Étude d'une variable statistique à deux dimensions

Dans les chapitres précédents, nous avons présenté les méthodes qui permettent de résumer et représenter les informations relatives à une variable. Un même individu peut être étudié à l'aide de plusieurs caractères (ou variables). Par exemple, les salaries en regardant leur ancienneté et leur niveau d'étude, la croissance d'un enfant en regardant son poids et sa taille. Dans la suite, nous introduisons l'étude globale des relations entre deux variables (en nous limitant au cas de deux variables). Donc, soit Ω une population et

$$\begin{aligned} Z : \Omega &\to \mathbb{R}^2, \\ w &\mapsto Z(w) = (X(w), Z(w)), \end{aligned}$$

ou directement

$$(X, Y) : \Omega \to \mathbb{R}^2,$$

Dans ce cas, Z est dite variable statistique à deux dimensions avec $\text{Card}(\Omega) = N$, avec N un entier fini. Le couple (X, Y) est appelé le couple de la variable statistique.

Exemple 20

- *On observe simultanément sur un échantillon de* 200 *foyers, le nombre d'enfants* X *et le nombre de chambre* Y*.*
- *On observe sur un échantillon de* 20 *foyers, le revenu mensuel* X *en Da et les dépenses mensuelles* Y*.*
- *Au près des étudiants pris au hasard parmi une section de L2 génie civil, on observe les notes de math3* X *et de statistique* 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.*

4.1 Représentation des séries statistiques à deux variables

Les séries statistiques à deux variables peuvent être présentées de deux façons.

Présentation 1

A chaque w_i, on associé (x_i, y_i), c'est à dire,

$$w_i \longrightarrow (x_i, y_i).$$

On rassemblera les données comme dans le tableau suivant

w_i	w_1	w_2	...	w_N
Variable X	$X(w_1)$	$X(w_2)$	...	$X(w_N)$
Variable Y	$Y(w_1)$	$Y(w_2)$	...	$Y(w_N)$

Cette représentation on la notera "présentation 1". Nous allons utiliser toujours les notations suivantes :

$$x_i := X(w_i)$$

et $y_i := Y(w_i)$.

Exemple 21

Soit Ω l'ensemble de 8 étudiants. Nous avons le tableau suivant

w_i	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
$X(w)$	8	2	6	6	11	10	7	2
$Y(w)$	9	10	11	7	14	16	12	5

avec X représente le nombre d'heures passées à préparer l'examen de statistique par étudiant et Y représente la note sur 20 obtenue à l'examen par l'étudiant.

Lors de cette représentation, nous pouvons traduire le tableau associe dans une figure appelée "le nuage de points" ou "diagramme de dispersion" (voir Figure 4.1). Cette représentation est obtenue en mettant dans un repère cartésien chaque couple d'observation (x_i, y_j) par un point.

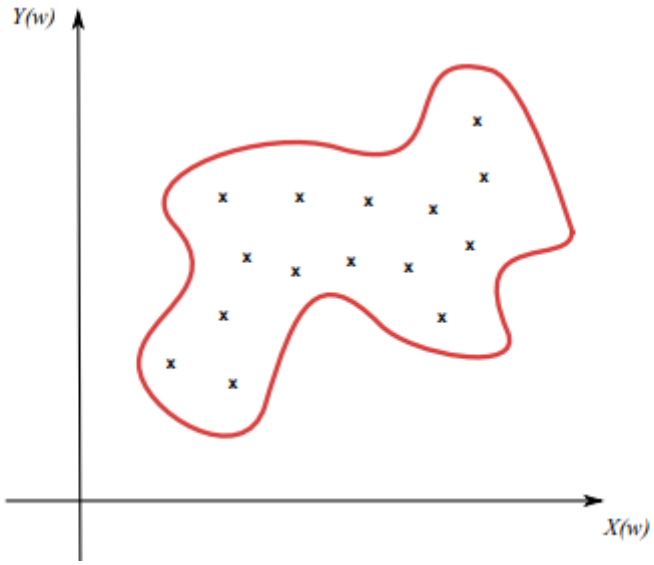


FIGURE 4.1: Représentation sous forme de nuage de points.

