Statistiques Flux Portuaires

Master 1 Management Portuaire et Maritime

Université du Littoral - Côte d'Opale, Pôle Lamartine

Laurent SMOCH

Septembre 2013

Laboratoire de Mathématiques Pures et Appliquées Joseph Liouville
Université du Littoral, zone universitaire de la Mi-Voix, bâtiment H. Poincaré
50, rue F. Buisson, BP 699, F-62228 Calais cedex

Table des matières

1 Séries statistiques à une variable **1**
1.1 Introduction . . . 1
1.2 Méthodes de représentation . . . 1
1.2.1 Vocabulaire . . . 1
1.2.2 Les tableaux . . . 2
1.2.3 Les graphiques . . . 3
1.3 Caractéristiques de position . . . 8
1.3.1 Le mode (ou dominante) . . . 8
1.3.2 La moyenne . . . 9
1.3.3 La médiane . . . 10
1.3.4 Les quartiles . . . 13
1.4 Caractéristiques de dispersion . . . 15
1.4.1 L'étendue . . . 15
1.4.2 L'écart absolu moyen . . . 15
1.4.3 La variance et l'écart-type . . . 16
1.5 Paramètres de concentration . . . 18
1.5.1 Définitions . . . 18
1.5.2 La courbe de Gini ou de Lorenz . . . 19
1.5.3 L'indice de la concentration ou indice de Gini . . . 19
1.5.4 Calcul du coefficient de Gini . . . 20
1.5.5 La médiale . . . 20
1.6 Exercices . . . 21

2 Séries statistiques à deux variables **33**
2.1 Introduction . . . 33
2.2 Tableaux de données. Nuages de points . . . 33
2.2.1 Tableaux de données . . . 33
2.2.2 Nuages de points . . . 34
2.3 Calcul des paramètres de position et de dispersion . . . 34
2.3.1 Le point moyen . . . 36
2.3.2 Les variances . . . 37
2.4 Vocabulaire, définitions . . . 37
2.4.1 La covariance . . . 37
2.4.2 Le coefficient de corrélation linéaire . . . 39
2.5 Ajustement linéaire (ou affine) . . . 39
2.5.1 Ajustement graphique . . . 39
2.5.2 Ajustement analytique - Méthode des moindres carrés . . . 40
2.6 Exercices . . . 42

Chapitre 2

Séries statistiques à deux variables

2.1 Introduction

Dans certains cas, il semble exister un lien entre les deux caractères d'une **série statistique à deux variables**, par exemple entre le poids et la taille d'une personne, les heures de révision pour un examen et la note obtenue, etc... Il est alors intéressant d'étudier simultanément deux caractères X et Y d'une même population E.

2.2 Tableaux de données. Nuages de points

On peut représenter les résultats sous forme de tableaux ou de graphiques.

2.2.1 Tableaux de données

On se donne plusieurs exemples ci-dessous impliquant indifféremment des caractères X et Y discrets ou continus.

Exemple 2.2.1 Au cours du troisième trimestre 2008, une marque automobile a lancé la commercialisation d'une nouvelle voiture avec deux motorisations distinctes de puissances respectives 138 chevaux DIN et 177 chevaux DIN. On dispose des quantités de voitures vendues par zones :

Zones	Nombre d'unités 138 CV vendues x_i	Nombre d'unités 177 CV vendues y_i
1	400	240
2	200	120
3	600	300
4	300	150
5	300	150
6	600	270

Exemple 2.2.2 Un même produit est vendu conditionné sous différentes formes et différents volumes. Le tableau suivant indique pour chaque type d'emballage le volume x_i et le prix y_i du produit.

x_i en cm^3	100	150	200	300	500	600	700	800	900	1000
y_i en euros	7	8	9,5	13	20	23	25	28,3	30,5	34

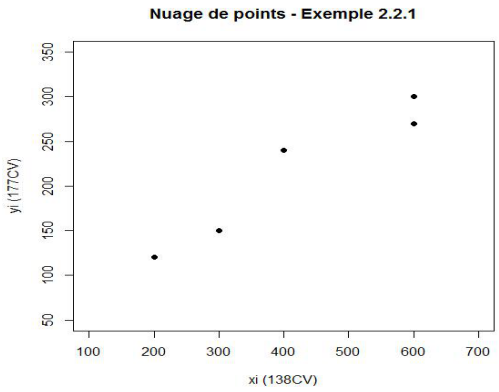
Exemple 2.2.3 Les chiffres d'affaires trimestriels d'une entreprise ont été pour les douze derniers trimestres :

Rang du trimestre x_i	1	2	3	4	5	6	7	8	9	10	11	12
Chiffre d'affaires (en milliers d'euros) y_i	300	450	130	200	280	410	200	250	320	500	210	250

