Statistiques

Table des matières

1 Statistique Descriptive pour Une Variable

1.1 Présentation

1.1.1 Étapes d'une statistique

Collecte des données Les observations sont effectuées au sein d'une population, relativement à un caractère, les résultats constituent une *série statistique.*
Par exemple les âges des élèves d'une classe, ou encore les nombres d'élèves du lycée reçus au BTS en 1997 ...

Analyse des données Il s'agit de la détermination de paramètres statistiques (effectifs, moyenne ...) qui permettent de caractériser la série statistique.

Interprétation des résultats À l'aide de propriétés mathématiques et en élaborant des tests on espère obtenir des indications suffisantes pour une exploitation des résultats (Études de marchés par exemple).

1.1.2 Vocabulaire statistique

Population La population est l'ensemble étudié, les éléments de cette population sont appelés *individus* ou encore *unités statistiques.*
Par exemple l'étude du parc automobile en France se fera sur un ensemble de véhicules, la population est cet ensemble, les individus sont les véhicules.

Échantillon Lorsque la population est importante on préfère prélever *au hasard* ou en tenant compte de certains critères, une partie ou sous-ensemble de cette population c'est l'échantillon. S'il est prélevé au hasard on dira que c'est un *échantillon aléatoire*.

Variable statistique

Définition 1.1 *La valeur statistique ou encore valeur du caractère est la mesure associée au caractère après avoir choisi une unité qui sera précisée.*

Les différentes valeurs obtenues constituent la *variable statistique*.
Par exemple les âges des automobiles dans un échantillon seront 5, 2, ... (années)
On distinguera deux types de variables statistiques suivant la nature mathématique de l'ensemble des valeurs que le caractère est susceptible de prendre.
Lorsque les valeurs du caractère sont isolées et appartiennent à un ensemble fini de nombres ou encore appartiennent à un ensemble infini tel que N, Z, D, Q on dira que la variable est *discrète*.
Par exemple les âges des élèves qui sont des nombres entiers positifs, mais ce pourrait être dans un autre contexte $\{20;\ 20,5;\ 21;\ 21,5\ldots\}$ qui sont des décimaux (dont le nombre de chiffres à droite de la virgule est fini : ici au plus un chiffre).
Ne pas confondre les décimaux et les réels qui ont des écritures décimales illimitées et ne sont pas tous décimaux.
On convient d'ordonner ces valeurs dans l'ordre croissant $x_1 < x_2 < x_3 < \ldots$
Au contraire, lorsque la variable peut prendre n'importe quelle valeur de R ou d'un intervalle réel, on dit qu'elle est *continue*

Dans le cas d'une variable continue on réalise une partition de R ou de l'intervalle de R contenant les valeurs de la variable en k classes qui sont notés $[a_0;\ a_1[$, $[a_1;\ a_2[$, ..., $[a_i;\ a_{i+1}[$,..., $[a_{k-1};\ a_k[$ ou bien $[a_1;\ a_2[$, ..., $[a_i;\ a_{i+1}[$,..., $[a_{k-1};\ a_k[$, $[a_k;\ a_{k+1}[$ selon le cas.
Les centres des classes $[a_i;\ a_{i+1}[$ sont les réels $c_i = \frac{a_i + a_{i+1}}{2}$.

Effectif, fréquence

Définition 1.2 *L'effectif de la valeur x_i d'une variable est le nombre n_i d'observations de la valeur x_i dans le cas discret ou le nombre d'observations dans la classe $[a_i;\ a_{i+1}[$ dans le cas continu.*

en biologie, l'effectif est appelé la *fréquence absolue.*
L'*effectif total* est le nombre total d'observations, c'est la somme de tous les effectifs.

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

Définition 1.3 *Les fréquences sont les quotients des effectifs des valeurs du caractère ou des effectifs des classes par l'effectif total.*

$$f_i = \frac{n_i}{N} = \frac{n_i}{\sum_{i=1}^{i=k} n_i}$$

Propriété 1.4 *Pour tout i on a* $\quad 0 \leq f_i \leq 1.$

En effet $0 \leq n_i \leq \sum_{i=1}^{i=k} n_i$ et donc en divisant par $N = \sum_{i=1}^{i=k} n_i$ on a $0 \leq \frac{n_i}{N} \leq \frac{N}{N} = 1$.

Propriété 1.5 *La somme des fréquences est 1,*

$$\sum_{i=1}^{i=k} f_i = 1.$$

$\sum_{i=1}^{i=k} f_i = \sum_{i=1}^{i=k} \frac{n_i}{N} = \frac{1}{N} \sum_{i=1}^{i=k} n_i = \frac{1}{N} N = 1.$

Histogramme : Les aires des rectangles de l'histogramme des effectifs sont proportionnelles aux effectifs des classes, on ne porte pas d'échelle sur le second axe lorsque les classes n'ont pas la même amplitude.

Exemple 1 *(Voir fig.1)*

S'il s'agit de l'histogramme des fréquences on agit de même.

