

Cours d'Eléments de Statistique

Jean Christophe Meunier

Module 4 Séries statistiques à deux variables

1^{ère} Bac, Commerce Extérieur
Année académique 2015-2016



A. Notion d'ajustement

Introduction

- Jusqu'à présent, travail sur une seule variable
 - Statistique descriptive : 'exploration' de l'échantillon sur une variable...
- Dans ce module, intérêt sur l'association entre deux variables :
 - Dans quelle mesure un changement sur une variable (x) est-il associé à un changement sur une seconde variable (y) ?
 - Ex :
 - Temps d'étude (x) - réussite (y)
 - Opinion sur un produit (x) – achat de ce produit (y)
 - ...

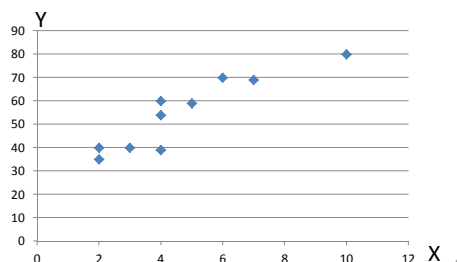
3

Illustration

- Temps d'études (x_i) et résultats à un examen (y_i)

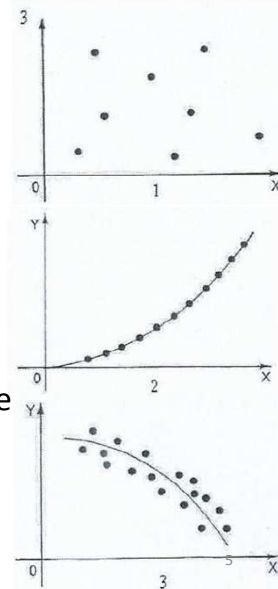
X_i (temps d'étude)	2	2	3	4	4	4	5	6	7	10
Y_i (résultats examen)	35	40	40	39	54	60	59	70	69	80

- Lorsque deux ou plusieurs variables sont récoltées sur les sujets, on peut voir s'il existe un lien entre ces variables
 - Dans un plan cartésien, chaque sujet est représenté par un point de coordonnées (x_i, y_i)
 - L'ensemble des points permet de visualiser le lien entre les variables



I. Diagramme de dispersion

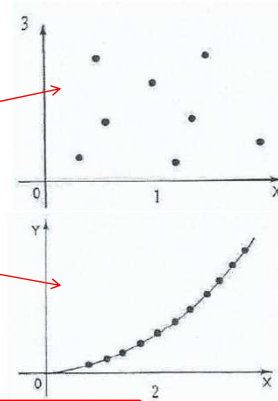
- Trois scénarii possibles
 - Les points se répartissent de manière quelconque
 - Indépendance
 - Les points suivent parfaitement une courbe qui est la représentation graphique d'une fonction mathématique
 - Dépendance fonctionnelle
 - Les points se répartissent le long d'une courbe assimilable à la représentation graphique d'une fonction mathématique
 - Dépendance statistique (corrélacionnelle)



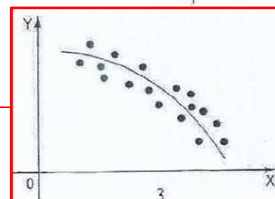
I. Diagramme de dispersion

Cas de figure hypothétique

- les points ne sont - pour ainsi dire - jamais répartis de manière totalement quelconque. Il est toujours possible d'estimer une courbe même si le lien est très faible
- Les points ne suivent - pour ainsi dire - jamais parfaitement une courbe

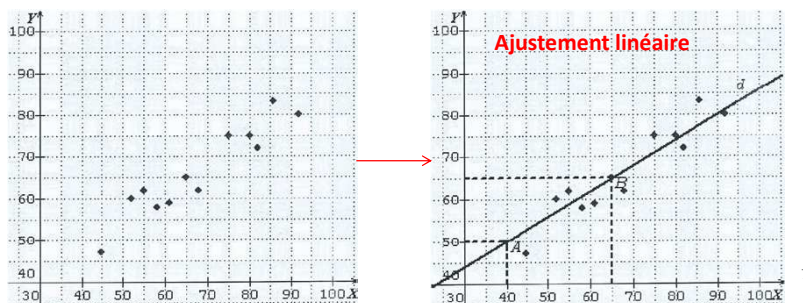


Cas de figure le plus fréquent



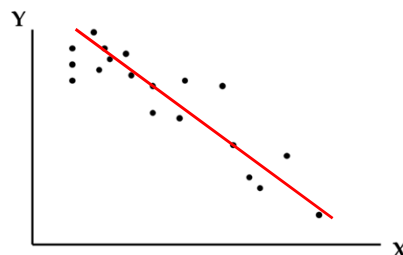
II. Notion d'ajustement

- Dans la plupart des cas : 'dépendance statistique'
 - Rechercher la fonction mathématique qui s'ajuste au mieux
 - La meilleure courbe peut être de formes diverses : parabole, exponentielle, hyperbole...
 - Le plus souvent, on se limite à trouver une droite