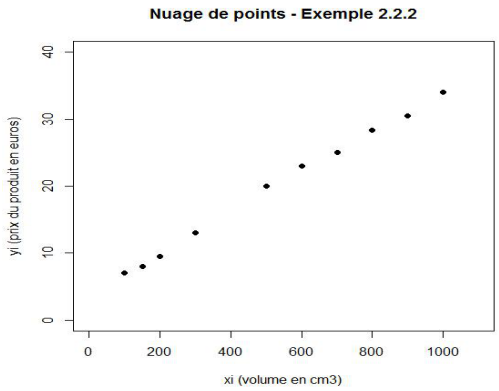
2.2.2 Nuages de points

Le plan étant muni d'un repère orthogonal, on peut associer au couple (x_i, y_i) de la série statistique double le point M_i de coordonnées x_i et y_i.
L'ensemble des points M_i obtenus constitue le **nuage de points** représentant la série statistique.

Dans l'exemple 2.2.1, on obtient le nuage ci-dessous :



Dans l'exemple 2.2.2, on obtient le nuage suivant :



Le nuage étant dessiné, on peut essayer de trouver une fonction f telle que la courbe d'équation $y = f(x)$ passe "le plus près possible" des points du nuage. C'est le problème de **l'ajustement**.
Lorsqu'il sera possible de tracer une droite D au voisinage des points, on parlera d'ajustement linéaire. Si l'ajustement linéaire ne convient pas, on peut penser à approcher le nuage à l'aide d'une parabole, d'une hyperbole, d'une fonction exponentielle...

2.3 Calcul des paramètres de position et de dispersion

Comme dans le chapitre sur les séries statistiques à une variable, il est possible de déterminer pour les séries statistiques à deux variables la moyenne arithmétique, la variance et tous les autres paramètres de position et de dispersion de chaque variable prise séparément. Il suffit pour cela de déterminer les distributions marginales des variables X et Y.

Exemple 2.3.1 On a réparti 1000 individus d'une population suivant
- le nombre x d'arrêts de travail suite à des accidents par an,
- l'âge y en années

de ces individus. On obtient les résultats synthétisés dans le tableau ci-dessous.

x_i \ y_j	$[20;30[$	$[30;40[$	$[40;50[$	$[50;60[$	$n_{i.}$
0	20	150	22	0	192
1	99	102	180	12	393
2	60	51	150	30	291
3	18	0	50	32	100
4	3	0	10	11	24
$n_{.j}$	200	303	412	85	1000

Ce tableau à double entrée est appelé **tableau de contingence** ou **d'effectifs** n_{ij}.

L'**effectif marginal** de la variable X est défini par :

$$\boxed{n_{i.} = \sum_{j=1}^{q} n_{ij} \text{ pour } i \in \{1, 2, \ldots, p\}}$$

où p et q sont respectivement les nombres de modalités (ou de catégories) de X et Y.

L'effectif marginal de la variable Y est défini par :

$$\boxed{n_{.j} = \sum_{i=1}^{p} n_{ij} \text{ pour } j \in \{1, 2, \ldots, q\}}$$

L'effectif total est donné par

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

On peut définir les **fréquences marginales**

$$\boxed{f_{i.} = \frac{n_{i.}}{n} \text{ et } f_{.j} = \frac{n_{.j}}{n}}$$

ainsi que la fréquence

$$\boxed{f_{ij} = \frac{n_{ij}}{n}}$$

pour $i \in \{1, 2, \ldots, p\}$ et $j \in \{1, 2, \ldots, q\}$.
Dans le cadre de l'exemple 2.3.1, on remarquera au préalable que les nombres de modalités pour X et Y sont respectivement $p = 5$ et $q = 4$. Les distributions marginales de X et Y en termes d'effectifs et de fréquences sont respectivement :

x_i	$n_{i.}$	$f_{i.}$
0	192	0,192
1	393	0,393
2	291	0,291
3	100	0,100
4	24	0,024
Total	1000	1

Classe	y_j	$n_{.j}$	$f_{.j}$
$[20;30[$	25	200	0,200
$[30;40[$	35	303	0,303
$[40;50[$	45	412	0,412
$[50;60[$	55	85	0,085
Total	–	1000	1

2.3.1 Le point moyen

Supposons qu'on souhaite calculer les moyennes arithmétiques de X et Y, il suffit pour cela d'utiliser les formules déja étudiées dans le chapitre 1 à savoir

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_{i.}x_i \qquad \overline{y} = \frac{1}{n}\sum_{j=1}^{q} n_{.j}y_j$$

Afin de calculer les sommes $\sum_i n_{i.}x_i$ et $\sum_j n_{.j}y_j$ on utilise le tableau des effectifs de X et Y qu'on complète avec une colonne supplémentaire, $n_{i.}x_i$ pour la variable X, $n_{.j}y_j$ pour la variable Y.
Considérons l'exemple 2.3.1 :

x_i	$n_{i.}$	$n_{i.}x_i$
0	192	0
1	393	393
2	291	582
3	100	300
4	24	96
Total	1000	1371

Classe	y_j	$n_{.j}$	$n_{.j}y_j$
$[20;30[$	25	200	5000
$[30;40[$	35	303	10605
$[40;50[$	45	412	18540
$[50;60[$	55	85	4675
Total	–	1000	38820

On en déduit que $\overline{x} = \dfrac{1371}{1000} = 1,371$ et $\overline{y} = \dfrac{38820}{1000} = 38,82$.