Série statistique

Définition 1.6 *Une série statistique est l'ensemble des couples $(x_i; n_i)$ ou $([a_i; a_{i+1}[;\ n_i)$.*

amplitude de la classe	5	5	10	20
classe	$[0;\ 5[$	$[5;\ 10[$	$[10;\ 20[$	$[20;\ 40[$
effectif	4	12	12	2
$2,5\times$ effectif / amplitude ($2,5$ est arbitraire)	2	6	3	$0,25$
hauteur en cm du rectangle	2 cm	6 cm	3 cm	$0,25$ cm

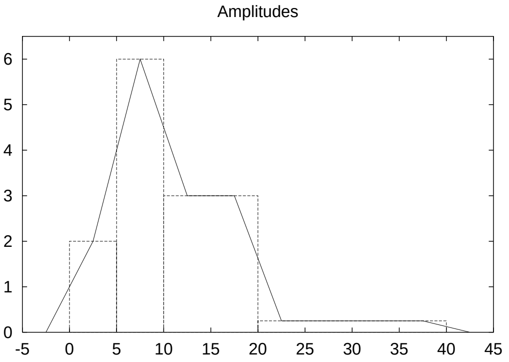

FIG. 1 – *Aires et Histogrammes.*

Effectifs cumulés croissants, fréquences cumulées croissantes

Définition 1.7 *Le tableau des effectifs cumulés croissants s'obtient en associant à chaque classe $[a_i;\ a_{i+1}[$, $1 \leq i \leq k$, la somme des effectifs $\nu_i = \sum_{t=1}^{t=i} n_t = n_1 + n_2 + ... + n_t + ... + n_i$.*

En général on dessine l'histogramme des effectifs cumulés croissants ou le polygone des effectifs cumulés croissants.
On définit de même les fréquences cumulées croissantes $\phi_i = \sum_{t=1}^{t=i} f_t = \frac{\nu_i}{N}$. On dessine aussi l'histogramme ou le polygone des fréquences cumulées croissantes.

Effectifs cumulés décroissants, fréquences cumulées décroissantes Se définissent d'une manière semblable aux effectifs ou aux fréquences cumulés croissants.

Exemple 2 *Dans le cas discret (Fig.2)*

Exemple 3 *Variable continue. À chaque classe $[a_i;\ a_{i+1}[$, $1 \leq i \leq k$, on associe la somme des effectifs $\nu'_i = \sum_{t=i}^{t=k} n_t = n_i + ... + n_t + ... + n_k$.*

1.1.3 Graphiques

Les principaux graphiques sont
Lorsque la variable est discrète ou discontinue : le diagramme en bâtons, le polygone des effectifs ou le polygone des fréquences.
Lorsque la variable est continue : l'histogramme (les aires des rectangles de l'histogramme des effectifs sont proportionnelles aux effectifs des classes).

1.2 Paramètres statistiques

1.2.1 Paramètres de position

Dominante ou mode

x_i	1	2	3	4	5	6	7	Totaux
n_i		2	3	4	8	6	2	25
f_i	0	$0,08$	$0,12$	$0,16$	$0,32$	$0,24$	$0,08$	1
fréquences cumulées croissantes	0	$0,008$	$0,2$	$0,36$	$0,60$	$0,92$	1	
fréquences cumulées décroissantes	1	$0,92$	$0,8$	$0,64$	$0,32$	$0,08$	0	

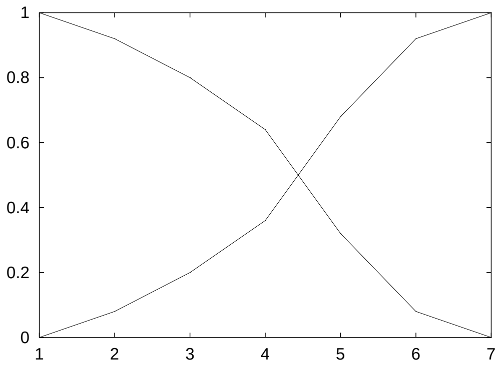

FIG. 2 – *Fréquences cumulées.*

Définition 1.8 *Lorsque la variable est discrète, une dominante ou mode est une valeur du caractère qui correspond à un effectif maximum, la série est unimodale, bimodale ... lorsque le nombre de modes est* 1, 2 *...*
Lorsque la variable est continue, une classe modale correspondra à un effectif maximum.

Remarque 1 *L'existence de plusieurs modes peut permettre de suspecter ou de mettre en évidence l'existence au sein de la population de plusieurs sous-populations d'origines différentes.*

Moyenne

Définition 1.9 *Lorsque la variable est discrète la moyenne $\bar{x}$ de la série statistique est la moyenne pondérée*

$$\bar{x} = \frac{\sum_{i=1}^{i=k} n_i x_i}{\sum_{i=1}^{i=k} n_i} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_k x_k}{N}.$$

Lorsque la variable est continue la moyenne est

$$\bar{x} = \frac{\sum_{i=1}^{i=k} n_i c_i}{\sum_{i=1}^{i=k} n_i}$$

où les $c_i = \frac{a_i + a_{i+1}}{2}$ sont les centres des classes.

