

Chapitre 4

Statistiques à deux variables

4.1 Introduction

Lors des années précédentes, vous avez rencontré des séries statistiques à 1 variable : il s'agissait d'une suite de nombres x_1, \dots, x_p (notes d'élèves, nombre de buts marqués, nombres de cigarettes par jour, etc).

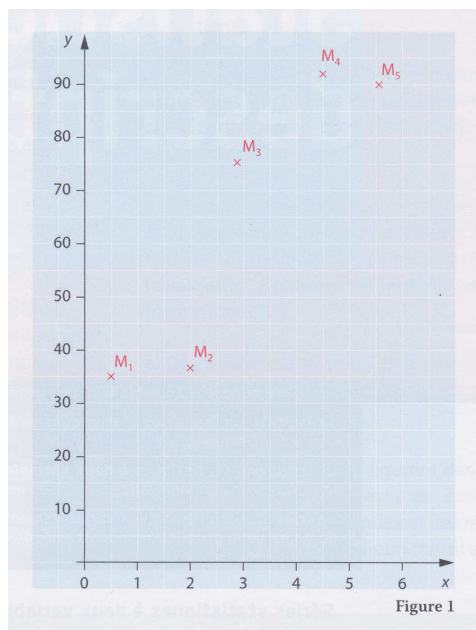
Dans ce chapitre nous allons étudier des séries statistiques à deux variables. Cela signifie que l'ensemble des valeurs obtenues sont des points $(x_1; y_1), \dots, (x_p; y_p)$ du plan. Dans certains cas, nous allons chercher à déterminer une relation entre les variables x_i et les variables y_i . Ce genre de relation pourra servir pour faire des estimations futures ; il faudra également trouver un moyen de quantifier la qualité de la relation obtenue : savoir si elle est proche de la réalité ou complètement irréaliste.

Débutons par un exemple pour fixer les idées.

Exemple 4.1.1. Une grande surface s'intéresse au lien entre ses dépenses publicitaires et son chiffre d'affaires : elle recueille les données suivantes, exprimées en millions d'euros, portant sur cinq périodes ; les dépenses sont notées x_1, \dots, x_5 et les chiffres d'affaires y_1, \dots, y_5 .

Dépenses publicitaires : x_i	0,5	2,0	2,9	4,5	5,6
Chiffres d'affaires : y_i	35	37	75	92	90

Il est bien entendu possible de placer les points $M(x_i; y_i)$ (pour $i = 1, \dots, 5$) dans un repère orthonormé. Le nuage de points est « allongé » de manière oblique (cf. figure suivante) et laisse penser qu'il existe une relation entre x_i et y_i



Définition 4.1.1. Le point moyen d'une série statistique est $(\bar{x}; \bar{y})$. Autrement dit, le point dont l'abscisse (resp. l'ordonnée) vaut la moyenne des abscisses (resp. des ordonnées).

Exemple 4.1.2. Dans l'exemple précédent, le point moyen a pour coordonnées $(3, 1; 65, 8)$.

Exercices à traiter : 36 et 39 page 270.

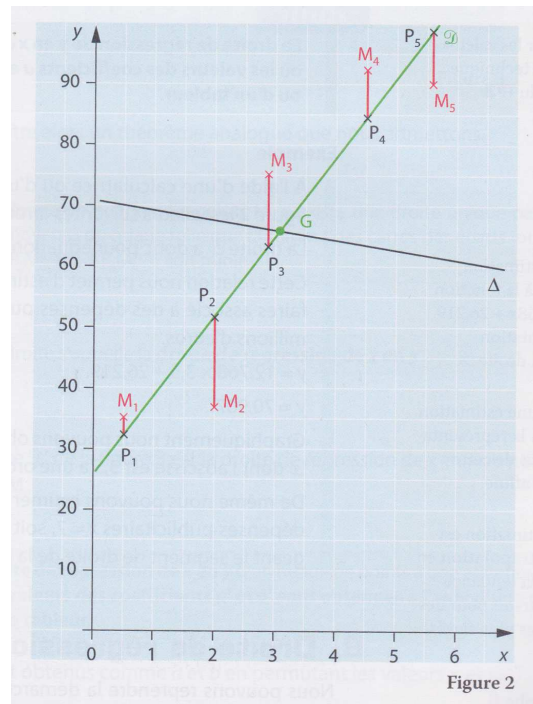
4.2 Ajustement affine par la méthode des moindres carrés

4.2.1 Régression de y en x

Au vu du nuage de points, nous allons chercher une relation affine entre les variables x_i et les variables y_i . Nous aimerions donc trouver une droite \mathcal{D} vérifiant les équations

$$y_i = ax_i + b \quad \text{pour tout } i = 1, \dots, p.$$

Pour trouver des coefficients a et b convenables, nous allons chercher à minimiser la distance (impliquant donc des carrés) entre la droite et les points $M(x_i; y_i)$. C'est pour cette raison que nous utilisons l'appellation « méthode des moindres carrés ». Graphiquement, nous avons l'impression d'avoir autant de points au dessus et en dessous de la droite \mathcal{D} (cf. figure ci-dessous).



Cette droite s'appelle la droite de régression de y en x car nous avons voulu expliquer la variable y en fonction de la variable x .

En pratique, les coefficients a et b sont déterminés à l'aide de la calculatrice :

1. Compléter deux listes, l'une avec les variables x_i , l'autre avec les variables y_i .
2. Utiliser la fonction « Stats 2 var » de la calculatrice.

