

## Chap2 / Séries statistiques à deux variables

### 2-1 / Définition d'une série statistique double

On considère une population d'effectif  $n$ , si on étudie deux caractères  $X$  et  $Y$  de cette population, on dit qu'on étudie une série stat double (SS-2D). Chaque individu de cette population est désigné par un nombre compris entre 1 et  $n$ .

A chaque individu  $i$  ( $1 \leq i \leq n$ ) correspond un couple  $(x_i, y_i)$ , où  $x_i$  est la modalité de  $X$  et  $y_i$  est la modalité de  $Y$  associé à l'individu.

L'ensemble des couples  $(x_i, y_i)$  définit une série stat à 2 variables.

N-B: Lorsque l'un des 2 caractères est une date (année, ...), on dit que la série statistique double est une série chronologique.

### 2-2 / Tableau de contingence, fréquences marginales, fréquences conditionnelles.

On s'intéresse à un couple de caractères noté  $(X, Y)$ , selon lesquels on va répartir l'ensemble de la population.

Notons  $x_1, x_2, \dots, x_k$  les  $k$  modalités de  $X$  et  $y_1, y_2, \dots, y_l$  les  $l$  modalités de  $Y$ . On désigne par  $n_{ij}$  l'effectif du groupe d'individus présentant la modalité  $x_i$  du caractère  $X$  et la modalité  $y_j$  de  $Y$ .

On rassemble toutes ces données dans un tableau de contingence.

la table de contingence et une matrice de dim =  $k \times l$

| $X \backslash Y$ | $y_1$    | ----- | $y_j$    | ----- | $y_l$    | Total    |
|------------------|----------|-------|----------|-------|----------|----------|
| $x_1$            | $n_{11}$ |       | $n_{1j}$ |       | $n_{1l}$ | $n_{1*}$ |
| $\vdots$         |          |       |          |       |          |          |
| $x_i$            | $n_{i1}$ |       | $n_{ij}$ |       | $n_{il}$ | $n_{i*}$ |
| $\vdots$         |          |       |          |       |          |          |
| $x_k$            | $n_{k1}$ |       | $n_{kj}$ |       | $n_{kl}$ | $n_{k*}$ |
| Total            | $n_{*1}$ | ----- | $n_{*j}$ | ----- | $n_{*l}$ | $n$      |

On définit :

(1) les effectifs marginaux de la série.

$$n_{i*} = \sum_{j=1}^l n_{ij} \quad ; \quad n_{*j} = \sum_{i=1}^k n_{ij}$$

(2) l'effectif total, noté  $n$  et :

$$n = \sum_{i=1}^k n_{i*} = \sum_{j=1}^l n_{*j} = \sum_{i=1}^k \sum_{j=1}^l n_{ij}$$

NB: les séries marginales  $[n_{i*}, n_{*j}]$  et

$[y_j, n_{*j}]$  sont des séries à une variable.  
 Chacune de ces séries a donc une moyenne  $\bar{x}$  et  $\bar{y}$  et un (2) écart-type  $\sigma_x$  et  $\sigma_y$

qu'on qualifie naturellement de moyenne marginale et d'écart-type marginal.

(3) les fréquences:

$$*1 \quad f_{ij} = \frac{n_{ij}}{n}$$

\*1 fréquences marginales:

$$\left\{ \begin{array}{l} f_{i*} = \sum_{j=1}^l f_{ij} = \frac{1}{n} \sum_{j=1}^l n_{ij} = \frac{n_{i*}}{n} \\ f_{*j} = \sum_{i=1}^k f_{ij} = \frac{1}{n} \sum_{i=1}^k n_{ij} = \frac{n_{*j}}{n} \end{array} \right.$$

$$\underline{\text{NB:}} \quad \sum_{i=1}^k f_{i*} = \sum_{j=1}^l f_{*j} = 1.$$

2-3/ Nuage de points et point moyen.

On définit également la notion de **fréquence**, sur le même principe que pour les séries à une variable :

$$\text{fréquence} = \frac{\text{effectif}}{\text{effectif total}} \quad \text{c'est-à-dire} \quad f_{i,j} = \frac{n_{i,j}}{N}.$$

Les **fréquences marginales** sont alors

$$f_{i,\bullet} = \sum_{j=1}^{\ell} f_{i,j} = \frac{1}{N} \sum_{j=1}^{\ell} n_{i,j} = \frac{n_{i,\bullet}}{N} \quad \text{et} \quad f_{\bullet,j} = \sum_{i=1}^k f_{i,j} = \frac{1}{N} \sum_{i=1}^k n_{i,j} = \frac{n_{\bullet,j}}{N}.$$

Bien entendu, la somme des fréquences marginales  $f_{i,\bullet}$  vaut 1, de même que la somme des fréquences marginales  $f_{\bullet,j}$  :

$$\sum_{i=1}^k f_{i,\bullet} = \sum_{j=1}^{\ell} f_{\bullet,j} = \sum_{i=1}^k \sum_{j=1}^{\ell} f_{i,j} = 1.$$