Médiane

Définition 1.10 *La médiane est la valeur du paramètre x telle que la moitié de l'effectif total correspond à la fois aux valeurs du paramètre inférieures et aux valeurs du paramètre supérieures à la médiane.*

Propriété 1.11 *La médiane est l'abscisse du point d'intersection des polygones des effectifs cumulés croissants et décroissants. L'ordonnée du point d'intersection des deux polygones est $\frac{N}{2}$.*

Lorsque la variable est continue et que la valeur de la médiane se trouve dans la classe $[a_i;\, a_{i+1}[$, les valeurs des fréquences cumulées croissantes sont ϕ_i pour l'ensemble des classes précédentes et ensuite $\phi_{i+1} = \phi_i + f_i$ (en ajoutant la fréquence f_i de la classe courante $[a_i;\, a_{i+1}[$), alors $\phi_i \leq 0,5 \leq \phi_{i+1}$ et la médiane m se calcule par interpolation linéaire : $m - a_i = \frac{(a_{i+1}-a_i)(0,5-\phi_i)}{\phi_{i+1}-\phi_i}$ qui se simplifie en

$$m = a_i + \frac{e_i(0,5 - \phi_i)}{f_i}$$

où $e_i = a_{i+1} - a_i$ est l'amplitude de la classe $[a_i;\, a_{i+1}[$.
En utilisant les effectifs cumulés on aurait

$$m = a_i + \frac{e_i}{n_i}\left(\frac{N}{2} - \nu_i\right).$$

1.2.2 Paramètres de dispersion

Variance

Définition 1.12 *Lorsque la variable est discrète, la variance est*

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

Lorsque la variable est continue, la variance est

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

où les c_i sont les centres des classes.

Propriété 1.13 *La variance est positive ou nulle : $V \geq 0$.*

Propriété 1.14 *Lorsque la variable est discrète, la variance est*

$$V = \frac{\sum_{i=1}^{i=k} n_i x_i^2}{\sum_{i=1}^{i=k} n_i} - \bar{x}^2.$$

Lorsque la variable est continue, la variance est

$$V = \frac{\sum_{i=1}^{i=k} n_i c_i^2}{\sum_{i=1}^{i=k} n_i} - \bar{x}^2$$

où les c_i sont les centres des classes.

Écart-type

Définition 1.15 *L'écart-type de la série statistique est $\sigma = \sqrt{V}$.*

Lorsque l'étude porte sur un échantillon de la population, dont la moyenne $\bar{x}$ n'est pas connue, on constate que la variance et l'écart-type calculés par les formules précédentes sont inférieurs aux valeurs réelles et on utilise la définition suivante de l'écart-type d'un échantillon de taille N d'une population de grande taille (par rapport à N).

Définition 1.16 *l'écart-type d'échantillon est*

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

où N est la taille de l'échantillon.

Remarque 2 *Sur la calculatrice on peut vérifier que*

$$\sigma_{n-1} = \sqrt{\frac{n}{n-1}}\,\sigma_n$$

où n est l'effectif total.

1.2.3 Changement d'origine et d'échelle

Soit une variable statistique x dont les valeurs du caractère sont notées x_i, $1 \leq i \leq k$, avec des effectifs n_i de somme N et soient deux réels α et β, on peut alors définir une variable statistique y dont les valeurs du caractère sont les $y_i = \alpha x_i + \beta$ avec les mêmes effectifs n_i.

Propriété 1.17 *Avec les notations ci-dessus, on a :*

$$\bar{y} = \alpha\bar{x} + \beta, \qquad \sigma_y = |\alpha|\sigma_x$$

où $\bar{x}$, $\bar{y}$ sont les valeurs moyennes et σ_x, σ_y les écarts-types des deux variables.

$\bar{y} = \frac{1}{N}\sum n_i y_i = \frac{1}{N}\sum(\alpha n_i x_i + \beta) = \frac{1}{N}\left((\alpha \sum n_i x_i + N\beta)\right) = \alpha\frac{1}{N}\sum n_i x_i + \beta = \alpha\bar{x} + \beta.$

$\sigma_y^2 = \frac{1}{N}\sum n_i(y_i - \bar{y})^2 = \frac{1}{N}\sum n_i(\alpha x_i + \beta - \alpha\bar{x} - \beta)^2 = \frac{1}{N}\sum \alpha^2 n_i(x_i - \bar{x})^2 = \alpha^2\frac{1}{N}\sum n_i(x_i - \bar{x})^2 = \alpha^2\sigma_x^2.$

Exemple 4 *Utilisation du changement de variable pour le calcul d'une moyenne.*

x_i	7.1	7.2	7.5	8.4	8.6	8.9	9
n_i	1	2	3	6	4	3	1
$y_i = 10x_i - 80$	-9	-8	-5	4	6	9	10