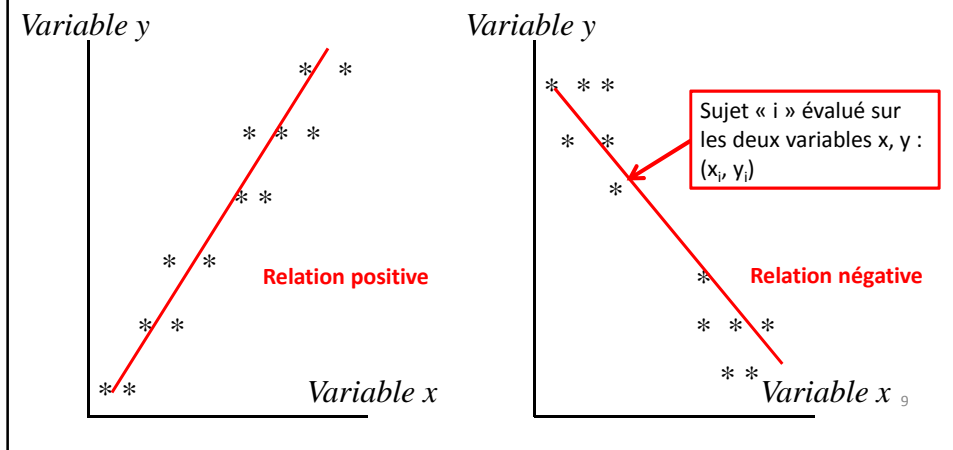
II. Notion d'ajustement

- On parle d'association linéaire :
 - la relation entre les variables x et y peut être décrite/estimée par une ligne droite
 - Pour variables d'échelles intervalle ou de rapport
 - Les variables d'échelles nominale ou ordinale n'ont pas de graduation à intervalles réguliers (association avec d'autres variables impossible ou difficile à établir)



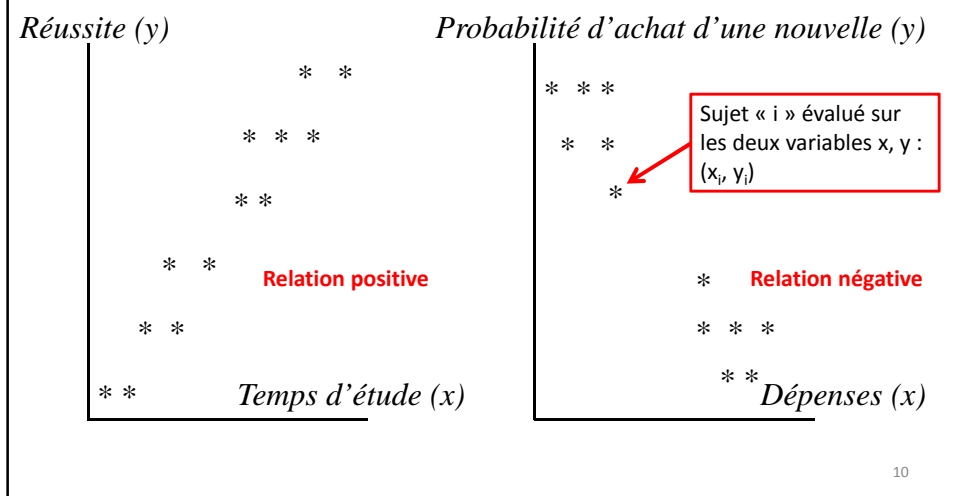
II. Notion d'ajustement

- L'inclinaison de la droite permet de voir le sens de la relation entre les deux variables



II. Notion d'ajustement

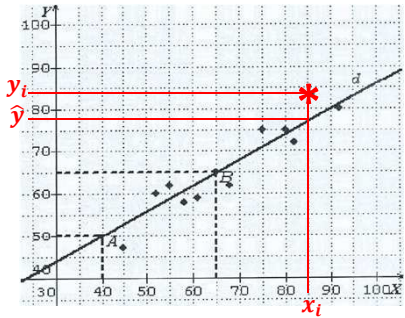
- Exemples :