► **EXEMPLE 1.** On s'intéresse à la répartition des salaires des employés d'une grande entreprise, répartis par tranches d'âge :

| âge (en années) \ salaire (en €) | [18;25[ (A) | [25;35[ (B) | [35;45[ (C) | [45;55[ (D) | [55;65[ (E) | total |
|----------------------------------|-------------|-------------|-------------|-------------|-------------|-------|
| [800;900[ (1)                    | 122         | 107         | 58          | 21          | 6           | 314   |
| [900;1000[ (2)                   | 125         | 131         | 86          | 24          | 11          | 377   |
| [1000;1200[ (3)                  | 51          | 66          | 65          | 54          | 37          | 273   |
| [1200;1500[ (4)                  | 38          | 27          | 70          | 73          | 31          | 239   |
| [1500;2000[ (5)                  | 6           | 10          | 32          | 29          | 44          | 121   |
| [2000;3000[ (6)                  | 0           | 5           | 18          | 15          | 20          | 58    |
| [3000;5000[ (7)                  | 0           | 3           | 7           | 10          | 5           | 25    |
| total                            | 342         | 349         | 336         | 226         | 154         | 1407  |

effectifs marginaux

Dans cet exemple, les modalités du caractère  $x$  (le salaire) sont désignées par des nombres (1 pour la classe [800;900[, 2 pour la classe [900;1000[, ..., 7 pour la classe [3000;5000]) et celles du caractère  $y$  (l'âge) par des lettres (A pour la classe [18;25[, B pour [25;35[, ..., E pour la classe [55;65]).

$n_{1,A} = 122$  indique le nombre de salariés dont le salaire fait partie de l'intervalle [800;900[ (modalité 1 du caractère  $x$ ) et dont l'âge est dans l'intervalle [18;25[ (modalité A du caractère  $y$ ). Les effectifs marginaux sont les effectifs  $n_{1,\bullet}, n_{2,\bullet}, \dots, n_{7,\bullet}$  des modalités 1, 2, ..., 7 du caractère  $x$  et les effectifs  $n_{\bullet,A}, n_{\bullet,B}, \dots, n_{\bullet,E}$  des modalités A, B, ..., E; ainsi,

$$n_{3,\bullet} = n_{3,A} + n_{3,B} + n_{3,C} + n_{3,D} + n_{3,E} = 273$$

est le nombre de salariés dont le salaire mensuel est dans l'intervalle [1000;1200[, et

$$n_{\bullet,B} = n_{1,B} + n_{2,B} + n_{3,B} + n_{4,B} + n_{5,B} + n_{6,B} + n_{7,B} = 349$$

est le nombre de salariés dont l'âge (en années) fait partie de l'intervalle [25;35[.

La fréquence marginale  $f_{4,\bullet}$  est la fréquence de la modalité 4 du caractère  $x$  :

$$f_{4,\bullet} = \frac{n_{4,\bullet}}{N} = \frac{239}{1407} \approx 0,1699 \approx 16,99\%.$$

La **fréquence conditionnelle de la modalité  $x_i$  (du caractère  $x$ ) par rapport à la modalité  $y_j$  (du caractère  $y$ )** est le nombre

$$f_{y_j}(x_i) = \frac{n_{i,j}}{n_{\bullet,j}} = \frac{f_{i,j}}{f_{\bullet,j}}.$$

Il donne la proportion d'individus présentant la modalité  $x_i$  parmi ceux qui présentent la modalité  $y_j$  ; on la qualifie d'ailleurs parfois de **fréquence de  $x_i$  sachant  $y_j$**  et on la note  $f(x_i | y_j)$ .

De même, la **fréquence conditionnelle de la modalité  $y_j$**  (du caractère  $y$ ) **par rapport à la modalité  $x_i$**  (du caractère  $x$ ) est le nombre

$$f_{x_i}(y_j) = \frac{n_{i,j}}{n_{i,\bullet}} = \frac{f_{i,j}}{f_{i,\bullet}},$$

noté aussi  $f(y_j | x_i)$  (**fréquence de  $y_j$  sachant  $x_i$** ), indiquant la proportion des individus présentant la modalité  $y_j$  au sein du groupe de ceux qui présentent la modalité  $x_i$ .

Illustrons cette notion de fréquence conditionnelle par quelques calculs relatifs à l'exemple 1 :

- parmi les employés dont le salaire est compris entre 900 € et 1000 €, la proportion de ceux dont l'âge est situé dans l'intervalle [45 ; 55[ est la fréquence conditionnelle de D par rapport à la modalité 2 :

$$f_2(D) = \frac{n_{2,D}}{n_{2,\bullet}} = \frac{24}{377} \approx 0,064 \approx 6,4 \% ;$$

- la proportion d'employés dont le salaire est compris entre 900 € et 1000 € parmi ceux de la tranche 45-55 ans est la fréquence conditionnelle de 2 par rapport à C :

$$f_D(2) = \frac{n_{2,D}}{n_{\bullet,D}} = \frac{24}{226} \approx 0,106 \approx 10,6 \%.$$

## 2 Nuage de points, point moyen

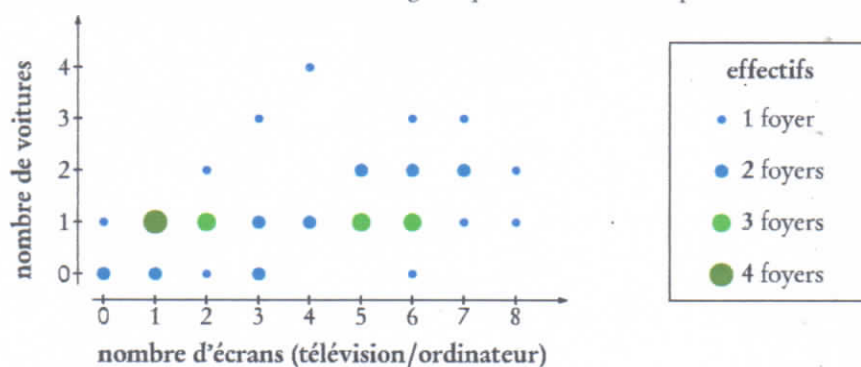
Il est souvent très difficile de représenter graphiquement les séries bivariées, spécialement dans le cas continu ; il peut alors être commode d'avoir recours à des schémas à 3 dimensions (de la même façon que des schémas plans constituent d'excellentes représentations des séries univariées).