$\bar{y} = \frac{-9-16-15+24+24+27+10}{20} = \frac{45}{20} = 2,25$ *donc* $\bar{x} = \frac{2,25+80}{10} = 8,225.$

Exemple 5 *Série particulière*

x_i	3	4	5	6	7
n_i	1	2	4	2	1
$y_i = x_i - 5$	-2	-1	0	1	2

Le calcul $\bar{y} = 0$ est immédiat d'où $\bar{x} = 5$.
$\sigma_y^2 = \frac{1}{10}(4+2+2+4) - 0^2 = 1,6$ *donc* $\sigma_x = \sigma_y = \sqrt{1,6} \approx 1,26.$

1.2.4 Centrage et réduction d'un caractère

Par un changement d'origine $y = x - \bar{x}$, il est possible d'obtenir un caractère y de moyenne nulle $\bar{y} = 0$.

Par un changement d'échelle (ou d'unité de mesure), si $\sigma_x \neq 0$, et $y = \frac{x}{\sigma_x}$ on obtient un caractère y d'écart-type 1.

En combinant les deux et en prenant

$$y = \frac{x - \bar{x}}{\sigma_x}$$

on obtient un caractère y centré et réduit, c'est-à-dire de moyenne nulle et d'écart-type 1.

Exemple 6 *Série centrée et réduite.*

x_i	3	5	6	8	10
n_i	1	3	4	1	1
$y_i = \frac{x_i - 6}{\sigma_x}$	$-1,68$	-0.56	0	$1,12$	$2,24$

$\bar{x} = 6$ et $\sigma_x = \sqrt{3,2} \approx 1,79$ d'où $y = \frac{x-\bar{x}}{\sigma_x} \approx \frac{x-6}{1,79}$ est centrée et réduite (du moins approximativement, dans le tableau, étant donné les erreurs d'arrondis).

1.2.5 Écart moyen à la moyenne

Avec les notations habituelles

Définition 1.18 *Soit la variable statistique x, l'écart moyen à la moyenne est*

$$Em = \frac{1}{N} \sum n_i |x_i - \bar{x}|$$

Exemple 7 *Calcul d'un écart moyen à la moyenne*

x_i	3	5	6	8	10
n_i	1	3	4	1	1

$\bar{x} = 6$ et $Em = \frac{1\times3+3\times1+4\times0+1\times2+1\times4}{10} = \frac{12}{10} = 1,2$
En moyenne les valeurs observées sont, en plus ou en moins, écartées de $1,2$ de la valeur moyenne 6 de la série.

2 Étude Conjointe de Deux Variables

2.1 Série statistique double

2.1.1 Deux caractères

On étudie pour une même population P deux caractères qualitatifs ou quantitatifs et on définit une série statistique sur l'ensemble des couples $(x; y)$ de valeurs des deux caractères. Dans ce qui suit, sauf indication contraire, chaque couple a pour effectif 1 et on n'utilisera pas de notation n_i pour ces effectifs, l'effectif total N est alors le nombre des couples $(x_i; y_i)$ de la série et i varie donc de 1 à N.

Âge en années x	36	42	48	54	60	66
Tension maximale y	$11,6$	$13,2$	14	$14,4$	$15,5$	$15,1$

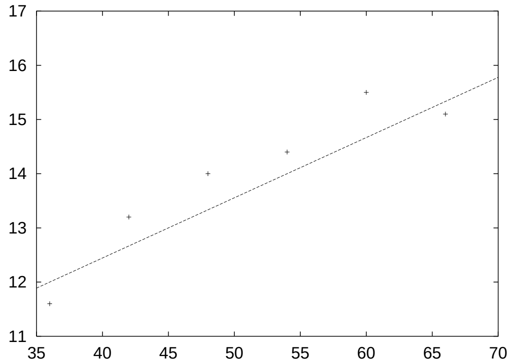

FIG. 3 – *Tension maximale.*

2.1.2 Nuage de points

Dans un repère orthogonal on représente les points $M\ (x;\, y)$ dont les coordonnées sont les couples de valeurs des deux caractères, l'ensemble de ces points est communément appelé le *nuage de points.*

2.1.3 Point moyen

Définition 2.1 *Le point moyen G du nuage de points de la série statistique double est le point de coordonnées $x = \frac{1}{N}\sum x_i$ et $y = \frac{1}{N}\sum y_i$, où N est le nombre de points du nuage.*

Exemple 1 *Le tableau (voir Fig. 3) donne, dans une population féminine, la moyenne de la tension artérielle maximale en fonction de l'âge.*
La droite tracée est la droite de Mayer, celle-c passe par le point moyen G du nuage (le placer).

2.2 Ajustement affine par une méthode graphique

2.2.1 Ajustement à la règle

Lorsque les points du nuage semblent presque alignés il peut être envisageable de rechercher une relation $y = ax + b$ ou $x = a'y + b'$ entre les deux caractères.
En traçant une droite L la plus proche possible de tous les points, entre ces points et dans la direction qu'ils suggèrent on obtient rapidement une assez bonne approximation de la relation $y = ax+b$. Si besoin est, les coefficients a et b se calculent à partir des coordonnées de deux points de L, mais généralement la méthode n'est utilisée que pour des lectures graphiques (interpolation ou extrapolation).

