Année Probatoire A Distance
Module Probabilités, Statistiques
Chapitre : Statistique descriptive

J. Gergaud / ENSAT

Mars 2004

Table des matières

Chapitre 1

Statistique descriptive

1 Introduction

Le but de la statistique descriptive est de représenter les données observées sous la forme de tableaux, graphiques, courbes et (ou) à l'aide de quelques paramètres afin d'en avoir une vue plus synthétique. On peut d'ailleurs considérer l'analyse des données comme une extension de la statistique descriptive.

Ce chapitre ne sera donc que descriptif et ne concernera que des données observées. A aucun moment nous n'aurons de généralisations à partir de ces données.

2 Types de données traitées

2.1 Notion de caractère

Définition 2.1.1 (Caractère). On appelle caractère tout critère sur lequel repose une étude statistique.

Exemple 2.1.2. La taille d'un individu, le poids d'un objet, la concentration d'une substance.

Définition 2.1.3 (Caractère quantitatif, variable statistique). On appelle caractère quantitatif ou variable statistique tout caractère directement représentable par des nombres.

Exemple 2.1.4. La taille, l'age d'un individu, le nombre de particules.

Définition 2.1.5 (Caractère qualitatif). On appelle caractère qualitatif tout caractère non quantitatif

Exemple 2.1.6. La couleur des yeux, pile ou face.

Remarque 2.1.7. On pourrait très bien coder pile ou face par 0 et 1, mais nous aurions tout de même un caractère qualitatif d'où le mot directement dans la définition. On peut aussi dire qu'une variable statistique est un caractère mesurable. Les opérations comme l'addition ont donc un sens sur un caractère quantitatif, ce qui n'est pas le cas sur un caractère qualitatif.

2.2 Types de caractères qualitatifs

On range les caractères qualitatifs en plusieurs catégories :
- Les caractères qualitatifs ordonnés (i.e. que l'on peut les classer) comme le niveau d'un élève (bon, moyen, mauvais).
- Les caractères qualitatifs non ordonnés comme la couleur des yeux.
- Les caractères dichotomiques (i.e. qui ne peuvent prendre que deux valeurs différentes) comme le sexe, pile ou face.

2.3 Types de variables statistiques

Définition 2.3.1 (Variable discrète). On appelle variable discrète toute variable qui ne peut prendre qu'un nombre fini ou dénombrable de valeurs.

Exemple 2.3.2.
- Nombre de points sur la face supérieur d'un dé.
- Nombre de lancés d'une pièce de monnaie avant d'obtenir face.

Définition 2.3.3 (Variable continue). On appelle variable continue toute variable pouvant prendre un nombre infini non dénombrable de valeurs.

Exemple 2.3.4. – Poids d'un individu.
– Taille d'un individu.
– Concentration d'une substance.

3 Statistique descriptive à une dimension

3.1 Introduction

Nous allons nous intéresser dans cette section au cas d'un seul caractère quantitatif. Nous avons donc au départ une suite de n nombres :$y_1, y_2, \ldots, y_n$. Nous pouvons bien évidemment avoir dans cette suite plusieurs fois la même valeur.

Définition 3.1.1 (Série statistique). On appelle série statistique la suite $y_1, y_2, \ldots, y_n$.

Exemple 3.1.2. Notes sur 10 de 10 élèves à un devoir de français.

$$10; 05; 01; 09; 02; 05; 01; 09; 09; 01$$

3.2 Les distributions de fréquences

Lorsque la série est trop grande mais que les valeurs prises par la variable ne sont pas trop nombreuses nous pouvons condenser les résultats sous la forme d'une distribution de fréquences. Notons x_i les différentes valeurs du caractère étudié obtenues $i = 1, \ldots, p$.

Définition 3.2.1 (Fréquence absolue ou fréquence). On appelle fréquence absolue le nombre d'occurrences d'une même valeur observée x_i, c'est-à-dire le nombre de fois où la valeur x_i est observée. On note n_i cette fréquence liée à la valeur x_i.

Remarque 3.2.2. On a toujours $n = \sum_{i=1}^{p} n_i$

Notation 3.2.3. On note aussi $n_{\cdot} = n$
Le point signifie que l'on a fait une sommation sur l'indice i.

Définition 3.2.4 (Fréquence relative). On appelle fréquence relative associée à x_i la quantité :

$$f_i = \frac{n_i}{n}$$

Remarque 3.2.5. On a toujours :

$$\sum_{i=1}^{p} f_i = \sum_{i=1}^{p} \frac{n_i}{n} = \frac{1}{n} \sum_{i=1}^{p} n_i = 1$$

Définition 3.2.6 (Fréquences cumulées absolues). Les fréquences cumulées absolues sont données par :

$$\begin{aligned} N_0 &= 0 \\ N_1 &= n_1 \\ &\vdots \\ N_k &= \sum_{i=1}^{k} n_i \text{ si } k \in \{1, \ldots, p\} \\ &\vdots \\ N_k &= n \text{ si } k > p \end{aligned}$$

Définition 3.2.7 (Fréquences cumulées relatives). Les fréquences cumulées relatives sont données par :

$$\begin{aligned} F_0 &= 0 \\ F_1 &= f_1 \\ &\vdots \\ F_k &= \sum_{i=1}^{k} f_i \text{ si } k \in \{1, \ldots, p\} \\ &\vdots \\ F_k &= 1 \text{ si } k > p \end{aligned}$$

Exemple 3.2.8. Nous donnons dans le tableau ci-dessous les valeurs des différentes fréquences définies précédemment pour l'exemple (3.1.2).

Notes	Fréquences absolues	Fréquences relatives	Fréquences cumulées absolues	Fréquences cumulées relatives
0	0	0	0	0
1	3	0,3	3	0,3
2	1	0,1	4	0,4
3	0	0	4	0,4
4	0	0	4	0,4
5	2	0,2	6	0,6
6	0	0	6	0,6
7	0	0	6	0,6
8	0	0	6	0,6
9	3	0,3	9	0,9
10	1	0,1	10	1

Nous avons étudié le cas où la variable ne pouvait prendre que peu de valeurs différentes. Il se pose donc la question de savoir ce que l'on fait lorsque l'on a des valeurs observées distinctes en grand nombre (ce qui est le cas en particulier lorsque l'on étudie des variables continues). Dans ce cas nous condensons les données en groupant les observations en classes. Le nombre de classes est en général compris entre 10 et 20 et l'intervalle de classe est constant (mais ceci n'est pas obligatoire). Une classe est définie par ses limites. La limite supérieure d'une classe étant la limite inférieure de la classe suivante. Quant à la valeur de la classe, on choisit souvent le milieu de la classe. Une fois que les classes ont été définies nous pouvons comme précédemment calculer les fréquences absolues, relatives, cumulées absolues et cumulées relatives.

Exemple 3.2.9. Distribution de fréquence des étendues des exploitations agricoles belges (ces donnée proviennent de l'ouvrage de Dagnélie "Théorie et méthodes statistiques" volume 1).

Etendues des exploitations	Valeurs des classes	Fréquences absolues	Fréquences relatives	Fréquences cumulées relatives
de 1 à 3ha	2ha	58122	0,2925	0,2925
de 3 à 5ha	4ha	38221	0,1924	0,4849
de 5 à 10ha	7,5ha	52684	0,2651	0,75
de 10 à 20ha	15ha	35188	0,1771	0,9271
de 20 à 30ha	25ha	8344	0,0420	0,9691
de 30 à 50ha	40ha	3965	0,0199	0,9890
de 50 à 100ha	75ha	1873	0,0094	0,9984
plus de 100ha	?	309	0,0016	1,000

Remarque 3.2.10. Dans l'exemple ci-dessus la dernière classe n'a pas de limite supérieure. On dit que la classe est ouverte.

3.3 Représentations graphiques

Dans le paragraphe précédent nous avons travaillé directement avec des nombres, mais un tableau de chiffres (même en quantité restreinte) n'est jamais très lisible aussi nous allons maintenant étudier les représentations graphiques des fréquences. Dans tous les cas nous aurons ici en abscisse les variables et en ordonnées les fréquences.

Considérons tout d'abord le cas des fréquences non cumulées. Deux cas se présentent suivant que les données sont groupées (i.e. mises en classes) ou non. Lorsque celles-ci sont non groupées, nous utiliserons des diagrammes en bâtons : Pour chaque valeur de x_i, nous traçons un segment de droite de longueur égale à la fréquence (absolue ou relative suivant les cas) associée à x_i.

Exemple 3.3.1. Reprenons les données de l'exemple (3.1.2), la figure (1.1) est le diagramme en bâtons relatif aux fréquences relatives.

Lorsque les données sont groupées, nous représentons ces fréquences par des rectangles contigus dont les intervalles de classes sont les bases et les hauteurs des quantités telles que l'aire de chaque rectangle soit proportionnelle à la fréquence de la classe correspondante.

