Cours : Statistique descriptive & introduction aux Probabilités

Pr : OUAISSA HAMID

Licence Education
Spécialité : Enseignement secondaire Mathématiques

Année universitaire :
2023-2024

1

Statistique à deux dimensions

1.1 Introduction

Dans le chapitre précédant, Nous nous intéressons aux méthodes qui permettent de résumer et représenter les informations relatives à une variable (analyse univarié). Cependant, dans plusieurs phénomènes, on peut envisager, pour une même population, l'étude de deux ou plusieurs caractères en même temps. C'est à dire, Un même individu peut être étudié à l'aide de plusieurs caractères (ou variables). Par exemple, les salariés en regardant leur ancienneté et leur niveau d'étude, la croissance d'un enfant en regardant son poids et sa taille etc.

Dans ce chapitre, nous allons étudié la statistique à deux dimensions (ou analyse bivariée) qui consiste à étudier deux variables statistiques d'une même population afin de déterminer une probable liaison, via des techniques descriptives ou probabilistes. Dans la suite, nous introduisons l'étude globale des relations entre deux variables (en nous limitant au cas de deux variables). Donc, soit une population

Il y a plusieurs objectifs pour réaliser ces études tels que :

- Étendre les notions de la statistique descriptive à une variable au cas de deux variables.
- La mise en évidence d'un lien ou d'une absence de lien entre les deux variables étudiées.
- Caractériser et commenter les relations qui peuvent exister entre deux séries d'observations considérées simultanément (sens, intensité).

Dans la suite, on considère une population Ω de N individus pour laquelle on présente deux observations relatives à deux caractères (variables statistiques) X et Y pouvant être de nature différente (qualitatif, quantitatif discret ou continu). Comme on a vu en statistique descriptive à une variable, on envisagera trois aspects pour présenter et analyser les données statistiques relatives à un couple de

caractères $(X;Y)$:
• Tableaux statistiques (à deux variables),
• Représentations graphiques (à deux variables).
• Les paramètres indiquant la relation entre les variables étudiés.

Exemple 1.1.1.
- *On observe simultanément sur un échantillon de personnes le nombre des infectés du virus Corona et leur âge au Maroc.*
- *On observe sur un échantillon de 20 foyers, le revenu mensuel X en Dh et les dépenses mensuelles Y .*
- *Une entreprise mène une étude sur la liaison entre les dépenses mensuelles en publicité X et le volume des ventes Y qu'elle réalise.*

On peut alors présenter les données observées dans une population relatives à deux variables X et Y, considérées simultanément, sous forme d'une distribution d'effectifs ou de fréquences à deux dimensions, appelées distribution conjointe du couple (X, Y). Pour cela on a recours à un tableau statistique, appelé Tableau de contingence.

1.2 Tableau de contingence

Soient X et Y deux variables statistiques qualitatives ou quantitatives définies sur la même population finie Ω. Le couple $(X;Y)$ est appelé une distribution statistique conjointe.

Soient $X(\Omega) = \{x_1, ..., x_r\}$ et $Y(\Omega) = \{y_1, ..., y_s\}$ l'ensemble des modalités ou valeurs prises par chacune des variables X et Y.

Définition 1.2.1. On appelle effectif partiel n_{ij} du couple $(x_i; y_j)$ le nombre des individus de la population Ω pour lesquels le caractère X prend la valeur x_i et le caractère Y prend la valeur y_j. C-à-d, Chaque case (ligne i et colonne j) du tableau indique l'effectif n_{ij} des individus présentant la modalité $(x_i; y_j)$ du couple $(X;Y)$.