2.2.2 Droite de Mayer

La droite de Mayer passe par le point moyen G du nuage et par deux autres points moyens G_1 et G_2 de deux moitiés du nuage obtenus en prenant les N_1 premiers points et les $N_2 = N - N_1$ autres.
Lorsque $N_1 = N_2$ le point G est le milieu du segment $[G_1G_2]$.

2.3 Méthodes utilisant des moyennes, lissage

2.3.1 méthode des moyennes mobiles

Consiste à remplacer M_1, M_2, ..., M_p par M'_1 $(x_p; \frac{y_1+\cdots+y_p}{p})$, puis les p points suivants par un point $M'_2(x_2p; \frac{y_{p+1}+\cdots+y_{2p}}{p})$ et aisi de suite.

2.3.2 méthode des moyennes échelonnées

Consiste à remplacer M_1, M_2, ..., M_p par leur point moyen M'_1 $(\frac{x_1+\cdots+x_p}{p}; \frac{y_1+\cdots+y_p}{p})$ et de recommencer avec les p points suivants ...

2.3.3 méthode des moyennes discontinues

Consiste à remplacer plusieurs points de même valeur du caractère x par un seul point moyen (ce point aura donc la même abscisse x que ceux qu'il remplace).

2.4 Ajustement affine par la méthode des moindres carrés

2.4.1 Covariance d'une série statistique double

Définition 2.2 *La covariance de la série double* $(x; y)$ *est*

$$\sigma_{xy} = \frac{1}{N} \sum_{i=1}^{i=N} [(x_i - \bar{x})(y_i - \bar{y})]$$

Les variances de x et de y se notent $\sigma_x^2 = \frac{1}{N} \sum[(x_i - \bar{x})^2]$ et $\sigma_y^2 = \frac{1}{N} \sum[(y_i - \bar{y})^2]$.
les écarts-types de x et de y sont σ_x et σ_y.

2.4.2 Régression linéaire de y en x

Soient le nuage des points M_i $(x_i; y_i)$ et une droite quelconque D d'équation $y = ax + b$, non parallèle au second axe.
Soient alors les points P_i $(x_i; ax_i + b)$ de mêmes abscisses que les points M_i et situés sur la droite D.
Si la droite D passait par tous les points M_i, ont aurait $P_i = M_i$ et les distances M_iP_i seraient nulles, mais en général une telle droite D n'existe pas.
En déterminant une droite D telle que la somme $S = \sum M_iP_i{}^2$ des carrés des distances M_iP_i soit minimale on obtiendra la droite appelée « droite de régression linéaire de y en x ». (Voir Fig. 4).

Le calcul ci-dessous met en évidence les variances σ_x^2, σ_y^2 et la covariance σ_{xy}, il n'est pas une démonstration complète des propriétés :
$S = \sum M_iP_i{}^2 = \sum(y_i - ax_i - b)^2 = \sum(y_i - \bar{y} - a(x_i - \bar{x}) + \bar{y} - a\bar{x} - b)^2$
en prenant $\bar{y} - a\bar{x} - b = 0$ on obtient une droite passant par le point moyen G du nuage et la somme devient
$S = \sum(y_i - \bar{y} - a(x_i - \bar{x}))^2 = \sum [(y_i - \bar{y})^2 - 2a(y_i - \bar{y})(x_i - \bar{x}) + a^2(x_i - \bar{x})]$
$S = N(\sigma_y^2 - 2a\sigma_{xy} + a^2\sigma_x^2)$
En prenant $a = \frac{\sigma_{xy}}{\sigma_x^2}$
$S = N(\sigma_y^2 - 2\sigma_{xy}^2 + \sigma_{xy}^2) = N(\sigma_y^2 + \sigma_{xy}^2)$.
On montre que c'est la valeur minimale de la somme et on admettra les propriétés suivantes.

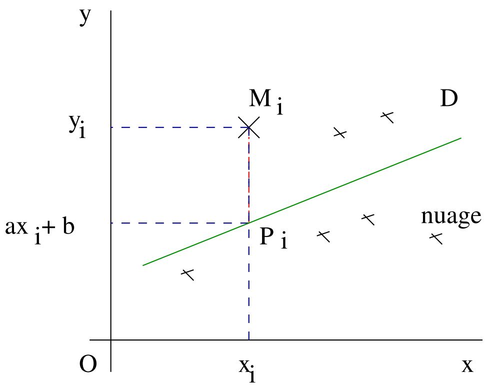

FIG. 4 – *Droite de régression linéaire de y en x.*

Définition 2.3 *La droite de régression de y en x est la droite D : $y = ax + b$ passant par le point moyen G du nuage et de coefficient directeur*

$$a = \frac{\sigma_{xy}}{\sigma_x^2}$$

Le calcul du coefficient b de l'équation se fait à l'aide des coordonnées du point G $(\bar{x}; \bar{y})$ de la droite D,

$$b = \bar{y} - a\bar{x}$$

Le signe de a est le signe de la covariance σ_{xy}.