Remarque 3.3.2. Si les classes sont équidistantes nous pouvons alors prendre comme hauteur les fréquences.

Exemple 3.3.3. Représentons les fréquences relatives des étendues des exploitations agricoles belges (exemple (3.2.9))

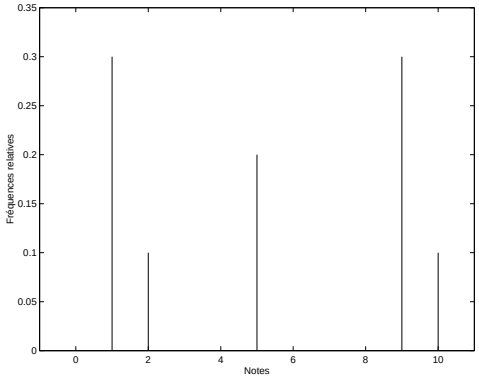

FIG. 1.1 – Diagramme en bâtons

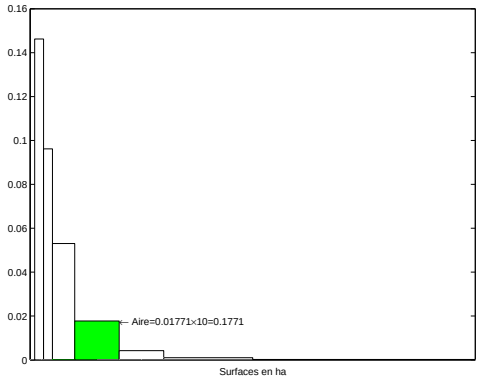

FIG. 1.2 – Histogramme

Définition 3.3.4 (Histogramme). On appelle histogramme un diagramme du type précédent.

Remarque 3.3.5. (i) Lorsque nous étudions une variable continue nous avons dans la pratique un grand nombre de mesures, certaines étant très proches les unes des autres, d'autres étant plus éloignées. Si nous représentions ces données sous la forme d'un diagramme en bâtons nous aurions un graphique du type suivant :

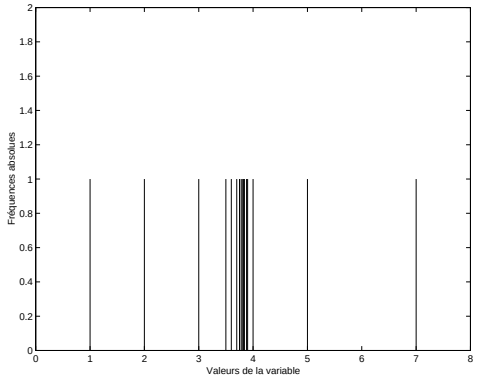

FIG. 1.3 – "Densité"

La densité d'une zone indiquerait alors que beaucoup de données seraient dans cette zone. Mais un tel graphique n'est pas très lisible et une idée est donc de représenter cette densité en ordonnées. Celle-ci est obtenue en divisant le nombre de mesures obtenues dans une classe (i.e. la fréquence absolue) par la longueur d'intervalle de classe. C'est bien ceci que nous représentons dans un histogramme.

(ii) Les fréquences relatives sont en fait dans la pratique des estimations de probabilités. On verra que dans le cas continu la probabilité qu'une variable aléatoire X appartienne à un intervalle $]x_i, x_{i+1}[$ est donnée par l'aire A délimitée par cet intervalle et la fonction de densité :

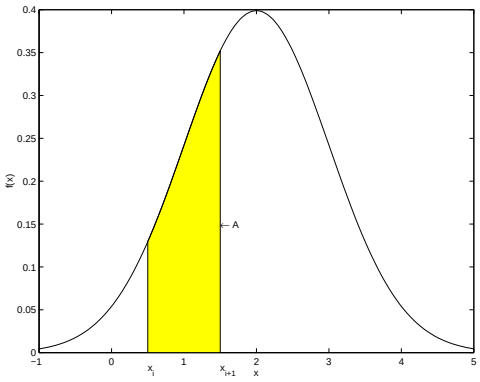

FIG. 1.4 – Fonction de densité

L'histogramme des fréquences relatives n'est alors qu'une approximation empirique de cette fonction de densité (si le facteur de proportionnalité est 1).

(iii) Si l'on veut mettre sur un même graphique une loi théorique de distribution de probabilités, il faut impérativement travailler avec les fréquences relatives, et un facteur de proportionnalité de 1 pour l'histogramme.

Remarque 3.3.6. Attention, dans un logiciel comme Excel , le terme histogramme n'a pas le sens ci-dessus.

Il nous reste maintenant à étudier le cas des fréquences cumulées. Celles-ci sont représentés par des polygônes de fréquences cumulés, mais nous avons encore ici une distinction suivant que les données soient groupées ou non.

Lorsque les données sont non groupées nous obtenons un polygône en escalier : la valeur de la fonction en un point x est le nombre d'observations (absolues ou relatives) qui sont inférieures ou égales à x.

Exemple 3.3.7. Reprenons encore l'exemple (3.1.2)

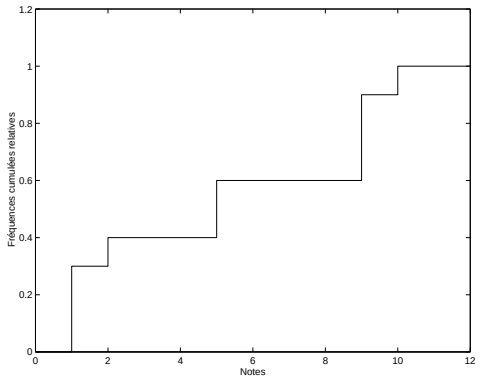

FIG. 1.5 – Fréquences cumulées relatives : données (3.1.2)

Quant aux données groupées, on joint par une ligne brisée les points obtenus en portant, pour les limites de classes supérieures des ordonnées égales aux fréquences cumulées.

Exemple 3.3.8. Fréquences cumulées relatives des étendues des exploitations agricoles belges (exemple (3.2.9)).

Remarque 3.3.9. Les polygônes de fréquences relatives sont une représentation empirique des fonctions de répartitions comme les histogrammes sont une représentation empirique des fonctions de densité.

3.4 Réduction des données

Le but est ici de caractériser les données à l'aide de quelques paramètres. Il y a deux grands types de paramètres : les paramètres de position ou de tendance centrale que nous étudierons en premier et les paramètres de dispersion que nous verrons ensuite.

Nous donnerons pour chaque paramètre que nous définirons la valeur numérique correspondant à l'exemple suivant :

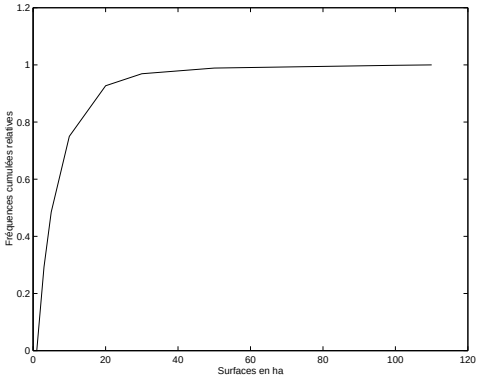

FIG. 1.6 – Fréquences cumulées relatives : données (3.2.9)

Exemple 3.4.1. Nous considérons 11 mesures faites de la hauteur du maître-brin d'une céréale donnée (en cm). Nous avons obtenu la série statistique suivante (mise en ordre croissant) :

$$59; 62; 63; 63; 64; 66; 66; 67; 69; 70; 70.$$

Les paramètres de position que nous allons étudier maintenant permettent de caractériser l'ordre de grandeur des observations. Le paramètre le plus utilisé dans la pratique est la moyenne arithmétique ou moyenne.

Définition 3.4.2 (Moyenne arithmétique). On appelle moyenne arithmétique ou moyenne la quantité donnée par :

- Si les observations sont données par une série statistique

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Si les observations sont données par leurs fréquences absolues

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i$$

Exemple 3.4.3. Pour les données de l'exemple (3.4.1), nous avons :

$$\bar{x} = 65,3636cm$$

Remarque 3.4.4. Lorsque les données sont groupées x_i est la valeur de la classe i.

Définition 3.4.5 (Médiane). La médiane est la valeur de part et d'autre de laquelle se trouve un nombre égal d'observations.

Remarque 3.4.6. (i) Pour les séries statistiques monotones (c'est-à-dire croissante ou décroissante) :
- si le nombre d'observations est pair la médiane est toute quantité comprise entre $x_{n/2}$ et $x_{n/2+1}$ (en général on prend $\tilde{x} = 1/2(x_{n/2} + x_{n/2+1})$)
- si le nombre d'observations est impair la médiane est $\tilde{x} = x_{n/2+1}$.