Le tableau de contingence est donc un tableau à deux dimensions (double entrée 1.2) présenté sous la forme suivante de telle sorte que :

- i désigne l'indice d'une ligne et j désigne l'indice d'une colonne.
- n_{ij} désigne l'effectif partiel.
- Effectif partiels marginaux :
  On note par $n_{i.}$ **l'effectif marginal de X** (effectif total en lignes) définit par

$$n_{i.} = \sum_{j=1}^{s} n_{ij}$$

X\ Y	y_1	y_2	...	y_j	...	y_s	Total
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1s}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2s}	$n_{2.}$
⋮	⋮	⋮	...	⋮	...	⋮	⋮
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{is}	$n_{i.}$
⋮	⋮	⋮	...	⋮	...	⋮	⋮
x_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.s}$	N

Table 2.1: Tableau de contingence

On note par $n_{.j}$ l'effectif marginal de Y (effectif total en colonnes) définit par

$$n_{.j} = \sum_{i=1}^{r} n_{ij}$$

- L'effectif total est donné par la relation suivante :

$$N = \sum_{j=1}^{s} n_{.j} = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{s} \sum_{i=1}^{r} n_{ij}$$

Définition 1.2.2. Si n_{ij} est l'effectif partiel d'un couple de valeurs $(x_i; y_j)$, et N l'effectif total de la population, le rapport

$$f_{ij} = \frac{n_{ij}}{N},$$

s'appelle la fréquence partielle du couple (x_i, y_j).
On appelle fréquence marginale $f_{.j}$ de la valeur y_j de la variable Y le rapport

$$f_{.j} = \frac{n_{.j}}{N}$$

On appelle fréquence marginale $f_{i.}$ de la valeur x_i de la variable X le rapport

$$f_{i.} = \frac{n_{i.}}{N}$$

Définition 1.2.3. La distribution conjointe des effectifs (ou fréquences) d'une série statistique double $(X; Y)$ définie sur une population Ω, n'est rien que l'application

définie par :

$$f : X(\Omega) \times Y(\Omega) \longrightarrow \mathbb{N}$$
$$(x_i; y_j) \longrightarrow n_{ij} \text{ ou } f_{ij}$$

Donc chaque couple $(x_i; y_j)$ associé son effectif n_{ij} (ou à sa fréquence f_{ij}), C'est exactement l'intérieur du tableau de contingence.

Définition 1.2.4. On appelle distribution marginale de X la distribution à une dimension des observations relatives au caractère X indépendamment des observations relatives au caractère de Y. Elle correspond aux effectifs $n_{i.}$ (ou aux fréquences $f_{i.}$) de la dernière colonne du tableau de contingence.
De la même façon, on définit la distribution marginale de Y. Elle correspond aux effectifs $n_{.j}$ (ou aux fréquence $f_{.j}$) de la dernière ligne du tableau de contingence.

Exemple 1.2.1. *Dans une société de 200 employés, on étudie les variables continues X : âge et Y : salaires, (regroupés en classes) le tableau de contingence est donné comme suit: On note I le nombre de modalités de X (ici $I = 3$) et J le*

X\ Y	[800,1000[(j=1)	[1000,1200[(j=2)	Total
[20,22[(i=1)	14	6	20
[22,24[(i=2)	28	46	74
[24,26[(i=3)	20	86	106
Total	62	138	200

nombre de modalités de Y (ici $J = 2$).
On désigne par $n_{21} = 28$ l'effectif partiel des salariés sont âgés entre 22 et 24 ans et ont un salaire compris entre 800 et 1000 dirhams. Tandis que $n_{1.}$ l'effectif marginal des salariés âgés entre 20 et 22 ans, et $n_{.2}$ l'effectif marginal des salariés qui ont un salaire entre 1000 et 1200.

X\ Y	[800,1000[(j=1)	[1000,1200[(j=2)	La distribution marginale de X
[20,22[(i=1)	7%	3%	10%
[22,24[(i=2)	14%	23%	37%
[24,26[(i=3)	10%	43%	53%
La distribution marginale de Y	31%	69%	100 %

1.2.1 Distributions conditionnelles

La notion de série conditionnelle est essentielle pour comprendre l'analyse de la régression. Un tableau de contingence se compose en autant de séries conditionnelles suivant chaque ligne et chaque colonnes. Alors, on appelle distribution

conditionnelle de X sous la condition $Y = y_j$, la distribution à une dimension des individus de la population relativement au caractère X sachant que la modalité de Y est égale à y_j. Elle correspond aux effectifs n_{ij} de la colonne j du tableau de contingence. De façon analogue, on définit la distribution conditionnelle de Y sous la condition $X = x_i$. Elle correspond aux effectifs n_{ij} de la ligne i du tableau de contingence.