2.4.3 Régression linéaire de x en y

On peut remarquer que $\sigma_{yx} = \sigma_{xy}$.
Par symétrie, en échangeant les lettres x et y dans les explications et résultats du paragraphe précédent :
On cherche une droite D' : $x = a'y + b'$ non parallèle au premier axe, telle que la somme des carrés des longueurs M_iQ_i soit minimale.
(Les points Q_i sont sur D' et ont pour ordonnées les mêmes y_i que les M_i correspondants).

Définition 2.4 *La droite de régression de x en y est la droite D' : $x = a'y + b'$ passant par le point moyen G du nuage et de coefficient directeur*

$$a' = \frac{\sigma_{xy}}{\sigma_y^2}$$

Le calcul du coefficient b' de l'équation se fait encore à l'aide des coordonnées du point G $(\bar{x}; \bar{y})$ de la droite D',

$$b' = \bar{x} - a'\bar{y}$$

(Voir Fig. 5).

Le signe de a est le signe de la covariance σ_{xy}.
On peut voir que $a' = a\frac{\sigma_x^2}{\sigma_y^2}$.

Lorsque $a' \neq 0$, l'équation $x = a'y + b'$ peut s'écrire $y = \frac{1}{a'} - \frac{b}{a'}$ et D' a pour coefficient directeur $\frac{1}{a'}$.

Le produit des coefficients directeurs des droites D et D' est donc égal à $\frac{a}{a'} = \frac{\sigma_y^2}{\sigma_x^2}$

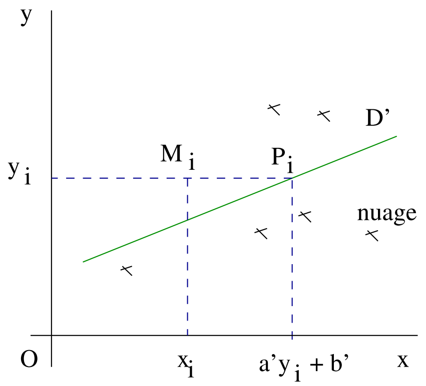

FIG. 5 – *Droite de régression linéaire de x en y.*

3 Corrélation linéaire

3.1 Coefficient de corrélation linéaire

Définition 3.1

Le coefficient de corrélation linéaire entre les valeurs des caractères x et y d'une série statistique double est

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Les droites de régression sont $D: \ y = ax + b$ et $D' : x = a'y + b'$ avec $a = \frac{\sigma_{xy}}{\sigma_x^2}$ et $a' = \frac{\sigma_{xy}}{\sigma_y^2}$.

3.2 Propriétés du coefficient de corrélation linéaire

Propriété 3.2 *On a les relations suivantes entre le coefficient de corrélation linéaire r et les coefficients a et a' des droites de régression de y en x et de x en y*

$$aa' = r^2, \quad r^2 = a^2 \frac{\sigma_x^2}{\sigma_y^2} = a^2 \frac{V_x}{V_y}, \quad r = a \frac{\sigma_x}{\sigma_y}$$

$aa' = \frac{\sigma_{xy}}{\sigma_x^2} \frac{\sigma_{xy}}{\sigma_y^2} = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2}$ et $r^2 = \left(\frac{\sigma_{xy}}{\sigma_x \sigma_y}\right)^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2}$.

$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = \frac{\sigma_{xy}^2}{\sigma_x^4} \frac{\sigma_x^2}{\sigma_y^2} = \left(\frac{\sigma_{xy}}{\sigma_x^2}\right)^2 \frac{\sigma_x^2}{\sigma_y^2} = a^2 \frac{\sigma_x^2}{\sigma_y^2} = a^2 \frac{V_x}{V_y}$.

Propriété 3.3 *On a*

- $-1 \leq r \leq 1$, *le coefficient de corrélation linéaire est compris entre* -1 *et* 1
- *Les points du nuage sont alignés si et seulement si* $r = -1$ *ou* $r = 1$
- *le coefficient de corrélation linéaire a même signe que les coefficients* a *et* a' *et* $r^2 = aa'$

Si $|r| = 1$, l'ajustement affine est parfait.
Si $|r| < 0,7$, l'ajustement affine n'est pas justifié.
Si $|r| > 0,7$ l'ajustement affine est envisageable et selon le domaine sur lequel porte la statistique et le problème étudié, on décide d'un seuil au-delà duquel la corrélation est suffisante pour justifier un ajustement affine.

x_i	5.5	9.7	8.7	11.8	19.0	5.9	9.5	17.3	13.3	11.0	18.0	7.8	1.5	1.3	1.8	12.0	2.7	15.4	12.9	6.2
y_i	8.5	13.2	8.7	11.1	3.8	6.5	7.4	5.6	6.5	5.9	6.7	4.9	0.8	7.4	18.1	4.7	10.2	17.8	11.2	9.0