Dans certains cas très particuliers (notamment lorsque les effectifs sont très faibles), on peut tout de même présenter de manière claire des séries bivariées à l'aide de diagrammes à 2 dimensions, comme le montrent les exemples suivants :

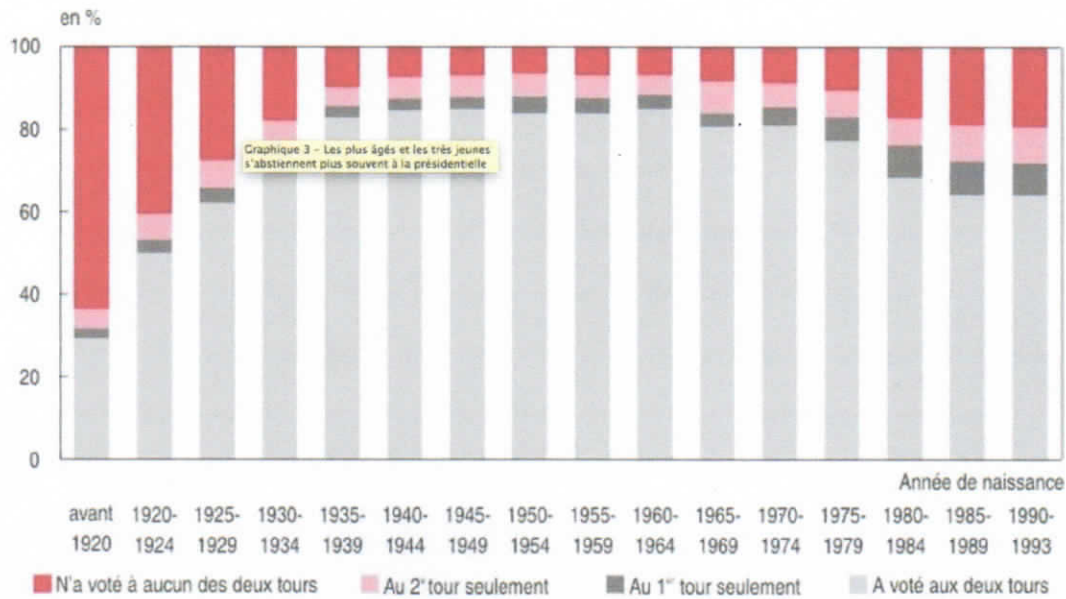
- **EXEMPLE 2.** On interroge 40 foyers sur leurs équipements automobile et audiovisuel, autrement dit sur le nombre de voitures et d'écrans (de télévision et d'ordinateur — on a choisi d'exclure les smartphones et les tablettes) en leur possession.

| nombre de voitures \ nombre d'écrans | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|--------------------------------------|---|---|---|---|---|---|---|---|---|
| 0                                    | 2 | 2 | 1 | 2 | 0 | 0 | 1 | 0 | 0 |
| 1                                    | 1 | 4 | 3 | 2 | 2 | 3 | 3 | 1 | 1 |
| 2                                    | 0 | 0 | 1 | 0 | 0 | 2 | 2 | 2 | 1 |
| 3                                    | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| 4                                    | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

Les résultats obtenus sont donnés dans le tableau de contingence précédent et sont représentés sur le schéma ci-dessous.



- **EXEMPLE 3.** Le schéma ci-dessous indique les taux de participation des Français aux deux tours de l'élection présidentielle de 2012, en fonction de leur année de naissance (source : INSEE).



Ici, les deux caractères considérés sont :

- l'année de naissance (caractère quantitatif discret, mais nécessitant un regroupement en classes, car le nombre de modalités est trop grand),
- la participation aux tours de l'élection (caractère qualitatif présentant 4 modalités).

Ce type de diagramme est appelé **en tuyaux d'orgue**.

Désormais, nous restreindrons encore davantage le cadre de notre étude, en nous plaçant dans le cas où les deux caractères considérés sont quantitatifs, avec des effectifs tous inférieurs ou égaux à 1. On obtient alors une série de couples — on ne conserve que ceux dont l'effectif est non nul — qui peut être indexée par un seul paramètre. Cette série sera ainsi notée  $(x_i; y_i)$ , pour  $i = 1, 2, \dots, N$ . Graphiquement, ces couples correspondent à des points du plan, rapporté à un repère ; on dit que l'ensemble de ces points est le **nuage de points** associé à la série.