Une distribution conditionnelle est une distribution statistique soumise à une condition sur la population à un événement particulier (une classe par exemple). On revient alors à l'exemple précédent :

Exemple 1.2.2. *On a le nombre de colonne est deux ($I = 2$) alors, on peut parler de deux distributions conditionnelles de X par rapport à Y. On a par exemple :*

1. *La distribution de X sachant* $Y \in [800, 1000[$.
2. *La distribution de X sachant* $Y \in [1000, 1200[$.

Pour la distribution de Y sous la condition $X = x_i$, on a trois lignes ($I = 3$)) il y a alors trois distributions conditionnelles de Y par rapport à $X = x_i$, à savoir :

1. *la distribution de Y sachant* $X \in [20; 22[$
2. *la distribution de Y sachant* $X \in [22; 24[$
3. *la distribution de Y sachant* $X \in [24; 26[$

On peut penser également aux fréquences conditionnelles :

Fréquences conditionnelles de X sachant Y

Définition 1.2.5. On désigne par la fréquence conditionnelle de la modalité x_i du caractère X sachant que $Y = y_j$, notée f_{x_i/y_j}, la proportion d'individus présentant la modalité x_i du caractère X par rapport au total des individus présentant la modalité y_j du caractère Y :

$$f_{x_i/y_j} = \frac{n_{ij}}{n_{.j}} = \frac{f_{ij}}{f_{.j}}$$

On présentera alors ces fréquences sur le tableau de contingence sous forme de pourcentage (Figure 1.1), exemple : On calcule les fréquences des âgés en considérant la condition sur le salaire des individus entre 800 et 1000, puis à la sous-population des individus ayant un salaire entre 1000 et 1200.
Interprétation : • On a $f_{x_1/y_1} = \frac{n_{11}}{n_{.1}} = \frac{7}{31} = 22.6\%$ des employés ayant un salaire entre 800 et 1000 sont âgés entre 20 et 22 ans.
• Parmi les employés ayant un salaire entre 1000 et 1200, on a $62.4\% = \frac{n_{32}}{n_{.2}} = \frac{86}{138}$ d'entre eux sont âgés entre 24 et 26 ans.

X\ Y	[800,1000[(j=1)	[1000,1200[(j=2)	Total
[20,22[(i=1)	22.6%	4.3%	.
[22,24[(i=2)	45.2%	33.3%	.
[24,26[(i=3)	32.2%	62.4%	.
Total	100%	100%	.

Figure 1.1: Fréquences conditionnelles de X sachant Y

Fréquences conditionnelles de Y sachant X

Définition 1.2.6. La fréquence conditionnelle de la modalité y_j du caractère Y sachant que $X = x_i$, notée f_{y_j/x_i}, est la proportion d'individus présentant la modalité y_j du caractère Y par rapport au total des individus présentant la modalité x_i du caractère X :

$$f_{y_j/x_i} = \frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{f_{i.}}$$

On présentera alors ces fréquences sur le tableau de contingence sous forme de pourcentage (Figure 1.2):

X\ Y	[800,1000[(j=1)	[1000,1200[(j=2)	Total
[20,22[(i=1)	70%	30%	100%
[22,24[(i=2)	37.8%	62.2%	100%
[24,26[(i=3)	18.9%	81.1%	100%
Total	.	.	.

Figure 1.2: Fréquences conditionnelles de Y sachant X

Interprétation :
- 70% des employés âgés entre 20 et 22 ans ont un salaire compris entre 800 et 1000.
- Parmi les employés âgés entre 22 et 24 ans, 62.2% d'entre eux ont un salaire compris entre 1000 et 1200.

Représentation graphique

Sur les distributions partielles (marginales et conditionnelles), on peut tracer des graphes (diagramme en tuyau d'orgue, diagramme en bâton, histogramme, fréquences cumulées) et on peut calculer des paramètres statistiques de la même façon que pour les distributions d'une seule variable (vus dans le chapitre précédant). Alors que pour représenter graphiquement une distribution conjointe de deux variables qualitatives, plusieurs possibilités sont envisageables. On peut citer par exemple :

- Histogramme groupé.
- Histogramme empilé.