Lorsqu'on pense pouvoir réaliser un ajustement linéaire (ou affine) d'un nuage, il semble intéressant, avant de tracer la droite, de placer le point dont l'abscisse est la moyenne des abscisses x_i et l'ordonnée la moyenne des ordonnées y_j.

Définition 2.3.1 *On appelle* ***point moyen*** *G d'un nuage de n points M_i de coordonnées (x_i, y_i) le point de coordonnées :*

$$(x_G, y_G) = (\overline{x}, \overline{y}) = \left(\frac{1}{n}\sum_{i=1}^{p} n_{i.}x_i, \frac{1}{n}\sum_{j=1}^{q} n_{.j}y_j\right)$$

On vérifie ainsi que le point moyen dans l'exemple 2.3.1 est $G(1,371; 38,82)$

2.3.2 Les variances

Comme on l'a vu précédemment, le calcul de la variance suivi de celui de l'écart-type nous permet de mesurer la dispersion des valeurs de la série statistique autour de la moyenne arithmétique. On peut calculer indépendamment les variances des deux variables X et Y à l'aide de formules déja utilisées dans le chapitre 1 dans le cadre des séries statistiques à deux variables :

$$V(X) = \frac{1}{n}\sum_{i=1}^{p} n_{i.}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i.}x_i^2 - \overline{x}^2 \qquad V(Y) = \frac{1}{n}\sum_{j=1}^{q} n_{.j}(y_j - \overline{y})^2 = \frac{1}{n}\sum_{j=1}^{q} n_{.j}y_j^2 - \overline{y}^2$$

où p et q sont respectivement le nombre de modalités (ou de catégories) de X et Y.
On en déduit que

$$\sigma(X) = \sqrt{V(X)} \qquad \sigma(Y) = \sqrt{V(Y)}$$

Appliquons ces formules dans le cadre de l'exemple 2.3.1. Une colonne supplémentaire est ajoutée au tableau, $n_{i.}x_i^2$ pour la variable X, $n_{.j}y_j^2$ pour la variable Y.

x_i	$n_{i.}$	$n_{i.}x_i$	$n_{i.}x_i^2$
0	192	0	0
1	393	393	393
2	291	582	1164
3	100	300	900
4	24	96	384
Total	1000	1371	2841

Classe	y_j	$n_{.j}$	$n_{.j}y_j$	$n_{.j}y_j^2$
$[20;30[$	25	200	5000	125000
$[30;40[$	35	303	10605	371175
$[40;50[$	45	412	18540	834300
$[50;60[$	55	85	4675	257125
Total	–	1000	38820	1587600

Par conséquent,

. $V(X) = \dfrac{2841}{1000} - (1,371)^2 \simeq 0,961$ et $\sigma(X) = \sqrt{0,961} \simeq 0,98$

. $V(Y) = \dfrac{1587600}{1000} - (38,82)^2 \simeq 80,61$ et $\sigma(Y) = \sqrt{80,61} \simeq 8,98$

2.4 Vocabulaire, définitions

2.4.1 La covariance

Il est possible comme lors de l'étude sur les séries à une variable de définir une variance sur les deux variables simultanément, c'est la covariance.

Définition 2.4.1 *La **covariance** d'une série statistique à deux variables X et Y est donnée par la formule*

$$Cov(X,Y) = \sigma_{XY} = \frac{\sum_{i,j} n_{ij}(x_i - \overline{x})(y_j - \overline{y})}{n}$$

Comme pour la variance, il existe une écriture de la covariance plus adaptée au calcul, obtenue simplement en développant la formule précédente.