(ii) Pour les données groupées la classe médiane est celle qui contient la médiane. En admettant que les observations appartenant à cette classe sont réparties uniformément, la médiane aura pour expression :

$$\tilde{x} = x'_i + \Delta x_i \frac{1/2 - F_i}{n_i}$$

où
i est l'indice de la classe médiane.
x'_i est la limite inférieure de cette classe.
Δx_i est l'intervalle de la classe i.
F_i est la fréquence cumulée relative de la classe i.

Exemple 3.4.7. Pour les données de l'exemple (3.4.1), nous avons :

$$\tilde{x} = 66$$

Définition 3.4.8 (Quartiles). On définit de façon similaire les quartiles i.e les 3 quantités qui séparent les données en 4 groupes contenant le même nombre de données. On notera Q_1, Q_2 et Q_3 les trois quartiles.

Exemple 3.4.9. Considérons les 24 données suivantes :
8 13 27 32 25 16 32 27 8 28 79 25 35 25 38 29 80 50 38 30 20 20 49 9
Ces données mises en ordre croissant sont :
8 8 9 13 16 20 20 25 25 25 27 27 28 29 30 32 32 35 38 38 49 50 79 80
Les quartiles sont alors : $Q_1 = 20$, $Q_2 = \tilde{x} = 27,5$ et $Q_3 = 36,5$.

Remarque 3.4.10. Le deuxième quartile est égale à la médiane.

Définition 3.4.11 (Mode). On appelle mode d'une distribution non groupée toute valeur rendant maximale la fréquence. On appelle classe modale d'une distribution groupée toute classe rendant maximale le rapport :

$$\frac{\text{Fréquence}}{\text{Intervalle de classe}}$$

Exemple 3.4.12. Pour les données de l'exemple (3.4.1), il y a 3 modes : 63,66,70.

Remarque 3.4.13. (i) Le mode est une valeur qui rend maximum la représentation graphique des fréquences non cumulées.

(ii) Dans le cas d'une distribution théorique d'une variable aléatoire continue le mode est toute valeur qui maximise la fonction de densité. C'est la valeur "la plus probable".

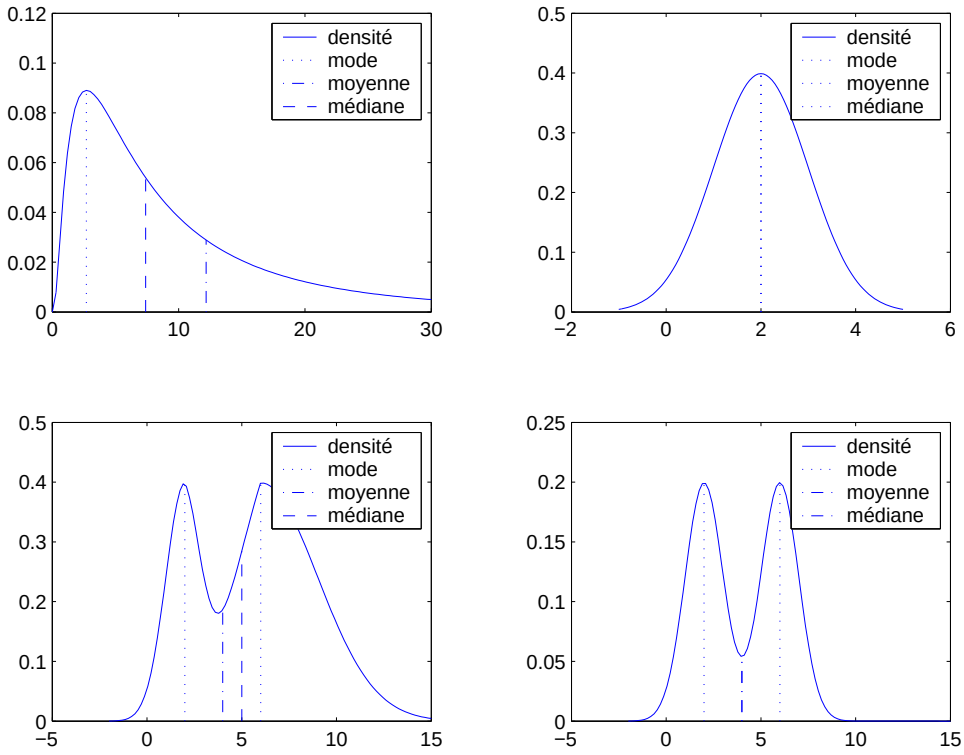

FIG. 1.7 – Différences entre le mode, la moyenne et la médiane

Les paramètres de position sont très insuffisants pour caractériser des données ; aussi nous avons besoin de savoir si les observations sont concentrées ou non autour d'un paramètre de position. C'est ce critère que l'on qualifie à l'aide des paramètres de dispersion. Le paramètre le plus connu et le plus utilisé est la variance d'un échantillon.

Définition 3.4.14 (Variance d'un échantillon). On appelle variance de l'échantillon la quantité :
– Si les données sont sous la forme d'une série statistique

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

– Si les données sont sous la forme d'une distribution de fréquences absolues

$$s^2 = \frac{1}{n}\sum_{i=1}^{p} n_i(x_i - \bar{x})^2$$

Remarque 3.4.15. (i) On note souvent $SCE = \sum_{i=1}^{n}(x_i - \bar{x})^2$. SCE est la Somme des Carrés des Ecarts, sous entendu à la moyenne.

(ii) On peut aussi écrire :

$$SCE = \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \tag{1.1}$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2 \tag{1.2}$$

$$= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \tag{1.3}$$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \tag{1.4}$$

Lorsque l'on effectue les calculs à la main, c'est la formule (1.4) que l'on utilise.

Exemple 3.4.16. Pour les données de l'exemple (3.4.1), nous avons :

$$s^2 = 11,3223\text{cm}^2$$

Remarque 3.4.17. On démontre que l'on a toujours :

$$\sum_{i=1}^{n}(x_i - a)^2 > \sum_{i=1}^{n}(x_i - \bar{x})^2 = ns^2 \text{ si } a \neq \bar{x}$$

Définition 3.4.18 (Ecart-type[1]). L'écart-type est la racine carré de la variance.

Exemple 3.4.19. Pour les données de l'exemple (3.4.1), nous avons : $s = \sqrt{s^2} = 3,3649\text{cm}$

Remarque 3.4.20. L'écart-type a la même dimension que les données (ce qui n'est pas le cas de la variance).

Définition 3.4.21 (Cœfficient de variation). On appelle cœfficient de variation l'indice de dispersion relatif exprimé en pourcentage :

$$cv = \frac{s}{\bar{x}} \times 100$$

On suppose bien évidemment que $\bar{x}$ est différent de 0.

Exemple 3.4.22. Pour les données de l'exemple (3.4.1), nous avons :

$$cv = 5,148\%$$

Définition 3.4.23 (Amplitude). On appelle amplitude l'écart entre les valeurs extrèmes des données

Exemple 3.4.24. Pour les données de l'exemple (3.4.1), nous avons :

$$w = 11$$

Définition 3.4.25 (Ecart interquartile). On appelle écart interquartile la différence entre le troisième et le premier quartile : $Q_3 - Q_1$

Exemple 3.4.26. Pour les données de l'exemple (3.4.9), nous avons :

$$Q_3 - Q_1 = 16,5$$

Définition 3.4.27 (boîte à moustaches[2]). Le diagramme en boîte à moustaches ou *box-plot* représente schématiquement les principales caractéristiques d'une variable numérique en utilisant les quartiles. On représente la partie centrale de la distribution par une boîte de largeur quelconque et de longueur l'intervalle interquartile. On trace à l'intérieur la position de la médiane et on complète la boîte par des "moustaches" de valeurs :

[1] standard deviation en anglais

[2] boxplot en anglais

– Pour la "moustache supérieure" : la plus grande valeur inférieure à $Q_3 + 1,5(Q_3 - Q_1)$.
– Pour la "moustache inférieure" : la plus petite valeur supérieure à $Q_1 - 1,5(Q_3 - Q_1)$.
Les valeurs extérieures représentées par des * sont celles qui sortent des " moustaches".