B. Ajustement linéaire : méthodes d'estimation d'une courbe

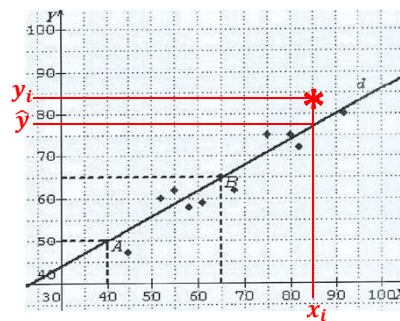
Introduction

- A partir du nuage de point
 - Estimer la droite d qui s'ajuste le mieux au données
 - L'équation d'une droite est $y = ax + b$
- Résolution du problème d'ajustement
 - trouver les valeurs 'a' et 'b'
 - détermine la position et l'orientation de la droite
 - Une fois la droite 'estimée'
 - Possibilité de faire des prédictions :
 - retrouver la valeur de y pour toute valeur de x



The scatter plot displays a set of data points (black dots) on a grid. A solid black line, labeled 'd', represents the linear regression. A point x_i is marked on the x-axis with a red asterisk. A vertical red line extends from x_i to the regression line. From that intersection, a horizontal red line extends to the y-axis, where the predicted value \hat{y}_i is marked with a red asterisk. Another horizontal red line is shown at a higher y-value, also marked with a red asterisk. Dashed lines illustrate the coordinates of some data points, such as one at (40, 50) and another at (70, 65).

12



Introduction

- Concrètement

- A partir de la droite $y = ax + b$, nous souhaitons retrouver la valeur de y pour toute valeur de x

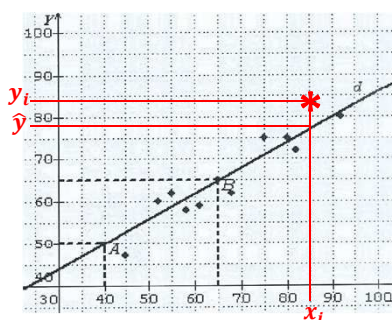
- Ex (pour $a=0,5$ et $b = 5$) :

- lorsque le score du sujet sur la variable $x = 10$; on peut estimer que son score le plus probable sur la variable y sera
- $$y = ax + b = 0,5 * 10 + 5 = 10$$

$$\hat{y} = ax_i + b$$

Où x_i est la valeur observée de la variable X et \hat{y} est la valeur estimée de la variables Y à partir de l'équation et pour la valeur x_i

On distingue y_i , valeur observée de Y de \hat{y} , valeur estimée de Y



13

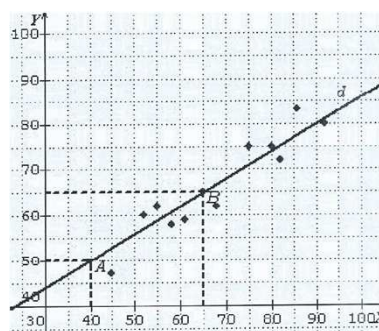
I. Méthode graphique

- Première étape

- Tracer manuellement la droite d qui semble s'ajuster au mieux au nuage de points
 - Choisir deux points – A et B – appartenant à la droite d et de leurs coordonnées (x_i, y_i)

- Seconde étape

- Sachant que
 - d est défini par $y = ax + b$
 - A et B $\in d$
 - On peut retrouver a et b
 - Système de deux équations à deux inconnues
 - A $\in d$ d'où : $50 = a * 40 + b$
 - B $\in d$ d'où : $65 = a * 65 + b$



14

I. Méthode graphique

- Résolution par substitution

- Deux équations à deux inconnues

- Soit A, $50 = a * 40 + b$
- Soit B, $65 = a * 65 + b$

- On exprime une des inconnues en fonction de l'autre (équation A)

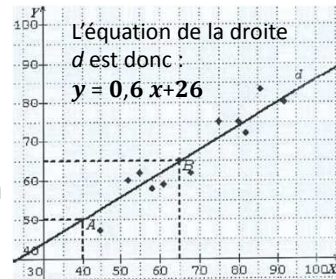
- Pour A, $b = 50 - 40a$

- On substitue le résultat dans l'équation B

- Dans B, $65 = 65a + 50 - 40a \rightarrow 65a - 40a = 65 - 50 \rightarrow 25a = 15 \rightarrow a = \frac{15}{25} = 0,6$