Pour le histogramme groupé, son principe est simple, on fixe une variable sur l'axe des abscisses et on trace des rectangles dont la hauteur est proportionnelle au effectif (ou fréquence) de l'autre variable

Exemple 1.2.3. *On traite la relation entre les distributions des observations de Maths et de Chimie des étudiants de LEM en première année à l'ENS de Tétouan. On traite alors deux caractères qualitatifs ordinales. Les modalités sont alors comme suit : Médiocre (M), passable (P), bien (B), très bien (TB) pour le module de Maths et (Non validé)NV, Rattrapage (R), Validé (V) pour le module de Chimie. On remarque que les modalités sont présentés de façon différente. Les données sont présentées comme suit (Figure 1.2.3) : Le histogramme groupé des*

Maths\ Chimie	*NV (j=1)*	*R (j=2)*	*V (j=3)*
M (i=1)	*10*	*20*	*5*
P (i=2)	*5*	*10*	*15*
B (i=3)	*5*	*3*	*10*
TB (i=4)	*20*	*7*	*10*

effectif associé à cette distribution est donné par (Figure 1.3):

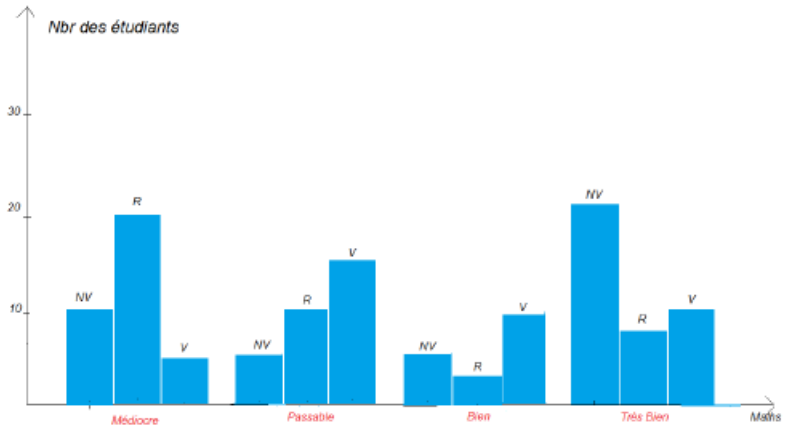

Figure 1.3: Le histogramme groupé associé aux observations des notes de Maths en fonction des observations de Chimie.

Pour les caractères de type quantitatif, on envisagera une étude de paramètre. En effet, on distingue deux types de paramètres :

— Les paramètres qui concernent une seule variable, ils caractérisent les distributions partielles (marginales et conditionnelles).
— Les paramètres qui décrivent les relations qui existent entre les deux variables considérées simultanément, ils caractérisent la distribution conjointe. Dans la suite, nous nous intéresserons aux paramètres décrivant la relation entre deux variables, plus précisément, nous nous focaliserons sur la covariance.

1.2.2 Covariance

La covariance est un paramètre qui donne la variabilité de X par rapport à Y (voir la figure 1.4)

Définition 1.2.7. On désigne par (X, Y) le couple de séries statistiques quantitatives définies sur une même population Ω, prenant respectivement les valeurs $x_1, ..., x_r$ et $y_1, ..., y_s$, et si n_{ij} désigne l'effectif partiel du couple $(x_i; y_j)$, on appelle covariance du couple (X,Y) et on note $Cov(X, Y)$ le nombre

$$Cov(X,Y) = \overline{XY} - \bar{X}\bar{Y} = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(x_i - \bar{X})(y_j - \bar{Y}) \quad (\text{ où } \overline{xy} = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}x_i y_j)$$

avec $\bar{X}$ est la moyenne de X et $\bar{Y}$ est est la moyenne de Y. La covariance peut s'écrire en fonction de fréquence comme suit :

$$Cov(X,Y) = \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}(x_i - \bar{X})(y_j - \bar{Y})$$