Présentation 2

Soit la variable statistique Z donnée par le couple (X, Y). Soient x_1, ..., x_k et y_1,...,y_l les valeurs prises respectivement par X et Y. Dans ce cas, nous définissons les valeurs de Z comme suite, pour i allant de 1 à k et pour j allant de 1 à l,

$$z_{ij} := (x_i, y_j).$$

La variable statistique Z prend $k \times l$ valeurs. Lors de cette étude, nous avons le tableau à double entrée (ou tableau de contingence) suivant (discrète ou continue)

$$n_{ij} := \text{Card}\{w \in \Omega \ : Z(w) = z_{ij}\}.$$

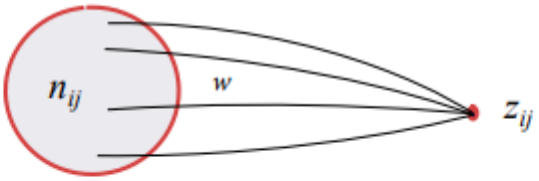


FIGURE 4.2: Le nombre d'individus qui prennent en même temps la valeur x_i et y_i.

Nous notons par f_{ij} la fréquence du coulpe (x_i, y_i). Cette fréquence est donnée par

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

avec

$$\begin{aligned} N &= \text{Card}(\Omega), \\ &= \sum_{j=1}^{l}\sum_{i=1}^{k} n_{ij}, \\ &= \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}. \end{aligned}$$

Remarque 16

Nous avons la propriété suivante,

$$\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} = 1.$$

Lois marginales

Sur la marge du tableau de contingence, on peut extraire les données seulement par rapport à X et seulement par rapport à Y (voir le tableau de contingence établi auparavant).

1. Effectifs et fréquences marginale par rapport à Y : nous avons, pour $j = 1...l$,

$$n_{\bullet j} := \sum_{i=1}^{k} n_{ij},$$

et

$$f_{\bullet j} := \frac{n_{\bullet j}}{N} = \sum_{i=1}^{k} f_{ij}.$$

2. Effectifs et fréquences marginale par rapport à X : nous avons, pour $i = 1...k$,

$$n_{i\bullet} := \sum_{j=1}^{l} n_{ij},$$

et

$$f_{i\bullet} := \frac{n_{i\bullet}}{N} = \sum_{j=1}^{l} f_{ij}.$$

Remarque 17

Nous avons les propriétés suivantes

$$\sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{l} n_{\bullet j} = N \qquad et \qquad \sum_{i=1}^{k} f_{i\bullet} = \sum_{j=1}^{l} f_{\bullet j} = 1.$$

Exercice 23

Nous considérons 10 *salariés qui sont observés à l'aide de deux variables "âge" et "salaire". Les informations brutes (pas encore traitées ou façonnées) sont données dans le tableau suivant,*

Salaire	6000	7400	7500	8200	8207	8900	9100	9900	9950	10750
Age	*15*	*26*	*20*	*43*	*47*	*37*	*52*	*34*	*50*	*44*

1. *Déterminer le tableau de contingence (X : âge, Y : salaire). Pour l'âge et pour le salaire, former respectivement des classes de pas de* 10 *ans et de* 1000 *Da.*
2. *Calculer f_{21}, f_{12}, f_{45} et f_{33}.*
3. *Déterminer les effectifs marginaux de X et de Y. Tracer le nuages de points.*
4. *Déterminer le tableau statistique des deux séries marginales X et Y.*