- Lorsque a est connu, on retrouve facilement b,

- à partir de l'équation A ($50 = 0,6 * 40 + b$) ou de l'équation B ($65 = 0,6 * 65 + b$) $\rightarrow b = 26$



15

I. Méthode graphique

- Méthode rapide mais pas objective

- L'estimation de la courbe dépend de la manière dont la droite a été tracée 'manuellement'

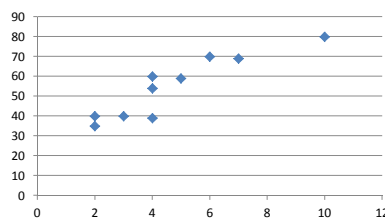
- Ex : deux observateurs peuvent obtenir des droites et équations différentes

16

Exercice : méthode graphique

- Donnez l'équation de la droite par la méthode graphique
 - A partir des données x_i et y_i suivantes,

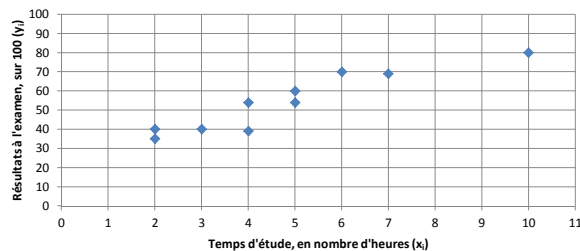
X_i (temps d'étude)	2	2	3	4	4	5	5	6	7	10
Y_i (résultats examen)	35	40	40	39	54	60	54	70	69	80



17

Exercice : méthode graphique

Résultats à l'examen (y_i) en fonction du temps d'étude (x_i)



1. Tracez une droite qui s'ajuste au mieux aux données
2. Choisissez deux points – A et B – appartenant à la droite d et retrouvez en les coordonnées x_i , y_i
 1. Soit A appartenant à d , $x_i = \dots\dots\dots$, $y_i = \dots\dots\dots$
 2. Soit B appartenant à d , $x_i = \dots\dots\dots$, $y_i = \dots\dots\dots$

18

Exercice : méthode graphique

3. A l'aide de la fonction d'une droite - ($y = ax + b$) - établissez le système de deux équations à deux inconnues pour les points A et B

1. Soit A appartenant à d et sachant que $y = ax + b$:

..... =

2. Soit B appartenant à d et sachant que $y = ax + b$:

..... =

19

Exercice : méthode graphique

4. Par la méthode de substitution, résolvez le système d'équations et retrouvez les valeurs des coefficients a et b

1. Dans une des deux équations, isolez b du reste de l'équation

.....

2. Dans la deuxième équation, substituez b par sa valeur retrouvée au point précédent

.....

3. Isolez maintenant a du reste de l'équation pour retrouver sa valeur

.....

4. Retrouvez les valeurs des coefficients a et b à partir des valeurs retrouvées

$a =$ $b =$

20

Exercice : méthode graphique

5. Donnez, à partir des coefficients a et b , l'équation de la droite que vous venez de tracer

Soit, $y = ax + b \rightarrow y = \dots x + \dots$

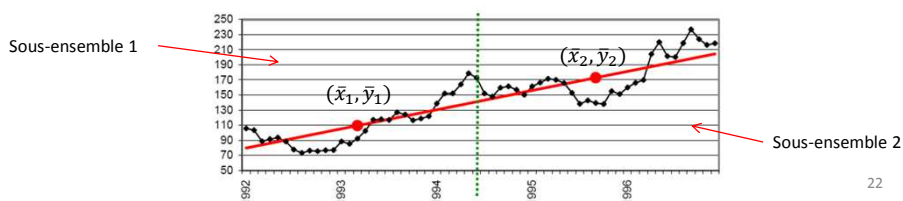
6. A l'aide de la fonction, faites une prédiction sur le résultat à l'examen (y_i) pour un temps d'étude (x_i) de 5 heures