Propriété 2.4.1

$$\boxed{Cov(X,Y) = \sigma_{XY} = \frac{1}{n}\sum_{i,j} n_{ij}x_i y_j - \overline{x}.\overline{y}}$$

Preuve :

$$Cov(X,Y) = \frac{\sum_{i,j} n_{ij}(x_i - \overline{x})(y_j - \overline{y})}{n} = \frac{1}{n}\sum_{i,j} n_{ij}(x_i y_j - \overline{x}y_j - \overline{y}x_i + \overline{x}.\overline{y})$$

$$= \frac{1}{n}\left(\sum_{i,j} n_{ij}x_i y_j - \overline{x}\sum_{i,j} n_{ij}y_j - \overline{y}\sum_{i,j} n_{ij}x_i + \overline{x}.\overline{y}\sum_{i,j} n_{ij}\right)$$

$$= \frac{1}{n}\sum_{i,j} n_{ij}x_i y_j - 2\overline{x}.\overline{y} + \overline{x}.\overline{y} = \frac{1}{n}\sum_{i,j} n_{ij}x_i y_j - \overline{x}.\overline{y}.$$

Remarque 2.4.1

- $Cov(X,Y) = \overline{xy} - \overline{x}.\overline{y}$
- $Cov(X,X) = V(X)$ (la covariance est en quelque sorte le "dédoublement" de la variance)

Considérons l'exemple 2.3.1. On doit tout d'abord déterminer $\sum_{i,j} n_{ij}x_i y_j$, tâche qui peut être réalisée en utilisant le tableau de contingence :

x_i \ y_j	$[20;30[$	$[30;40[$	$[40;50[$	$[50;60[$	$n_{i.}$
0	20 (0)	150 (0)	22 (0)	0 (0)	192
1	99 (2475)	102 (3570)	180 (8100)	12 (660)	393
2	60 (3000)	51 (3570)	150 (13500)	30 (3300)	291
3	18 (1350)	0 (0)	50 (6750)	32 (5280)	100
4	3 (300)	0 (0)	10 (1800)	11 (2420)	24
$n_{.j}$	200	303	412	85	1000

Par exemple, $n_{11}x_1y_1 = 20 \times 0 \times 25 = 0$, $n_{23}x_2y_3 = 180 \times 1 \times 45 = 8100$,...
On obtient $\sum_{i,j} n_{ij}x_i y_j = 56075$ ce qui permet de déterminer la covariance entre X et Y :

$$Cov(X,Y) = \frac{56075}{1000} - (1,371 \times 38,82) \simeq 2,853$$

2.4.2 Le coefficient de corrélation linéaire

Pour mesurer l'intensité de la relation linéaire entre X et Y (autrement que par interprétation graphique du nuage de points), on définit le coefficient de corrélation linéaire $r(X,Y)$.

Définition 2.4.2 *Le coefficient de corrélation linéaire d'une série statistique double de variables X et Y est le nombre r défini par*

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

Propriété 2.4.2 *Le coefficient de régression vérifie :*

$$\boxed{-1 \leq r(X,Y) \leq 1}$$

Considérons l'exemple 2.3.1

$$r(X,Y) = \frac{2,853}{0,98 \times 8,98} \simeq 0,324$$

Commentaires :

- $r = 1$ ou $r = -1$ si et seulement si les points $M_i(x_i, y_i)$ sont alignés.
- Si r est voisin de 1 ou -1, la corrélation linéaire entre X et Y est très forte.
- Si r est proche de 0, il n'existe pas de corrélation linéaire entre X et Y. Les variables X et Y sont linéairement indépendantes ; il peut néanmoins exister une autre relation fonctionnelle entre X et Y, par exemple $Y = aX^2 + bX + c, \ldots$
- On peut présumer d'une corrélation linéaire pour $|r| \geq 0,866$ de sorte que la présomption de corrélation linéaire commence à partir de la valeur $\boxed{|r| \simeq 0,87}$

2.5 Ajustement linéaire (ou affine)

Avant de préciser ces notions par le calcul, il est bon de comprendre comment se pose le problème de la corrélation linéaire. Le problème consiste à déterminer dans quelle mesure les deux variables X et Y sont liées (c'est-à-dire dépendent l'une de l'autre). Par exemple, on peut intuitivement penser que la taille et le poids des individus d'une population sont liés, que par contre il est plus improbable que la taille et le revenu mensuel des habitants d'un pays donné soient liés. Si on arrive, à l'aide des données dont on dispose, à déterminer s'il existe une certaine fonction f telle que $\forall k \in \{1, 2, \ldots, n\}$, $y_k = f(x_k)$ on pourra répondre avec plus de précision à cette idée de lien entre X et Y.

Lorsque le nuage de points est nettement longiligne, les points étant disposés suivant une direction privilégiée, la corrélation est dite affine. Il est utile alors, dans un but d'extrapolation, de déterminer une droite rendant compte le mieux possible de la tendance observée. On dit qu'on effectue un **ajustement affine**.
On distingue deux types d'ajustement : les ajustements graphiques et les ajustements analytiques nécessitant un calcul spécifique

2.5.1 Ajustement graphique

On peut utiliser trois techniques :

– Ajustement direct à la règle
On utilise une règle transparente qu'on dispose de façon à l'ajuster le mieux possible suivant la direction privilégiée constatée et on s'efforce d'équilibrer le nombre de points situés de part et d'autre.