Solution : En utilisant les hypothèses, nous considérons les classes suivantes,

$$[15, 25[,\ [25, 35[,\ [35, 45[,\ [45, 55[,$$

pour l'âge et

$$[6, 7[,\ [7, 8[,\ [8, 9[,\ [9, 10[,\ [10, 11[,$$

pour le salaire ($\times 1000$). De plus, nous avons

$$Nombre\ de\ classe = \frac{e}{a_{\hat{a}ge}} = \frac{x_{\max} - x_{\min}}{a_{\hat{a}ge}} = \frac{52 - 15}{10} = 3.7 \simeq 4\ classes,$$

pour l'âge et

$$Nombre\ de\ classe = \frac{e}{a_{sal}} = \frac{y_{\max} - y_{\min}}{a_{sal}} = \frac{10750 - 6000}{1000} = 4.75 \simeq 5\ classes,$$

pour le salaire. Cette série statistique est représentée par le tableau suivant,

Age \ Salaire	$[6,7[$	$[7,8[$	$[8,9[$	$[9,10[$	$[10,11[$	$n_{i\bullet}$	$f_{i\bullet}$
$[15,25[$	1	1	0	0	0	0	0.2
$[25,35[$	0	1	0	1	0	2	0.2
$[35,45[$	0	0	2	0	1	3	0.3
$[45,55[$	0	0	1	2	0	3	0.3
$n_{\bullet j}$	1	2	3	3	1	10	1
$f_{\bullet j}$	0.1	0.2	0.3	0.3	0.1	1	$\varnothing$

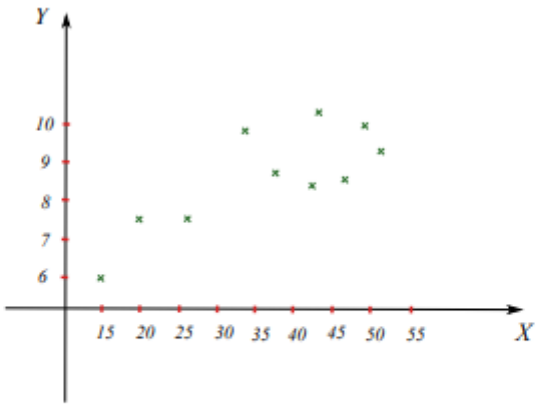
De ce fait, nous avons

$$f_{12} = \frac{n_{12}}{N} = \frac{1}{10} = 0.1, \qquad f_{21} = \frac{n_{21}}{N} = \frac{0}{10} = 0,$$

et

$$f_{45} = \frac{n_{45}}{N} = \frac{0}{10} = 0, \qquad f_{33} = \frac{n_{33}}{N} = \frac{2}{10} = 0.2.$$

Le nuage de points est tracé, à partir des données brutes, dans la figure suivante.



Enfin, les deux tableaux statistiques de X et de Y sont donnés, respectivement, par

4.2 Description numérique

4.2.1 Caractéristique des séries marginales

Dans le cas d'une variable statistique à deux dimensions X et Y, les moyennes sont données respectivement par

$$\overline{x} := \frac{1}{N}\sum_{i=1}^{k} n_{i\bullet}x_i = \sum_{i=1}^{k} f_{i\bullet}x_i \quad \text{(moyenne } de\ X\text{)},$$

et

$$\overline{y} := \frac{1}{N}\sum_{j=1}^{l} n_{\bullet j}y_j = \sum_{j=1}^{l} f_{\bullet j}y_j \quad \text{(moyenne } de\ Y\text{)}.$$

Remarque 18

Dans le cas continu, x_i et y_j représentent respectivement le centre des classes de X et Y, c'est à dire,

$$x_i = \frac{L_{i+1} + L_i}{2} \qquad et \qquad y_j = \frac{L_{j+1} + L_j}{2}.$$

Exemple 22

Nous calculons $\overline{x}$ et $\overline{y}$ pour l'exercice traité précédemment. Nous avons la moyenne d'âge

$$\overline{x} = \frac{1}{10}(40 + 60 + 120 + 150) = 37 \ ans.$$

et la moyenne du salaire

$$\overline{y} = \frac{1}{10}(6.5 + 15 + 25.5 + 28.5 + 10.5) \times 100 = 8600 \ Da.$$