Pour $x_i = 5$, $y_i = \dots\dots\dots$

21

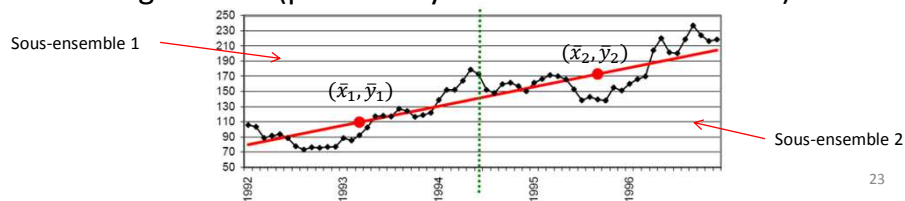
II. Méthode par la droite de Mayer

- On découpe le nuage de points en deux sous-ensembles de même effectif
 - Pour chacun des deux sous-ensembles, on calcule la moyenne des x_i et la moyenne des y_i .
 - On obtient ainsi deux points (\bar{x}_1, \bar{y}_1) et (\bar{x}_2, \bar{y}_2) , appelés points moyens



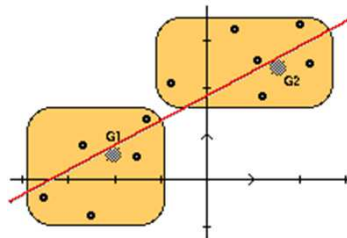
II. Méthode par la droite de Mayer

- Il reste à tracer la droite passant par ces deux points (\bar{x}_1, \bar{y}_1) et (\bar{x}_2, \bar{y}_2)
- L'équation de la droite s'obtient de la même manière que pour la méthode graphique, mais
 - Contrairement à la méthode graphique, les deux points ne sont pas obtenus à partir d'une droite tracée manuellement mais sont estimés de manière plus rigoureuse (points moyens des sous-ensembles)



II. Méthode par la droite de Mayer

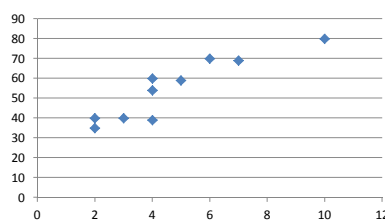
- Méthode plus rigoureuse que la méthode graphique, mais manque de précision également
 - Estimation des points moyens à partir de sous-ensembles dont les dispersions ne sont pas forcément les mêmes : biais possibles



Exercice : droite de Mayer

- Donnez l'équation de la droite par la méthode de la droite de Mayer
 - A partir des données x_i et y_i suivantes,

X_i (temps d'étude)	2	2	3	4	4	5	5	6	7	10
Y_i (résultats examen)	35	40	40	39	54	60	54	70	69	80



25

Exercice : droite de Mayer

	Sous-ensemble 1					Sous-ensemble 2				
X_i (temps d'étude)	2	2	3	4	4	5	5	6	7	10
Y_i (résultats examen)	35	40	40	39	54	60	54	70	69	80

- Calculez la moyenne des x_i et des y_i pour les sous-ensembles 1 et 2 et reportez les points moyens (\bar{x}_1, \bar{y}_1) et (\bar{x}_2, \bar{y}_2) sur le graphe. Tracez la droite d passant par ces points

Sous-ensemble 1 : $\bar{x}_1 = \dots\dots\dots$ $\bar{y}_1 = \dots\dots\dots$

Sous-ensemble 2 : $\bar{x}_2 = \dots\dots\dots$ $\bar{y}_2 = \dots\dots\dots$

26

Exercice : droite de Mayer

2. A l'aide de la fonction d'une droite ($y = ax + b$) établissez le système de deux équations à deux inconnues pour les points moyens des sous-ensembles 1 et 2 :

Soit (\bar{x}_1, \bar{y}_1) appartenant à d et sachant que $y = ax + b$:

..... =

Soit (\bar{x}_2, \bar{y}_2) appartenant à d et sachant que $y = ax + b$:

..... =

27

Exercice : droite de Mayer

3. Par la méthode de substitution, résolvez le système d'équations et retrouver les valeurs des coefficients a et b

1. Dans une des deux équations, isolez b du reste de l'équation

.....

2. Dans la deuxième équation, substituez b par sa valeur retrouvée au point précédent

.....

3. Isolez maintenant a du reste de l'équation pour retrouver sa valeur

.....