$n = 20$
$\bar{x} = 9.57, V_x = 28.90, \sigma_x = 5.38$
$\bar{y} = 8.40, V_y = 17.77, \sigma_y = 4.22$
$\sigma_{xy} = -1.79, r = -0.08$
$D: \quad y = ax + b, a = -0.06, b = 8.99$
$D': \quad x = a'y + b', a' = -0.10, b' = 10.41$

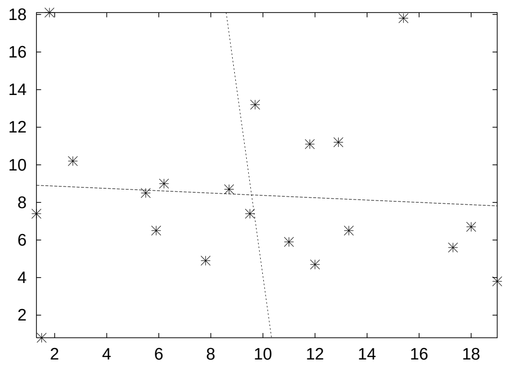

FIG. 6 – *Corrélation Proche de* 0.

Remarque 1 *Une interprétation géométrique du coefficient de corrélation linéaire r que nous ne détaillerons pas ici montre que $r = cos(\theta)$ où θ est un angle de deux vecteurs, le tableau ci-dessous donne quelques valeurs de l'angle :*

$r = cos\theta$	*Angle θ radians et degrés*	*Interprétation*
1	$0, 0°$	*même direction, alignement*
$0,87$	$\frac{\pi}{6}, 30°$	
$0,7$	$\frac{\pi}{4}, 45°$	
$0,5$	$\frac{\pi}{3}, 60°$	
0	$\frac{\pi}{2}, 90°$	*orthogonalité, indépendance*

3.3 Exemples de quelques cas possibles

Exemple 1 *Coefficient de corrélation nul ou presque*
(Voir Fig. 6)

Exemple 2 *Bonne corrélation*
(Voir Fig. 7)

Exemple 3 *Très forte corrélation*
(Voir Fig. 8)

3.4 Exemple de régression exponentielle

Exemple 4 *Allure exponentielle du nuage (Voir Fig. 9)*
On remarque que les points du nuage ne paraissent pas alignés et qu'au contraire ils semblent

x_i	22.1	32.5	41.1	51.3	61.0	71.9	82.4	92.2	101.2	110.6	122.8	130.6
y_i	212.8	239.8	193.4	199.5	199.2	179.8	177.4	172.5	191.0	184.9	158.9	173.0
x_i	141.4	152.6	162.0	172.6	182.9	190.6	201.8	210.8	222.1	232.7	240.4	251.4
y_i	151.1	147.5	142.4	139.9	132.5	150.4	100.6	142.3	81.2	76.3	92.1	71.4

$n = 24$
$\bar{x} = 136.71, V_x = 4785.19, \sigma_x = 69.18$
$\bar{y} = 154.58, V_y = 1940.84, \sigma_y = 44.05$
$\sigma_{xy} = -2868.31, r = -0.94$
$D: \quad y = ax + b, a = -0.60, b = 236.52$
$D': \quad x = a'y + b', a' = -1.48, b' = 365.16$

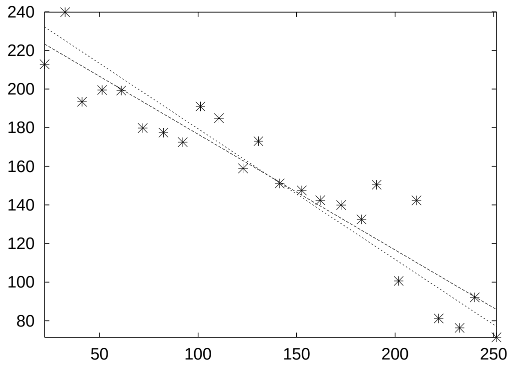

FIG. 7 – *Bonne Corrélation.*

x_i	5.5	6.4	7.9	8.6	9.6	10.3	11.7	12.0	13.7	14.3	15.2	16.3
y_i	3.4	3.9	4.6	5.1	6.1	7.1	7.2	8.5	8.5	9.1	10.7	11.2

$n = 12$
$\bar{x} = 10.96, V_x = 11.22, \sigma_x = 3.35$
$\bar{y} = 7.12, V_y = 6.07, \sigma_y = 2.46$
$\sigma_{xy} = 8.14, r = 0.99$
$D: \quad y = ax + b, a = 0.73, b = -0.84$
$D': \quad x = a'y + b', a' = 1.34, b' = 1.41$