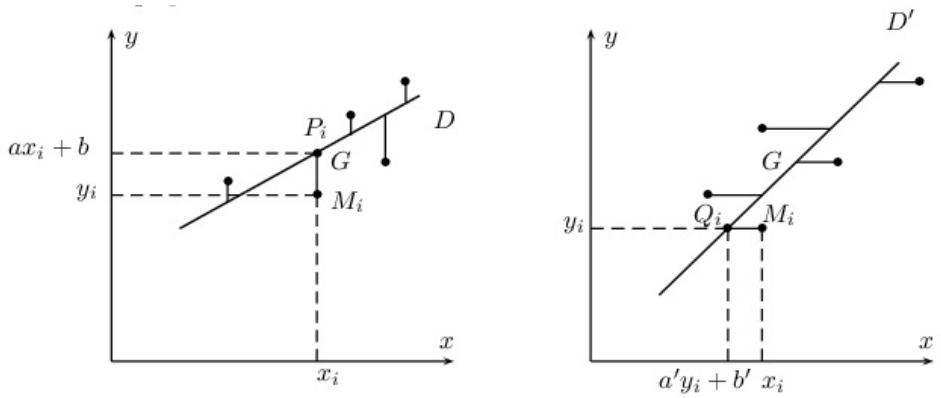
– Utilisation du point moyen
On montre par le calcul que, pour obtenir le meilleur ajustement affine, il convient de prendre une droite passant par le point moyen G.

– Fractionnement du nuage, méthode de Mayer.

2.5.2 Ajustement analytique - Méthode des moindres carrés

On considère une série statistique à deux variables représentée par un nuage justifiant un ajustement affine. La méthode d'ajustement linéaire des moindres carrés conduit à obtenir 2 droites de régression $D_{Y/X}$ et $D_{X/Y}$ concourantes au point moyen $G(\overline{x}, \overline{y})$ mais il est nécessaire de remarquer que ces 2 droites existent quel que soit le lien existant entre X et Y (même s'il n'y a aucune dépendance entre X et Y, on peut toujours tracer la droite telle que la somme des carrés des distances des points M_i du nuage aux points de la droite de mêmes abscisses soit minimale, ce minimum pouvant d'ailleurs être grand).
Le problème peut donc être posé ainsi : "à l'aide des deux droites de régression $D_{Y/X}$ et $D_{X/Y}$, quel critère numérique permet de dire si X et Y sont plus ou moins dépendantes l'une de l'autre par l'intermédiaire d'une fonction f affine ?"

On considère les deux graphes suivants :



Définition 2.5.1 *On appelle* ***droite d'ajustement par la méthode des moindres carrés de*** Y ***par rapport à*** X *la droite*
- *passant par* G
- *qui minimise la somme des carrés des écarts* M_iP_i *entre les ordonnées des points du nuage et les ordonnées des points de la droite ayant même abscisse.*

On dit aussi ***droite de régression de*** Y ***en*** X

Propriété 2.5.1 *La droite des moindres carrés de* Y *par rapport à* X *est notée* $D_{Y/X}$. *Son équation est*

$$\boxed{y = ax + b}$$

avec

$$\left\{ \begin{array}{rcl} a & = & \dfrac{Cov(X,Y)}{V(X)} \text{ coefficient directeur de } D_{Y/X} \\ b & = & \overline{y} - a\overline{x} \end{array} \right.$$

Définition 2.5.2 *On appelle* ***droite d'ajustement par la méthode des moindres carrés de*** X ***par rapport à*** Y *la droite*
- *passant par* G
- *qui minimise la somme des carrés des écarts* Q_iM_i *entre les abscisses des points du nuage et les abscisses des points de la droite ayant même ordonnée.*

On dit aussi ***droite de régression de*** X ***en*** Y

Propriété 2.5.2 *La droite des moindres carrés de X par rapport à Y est notée $D_{X/Y}$. Son équation est*

$$\boxed{x = a'y + b'}$$

avec

$$\boxed{\begin{cases} a' &= \dfrac{Cov(X,Y)}{V(Y)} \text{ coefficient directeur de } D_{X/Y} \\ b' &= \overline{x} - a'\overline{y} \end{cases}}$$

Remarque 2.5.1

- Les deux droites $D_{Y/X}$ et $D_{X/Y}$ passent par le point moyen du nuage.
- Les points du nuage sont alignés si et seulement si les droites $D_{Y/X}$ et $D_{X/Y}$ sont confondues.