- **EXEMPLE 4.** Lors d'un examen d'une durée de 4 heures, on a relevé la durée de composition (c'est-à-dire au bout de combien de temps chacun a rendu sa copie) et la note (sur 20) des 12 étudiants qui se sont présentés.

| numéro étudiant $i$        | 1      | 2      | 3      | 4   | 5      | 6   | 7      | 8   | 9   | 10     | 11     | 12     |
|----------------------------|--------|--------|--------|-----|--------|-----|--------|-----|-----|--------|--------|--------|
| note $x_i$                 | 8      | 14     | 9      | 17  | 19     | 3   | 7      | 15  | 12  | 11     | 9      | 16     |
| durée de composition $y_i$ | 2 h 45 | 3 h 25 | 2 h 15 | 4 h | 3 h 50 | 1 h | 3 h 30 | 4 h | 4 h | 3 h 15 | 2 h 50 | 3 h 30 |

Le nuage de points correspondant figure sur le schéma ci-contre.

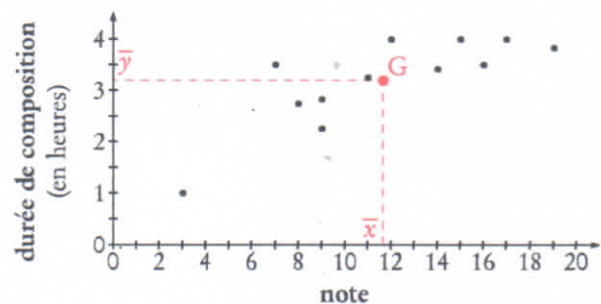
Le **point moyen** du nuage est le point (noté habituellement  $G$ ) dont les coordonnées sont les moyennes marginales de la série :

$$G(\bar{x}; \bar{y}).$$

Pour l'exemple 3, on trouve, après calculs, les moyennes marginales suivantes :

$$\bar{x} \approx 11,67 \quad \text{et} \quad \bar{y} = 3 \text{ h } 11 \text{ min } 40 \text{ s};$$

on en déduit la position du point moyen sur le graphique ci-contre.



### 3 Ajustements affines

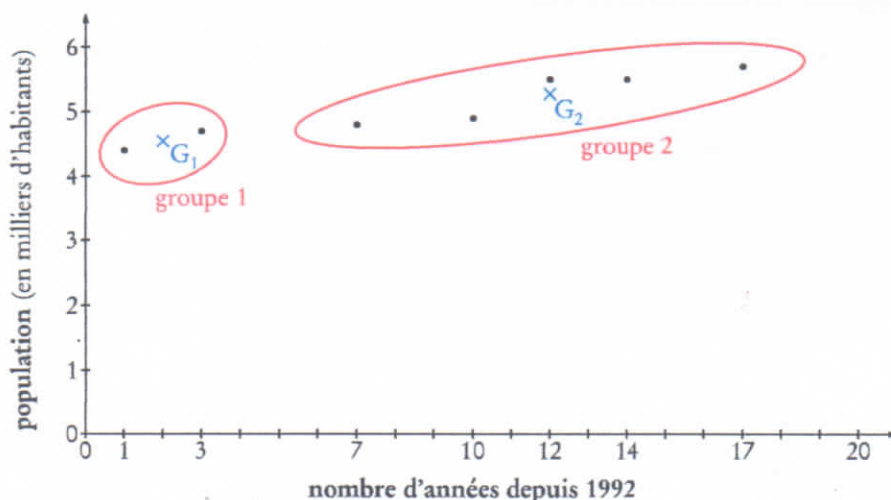
On peut avoir besoin d'effectuer une approximation du nuage par une courbe, dépendant de sa *distribution* (sa « forme »), afin d'obtenir, par exemple, une prévision d'un phénomène économique. C'est cette approximation que l'on appelle **ajustement**. Dans un premier temps, nous allons nous intéresser à des ajustements par des *fonctions affines*, ce qui n'a un sens que lorsque le nuage de points semble distribué le long d'une droite.

#### 3.1 Ajustement affine par la droite de Mayer

Une des méthodes les plus simples consiste à partager le nuage de points en deux groupes (choisis en fonction de la forme du nuage, mais, très souvent, on partage le nuage en deux groupes de points de même effectif). Chacun de ces nouveaux nuages admet un point moyen ( $G_1$  pour le premier nuage,  $G_2$  pour le second); la droite ( $G_1G_2$ ), appelée **droite de Mayer**, constitue alors une approximation du nuage.

► **EXEMPLE 5.** Les derniers recensements de la population de la ville de Carfain ont abouti aux données suivantes :

| année                                      | 1993 | 1995 | 1999 | 2002 | 2004 | 2006 | 2009 |
|--------------------------------------------|------|------|------|------|------|------|------|
| nombre d'années $x_i$ depuis 1992          | 1    | 3    | 7    | 10   | 12   | 14   | 17   |
| population $y_i$ (en milliers d'habitants) | 4,4  | 4,7  | 4,8  | 4,9  | 5,5  | 5,5  | 5,7  |



Compte tenu de la distribution des points du nuage ci-contre (ils sont « plus ou moins » alignés), un ajustement affine est ici approprié.

Nous allons partager les points du nuage représenté ci-contre en deux groupes : le premier sera formé des deux premiers points (correspondant aux modalités 1 et 3 de  $x$ ) et le second des cinq derniers points (correspondant aux autres modalités de  $x$ ).