Exemple 3.4.28. Reprenons l'exemple (3.4.9). Nous avons $Q_1 = 20$, $\tilde{x} = 27,5$ $Q_3 = 36,5$ et $Q_3 - Q_1 = 16,5$. Par suite :
– la plus grande des données inférieure à $Q_3 + 1,5(Q_3 - Q_1)$ est 50 ;
– la plus patite des données supérieure à $Q_1 - 1,5(Q_3 - Q_1)$ est 8.
D'où le schéma suivant :

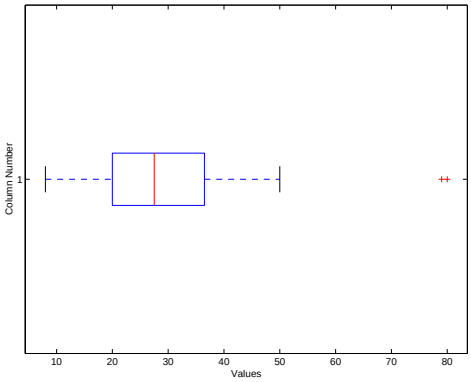

FIG. 1.8 – Boîte à moustaches

Définition 3.4.29 (Moment d'ordre k par rapport à un point c). On appelle moment d'ordre k par rapport à un point c la quantité :
– Si les données sont sous la forme d'une série statistique

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^k$$

– Si les données sont sous la forme d'une distribution de fréquences

$$\frac{1}{n}\sum_{i=1}^{p} n_i(x_i - c)^k$$

Notation 3.4.30. (i) Lorsque $c = 0$ le moment d'ordre k s'appelle moment par rapport à l'origine et on le note a_k.
(ii) Lorsque $c = \bar{x}$ le moment d'ordre k s'appelle moment centré et on le note m_k.

Remarque 3.4.31. $a_1 = \bar{x}$, $m_1 = 0$ et $m_2 = s^2$.

Remarque 3.4.32. (i) Les moments centrés d'ordre k pairs sont des paramètres de dispersion.
(ii) Les moments centrés d'ordre k impairs sont des indices de dissymétrie ou d'obliquité : Ils sont nuls pour les distributions symétriques et différentes de 0 pour les distributions dissymétriques.

Définition 3.4.33 (Cœfficients de Pearson). Les cœfficients de Pearson sont :
(i) Le degré de symétrie donné par

$$b_1 = \frac{m_3^2}{m_2^3} = \frac{m_3^2}{s^6}$$

(ii) Le degré d'aplatissement[3] donné par :

$$b_2 = \frac{m_4}{m_2^2} = \frac{m_4}{s^4}$$

Exemple 3.4.34. Pour les données de l'exemple (3.4.1), nous avons :

$$b_1 = 0,0298 \quad b_2 = 2,12$$

[3]kurtosis en anglais, attention le terme kurtosis est parfois aussi utilisé pour désigner le cœfficient g_2 de Fisher ci-après

Définition 3.4.35 (Cœfficient de Fisher). Les cœfficients de Fisher sont :

(i) Le degré de symétrie[4] donné par :

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{m_3}{s^3} = \sqrt{b_1}$$

(ii) Le degré d'aplatissement donné par :

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{m_4}{s^4} - 3 = b_2 - 3$$

Exemple 3.4.36. Pour les données de l'exemple (3.4.1), nous avons :

$$g_1 = 0,1726 \quad g_2 = -0,88$$

Remarque 3.4.37. Pour la loi normale réduite (cf. chapitre sur les probabilités) on a : $g_1 = 0$ et $g_2 = 0$.

Les figures (1.9,1.10) donnent des exemples de distributions théoriques avec différentes valeurs des cœfficients de symétrie et d'aplatissement.

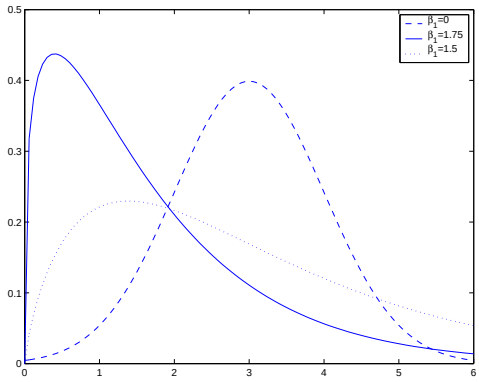

FIG. 1.9 – Différentes fonctions de densité pour différentes valeur du cœfficient de symétrie

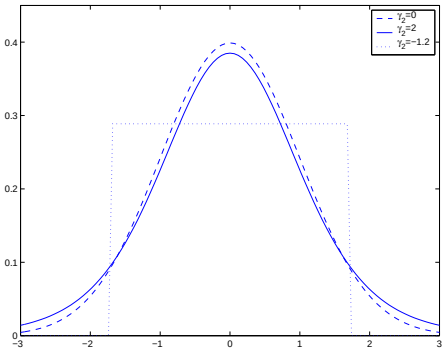

FIG. 1.10 – Différentes fonctions de densité pour différentes valeur du cœfficient d'applatissement

3.5 Exemples

Exemple 3.5.1. Les données de la table (1.1) sont des longueurs de la rectrice centrale de la gélinotte huppée mâle, juvénile. La figure (1.11) donne les différentes représentations graphiques de ces données.

Exemple 3.5.2. Les figures (1.12) et (1.13) donnent pour les mêmes données respectivement les histogrammes et les boîtes à moustaches pour les longueurs d'ailes de mésanges noires selon leur âges et leurs sexes.

[4] skewness en anglais

153	165	160	150	159	151	163
160	158	149	154	153	163	140
158	150	158	155	163	159	157
162	160	152	164	158	153	162
166	162	165	157	174	158	171
162	155	156	159	162	152	158
164	164	162	158	156	171	164
158						

TAB. 1.1 – Longueurs de la rectrice centrale de la gélinotte huppée mâle, juvénile

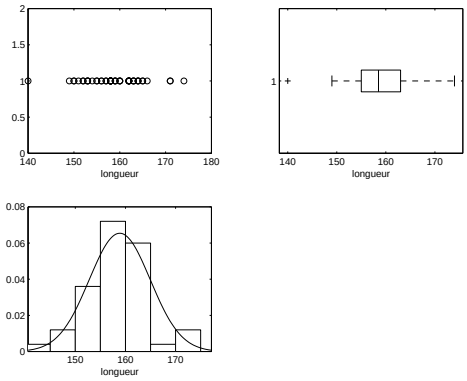

FIG. 1.11 – Données, boîte à moustaches et histogramme

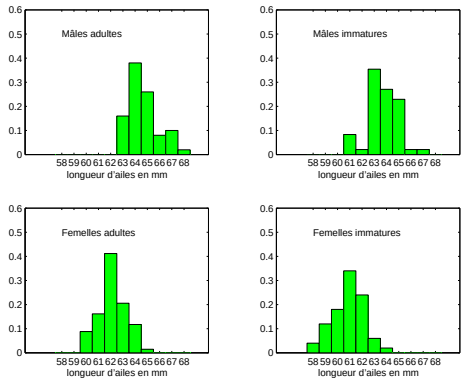

FIG. 1.12 – Distributions des longueurs d'ailes de mésanges noires selon leur âge et sexe

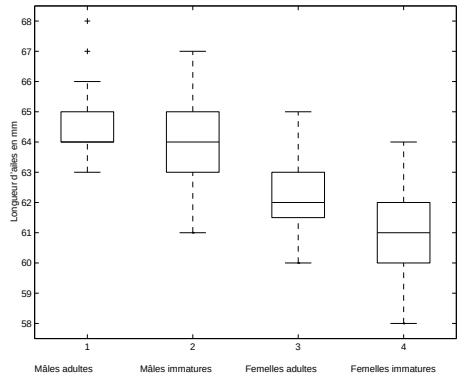

FIG. 1.13 – Distributions des longueurs d'ailes de mésanges noires selon leur âge et sexe

4 Statistique descriptive à 2 dimensions

4.1 Introduction

De même qu'en dimension 1 nous désirons représenter les données sous la forme de tableaux ou de graphiques ou de réduire les données à quelques paramètres. La grande différence avec la section précédente est que nous

pouvons essayer de mettre en évidence les relations qui peuvent exister entre deux caractères.

Comme en dimension 1 nous nous intéressons à des variables quantitatives et nous aurons comme données initiales une suite double :

$$x_1, x_2, \ldots, x_n$$

$$y_1, y_2, \ldots, y_n$$

La valeur du caractère 1 pour l'individu i est x_i La valeur du caractère 2 pour l'individu i est y_i

Définition 4.1.1 (Série statistique double). On appelle série statistique double la suite de n couples de valeurs (x_i, y_i).

Exemple 4.1.2. Poids des feuilles et poids des racines (en grammes) de 1000 individus de Cichorium intybus (cet exemple provient de l'ouvrage de Dagnélie).

feuilles :	71	76	106	108	109	111	111	112	...	662	673	679	741
racines :	56	51	40	174	62	59	84	94	...	174	290	290	230

4.2 Les distributions en fréquences

Comme dans le cas monodimensionnel lorsque le nombre de données est trop important nous condensons des données en une distribution de fréquences. Pour cela nous construisons un tableau à double entrée ; le nombre d'individus n_{ij} ayant les occurrences x_i et y_j des caractères x et y se trouve à l'intersection de la ligne i et de la colonne j. Dans ce paragraphe les indices i et j qualifient les occurrences des caractères pour des variables discrètes et les classes pour des variables continues et non pas des individus : $x_i \neq x_{i'}$ si $i \neq i'$ et $y_j \neq y_{j'}$ si $j \neq j'$. Le tableau que l'on construit a donc la structure suivante :

$x : y$	y_1	y_2	...	y_j	...	y_q	$Totaux$
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	$n_{1.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_p	n_{p1}	n_{p2}	...	n_{pj}	...	n_{pq}	$n_{p.}$
$Totaux$	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.q}$	$n_{..}$

Définition 4.2.1 (Fréquence marginale). On appelle fréquence marginale les quantités définies par :

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

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

Notation 4.2.2. Nous rappelons que le point en indice signifie que l'on a sommé sur cet indice. Avec cette notation, nous avons donc aussi :

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

Remarque 4.2.3. (i) Nous avons pris ici le cas des fréquences absolues mais nous pouvons bien évidemment construire des tableaux de fréquences relatives :

$$n'_{ij} = \frac{n_{ij}}{n}$$

(ii) Nous ne construisons pas en général de tableau de fréquences cumulées.