La démarche à suivre est résumée dans le tutoriel

<https://www.youtube.com/watch?v=1C11fgYE51s&list=PLVUDmbpupCapZdNo8QaVDkVz3equeIYK1&index=35&t=242>

Exemple 4.2.1. Dans l'exemple précédent, la calculatrice fournit les valeurs suivantes :

$$a = 12,768 \quad \text{et} \quad b = 26,219$$

La droite \mathcal{D} a donc pour équation $y = 12,768x + 26,219$.

Remarque. Bien entendu, il est possible de faire le contraire et chercher à expliquer la variable x en fonction de la variable y . En reprenant ce qui précède, nous obtenons alors une nouvelle droite \mathcal{D}' d'équation

$$x = a'y + b'$$

avec des coefficients a' et b' obtenues à l'aide d'une calculatrice (en échangeant les colonnes x_i et y_i).

4.2.2 Pourquoi faire ceci ?

Comme nous l'avons dit plus tôt, L'intérêt derrière ces calculs est d'obtenir une relation ($y = ax + b$ ou $x = a'y + b'$) permettant de prédire/estimer ce qui pourrait se produire à l'avenir. Ce point sera plus clair grâce aux exercices.

Exercices à traiter : 45 et 46 page 271.

4.3 Qualité de l'ajustement

Lorsque nous avons des nuages de points de forme « allongés », il est naturel de chercher à obtenir une relation affine entre les variables. Il est ensuite naturel de s'interroger sur la qualité de l'ajustement qui a été fait (via les droites \mathcal{D} ou \mathcal{D}'). Pour cela, on utilise un coefficient de corrélation, noté r .

Ce coefficient est obtenu à l'aide de la calculatrice et vérifie les propriétés suivantes :

- $r^2 = aa'$
- $r \in [-1; 1]$

Pour mieux appréhender ce que signifie ce nombre voici quelques commentaires :

1. Lorsque $r = 1$ ou $r = -1$, cela signifie que les deux droites \mathcal{D} ou \mathcal{D}' possèdent le même coefficient directeur, l'alignement des points est parfait (cf. figure ci-dessous).

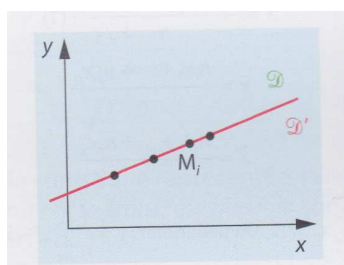


Figure 5

$$r = 1$$

$$a > 0 \text{ et } a' > 0$$

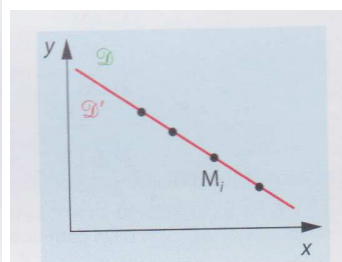


Figure 6

$$r = -1$$

$$a < 0 \text{ et } a' < 0$$

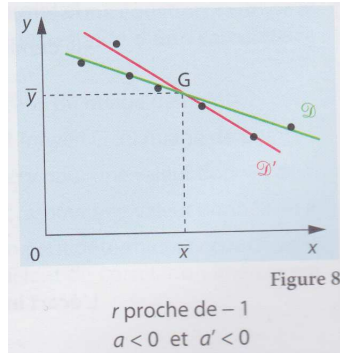
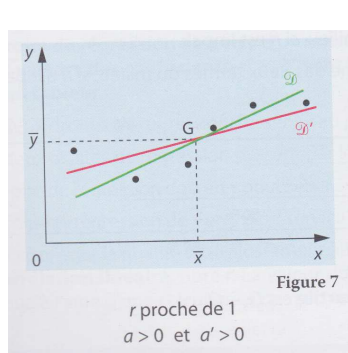
2. Plus r est proche de 1 ou de -1 , plus l'ajustement affine est de qualité (cf. figure ci-dessous).

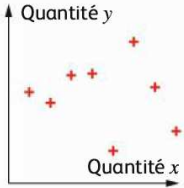
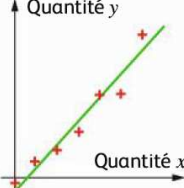
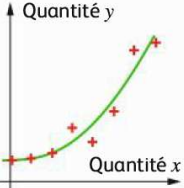
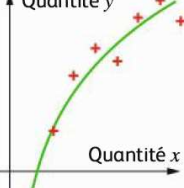
Remarque. Attention à ne pas confondre corrélation et liaison de cause à effets. Par exemple, les variables x_i et y_i peuvent mesurer deux effets d'une même cause.

Exercices à traiter : 44 page 270 et 59 page 276 (partie A).

Il est naturel de s'interroger vis-à-vis de la pertinence d'un ajustement affine dans certains cas de figures. La forme du nuage de points peut indiquer si cela sera utile ou non. Voici quelques indications à ce sujet.

Finalement, faire de la régression cela revient à trouver quelle courbe (définie par une fonction de référence) est la « plus proche » du nuage de points. Lorsqu'il faudra employer un ajustement non-affine, l'énoncé de l'exercice détaillera la marche à suivre (la plupart du temps, il suffira de



Pas de forme particulière	Nuage de forme « allongée », proche d'une droite
Pas d'ajustement possible 	Possibilité d'un ajustement affine 
Nuage à croissance « accélérée »	Nuage à croissance « ralentie »
Possibilité d'un ajustement de type parabolique ou exponentiel 	Possibilité d'un ajustement de type racine carrée ou logarithmique 

procéder à un changement de variables judicieux permettant de se ramener à un ajustement affine et d'employer les outils du chapitre).

Exercices à traiter : 59 page 276 (partie B).