Chacun de ces nuages admet un point moyen ( $G_1$  pour le premier,  $G_2$  pour le second), dont on calcule les coordonnées :

$$x_{G_1} = \frac{1+3}{2} = 2 \quad \text{et} \quad y_{G_1} = \frac{4,4+4,7}{2} = 4,55$$

pour  $G_1$ , et

$$x_{G_2} = \frac{7+10+12+14+17}{5} = 12 \quad \text{et} \quad y_{G_2} = \frac{4,8+4,9+5,5+5,5+5,7}{5} = 5,28$$

pour  $G_2$ . La droite de Mayer du nuage passe donc par les points  $G_1(2; 4,55)$  et  $G_2(12; 5,28)$ . Déterminons à présent son équation.

Elle admet une équation de la forme  $y = ax + b$ . Les points  $G_1$  et  $G_2$  lui appartiennent ; on aboutit donc au système

$$\begin{cases} 2a + b = 4,55 \\ 12a + b = 5,28 \end{cases}$$

En soustrayant membre à membre ces deux équations, on trouve  $-10a = -0,73$ , d'où  $a = 0,073$ . Par suite :

$$b = 4,55 - 2 \times 0,073 = 4,404.$$

La droite de Mayer du nuage admet, par conséquent, pour équation  $y = 0,073x + 4,404$ . Elle nous permet d'obtenir, entre autres, une estimation de la population en 2012 ; pour ce faire, il suffit de calculer  $y$  pour  $x = 20$  (car 20 années séparent 1992 et 2012) :

$$y = 0,073 \times 20 + 4,404 = 5,864.$$

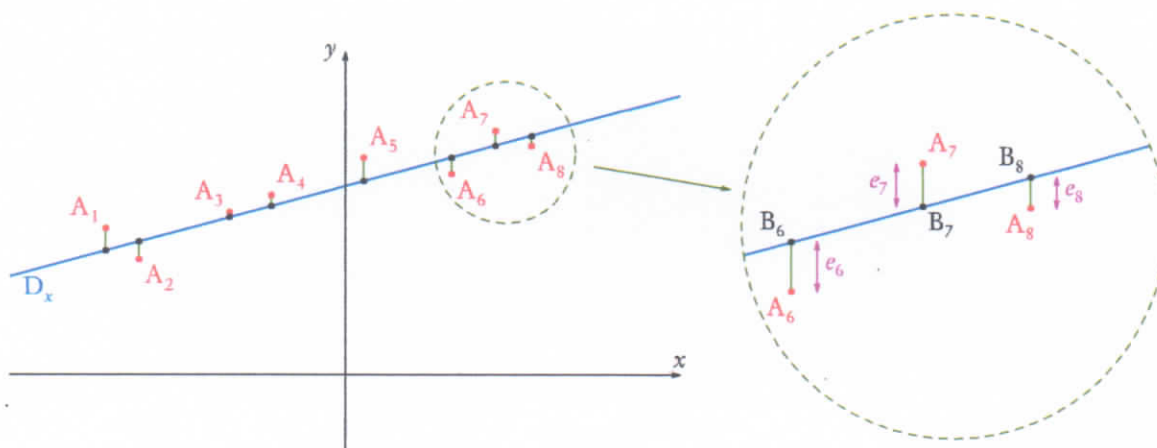
On peut donc estimer que la population de cette ville devrait s'élever, en 2012, à 5864 habitants.

Il est à noter que, bien entendu, la droite de Mayer d'un nuage dépend de la façon dont on forme les deux groupes de points, mais que le point moyen du nuage (G) se trouve sur la droite de Mayer, quel que soit le regroupement effectué.

### 3.2 Ajustement affine par la méthode des moindres carrés

Le moyen le plus couramment employé pour approcher un nuage de points par une droite est la méthode des moindres carrés. L'objectif est ici de choisir la droite de telle sorte que la somme des carrés des « distances » entre les points du nuage et la droite soit la plus petite possible. On va préciser dans ce qui suit ce que l'on entend par « distance » (car ce terme n'est pas à prendre au sens mathématique habituel).

La droite d'ajustement (**droite de régression de  $y$  en  $x$** ) que l'on cherche à obtenir, et que l'on notera  $D_x$ , a pour équation  $y = ax + b$ .



Plaçons-nous dans le cas général, en considérant le nuage formé par des points  $A_1(x_1; y_1), A_2(x_2; y_2), \dots, A_N(x_N; y_N)$ , et considérons les points  $B_1, B_2, \dots, B_N$ , de mêmes abscisses que  $A_1, A_2, \dots, A_N$ , mais situés sur la droite  $D_x$  (d'équation  $y = ax + b$ ) ; autrement dit, les ordonnées des points  $B_1, B_2, \dots, B_N$  sont respectivement  $ax_1 + b, ax_2 + b, \dots, ax_N + b$ .