(iii) Nous pouvons bien entendu étudier séparément les caractères x et y et notamment faire deux statistiques descriptives à une dimension. Cela revient alors à travailler avec les fréquences marginales.

Définition 4.2.4 (Fréquence conditionnelle relative). On appelle fréquence conditionnelle relative pour que $x = x_i$ (respectivement $y = y_j$) sachant que $y = y_j$ (respectivement $x = x_i$) la quantité :

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

(respectivement

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

Définition 4.2.5 (Profils lignes, profils colonnes). On appelle profils lignes (respectivement profils colonnes) le tableau des fréquences conditionnelles relatives $f_{j.i}$ (respectivement $f_{i/j}$).

Remarque 4.2.6. (i) Le tableau de fréquence relative est une représentation empirique de la fonction de probabilité d'un couple de variables aléatoires et les fréquences conditionnelles relatives représentent des probabilités conditionnelles.

(ii) le tableau des profils lignes est une représentation empirique les lois de distributions conditionnelles.

(iii) Si la tableau de contingence comporte en fait en ligne différentes populations et en colonne les différentes modalités d'un caractère qualitatif (c'est-à-dire les valeurs d'une variable aléatoire discrète), alors les profils lignes sont les lois de probabilités sur les différentes populations du caractère étudié.

Exemple 4.2.7. Avec les données de l'exemple (4.1.2) nous obtenons :

Feuilles :Racines	40 à 79	80 à 119	120 à 159	160 à 199	200 à 239	240 à 279	280 à 319	320 à 259	Totaux
0 à 79	2								2
80 à 159	49	46	5	2					102
160 à 239	86	137	46	11					280
240 à 319	27	153	89	25	7				301
320 à 399	5	45	91	40	6				187
400 à 479		10	33	21	16	1	1		82
480 à 559		1	4	11	10	3			29
560 à 639			2	1	2	4		1	10
640 à 719				1		3	2		6
720 à 799					1				1
Totaux	169	392	270	112	42	11	3	1	1000

Exemple 4.2.8. La table (4.2.8) donne l'évolution de l'âge de la population agricole familiale dans un canton du Loiret. La table (1.3) donne quant-à elle les profils lignes.

Année :Age	< à 25 ans	25 à 34 ans	35 à 44 ans	45 à 54 ans	55 à 64 ans	> à 65 ans	Total
1970	88	24	27	61	20	25	245
1979	63	17	20	39	27	25	191
1988	41	15	18	22	31	17	144
Total	192	56	65	122	78	67	580

TAB. 1.2 – Tableau de contingence, exploitations agricoles dans le Loiret

Année :Age	< à 25 ans	25 à 34 ans	35 à 44 ans	45 à 54 ans	55 à 64 ans	> à 65 ans
1970	0.3592	0.0980	0.1102	0.2490	0.0816	0.1020
1979	0.3298	0.0890	0.1047	0.2042	0.1414	0.1309
1988	0.2847	0.1042	0.1250	0.1528	0.2153	0.1181

TAB. 1.3 – Tableau des profils lignes

4.3 Représentations graphiques

Les séries statistiques doubles peuvent être représentées par un nuage de points (1.14).

Quant aux distributions de fréquences elles se représentent dans un espace à trois dimensions par un diagramme en bâtons si les variables sont discrètes et par un stéréogramme si la variable est continue. Un stéréogramme est un diagramme composé de parallélépipèdes rectangles de bases les rectangles correspondant aux cellules du tableau statistique et de hauteur les fréquences divisées par la surface de la base (ceci toujours pour avoir une estimation de la densité de probabilité).

Exemple 4.3.1. Avec les données de l'exemple (4.1.2) on obtient la figure (1.15)

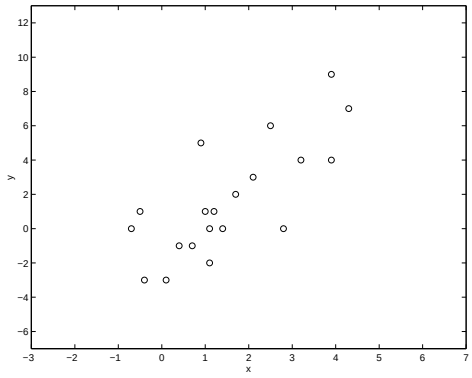

FIG. 1.14 – Nuage de points

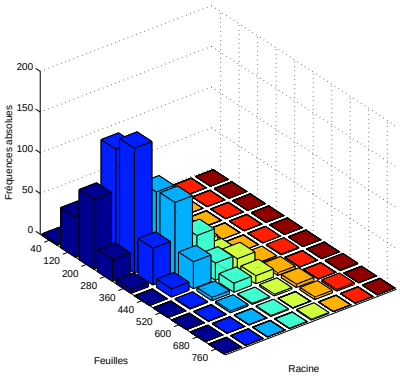

FIG. 1.15 – Stéréogramme

Exemple 4.3.2. Reprenons l'exemple (4.2.8) de l'évolution de l'âge de la population agricole familiale dans un canton du Loiret. On peut représenter les profils lignes (1.16). ceci nous permet de visualiser les fifférences de répartition des âges en fonction des année. Ici, nous avons l'ensemble des populations étudiées, les profils lignes sont donc exactement les lois de probabilités sur ces 3 populations. Dans le cas où nous n'aurions , pour chaque population que des échantillons, il faudrait effectuer un test statistique (test du χ^2 ici) pour savoir s'il y a réellement une différence dans les lois de distributions. Ceci est hors de notre programme.

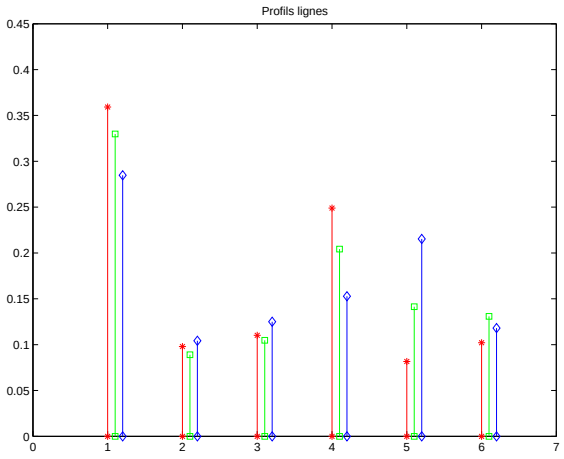

FIG. 1.16 – Profils lignes

4.4 Réduction des données

Nous avons ici deux types de paramètres, tout d'abord les paramètres liés à une seule variable qui caractérisent les fréquences marginales et conditionnelles. Nous avons dans ce cas les paramètres habituels de la statistique descriptive à une dimension qui sont principalement les moyennes marginales $\bar{x}$ et $\bar{y}$ et les variances marginales s_x^2 et s_y^2, ainsi que les moyennes conditionnelles $\bar{x_j}$ et $\bar{y_i}$ et les variances conditionnelles $s_{x/j}^2$ et $s_{i/y}^2$. Ensuite nous avons les paramètres permettant de décrire des relations existant entre les deux séries d'observations. Ce sont ces paramètres que nous allons étudier maintenant.

Définition 4.4.1 (Covariance d'un échantillon). On appelle covariance d'un échantillon la quantité :

- Si les données sont sous la forme d'une série statistique double

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Si les données sont sous la forme d'une distribution en fréquence

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(x_i - \bar{x})(y_j - \bar{y})$$

Remarque 4.4.2.

On note souvent $SPE = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$. SPE est la Somme des Produits des Ecarts, sous entendu aux moyennes.

On peut aussi écrire :

$$\begin{aligned} SPE &= \sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - y_i\bar{x} + \bar{x}\bar{y}) && (1.5)\\ &= \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\bar{x}\bar{y} && (1.6)\\ &= \sum_{i=1}^{n} x_i y_i - 2n\bar{x}\bar{y} + n\bar{x}\bar{y} && (1.7)\\ &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} && (1.8) \end{aligned}$$

Lorsque l'on effectue les calculs à la main, c'est la formule (1.8) que l'on utilise.