Propriété 2.5.3

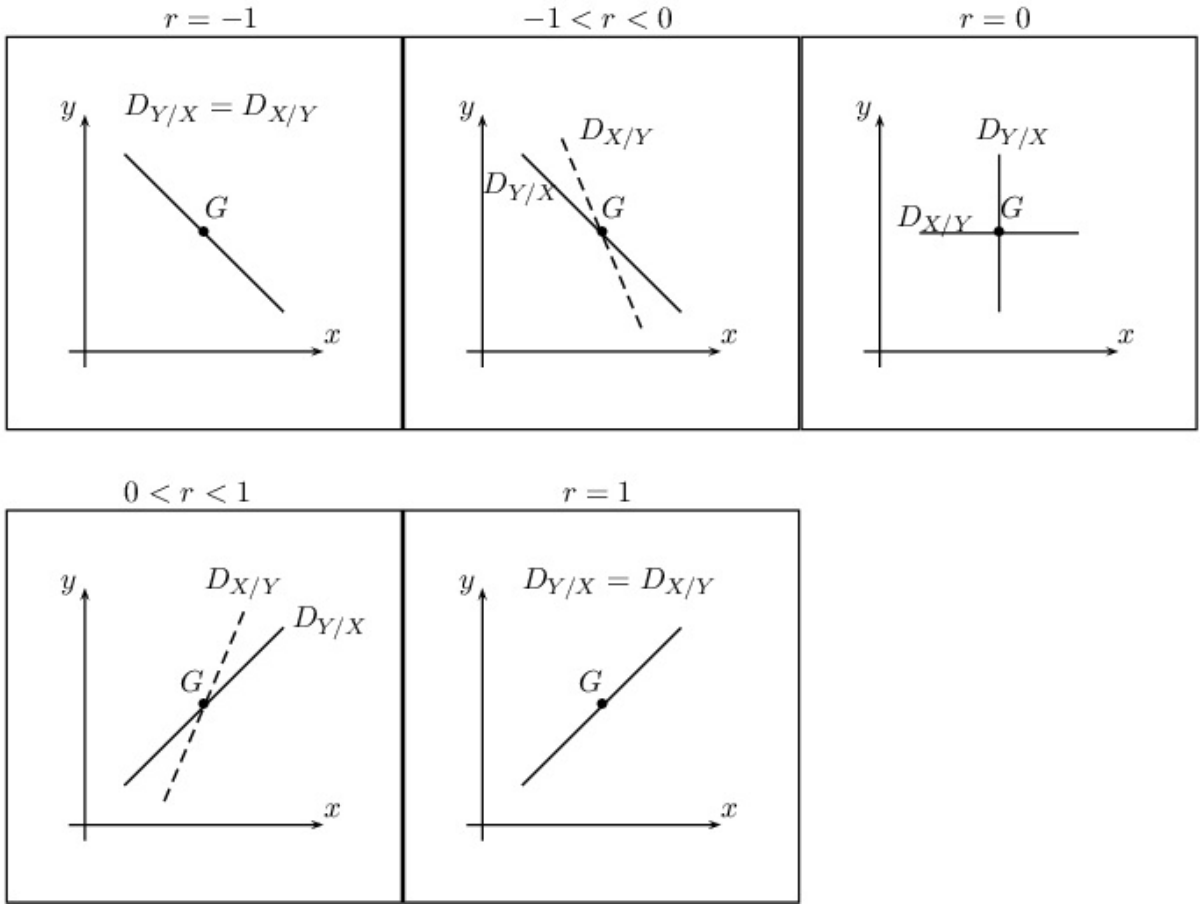
$$\boxed{D_{Y/X} = D_{X/Y} \Leftrightarrow r^2 = 1}$$

Preuve :

- L'équation de $D_{Y/X}$ peut se réécrire sous la forme $y = ax + (\overline{y} - a\overline{x}) \Leftrightarrow a(x - \overline{x}) = y - \overline{y}$
- De même, l'équation $D_{X/Y}$ peut s'écrire $x = a'y + (\overline{x} - a'\overline{y}) \Leftrightarrow x - \overline{x} = a'(y - \overline{y})$

Donc, $D_{Y/X} = D_{X/Y} \Leftrightarrow a.a' = 1 \Leftrightarrow \dfrac{Cov(X,Y)^2}{\sigma_X^2 \sigma_Y^2} = 1$. Or $r = \dfrac{Cov(X,Y)}{\sigma_X \sigma_Y}$ donc on obtient le résultat voulu.

