de -1 ou $+1$, les caractères x et y sont dit **bien corrélés**, ce qui signifie qu'ils sont liés entre eux par une relation presque affine, ou encore que le nuage de points est **presque aligné le long d'une droite** (croissante lorsque $r_{x,y}$ est proche de $+1$ et décroissante lorsque $r_{x,y}$ est proche de -1).
- Au contraire, lorsqu'il n'y a aucun lien affine entre x et y, le coefficient de corrélation est faible (c'est-à-dire proche de 0 en valeur absolue).



Remarque. Il faut prendre garde à la confusion fréquente entre **corrélation et causalité**. Que deux phénomènes soient corrélés n'implique en aucune façon que l'un soit la cause de l'autre !
Souvent, une forte corrélation indique seulement que les deux caractères dépendent d'un troisième, qui n'a pas été mesuré. Par exemple, la corrélation forte entre le rendement des impôts en France et la criminalité au Japon indique seulement que les deux sont liés à l'augmentation globale de la population.
Cependant, il arrive qu'une forte corrélation traduise bien une vraie causalité, comme entre le nombre de cigarettes fumées par jour et l'apparition d'un cancer du poumon. Mais ce n'est pas la statistique qui démontre la causalité, elle permet seulement de la détecter.

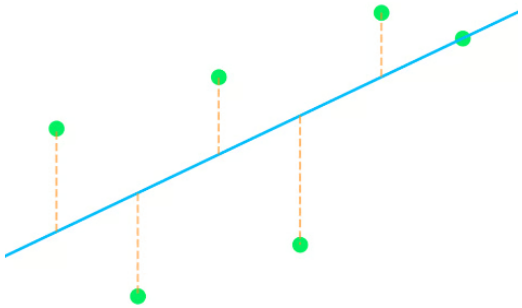
4. Ajustement affine par la méthode des moindres carrés

Dans ce paragraphe, on se place dans le cas où la série statistique est de la forme $(x_1, y_1), \dots, (x_N, y_N)$, avec les (x_i) et les (y_i) deux à deux distincts.
On représente alors cette série par un nuage où les points sont "sans épaisseurs" (effectif 1 pour chaque point), et jamais alignés verticalement ou horizontalement.

On a vu que lorsque le coefficient $r_{x,y}$ est proche de ± 1, le nuage de points est "bien aligné le long d'une droite". Il est donc légitime de chercher à obtenir une droite qui modélise le nuage de point de la meilleure manière possible. On appelle une telle droite un **ajustement affine** de la série statistique double.

Il existe plusieurs méthode d'ajustement affine. La méthode présentée ici est la **régression linéaire**.
Elle consiste à choisir, selon la **méthode des moindres carrées**, une droite dont la « distance » avec le nuage de points est la plus petite possible.
On choisit alors la droite qui minimise la somme des carrés des distances verticales qui séparent la droite de chacun des points du nuage :

$$S = \sum_{k=1}^{N} (y_k - (ax_k + b))^2$$



Proposition

La droite obtenue par la méthode des moindre carrée est la droite d'équation $Y = aX + b$ avec :

$$a = \frac{\text{cov}(x,y)}{s_x^2} = \frac{\text{cov}(x,y)}{V_x} \qquad \text{et} \qquad b = \overline{y} - a\overline{x}$$

Il s'agit donc de la droite de coefficient directeur $a = \dfrac{\text{cov}(x,y)}{s_x^2}$ qui passe par le point moyen G.
On l'appelle la **droite de régression linéaire de y en x**.

Remarque. De la même façon, on peut déterminer la droite de régression linéaire de x en y, qui minimise les distances horizontales de chacun des points du nuage à la droite. On obtient une droite d'équation $X = a'Y + b'$ avec $a' = \dfrac{\text{cov}(x,y)}{s_y^2}$ et $b' = \overline{x} - a'\overline{y}$.