4. Retrouvez les valeurs des coefficients a et b à partir des valeurs retrouvées

$a =$ $b =$

28

Exercice : droite de Mayer

4. Donnez, à partir des coefficients a et b, l'équation de la droite que vous venez de tracer

Soit, $y = ax + b \rightarrow y = \dots x + \dots$

5. A l'aide de la fonction, faites une prédiction sur le résultat à l'examen (y_i) pour un temps d'étude (x_i) de 5 heures

Pour $x_i = 5$, $y_i = \dots\dots\dots$

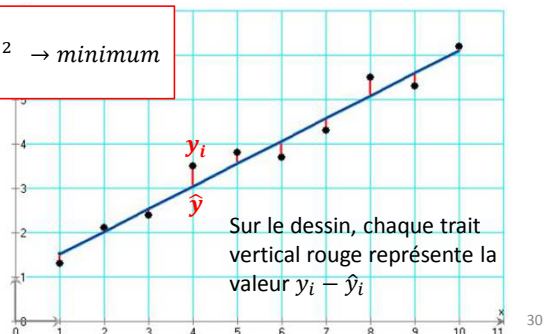
29

III. Méthode des moindres carrés

- Principe

- Faire passer la droite à travers le nuage de points de façon à ce que les carrés des différences ($y_i - \hat{y}_i$) soient les plus faibles possible pour l'ensemble des points \rightarrow 'moindres carrés'

- Soit,
$$\sum_{i=1}^{i=n} (y_i - \hat{y}_i)^2 \rightarrow \text{minimum}$$



30

III. Méthode des moindres carrés

- Développement mathématique

– On veut minimiser la quantité $q = \sum_{i=1}^{i=n} (y_i - \hat{y}_i)^2$

Considérant $\hat{y}_i = ax_i + b$

La fonction q peut également s'exprimer comme $\sum_{i=1}^{i=n} (y_i - ax_i - b)^2$

– La valeur minimale d'une fonction peut se calculer en posant la dérivée = 0

Dérivée de q par rapport à a : $\frac{dq}{da} = -2 \sum_{i=1}^{i=n} ((y_i - ax_i - b)x_i) = 0$

Dérivée de q par rapport à b : $\frac{dq}{db} = -2 \sum_{i=1}^{i=n} ((y_i - ax_i - b)1) = 0$

31

III. Méthode des moindres carrés

- A partir du calcul des dérivées
 - on peut obtenir le calcul des coefficients a et b de la droite des moindres carrés : $\hat{y}_i = ax_i + b$ (non développé ici)

- Coefficient a $a = \frac{\sum_{i=1}^{i=n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{i=n} (x_i - \bar{x})^2}$

- Coefficient b $b = \bar{y} - a\bar{x}$

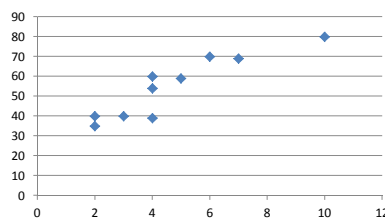
- Le calcul de a et de b suggère que la droite estimée (par les moindres carrés) passe nécessairement par le point moyen : \bar{x}, \bar{y} (aussi appelé centre de gravité du nuage de points)

32

Exercice : moindres carrés

- Donnez l'équation de la droite par la méthode des moindres carrés
 - A partir des données x_i et y_i suivantes,

X_i (temps d'étude)	2	2	3	4	4	5	5	6	7	10
Y_i (résultats examen)	35	40	40	39	54	60	54	70	69	80



33

Exercice : moindres carrés

- A l'aide du tableau ci-dessous, calculez les valeurs des coefficients a et b :

$$a = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad b = \bar{y} - a\bar{x}$$

x_i	$x_i - \bar{x}$	y_i	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
$\bar{x} = \sum x_i / N$	Total	$\bar{y} = \sum y_i / N$	Total	Total	Total
a=	b=				

34

Exercice : moindres carrés

2. Donnez l'équation de la droite d à partir des coefficients a et b que vous venez de calculer
Soit, $y = ax + b$, sachant $a = \dots$ et $b = \dots$
 $\rightarrow Y = \dots x + \dots$
3. A l'aide de l'équation de la droite d , retrouvez les valeurs estimées de y pour deux valeurs de x (par ex. pour $x_i = 1$ et pour $x_i = 9$)
Pour $x_i = 1$, $y = \dots$
Pour $x_i = 9$, $y = \dots$
4. Tracez la droite d passant par ces deux points

35

Conclusions

- Méthodes graphique et par la droite de Mayer
 - Permettent une estimation rapide de la droite d'ajustement... mais manque de précisions
- Méthode des moindres carrés
 - La plus utilisée car la plus précise et juste
 - Sert de base au calcul de mesures courantes d'association entre deux variables
 - Covariance, corrélation, régression ... seront vues ultérieurement (Cours Statistique Inférentielle)

36