Exemple 4.4.3. On considère la série statistique double suivante :

x	165,5	164,0	156,0	174,0	169,0	157,5	159,0	152,0	155,0	159,0
y	177,0	172,0	163,0	183,5	171,5	165,0	160,5	154,5	163,0	162,0

x (respectivement y) représente la taille (respectivement l'envergure) de 10 adolescents nés en 1947 (mensurations relevées en 1962). On a alors :

$$cov(x,y) = 49,68$$

Remarque 4.4.4. (i) La covariance peut-être positive ou négative. Une covariance positive (respectivement négative) indique une relation entre les données croissantes (respectivement décroissantes), i.e. que les valeurs élévées d'une série correspondent, dans l'ensemble, à des valeurs élevées (respectivement faibles) de l'autre.

(ii) L'existence de termes positifs et négatifs dans le calcul de la covariance justifie pour celle-ci l'absence de correction analogue aux corrections de Sheppard.

Théorème 4.4.5. *On a toujours la relation suivante :*

$$|cov(x,y)| \leq s_x s_y$$

L'égalité n'a lieu que si les points (x_i, y_i) sont alignés.

Démonstration

Développons l'expression positive suivante :

$$\frac{1}{n}\sum_{i=1}^{n}(\lambda(x_i - \bar{x}) - (y_i - \bar{y}))^2 = \lambda^2 s_x^2 - 2\lambda cov(x,y) + s_y^2 \geq 0$$

On sait qu'une condition nécessaire et suffisante pour qu'un trinôme soit toujours de même signe est que son discriminant Δ soit négatif ou nul. Par suite nous avons :

$$\begin{aligned} \Delta = 4cov^2(x,y) - 4s_x^2 s_y^2 &\leq 0 \\ \iff cov^2(x,y) &\leq s_x^2 s_y^2 \\ \iff |cov(x,y)| &\leq s_x s_y \end{aligned}$$

De plus nous avons l'égalité $|cov(x,y)| = s_x s_y$ si et seulement si $\Delta = 0$ et donc s'il existe $\lambda_1 = cov(x,y)/s_x^2$ tel que

$$\begin{aligned} \sum_{i=1}^{n} (\lambda_1(x_i - \bar{x}) - (y_i - \bar{y}))^2 = 0 &\iff \lambda_1(x_i - \bar{x}) = y_i - \bar{y}\ \forall i \\ &\iff \text{Les points } (x_i, y_i)_{i=1,\ldots,n} \text{ sont alignés} \end{aligned}$$

□

4.5 Droite de régression

Introduction

Exemple 4.5.1. [5] On désire savoir comment le taux de cholestérol sérique dépend de l'âge chez l'homme. Pour cela on a pris 5 échantillons d'hommes adultes d'âges bien déterminés 25, 35, 45, 55 et 65 ans. On a obtenu les données suivantes :

Ages	25	25	25	25	25	25	25	35	35	35
Taux	1.8	2.3	2	2.4	2	2.5	2.6	2.6	2.9	2.3
Ages	35	35	35	35	45	45	45	45	45	45
Taux	2.4	2.1	2.5	2.7	2.7	3	3.1	2.3	2.5	3
Ages	45	45	55	55	55	55	55	65	65	65
Taux	3.3	2.7	3.1	2.9	3.4	2.4	3.4	3.7	2.8	3.3
Ages	65	65	65							
Taux	3.5	3.3	2.6							

La figure 1.17 donne la représentation graphique de ces données.

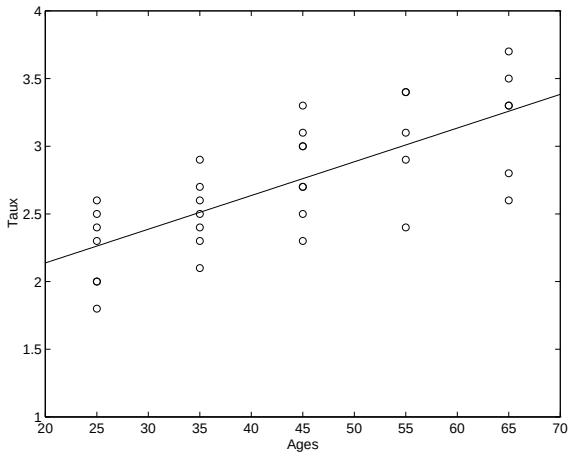

FIG. 1.17 – Taux de cholestérol en fonction de l'âge

Que peut-on conclure de ces données ?

En pratique nous sommes souvent amenés à rechercher une relation entre deux variables x et y. Pour cela, dans un premier temps, nous collectons des données $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Ensuite nous représentons graphiquement ces données. Nous pouvons par exemple avoir les cas suivants :

[5] Exemple provenant de l'ouvrage de Grémy et Salmon, "Bases statistiques", page 122.

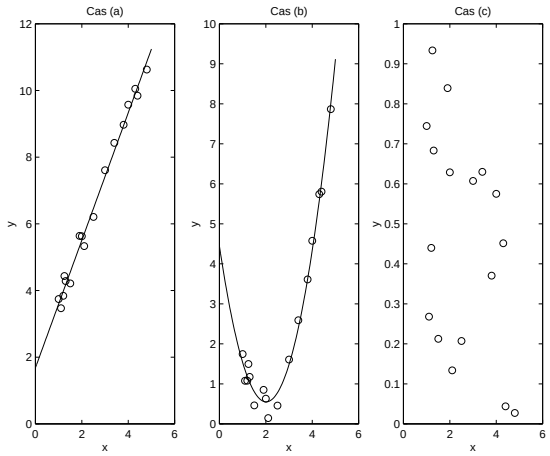

FIG. 1.18 – Différentes formes de graphes

Suivant les cas de la figure 1.18, nous pouvons penser aux modèles :

Cas (a) $y(x) = \beta_0 + \beta_1 x$;

Cas (b) $y(x) = \beta_0 + \beta_1 x + \beta_2 x^2$;

Cas (c) pas de modèle.

L'objet de la régression linéaire simple est l'étude du cas (a). L'un des buts de la régression linéaire simple est de prédire la "meilleure" valeur de y connaissant x (si le modèle linéaire est bien évidemment correct). L'objectif de cette section est uniquement descriptif, aussi nous n'allons étudier que l'estimation ponctuelle des paramètres.

Estimation des paramètres

Une droite sera d'autant plus proche des points $M_i(x_i, y_i)$ que les écarts entre ces points et la droite seront faibles. L'un des critères les plus utilisés est le critère des moindres carrés qui est la somme des carrés des écarts $r_i = y_i - \hat{y}_i$ (cf figure (1.19)).

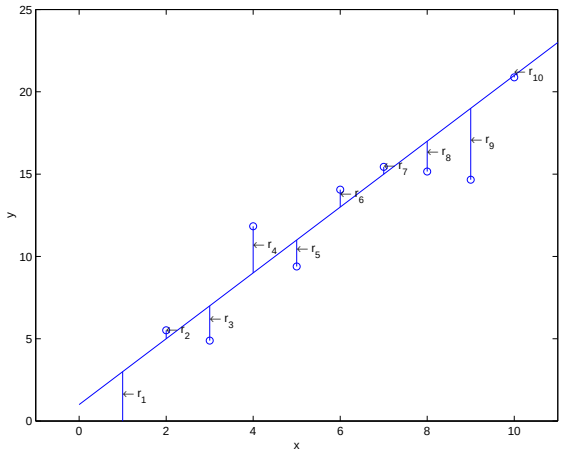

FIG. 1.19 – Moindres carrés.

Ici, les points $(x_i, y_i)_{i=1,\dots,n}$ sont connus, la question est de trouver les valeurs des paramètres β_0 et β_1 qui rendent la valeur du critère la plus faible possible. Nous sommes ainsi ramené au problème d'optimisation suivant :

$$(P) \left\{ \begin{array}{l} Min \quad f(\beta) = \frac{1}{2}\sum_{i=1}^{n} r_i^2 = \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = \\ \beta \in \mathbf{R}^2 \end{array} \right.$$

En effet, plus $f(\beta)$ sera proche de 0, plus les carrés des résidus, donc les résidus r_i seront "proches" de 0.

Théorème 4.5.2. *La solution du problème* (P) *est :*

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{1.9}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{SPE}{SCE_x} \tag{1.10}$$

Démonstration
On démontre qu'une condition nécessaire et suffisante de solution du problème d'optimisation est $f'(\beta) = (0, 0)$ (cf.5). D'où le système linéaire suivant :

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \tag{1.11}$$

$$\Leftrightarrow \begin{cases} \bar{y} = \beta_0 + \beta_1 \bar{x} \\ n\beta_0 \bar{x} + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

$$\Leftrightarrow \begin{cases} \bar{y} = \beta_0 + \beta_1 \bar{x} \\ n(\bar{y} - \beta_1 \bar{x})\bar{x} + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

$$\Leftrightarrow \begin{cases} \bar{y} = \beta_0 + \beta_1 \bar{x} \\ \beta_1 (\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \end{cases}$$

$$\Leftrightarrow \begin{cases} \bar{y} = \beta_0 + \beta_1 \bar{x} \\ \beta_1 = \dfrac{SPE}{SCE_x} \end{cases}$$

□

Remarque 4.5.3. On a supposé dans le calcul que $SCE_x \neq 0$, c'est-à-dire que tous les x_i ne sont pas identiques.