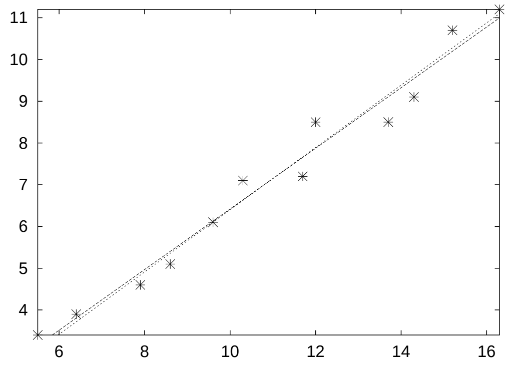

FIG. 8 – *Très Forte Corrélation.*

x_i	3.4	5.3	4.4	7.5	6.1	8.5	9.3	9.8	11.0	11.2	12.9	14.7
y_i	2.1	2.9	2.1	4.1	3.0	4.9	5.0	5.6	6.4	7.1	9.1	12.0
x_i	15.2	16.2	18.4	19.6	19.7	20.5	22.0	21.9	25.0	25.4	24.6	26.4
y_i	12.8	15.2	21.4	26.6	27.7	30.8	39.9	38.9	64.5	69.8	60.9	81.8

$n = 24$
$\bar{x} = 14.96, V_x = 51.11, \sigma_x = 7.15$
$\bar{y} = 23.11, V_y = 559.98, \sigma_y = 23.66$
$\sigma_{xy} = 153.23, r = 0.91$
$D: \quad y = ax + b, a = 3.00, b = -21.74$
$D': \quad x = a'y + b', a' = 0.27, b' = 8.64$

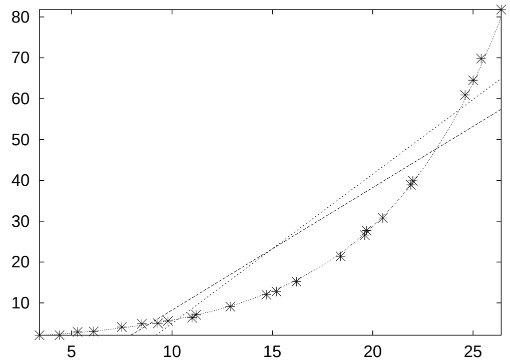

FIG. 9 – *Nuage en forme de courbe exponentielle.*

disposés sur la courbe représentative d'une fonction d'allure exponentielle.

Exemple 5 *Transformation par la fonction logarithme népérien*
Ou transformation de y en $z = \ln(y)$ à partir des données de l'exemple précédent.
(Voir Fig. 10)
On a un très bon coefficient de corrélation linéaire en étudiant la série double des caractères x et $z = \ln(y)$ au lieu de y (Voir l'exemple précédent Fig. 9).
Les calculs indiquent que $z = ax + b$ avec $a = 0.16$, $b = 0.16$ et on peut en déduire que $y = e^{0,16x+0,16}$, ou encore $y = e^{0,16x} \times e^{0,16} = 1,17e^{0,16x}$ car $e^{0,16} \approx 1,17$.
C'est justement la courbe C d'équation $1,17e^{0,16x}$ qui est tracée sur la figure de l'exercice précédent.
Cet exemple montre comment on peut utiliser une transformation $y \mapsto z = \ln(y)$ pour pouvoir effectuer une régression linéaire et enfin la transformation inverse $z \mapsto y = e^z$ pour obtenir la régression exponentielle.

x_i	3.4	5.3	4.4	7.5	6.1	8.5	9.3	9.8	11.0	11.2	12.9	14.7
$z_i = \ln(y_i)$	0.7	1.1	0.8	1.4	1.1	1.6	1.6	1.7	1.9	2.0	2.2	2.5

x_i	15.2	16.2	18.4	19.6	19.7	20.5	22.0	21.9	25.0	25.4	24.6	26.4
$z_i = \ln(y_i)$	2.6	2.7	3.1	3.3	3.3	3.4	3.7	3.7	4.2	4.2	4.1	4.4

$n = 24$
$\bar{x} = 14.96, V_x = 51.11, \sigma_x = 7.15$
$\bar{z} = 2.55, V_z = 1.31, \sigma_z = 1.15$
$\sigma_{xz} = 8.18, r = 1.00$
$D : \quad z = ax + b, a = 0.16, b = 0.16$
$D' : \quad x = a'z + b', a' = 6.24, b' = -0.98$

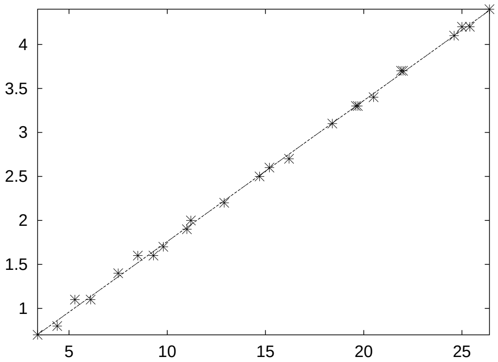

FIG. 10 – *Transformation* $z = \ln(y)$.

Index

Table des figures