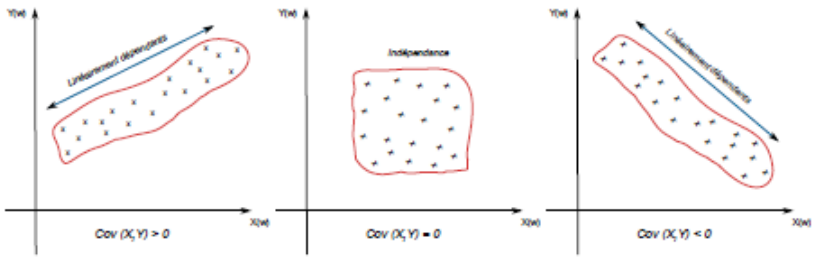

Figure 1.4: La covariance et la variabilité

Propriétés 1.2.1. On présente ici quelque propriétés intéressantes de la covariance :

- $Cov(X, X) = \sigma_X^2$ (où σ_X la variance de X)

- $Cov(Y,Y) = \sigma_Y^2$ (où σ_Y la variance de Y)
- $Cov(X,Y) = Cov(Y,X)$ (la covariance est symétrique)
- $Cov(X,Y) = \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_i y_j - \bar{X}\bar{Y}$ (en utilisant les fréquences)
- Si X et Y sont indépendantes alors $Cov(X,Y) = 0$.

Indépendenace de deux variables statistiques

Définition 1.2.8. On dit que deux variables statistiques X et Y sont indépendantes si et seulement si, pour tout i et j,

$$fij = f_{i.} \times f_{j.}$$

Il suffit que cette égalité ne soit pas vérifiée dans une seule cellule pour que les deux variables ne soient pas indépendantes.. De manière équivalente, pour tout i et j,

$$N \times n_{ij} = n_{i.} \times n_{j.}$$

Dans ce cas, si X et Y sont indépendantes alors $Cov(X,Y) = 0$. La réciproque est fausse

Remarque 1.2.1. La moyenne et la variance de la variable X est donnée par les formules suivantes

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{r} n_{i.}x_i = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}x_i$$

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{r} n_{i.}(x_i - \bar{X})^2 = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(x_i - \bar{X})^2$$

Corrélation

Le coefficient de corrélation permet de donner une mesure synthétique de l'intensité de la relation entre deux caractères et de son sens lorsque cette relation est monotone. Le coefficient de corrélation de Pearson permet d'analyser les relations linéaires (voir ci-dessous). Autrement dit, On dit qu'il y a corrélation entre deux variables lorsqu'elles ont tendance à varier soit toujours dans le même sens (X augmente, Y a tendance à augmenter aussi), soit en sens inverse (X augmente, Y a tendance à diminuer). Des questions se posent :
- Comment peut-on quantifier cette liaison ?
- Comment peut-on tester si cette liaison est statistiquement significative ?

Pour mesurer l'intensité de la relation entre deux variables, on définit le coefficient de corrélation de Pearson, connu par son efficacité et sa précision.

Coefficient de corrélation linéaire

Le coefficient de corrélation de Pearson ρ est une mesure du degré d'association linéaire entre deux variables quantitatives X et Y.

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

ρ est un nombre compris entre -1 et 1. ($|\rho| \leq 1$)

Remarque 1.2.2. Le coefficient $\rho(X,Y)$ mesure le degré de liaison linéaire entre X et Y. Nous avons les différentes caractéristiques suivantes (voir Figure 1.5).

• Plus le module de $\rho(X,Y)$ est proche de 1 plus X et Y sont liées linéairement.

• Plus le module de $\rho(X,Y)$ est proche de 0 plus il y a l'absence de liaison linéaire entre X et Y.

• Par définition, si $\rho(X,Y) = 0$, alors $Cov(X,Y) = 0$.