On a finalement les 5 cas graphiques suivants :



L'ajustement linéaire que nous venons d'étudier avait pour objet de remplacer le nuage de points $M_i(x_i, y_i)$ par une droite D d'équation $y = ax + b$ (ou $x = a'y + b'$), résumant en partie la liaison entre X et Y. Cette droite D permet *à partir des valeurs x_i observées d'obtenir une valeur ajustée $\hat{y}_i$ pour y_i.*

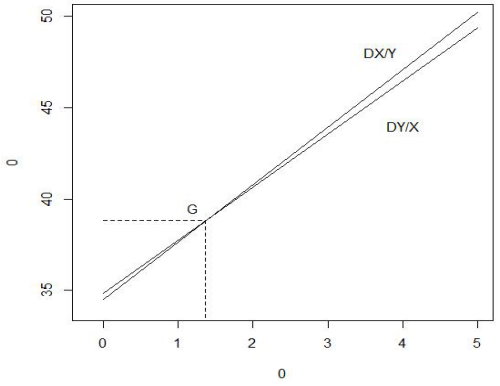
Considérons l'exemple 2.3.1 et déterminons les deux droites de régression $D_{Y/X}$ et $D_{X/Y}$:

- On a $a = \dfrac{Cov(X,Y)}{V(X)} = \dfrac{2,853}{0,98} \simeq 2,911$ et $b = \overline{y} - a\overline{x} \simeq 38,82 - 2,911 \times 1,371 \simeq 34,829$

  On déduit de ces résultats l'équation de $D_{Y/X}$: $y = 2,911x + 34,829$
- On a $a' = \dfrac{Cov(X,Y)}{V(Y)} = \dfrac{2,853}{8,98} \simeq 0,318$ et $b' = \overline{x} - a'\overline{y} \simeq 1,371 - 0,318 \times 38,82 \simeq -10,974$

  On déduit de ces résultats l'équation de $D_{Y/X}$: $x = 0,318y - 10,974$

Représentons graphiquement ces deux droites et comparons les résultats aux graphiques précédents.



On reconnaît immédiatement le cas $0 < r < 1$ (on rappelle d'ailleurs que $r \simeq 0,324$). Comme le coefficient de corrélation linéaire ne vérifie pas la condition $|r| > 0,87$, cela implique qu'il n'y ait pas de réelle corrélation linéaire entre X et Y. Il n'existe pas, par le biais des données disponibles, de relation linéaire entre l'âge d'un individu et le nombre d'arrêts de travail subi annuellement par cette personne. Il n'est pas donc possible d'extrapoler des résultats, c'est-à-dire de traiter des valeurs non contenues dans la série initiale et de répondre à des questions du type

- "Quelle estimation portant sur le nombre d'arrêts de travail peut-on donner si la personne est agée de 47 ans ?"
- "Quelle estimation de l'âge d'une personne peut on donner lorsque le nombre d'arrêts de travail qu'elle a subi est égal à 7 ?"

En supposant que le coefficient de corrélation vérifie l'inégalité $|r| > 0,87$, il est possible alors de répondre à de telles interrogations.

2.6 Exercices

Exercice 27 Après un examen de mathématiques, on a réparti 100 étudiants de 1ère année en économie selon le nombre x de minutes qu'ils ont passé lors des révisions la veille et la note y sur 20 qu'ils ont obtenue. Les résultats sont les suivants :

x \ y	$[0;4[$	$[4;8[$	$[8;12[$	$[12;16[$	$[16;20[$	$n_{i.}$
$[0;60[$	16	6	2	1	0	
$[60;120[$	5	10	5	3	0	
$[120;180[$	2	7	17	10	4	
$[180;240[$	1	1	1	5	4	
$n_{.j}$						

1. Compléter le tableau.
2. Combien d'étudiants ont entre 8 et 20 à leur examen ?
3. Quel est le pourcentage d'étudiants ayant eu une note inférieure à 8 et ayant révisé moins de 2 heures ?
4. Quel est le pourcentage d'étudiants ayant eu une note supérieure à 8 et ayant révisé plus de 2 heures ?
5. Présenter les histogrammes des séries x et y.
6. Tracer la courbe des effectifs cumulés croissants pour les deux séries statistiques.

Exercice 28 L'étude des hauteurs barométriques en fonction de l'altitude a permis d'établir le tableau suivant avec x l'altitude en km et y la hauteur barométrique en cm de mercure :

x_i	0	1	2	4	6	10
y_i	76	67	59	46	35	20

1. Représenter graphiquement la série statistique par un nuage de points.
2. Déterminer par la méthode des moindres carrés l'équation de la droite de régression de y en x. Tracer cette droite sur le graphique.
3. Sachant que la hauteur barométrique en un lieu est de 40 cm, calculer l'altitude.

Exercice 29 Le tableau ci-dessous donne l'évolution de la consommation de médicaments des ménages en France. x représente l'année, y le montant en milliards d'euros de la consommation de médicaments.

x_i	1970	1975	1980	1985	1990	1995	2002	2006
y_i	0,352	0,6660	1,073	2,026	3,369	6,420	9,592	10,89

1. Déterminer les coordonnées du point moyen $G(\overline{x}, \overline{y})$.
2. Calculer le coefficient de corrélation linéaire à $0,001$ près. Interpréter ce résultat.
3. Déterminer par la méthode des moindres carrés une équation de la droite de régression D de Y en X.
4. En supposant que l'évolution se poursuive de la même façon dans les années suivantes, donner une estimation de la consommation de médicaments des ménages en France en 2010.
   En quelle année la consommation dépassera-t-elle 12,5 milliards d'euros ?