L'écart absolu  $e_i = A_i B_i$ , ou **résidu** (qui mesure la distance entre les points  $A_i$  et  $B_i$ ), a donc pour carré  $e_i^2 = [y_i - (ax_i + b)]^2$ . La méthode des moindres carrés consiste à trouver les coefficients  $a$  et  $b$  pour lesquels la somme

$$\sum_{i=1}^N A_i B_i^2 = \sum_{i=1}^N e_i^2 = \sum_{i=1}^N [y_i - (ax_i + b)]^2$$

est minimale (la droite d'équation  $y = ax + b$  est alors la plus « proche » des points  $A_1, A_2, \dots, A_N$ ). Des calculs élémentaires montrent que, sous cette contrainte, le coefficient directeur de la droite  $D_x$  et son ordonnée à l'origine sont donnés par les formules

$$a = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2} \quad \text{et} \quad b = \bar{y} - a\bar{x},$$

la dernière égalité traduisant le fait que le point moyen du nuage se trouve sur la droite de régression  $D_x$ . Puisque la variance de  $x$  est

$$V(x) = \sigma_x^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2,$$

on est naturellement amené à définir la **covariance** de  $x$  et  $y$ , notée  $\text{Cov}(x, y)$  ou  $\sigma_{x,y}$ , par la formule

$$\text{Cov}(x, y) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}).$$

C'est une « généralisation » à deux variables statistiques de la notion de variance ; en effet,  $\text{Cov}(x, x) = V(x)$ . La covariance de  $x$  et  $y$  est l'expression qui apparaît naturellement lorsque l'on calcule la variance de la variable  $(x + y)$  :

$$V(x + y) = \sigma_{x+y}^2 = \sigma_x^2 + 2\sigma_{x,y} + \sigma_y^2 = V(x) + 2\text{Cov}(x, y) + V(y),$$

(qui rappelle l'identité remarquable  $(a + b)^2 = a^2 + 2ab + b^2$ ). De la même façon que

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

(cette dernière égalité constitue le *théorème de König-Huygens*), on prouve que

$$\text{Cov}(x, y) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N} \sum_{i=1}^N x_i y_i - \bar{x} \bar{y}.$$

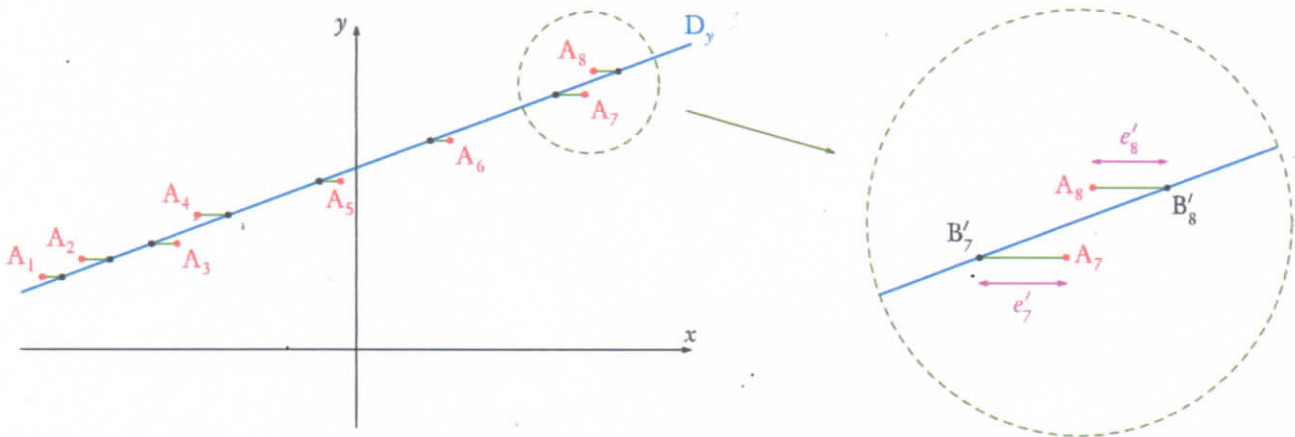
En résumé, la droite de régression  $D_x$ , de  $y$  en  $x$ , a pour équation  $y = ax + b$ , avec

$$a = \frac{\text{Cov}(x, y)}{V(x)} = \frac{\sigma_{x,y}}{\sigma_x^2} \quad \text{et} \quad b = \bar{y} - a\bar{x}.$$

On peut également déterminer la **droite de régression de  $x$  en  $y$**  du nuage : il s'agit de la droite d'équation  $x = ay + b$  pour laquelle la somme

$$\sum_{i=1}^N [x_i - (ay_i + b)]^2 = \sum_{i=1}^N e_i'^2 = \sum_{i=1}^N A_i B_i'$$

est minimale (les écarts absolus  $e_i'$  sont cette fois mesurés parallèlement à l'axe des  $x$ ).



On démontre que ce minimum est atteint pour la droite  $D_y$  d'équation  $x = ay + b$  (autrement dit,  $y = \frac{1}{a}(x - b)$ ), où les coefficients  $a$  et  $b$  sont donnés par les formules

$$a = \frac{\text{Cov}(x, y)}{V(y)} = \frac{\sigma_{x,y}}{\sigma_y^2} \quad \text{et} \quad b = \bar{x} - a\bar{y}.$$

On peut remarquer que ce sont les mêmes formules que pour la droite  $D_x$  ; seuls  $x$  et  $y$  ont été intervertis.