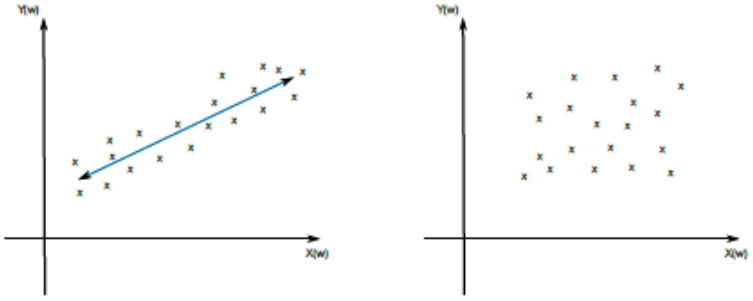

Figure 1.5: A gauche, le coefficient de corrélation est proche de 1. A droite le coefficient est proche de 0

Remarque 1.2.3. Si $\rho > 0$ cela signifie que les deux variables évoluent en même sens (sens positif).

— Si $\rho < 0$ cela signifie que les deux variables évoluent en sens contraire (sens négatif).

— Lorsque $\rho = 0$, on dit que X et Y sont non corrélés : il n'y a pas d'association linéaire entre X et Y.

— Si $\rho = \pm 1$ alors l'une des variable est une fonction affine de l'autre, (Y est une fonction affine de X i.e. Y = aX +b avec b du signe de ρ).

Lorsque X et Y sont indépendantes $\rho = 0$ mais la réciproque est en général fausse.

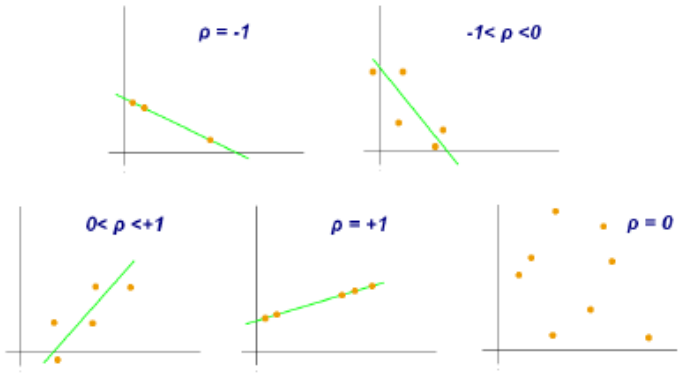

Figure 1.6: Exemples de diagrammes de dispersion avec différentes valeurs de coefficient de corrélation.

1.3 Régression linéaire

La régression linéaire permet d'étudier la relation entre deux variables quantitatives, en étudiant les variations de l'une en fonction des valeurs de l'autre. Dans le cas où tous les points $(x_i; y_i)$ du diagramme de régression sont alignés, on dit que la relation entre X et Y est une dépendance fonctionnelle linéaire. Si les points du diagramme de régression ne sont pas tous alignés, on doit mesurer le degré de dépendance linéaire entre les deux variables. Lorsque l'on a estimé la droite de régression, on doit se demander si cette estimation est de bonne qualité. Pour cela, on envisage un nouveau tableau de tel sorte que chaque couple de variables est sans pondération. On présente alors les données sous la forme suivante (Tableau 1.3):

Modalités de X	x_1	x_2	x_3	...	x_i	...	x_N
Modalités de Y	y_1	y_2	y_3	...	y_i	...	y_N

Afin de réaliser une bonne analyse de la relation entre ces deux variables, on se pose les questions suivantes :

• Quelle est la nature de la relation qui existe entre les deux variables X et Y?

• Si on connaît la valeur de l'une des variables, peut-on estimer la valeur de la deuxième variable ?

• Comment mesurer la précision de l'estimation ?

Pour répondre à ces questions, on commence par représenter le nuage de points $M_i(x_i; y_i)$ dans un repère cartésien. Ensuite, on cherchera s'il existe une droite ou une courbe qui passe par le plus possibles des points M_i pour qu'elle soit une bonne approximation du nuage de points. Ceci indiquera s'il y a une dépendance linéaire

entre les deux variables. Si c'est le cas, les points du nuage seront concentrés autour de la droite, et il devient possible de prévoir la seconde variable si on connaît la première. Ce nuage de points avec la courbe d'ajustement linéaire est appelé **Diagramme de régression**. La solution de ce problème est particulièrement simple lorsqu'on cherche une fonction linéaire