Si on trace ces deux droite $Y = aX + b$ et $Y = \dfrac{X - b'}{a'}$ dans un même repère, elles passent toutes les deux par le point médian $G(\overline{x}, \overline{y})$.
Ces deux droites sont presque confondues lorsque les coefficients directeurs sont proches. Or on a :

$$\begin{aligned} a \simeq \frac{1}{a'} &\iff aa' \simeq 1 \iff \frac{\text{cov}(x,y)}{s_x^2}\frac{\text{cov}(x,y)}{s_y^2} \simeq 1 \\ &\iff \left(\frac{\text{cov}(x,y)}{s_x s_y}\right)^2 \simeq 1 \\ &\iff (r_{x,y}^2) \simeq 1 \iff r_{x,y} \simeq \pm 1 \end{aligned}$$

On retrouve bien que plus le coefficient de corrélation est plus proche de ± 1, meilleur est l'ajustement affine.

⚠ La droite de régression existe toujours. Toutefois, lorsque la corrélation n'est pas bonne (c'est-à- dire lorsque $r_{x,y}$ n'est pas proche de ± 1), cette droite passe au travers du nuage de points mais en est une très mauvaise modélisation !

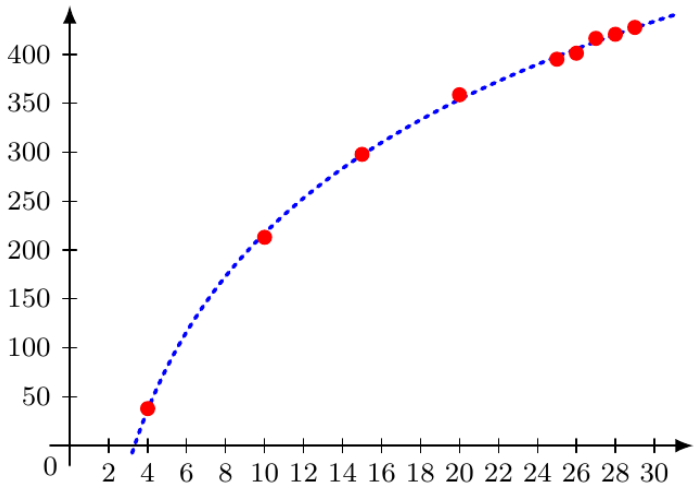
5. Autres ajustements

De nombreux autres ajustements peuvent être possibles lorsque le nuage de point semble "ressembler" à une courbe autre qu'une droite. On peut alors réaliser des ajustements polynomiaux, logarithmiques, exponentiels, etc.

Exemple

Le tableau ci-dessous donne la production d'électricité d'origine nucléaire en France, exprimée en milliards de kWh, entre 1979 et 2004.

x : nombre d'années après 1975	4	10	15	20	25	26	27	28	29
y : production électrique	37,9	213,1	297,9	358,8	395,2	401,3	416,5	420,7	427,7



La distribution du nuage de points suggère qu'ils se trouvent sur une courbe logarithmique.
Pour faire cet ajustement logarithmique, on se ramène à un ajustement affine en travaillant sur le couple de caractères $(\ln(x), y)$.
La nouvelle série statistique est :

$x' = \ln(x)$	1,386	2,303	2,708	2,996	3,219	3,258	3,296	3,332	3,367
y	37,9	213,1	297,9	358,8	395,2	401,3	416,5	420,7	427,7

et le nouveau nuage de points (à tracer !) laisse bien penser que l'on peut effectuer un ajustement affine.
On vérifie que la corrélation est bonne (elle est même excellente ici puisque $r_{\ln(x),y} \simeq 0,9997$) et on détermine (après calculs) que la droite de régression a pour équation : $Y = 197,2X' - 236,9$.
On en déduit que les points du nuage de la série de départ sont répartis autour de la courbe d'équation :

$$Y = 197,2\ln(X) - 236,9$$