Nous noterons dans la suite $\hat{\beta}_0$ et $\hat{\beta}_1$ ces solutions.

Exemple 4.5.4. Reprenons l'exemple (4.5.1). Lorsque l'on effectue les calculs à la main il est utile de calculer le tableau préliminaire 1.5.

	y	x	xy	y^2	x^2
1	y_1	x_1	$x_1 y_1$	y_1^2	x_1^2
⋮	⋮	⋮	⋮	⋮	⋮
i	y_i	x_i	$x_i y_i$	y_i^2	x_i^2
⋮	⋮	⋮	⋮	⋮	⋮
n	y_n	x_n	$x_n y_n$	y_n^2	x_n^2
Totaux	$Y_.$	$X_.$	$\sum_i x_i y_i$	$\sum_i y_i^2$	$\sum_i x_i^2$
Moyennes	$\bar{y}_.$	$\bar{x}_.$			

TAB. 1.4 – Calculs préliminaires

Les estimations ponctuelles sont alors

$$\hat{\beta}_1 = \frac{4103.5 - \dfrac{1445 \times 90.1}{33}}{69625 - \dfrac{1445^2}{33}} = \frac{158.2}{6351.5} = 0.025$$

$$\hat{\beta}_0 = 2.73 - 0.025 \times 43.79 = 1.64$$

	x	y	xy	x^2	y^2
1	25	1.8	45.0	625	3.24
2	25	2.3	57.5	625	5.29
3	25	2.0	50.0	625	4.00
4	25	2.4	60.0	625	5.76
5	25	2.0	50.0	625	4.00
6	25	2.5	62.5	625	6.25
7	25	2.6	65.0	625	6.76
8	35	2.6	91.0	1225	6.76
9	35	2.9	101.5	1225	8.41
⋮	⋮	⋮	⋮	⋮	⋮
33	65	2.6	169.0	4225	6.76
Totaux	1445	90.1	4103.5	69625	253.31
Moyennes	43.79	2.73			

TAB. 1.5 – Calculs préliminaires sur l'exemple

Remarque 4.5.5. Nous noterons r_i le résidu d'indice i :

$$r_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$$

On vérifie alors que

$$\begin{array}{rcl} \sum_{i=1}^n r_i & = & \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) \\ & = & \sum_{i=}^n y_i - \sum_{i=1}^n \hat{\beta}_0 - \sum_{i=1}^n \hat{\beta}_1 x_i \\ & = & n\bar{y} - n\hat{\beta}_0 - n\bar{x}\hat{\beta}_1 = 0 \end{array}$$

De la même façon que nous avons cherché à "exprimer" y en fonction de x, on peut essayer d'"exprimer" x en fonction de y et nous obtenons ainsi la droite de régression d'équation :

$$x = \beta_{1xy} y + \beta_{0xy}$$

Les estimations sont alors

$$\hat{\beta}_{1xy} = \frac{SPE}{s_y^2} \quad \text{et} \quad \hat{\beta}_{0xy} = \bar{x} - \hat{\beta}_{1xy}\bar{y}$$

Exemple 4.5.6. Si nous reprenons les données de l'exemple (4.5.1) nous obtenons :

$$\begin{array}{ll} \hat{\beta}_{1xy} = 21.64 & \hat{\beta}_{0xy} = -15,29 \\ \hat{\beta}_{1yx} = 0.025 & \hat{\beta}_{0yx} = 1.64 \end{array}$$

Définition 4.5.7 (Cœfficient de corrélation linéaire). On appelle cœfficient de corrélation linéaire le rapport de la covariance sur les produits des écart-types :

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

On peut aussi écrire

$$r = \frac{SPE}{\sqrt{SCE_x SCE_y}}$$

Notons $\vec{x}_c$ (respectivement $\vec{y}_c$) le vecteur des données centrées de la variable x (respectivement y). C'est-à-dire que $\vec{x}_c = (x_1 - \bar{x}, \ldots, x_n - \bar{x})^T$ et $\vec{y}_c = (y_1 - \bar{y}, \ldots, y_n - \bar{y})^T$. Ces vecteurs sont dans $\mathbf{R}^n$. Alors SPE est le produit scalaire entre ces deux vecteurs centrées et SCE_x et SCE_y sont les normes au carrés de ces vecteurs. Par suite le cœfficient de corrélation linéaire s'interprète comme le cosinus de l'angle de ces deux vecteurs de $\mathbf{R}^n$. On en déduit la remarque suivante.

Remarque 4.5.8. Le cœfficient de corrélation linéaire a les propriétés suivantes :

(i)

$$r \in [-1, +1]$$

.

(ii) $|r| = 1$ si et seulement si les points (x_i, y_i) sont alignés.

On montre que l'on a en fait les différents cas de figures suivant

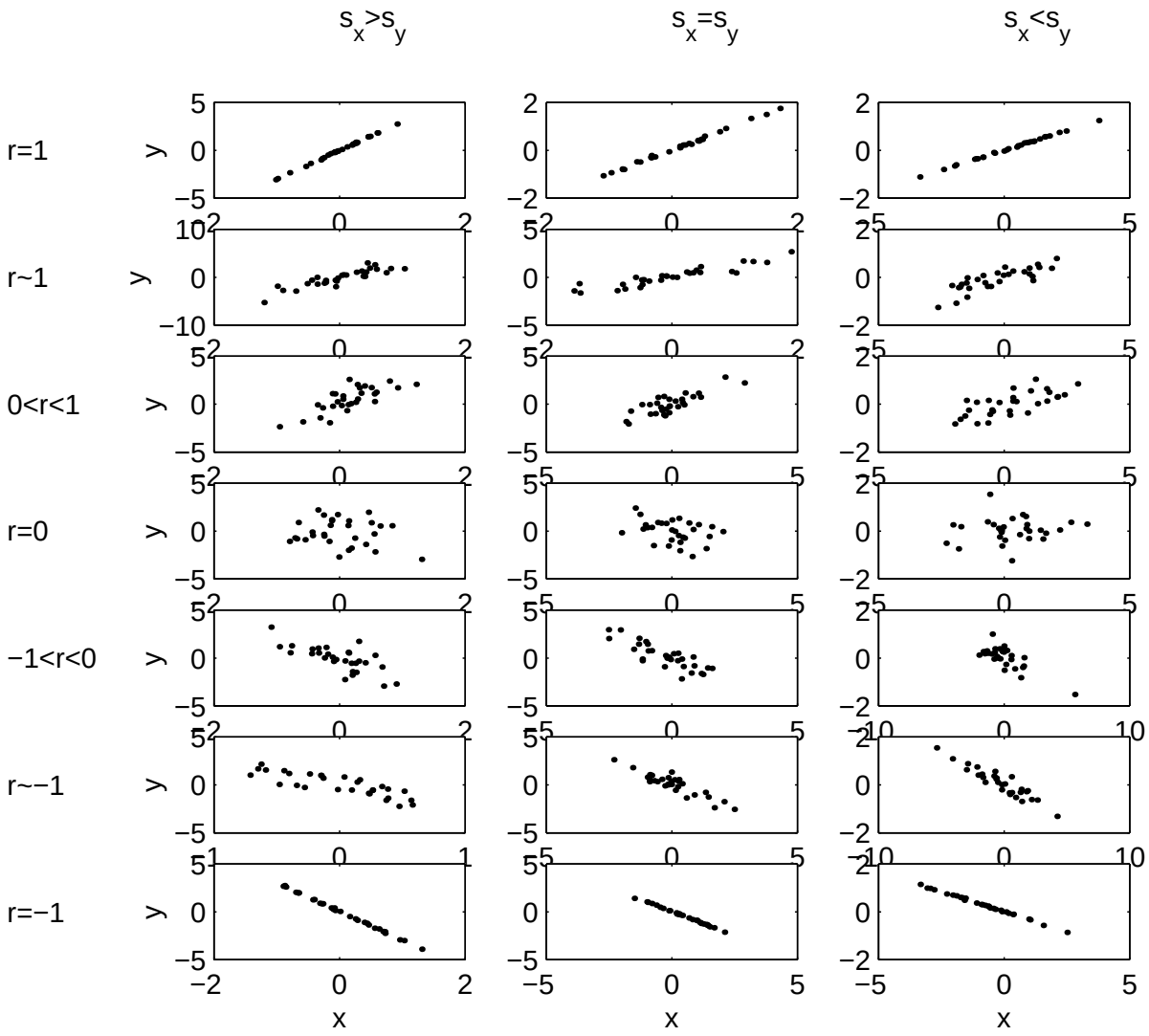

FIG. 1.20 – Liens entre les nuages de points et le cœfficient de corrélation linéaire

Remarque 4.5.9. Nous tenons à bien faire remarquer que le cœfficient de corrélation linéaire ne mesure qu'une liaison de nature liéaire. Pour les 5 graphiques de la figure (1.21), on a les mêmes valeurs de $\bar{x}, \bar{y}, s_x, s_y, r$ et donc la même droite de régression. Il est évident que les phénomènes sont trés différents :

(i) pour le graphique en haut à gauche, il y a une forte dispersion mais le modèle linéaire semble a priori approprié ;

(ii) pour le graphique en haut à droite, un modèle parabolique serait sans doute plus adapté ;

(iii) pour le graphique au milieu à gauche, il y a sans doute une donnée aberrante qui a un fort résidu ;

(iv) pour le graphique au milieu à droite, la dispersion des données semble augmenter quand x augmente ;

(v) pour le graphique en bas à gauche, il y a une donnée qui a une forte influence et un résidu nul.