Exercice 30 Une bibliothèque municipale établit le bilan de ses activités pour les 4 dernières années. Le tableau suivant donne en milliers pour chaque année,
- l'augmentation du nombre des prêts de livres x_i
- le nombre des nouveaux lecteurs inscrits y_i
- le nombre des nouveautés achetées z_i

	2003	2004	2005	2006
x_i	3	7	1	5
y_i	0,3	1,2	0,1	0,8
z_i	0,9	3,2	2,1	2,8

1. Calculer le coefficient de corrélation linéaire des séries (x_i) et (y_i).
2. Calculer le coefficient de corrélation linéaire des séries (x_i) et (z_i).
3. Déterminer quel élément (y_i ou z_i) est le plus susceptible d'avoir influencé l'augmentation des prêts de livres x_i.
4. Déterminer la droite d'ajustement par la méthode des moindres carrés $D_{Y/X}$.
5. En déduire une estimation du nombre de nouveaux lecteurs inscrits y_i, si l'on pense que l'augmentation du nombre des prêts sera de 9000 en 2007.

Exercice 31 50 étudiants d'une promotion ont effectué deux contrôles, l'un en méthodes quantitatives dont les notes sont x_i, l'autre en marketing dont les notes sont y_j.
On obtient la série statistique double donnée par le tableau ci-dessous :

y_j \ x_i	2	8	12	18
6	8	1	1	0
9	1	10	2	0
11	1	2	14	1
14	0	0	2	7

1. En précisant dans un tableau complet à double entrée les détails des calculs, déterminer une équation de régression de Y en X.
2. Déterminer une équation de la droite de régression de X en Y.
3. Calculer le coefficient de corrélation linéaire.
4. Si un étudiant obtient 15 au devoir de marketing, quelle note peut-on prévoir en méthodes quantitatives ?
5. Si un étudiant obtient 4 au devoir de méthodes quantitatives, quelle note peut-on prévoir en marketing ?

Exercice 32 Dans le cadre d'une enquête médicale, une étude est réalisée auprès de 60 patients concernant la taille et le poids de chacun. Les résultats sont donnés dans le tableau ci-dessous. On exprimera le poids x en kilogrammes et la taille y en centimètres. On arrondira les résultats des questions suivantes à 10^{-3} près.

y (cm) \ x (kg)	[50; 60[	[60; 70[	[70; 80[	[80; 90[	[90; 100[
[150; 160[	10	2	1	0	0
[160; 170[	1	12	3	1	0
[170; 180[	1	1	13	2	0
[180; 190[	0	0	2	6	1
[190; 200[	0	0	0	1	3

1. Calculer le pourcentage de personnes pesant moins de 80 kilos et mesurant plus de 180 centimètres.
2. Calculer la proportion de personnes pesant entre 60 et 90 kilos.

Dans la suite du problème, on utilisera les résultats suivants :

y (cm) \ x (kg)	[50; 60[	[60; 70[	[70; 80[	[80; 90[	[90; 100[	$n_{.j}$	$n_{.j}y_j$	$n_{.j}y_j^2$
[150; 160[	10	2	1	0	0	13	2015	312325
[160; 170[	1	12	3	1	0	17	2805	462825
[170; 180[	1	1	13	2	0	17	**2975**	520625
[180; 190[	0	0	2	6	1	9	1665	308025
[190; 200[	0	0	0	1	3	4	780	152100
$n_{i.}$	12	15	19	10	4	60	10240	1755900
$n_{i.}x_i$	660	975	1425	850	380	4290		
$n_{i.}x_i^2$	36300	63375	**106875**	72250	36100	314900		

y (cm) \ x (kg)	[50; 60[	[60; 70[	[70; 80[	[80; 90[	[90; 100[	
[150; 160[	85250	20150	11625	0	0	117025
[160; 170[	9075	128700	37125	14025	0	188925
[170; 180[	9625	11375	170625	29750	0	221375
[180; 190[	0	0	27750	**94350**	17575	139675
[190; 200[	0	0	0	16575	55575	72150
	103950	160225	247125	154700	73150	739150

3. Retrouver (en expliquant vos calculs) les trois valeurs encadrées dans les deux tableaux précédents.
4. Calculer les moyennes arithmétiques $\overline{x}$ et $\overline{y}$ des poids et tailles.
5. Calculer les variances $V(x)$ et $V(y)$. Déterminer alors les écart-types $\sigma(x)$ et $\sigma(y)$.
6. Calculer la covariance et en déduire le coefficient de corrélation linéaire entre le poids et la taille.
7. Un ajustement linéaire est-il justifié. Peut-on prévoir alors à l'aide de l'échantillon donné la taille d'une personne à partir de son poids ?