Revenons sur la situation présentée à l'exemple 5 et essayons de donner une prévision de la population de Carfain en 2012 par la méthode des moindres carrés. Le point moyen du nuage est  $G(\bar{x}; \bar{y})$ , avec

$$\bar{x} = \frac{1 + 3 + 7 + 10 + 12 + 14 + 17}{7} = \frac{64}{7} \approx 9,1429 \quad \text{et} \quad \bar{y} = \frac{4,4 + 4,7 + 4,8 + 4,9 + 5,5 + 5,5 + 5,7}{7} = \frac{35,5}{7} \approx 5,0714.$$

La droite de régression  $D_x$ , de  $y$  en  $x$ , a pour équation  $y = ax + b$ , avec

$$a = \frac{\text{Cov}(x, y)}{V(x)} \approx \frac{\frac{1}{7} \times (1 \times 4,4 + 3 \times 4,7 + 7 \times 4,8 + 10 \times 4,9 + 12 \times 5,5 + 14 \times 5,5 + 17 \times 5,7) - 9,1429 \times 5,0714}{\frac{1}{7} \times (1^2 + 3^2 + 7^2 + 10^2 + 12^2 + 14^2 + 17^2) - 9,1429^2} \approx 0,081$$

et

$$b = \bar{y} - a\bar{x} \approx 5,0714 - 0,081 \times 9,1429 \approx 4,331.$$

Par conséquent, pour  $x = 20$  (ce qui correspond à l'année 2012), on trouve  $y = 0,081 \times 20 + 4,331 = 5,951$ . Selon ce modèle, on peut estimer la population de Carfain, en 2012, à 5951 habitants. (Pour mémoire, avec la droite de Mayer, on avait obtenu 5864 habitants.)

La droite de régression  $D_y$  de  $x$  en  $y$ , quant à elle, a pour équation  $x = ay + b$ , avec

$$a = \frac{\text{Cov}(x, y)}{V(y)} \approx \frac{\frac{1}{7} \times (1 \times 4,4 + 3 \times 4,7 + 7 \times 4,8 + 10 \times 4,9 + 12 \times 5,5 + 14 \times 5,5 + 17 \times 5,7) - 9,1429 \times 5,0714}{\frac{1}{7} \times (4,4^2 + 4,7^2 + 4,8^2 + 4,9^2 + 5,5^2 + 5,5^2 + 5,7^2) - 5,0714^2} \approx 11,297$$

et

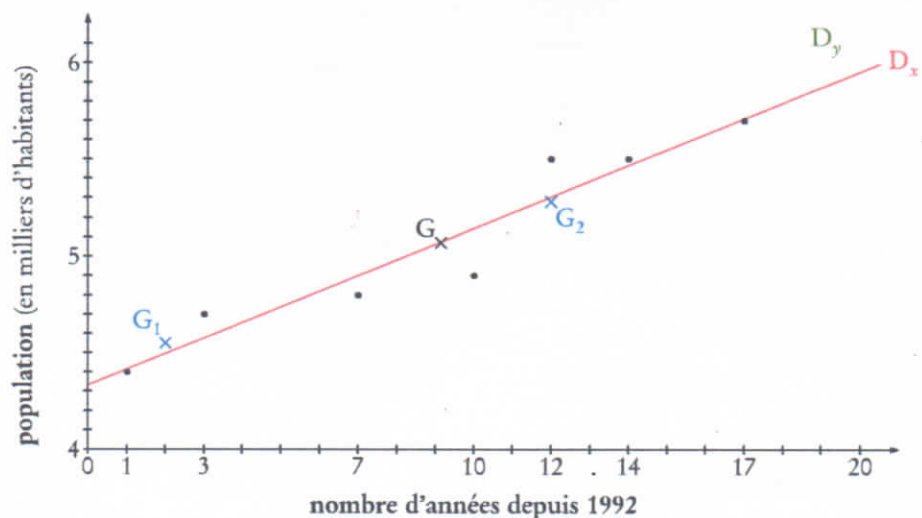
$$b = \bar{x} - a\bar{y} \approx -48,147.$$

Pour trouver la valeur de  $y$  correspondant à  $x = 20$ , on résout l'équation

$$11,297y - 48,147 = 20 \iff 11,297y = 20 + 48,147 = 68,147 \iff y = \frac{68,147}{11,297} \approx 6,033$$

d'où 6033 habitants selon ce modèle.

Pour mieux visualiser les différences obtenues entre les résultats, représentons le nuage de points avec les 3 droites déterminées jusque-là : ( $G_1, G_2$ ),  $D_x$  et  $D_y$ .



On définit le **coefficient de corrélation linéaire** de la série  $(x_i; y_i)$ ; il s'agit du nombre

$$r_{x,y} = \frac{\text{Cov}(x, y)}{\sqrt{V(x)V(y)}} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y}.$$

Ce réel  $r_{x,y}$ , exprimé **sans unité**, fait *toujours* partie de l'intervalle  $[-1; 1]$ . Lorsque  $r_{x,y}$  est proche de 1 ou de  $-1$ , cela traduit une forte dépendance (statistique) affine entre les valeurs de  $x$  et celles de  $y$  et motive un ajustement affine entre les deux variables. Plus  $r_{x,y}$  est proche de 0, moins les points du nuage sont répartis le long d'une ligne droite; il faut alors envisager un autre type d'ajustement (cf. paragraphe suivant), sous réserve que le nuage ait une forme « allongée ».

**Attention! Même une corrélation linéaire parfaite ( $r = \pm 1$ ) ne garantit pas l'existence d'un lien de cause à effet entre les variables  $x$  et  $y$ .**

Dans le cas de l'exemple 5, le calcul du coefficient de corrélation linéaire donne

$$r_{x,y} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y} \approx \frac{2,347}{5,383 \times 0,456} \approx 0,956.$$

Ceci indique qu'un ajustement affine est tout à fait adapté à la situation.