Nous définissions maintenant la variance de X et la variance de Y comme suit,

$$Var(X) := \overline{x^2} - (\overline{x})^2, \qquad \text{avec} \quad \overline{x^2} := \frac{1}{N}\sum_{i=1}^{k} n_{i\bullet}x_i^2 = \sum_{i=1}^{k} f_{i\bullet}x_i^2,$$

et

$$Var(Y) := \overline{y^2} - (\overline{y})^2, \qquad \text{avec} \quad \overline{y^2} := \frac{1}{N}\sum_{j=1}^{l} n_{\bullet j}y_j^2 = \sum_{j=1}^{l} f_{\bullet j}y_j^2.$$

Les écarts-type de X et de Y sont donnés, respectivement, par

$$\sigma_X := \sqrt{Var(X)} \qquad \text{et} \qquad \sigma_Y := \sqrt{Var(Y)}.$$

4.2.2 Série conditionnelle

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.

Série conditionnelle par rapport à X

Elle est notée par X/y_j (ou X_j) et on dit que c'est la série conditionnelle de X sachant que $Y = y_j$. Nous calculons dans ce cas la fréquence conditionnelle $f_{i/j}$ (f_i sachant j), pour $i = 1, ..., k$, par

$$f_{i/j} := \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}}.$$

Nous avons aussi la moyenne conditionnelle $\overline{x}_j$, c'est à dire la moyenne des valeurs de X sous la condition y_j, elle est définie par

$$\overline{x}_j := \sum_{i=1}^{k} f_{i/j}x_i = \frac{1}{n_{\bullet j}}\sum_{i=1}^{k} n_{ij}x_i.$$

Pour l'écart-type conditionnel, nous avons $\sigma_{X_j} := \sqrt{Var(X_j)}$ avec

$$Var(X_j) := \sum_{i=1}^{k} f_{i/j}(x_i - \overline{x}_j)^2 = \overline{x^2}_j - (\overline{x}_j)^2.$$

Remarque 19

Dans le cas où nous avons un tableau des données brutes "representation 1" (nous n'avons pas d'effectifs), nous avons les formules suivantes

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i \qquad et \qquad \overline{y} = \frac{1}{N}\sum_{i=1}^{n} y_i.$$

De plus, nous avons

$$\overline{xy} = \frac{1}{N}\sum_{i=1}^{n} x_i y_i.$$

Remarque 20

La covariance est une notion qui généralise la variance, En effet,

$$Cov(X,X) = Var(X) \qquad et \qquad Cov(Y,Y) = Var(Y).$$

Cela provient de la définition, c'est à dire,

$$Cov(X,X) = \overline{xx} - \overline{x}\ \overline{x} = \overline{x^2} - \overline{x}^2 = Var(X).$$

Définition 25

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

$$f_{ij} = f_{i\bullet} \times f_{\bullet j}.$$

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,

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

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

Cette définition donne une interprétation intéressante de d'indépendance ; elle signifie que dans ce cas, les effectifs des modalités conjointes peuvent se calculer uniquement à partir des distributions marginales, supposées « identiques » aux distributions de X et Y dans la population ; en d'autres termes, si X et Y sont indépendantes, les observations séparées de X et de Y donnent la même information qu'une observation conjointe.

4.3 Ajustement linéaire

Dans le cas où on peut mettre en évidence l'existence d'une relation linéaire significative entre deux caractères quantitatifs continus X et Y (la silhouette du nuage de points est étirée dans une direction), on peut chercher à formaliser la relation moyenne qui unit ces deux variables à l'aide d'une équation de droite qui résume cette relation. Nous appelons cette démarche l'ajustement linéaire.