$$Y = aX + b$$

Le critère utilisé pour optimiser l'approximation est le critère des moindres carrés. On parle alors de régression (ou ajustement) linéaire de Y en fonction de X. Mais il est possible d'avoir d'autres types de régression non linéaires : exponentielle, logarithmique, polynômiale, etc. En prenant compte des erreurs entre les points, le modèle linéaire s'écrit

$$Y = aX + b + \varepsilon$$

où ε est une variable aléatoire appelée erreur résiduelle. On considère par exemple deux séries statistiques quantitatives, X représentant les notes de Maths et Y, représentant les notes de Chimie sur la même population Ω. Soit $X(\Omega) = \{x_1, ..., x_r\}$ l'ensemble des valeurs prises par X. $Y(\Omega) = \{y_1, ..., y_r\}$ l'ensemble des valeurs prises par Y. Les données sont réparties et arrangées dans un ordre croissant dans le tableau de pondération (1.7) suivant :

Étudiant N	1	2	3	4	5	6	7	8	9	10
Note de Maths (X)	3	6	7	9	10	12	13	15	17	19
Note de Chimie (Y	2	4.5	5.5	7	8	10	12.5	14	16	17

Figure 1.7: Notes de maths et de chimie

On remarque d'après ces données que plus la note de Maths est élevée, celle de Chimie l'est aussi et inversement proportionnel. Il y donc certainement une corrélation dans le sens positive entre ces deux variables. On obtient le nuage de points suivant qui représente le Diagramme de régression dans la figure 1.8.

On remarque que d'après la répartition des points sur le graphe, une droite est une bonne approximation de ce nuage de points. Il existe plusieurs méthodes pour approcher cette droite, dans ce cours nous nous intéresserons à la méthode des moindres carrées, connue par sa bonne précision.

1.4 Méthode de Moindres carrés

La droite de régression a pour equation y = ax+b. On doit déterminer les deux paramètres a et b telle que la variation résiduelle des moindres carrées

$$\xi(a;b) = \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}(y_j - (ax_i + b))^2$$

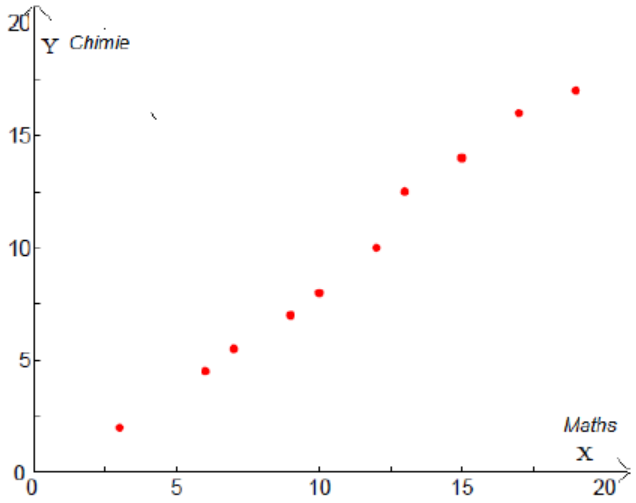

Figure 1.8: Le nuage de points $M_i(x_i; y_i)$.

soit minimal.
On va essayer de démontrer d'où vient la relation de cette approximation (cette démonstration est hors programme, c'est seulement pour avoir une idée). La méthode des moindres carrées consiste donc à minimiser la fonction $\xi(a; b)$ (la somme des erreurs commises). Nous avons les conditions d'optimalité,

$$\frac{\partial \xi}{a} = \frac{\partial \xi}{b} = 0.$$

C'est à dire

$$\frac{\partial \xi}{a} = -2\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_i(y_j - (ax_i + b)) = 0$$
$$\frac{\partial \xi}{b} = -2\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}(y_j - (ax_i + b)) = 0$$

En distribuant la somme Σ, il vient

$$\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_iy_j - a\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_i^2 - b\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_i = 0$$
$$\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}y_j - a\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_i - b\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} = 0$$

Par conséquent

$$a\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_i^2 + b\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_i = \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_i y_j \tag{1.1}$$

$$a\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_i + b = \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}y_j \tag{1.2}$$

La droite de régression estimée qui rend la distance entre elle et les points du nuage minimale, est donc $y = a^*x + b^*$ avec a^* est l'estimation de a et b^* est l'estimation de b qui sont exprimés comme suit (en utilisant le système 1.2)

$$a^* = \frac{Cov(X,Y)}{\sigma_X^2}, \qquad b^* = \bar{Y} - a^*\bar{X}$$

où $\bar{X}$ et $\bar{X}$ sont respectivement les moyennes de X et Y.