Il est à noter que, de plus en plus, pour laisser de côté le signe du coefficient  $r_{x,y}$  (qui n'a que peu d'importance), on préfère s'intéresser au **coefficient de détermination** de la série, qui n'est autre que le carré de  $r_{x,y}$ .

## 4 Autres ajustements

Un ajustement affine est pertinent lorsque les points du nuage semblent être répartis le long d'une ligne droite. Lorsque le nuage est « allongé » mais a une autre forme, on a recours à un ajustement différent ; il en existe de nombreux, reposant sur différentes familles de fonctions : les fonctions logarithmes, exponentielles, puissances, logistiques, etc.

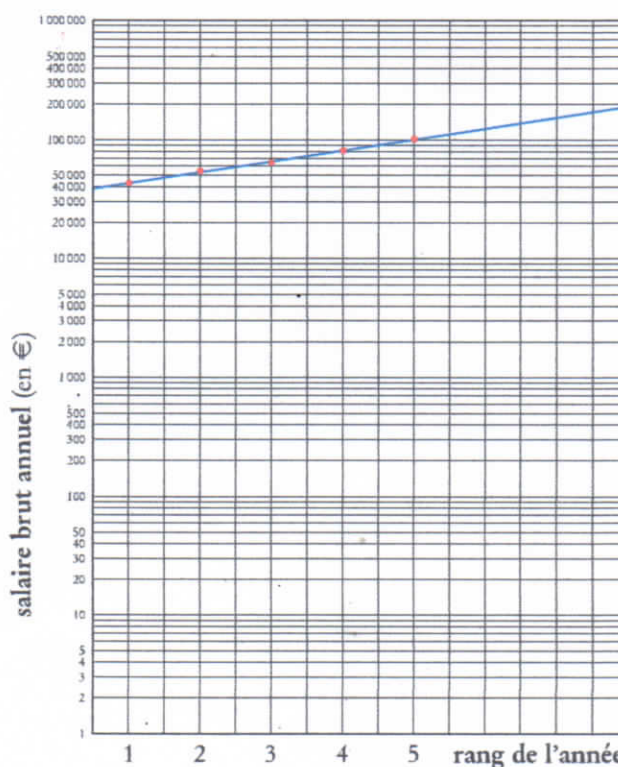
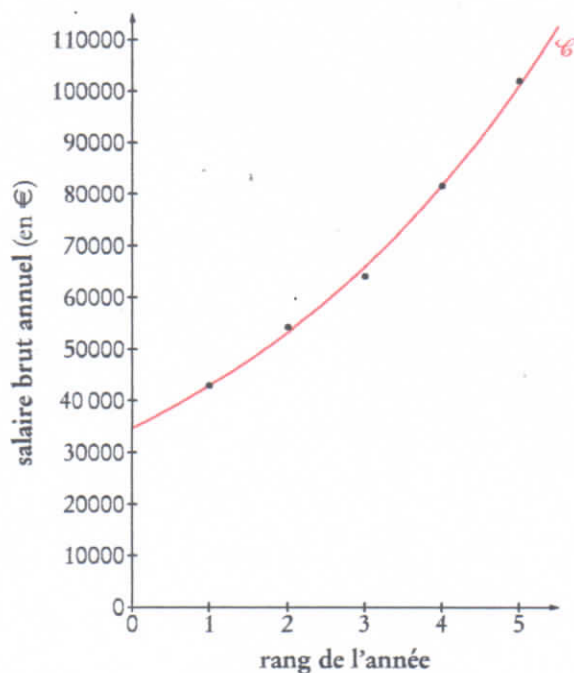
### 4.1 Ajustement exponentiel

Pour les phénomènes à croissance forte ou à décroissance rapide, il peut être pertinent d'approcher la forme générale du nuage par une fonction exponentielle de la forme  $x \mapsto e^{ax+b}$  ou  $x \mapsto \alpha e^x + \beta$ . Pour déterminer les valeurs de  $a$  et  $b$ , ou de  $\alpha$  et  $\beta$ , on effectue un ajustement affine (par exemple, par la méthode des moindres carrés) sur la série  $(x_i; \ln y_i)$  ou la série  $(e^{x_i}; y_i)$ .

► **EXEMPLE 6.** Le tableau ci-dessous indique le salaire brut annuel, en euros, perçu par un cadre.

Le coefficient de corrélation linéaire entre  $x$  et  $y$  est tout à fait honorable ( $r_{x,y} \approx 0,988$ ), mais la répartition des points au sein du nuage laisse à penser que l'augmentation du salaire de ce cadre est de plus en plus rapide au fil des années, phénomène que ne modéliserait pas une ligne droite.

| année                 | 2005   | 2006   | 2007   | 2008   | 2009    |
|-----------------------|--------|--------|--------|--------|---------|
| rang $x_i$ de l'année | 1      | 2      | 3      | 4      | 5       |
| salaire $y_i$ (en €)  | 42 900 | 54 200 | 64 100 | 81 600 | 102 000 |
| $z_i = \ln(y_i)$      | 10,666 | 10,900 | 11,068 | 11,310 | 11,533  |



Le coefficient de corrélation linéaire entre  $x$  et  $z = \ln(y)$  est très bon ( $r_{x,z} \approx 0,999$ ), ce qui justifie un ajustement affine. Par la méthode des moindres carrés, on trouve une droite de régression  $D_x$  de  $z$  en  $x$ , d'équation  $z = 0,21x + 10,45$ . On en déduit