4.3.1 Coefficient de corrélation

Les coefficients de corrélation permettent 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). Il existe d'autres coefficients pour les relations non-linéaires et non-monotones, mais ils ne seront pas étudiés dans le cadre de ce cours.

Proposition 3
Le coefficient ρ_{XY} est compris entre $[-1, 1]$, ou encore

$$| \rho_{XY} | \leq 1.$$

– Plus le module de ρ_{XY} est proche de 1 plus X et Y sont liées linéairement.

– Plus le module de ρ_{XY} est proche de 0 plus il y a l'absence de liaison linéaire entre X et Y.

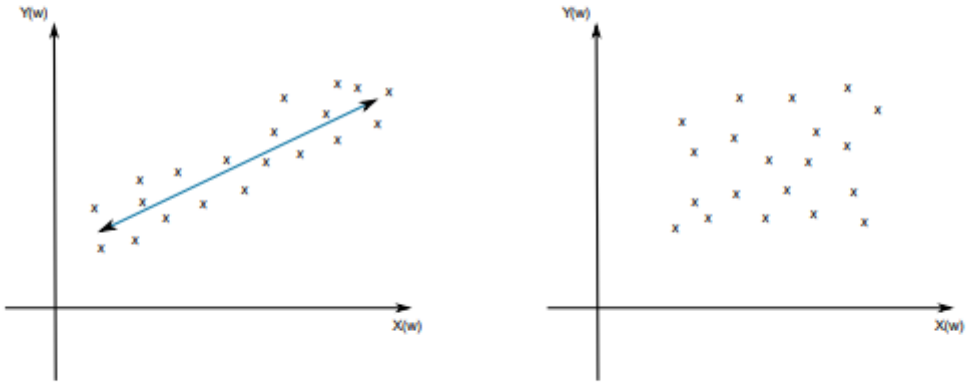


FIGURE 4.4: A gauche, le coefficient de corrélation est proche de 1. A droite, le coefficient de corrélation est proche de 0.

4.3.2 Droite de régression

L'idée est de transformer un nuage de point en une droite. Celle-ci doit être la plus proche possible de chacun des points. On cherchera donc à minimiser les écarts entre les points et la droite.

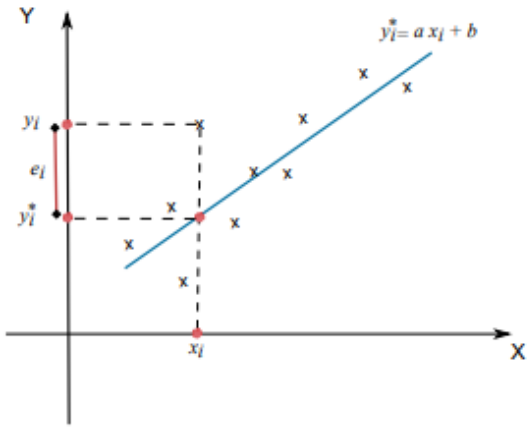


FIGURE 4.7: La droite la plus proche possible de chacun des points.

Pour cela, on utilise la méthode des moindres carrées. Cette méthode vise à expliquer un nuage de points par une droite qui lie Y à X, c'est à dire,

$$Y = aX + b,$$

telle que la distance entre le nuage de points et droite soit minimale. Cette distance matérialise l'erreur, c'est à dire la différence entre le point réellement observé et le point prédit par la droite. Si la droite passe au milieu des points, cette erreur sera alternativement positive et négative, la somme des erreurs étant par définition nulle. Ainsi, la méthode des moindres carrés consiste à chercher la valeur des paramètres a et b qui minimise la somme des erreurs élevées au carré.

On pose

$$\sum_{i=1}^{n} e_i^2 = U(a, b),$$

avec e_i est l'erreur commise sur chaque observation, c'est à dire,

$$\mid e_i \mid = \mid y_i - y_i^\star \mid = \mid y_i - ax_i - b \mid .$$