Tout cela , nous permet donc d'énoncer le théorème suivant

Théorème 1.4.1. *Étant donné un couple (X;Y) de séries quantitatives définies sur une même population Ω, il n'existe qu'une seule manière d'effectuer une régression de Y en X, à l'aide d'une fonction affine f, par la méthode des moindres carrés, cette droite est exprimée par l'équation $f(X) = aX + b$ où a et b sont définies par*

$$a^* = \frac{Cov(X,Y)}{\sigma_X^2}, \qquad b^* = \bar{Y} - a^*\bar{X}$$

D'une façon analogue, le régression de X en Y donne lieu à droite d'équation $Y = a'X + b'$ où

$$a' = \frac{Cov(X,Y)}{\sigma_Y^2}, \qquad b' = \bar{X} - a'\bar{Y}$$

Exemple 1.4.1. *La droite de régression de la note de Maths en fonction de la note de Chimie, par la méthode des moindres carrés, a pour équation :*

$$Y = 0.999X + 1.447;$$

car on a $\bar{X} = 11.1$, $\bar{Y} = 9,65$, $Cov(X,Y) = 23,08$ et $\sigma_X^2 = V(X) = 23,09$. On peut maintenant tracer la courbe de l'ajustement linéaire (voir la figure1.9)

Remarque 1.4.1. Le coefficient de corrélation $\rho(X,Y)$ permet de justifier le fait de l'ajustement linéaire. On peut adopter les critères numériques suivants (voir Figure 1.10),

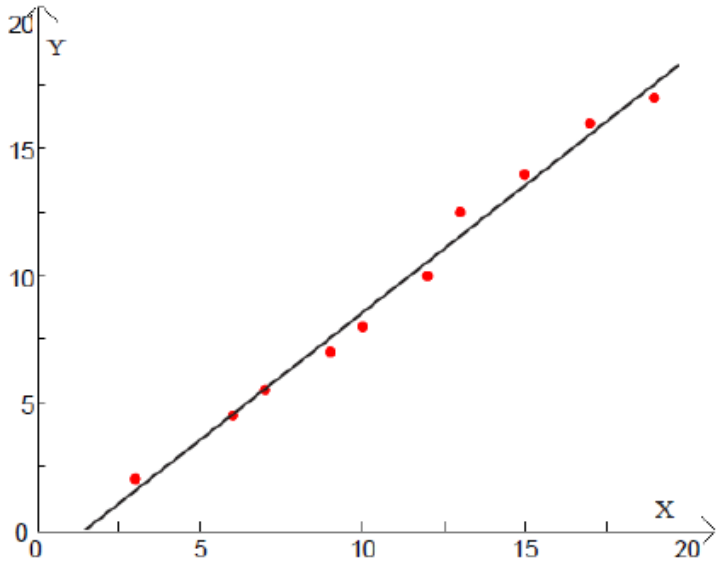

Figure 1.9: Ajustement linéaire de Y en X

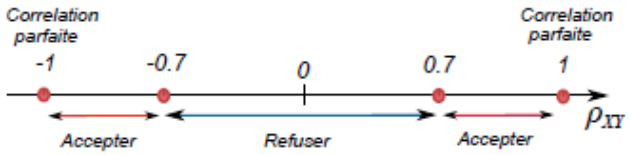

Figure 1.10: La zone d'acceptation ou de refus de l'ajustement linéaire

- Si $|\rho(X,Y)| < 0.7$, alors l'ajustement linéaire est refusé (droite refusée).
- Si $|\rho(X,Y)| \geq 0.7$, alors l'ajustement linéaire est accepté (droite acceptée).