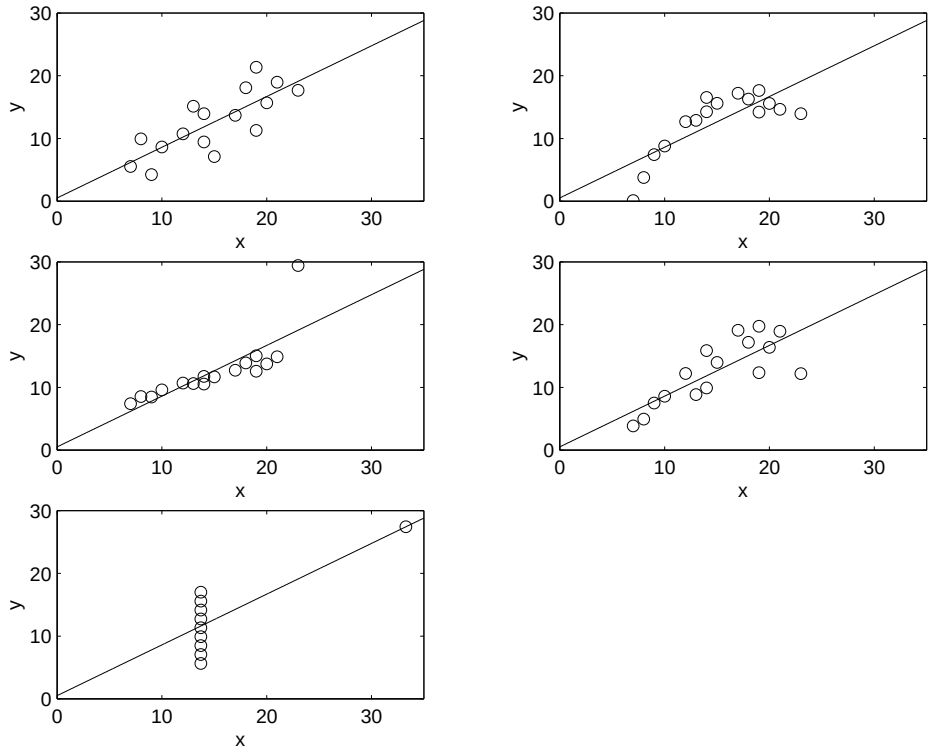

FIG. 1.21 – Exemple de données ayant les mêmes valeurs des paramètres $\bar{x}, \bar{y}, s_x, s_y$ et r et donc la même droite de régression

5 Compléments

5.1 Changement de variables

Nous allons tout d'abord voir que lorsque le modèle n'est pas au départ linéaire, on peut parfois s'y ramener par un bon changement de variable. Considerons l'exemple suivant :

Exemple 5.1.1. Le carbone radioactif ^{14}C est produit dans l'atmosphère par l'effet des rayons cosmiques sur l'azote atmosphérique. Il est oxydé en $^{14}CO_2$ et absorbé sous cette forme par les organismes vivants qui, par suite, contiennent un certain pourcentage de carbone radioactif relativement aux carbones ^{12}C et ^{13}C qui sont stables. On suppose que la production de carbone ^{14}C atmosphérique est demeurée constante durant les derniers millénaires. On suppose d'autre part que, lorsqu'un organisme meurt, ses échanges avec l'atmosphère cessent et que la radioactivité due au carbone ^{14}C décroit suivant la loi exponentielle suivante :

$$A(t) = A_0 e^{-\lambda t}$$

où λest une constante positive, t représente le temps en année et $A(t)$ est la radioactivité exprimée en nombre de désintégrations par minute et par gramme de carbone. On désire estimer les paramètres A_0 et λ par la méthode des moindres carrés. Pour cela on analyse les troncs (le bois est un tissu mort) de très vieux arbres *Sequoia gigantea* et *Pinus aristaca*. Par un prélèvement effectué sur le tronc, on peut obtenir :

- son age t en année, en comptant le nombre des anneaux de croissance,
- sa radioactivité A en mesurant le nombre de désintégration.

t	500	1000	2000	3000	4000	5000	6300
A	14.5	13.5	12.0	10.8	9.9	8.9	8.0

Posons $y(t) = \ln A(t), \beta_0 = \ln A_0, \beta_1 = -\lambda$ et $y_i = \ln(A_i)$. Le modèle s'écrit alors

$$y(t) = \beta_0 + \beta_1$$

Nous sommes donc ramené au cas de la régression linéaire simple.

5.2 Cas à plus d'une variable explicative

Cette section dépasse le cadre de la statistique descriptive puisque si nous avons par exemple 4 variables, nous ne pouvons plus faire de graphique. Mais nous allons voir cependant qu'en ce qui concerne l'estimation des paramètres, cela ne change pas grand chose.

Avant de passer au cas à p variables, nous allons réécrire le problème de la régression linéaire simple à 1 variable. Posons

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}; \qquad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}; \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{et} \quad r = \begin{pmatrix} r_1 \\ \vdots \\ r_i \\ \vdots \\ r_n \end{pmatrix}$$

alors

$$y_i = \beta_0 + \beta_1 x_i + r_i \quad \forall i = 1, \ldots, n \tag{1.12}$$

est équivalent à

$$y = X\beta + r \tag{1.13}$$

et le problème d'optimisation s'écrit alors

$$(P) \begin{cases} Min \quad f(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = \frac{1}{2}||X\beta - y||^2 \\ \beta \in \mathbf{R}^2 \end{cases}$$

La condition nécessaire de solution du premier ordre nous donne alors (dérivée des fonctions composées) ce que nous appelons l'équation normale

$$\nabla f(\beta) = X^T X - X^T y = 0 \tag{1.14}$$

La dérivée seconde de f est alors :

$$\nabla^2 f(\beta) = X^T X$$

On démontre en optimisation que dans ce cas si $\nabla^2 f(\beta)$ est semi-définie positive alors la fonction f est convexe (graphiquement c'est une cuvette pour une fonction de $\mathbf{R}^2$ à valeurs dans $\mathbf{R}$) et dans ce cas l'équation (1.14) est une condition nécesssaire et suffisante de solution de notre problème d'optimisation. Or ici

$$< XX^T\beta, \beta > = < X\beta, X\beta > \geq 0$$

ceci pour tout β, donc $\nabla^2 f(\beta)$ est bien semi-définie positive.

Remarque 5.2.1. Dans le cas de la régression linéaire simple, si on développe l'equation normale (1.14), on retrouve bien le système linéaire (1.11).

Nous allons maintenant étudier le cas où l'on a plus d'une variable explicative.

Considérons le modèle :

$$y(x) = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p \tag{1.15}$$

où $x = (x_1, \ldots, x_p)$.

On collecte n $(p+1)$-uplets $(y_i, x_{i1}, \ldots, x_{ip})_{i=1,\ldots,n}$. Notre problème d'optimisation pour estimer nos paramètres s'écrit alors

$$(P) \begin{cases} Min & f(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = \frac{1}{2}||X\beta - y||^2 \\ \beta \in \mathbf{R}^2 \end{cases}$$

avec ici

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}; \qquad X = \begin{pmatrix} 1 & x_{i1} & \ldots & x_{ip} \\ \vdots & \vdots & & \vdots \\ 1 & x_{i1} & \ldots & x_{ip} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}; \qquad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \quad \text{et} \quad r = \begin{pmatrix} r_1 \\ \vdots \\ r_i \\ \vdots \\ r_n \end{pmatrix}$$

Par suite l'estimation des paramètres est aussi donné par la résolution du système linéaire des équations normales

$$X^T X\beta = X^T y$$