Coefficient de corrélation linéaire

et

droite de régression

Remarque : Tu devrais regarder la présentation « Tableau à double entrée et nuage de points.ppt »avant de visionner celui-ci.

Dans une relation statistique, il ne s'agit pas de savoir si l'une des variables est la cause et l'autre, l'effet.

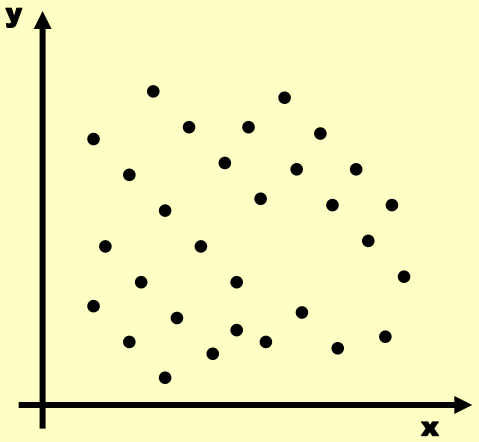
Il s'agit simplement de déterminer s'il existe un lien de nature quelconque entre ces variables.

Ce lien d'une variable envers l'autre s'appelle corrélation.

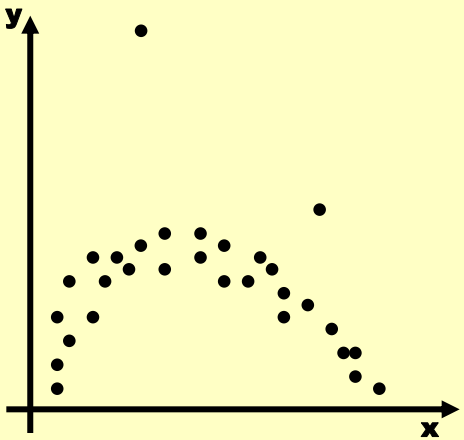
Le nuage de points (ou diagramme de dispersion) est une façon de représenter graphiquement ce lien.

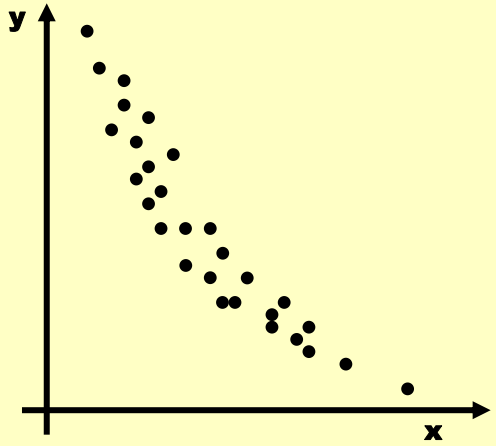
En représentant les données sous la forme d'un nuage de points, certaines formes peuvent apparaître nous permettant de caractériser qualitativement et quantitativement ce lien.

Lorsque les points sont dispersés de manière aléatoire, le nuage montre qu'il n'a a pas vraiment de corrélation.



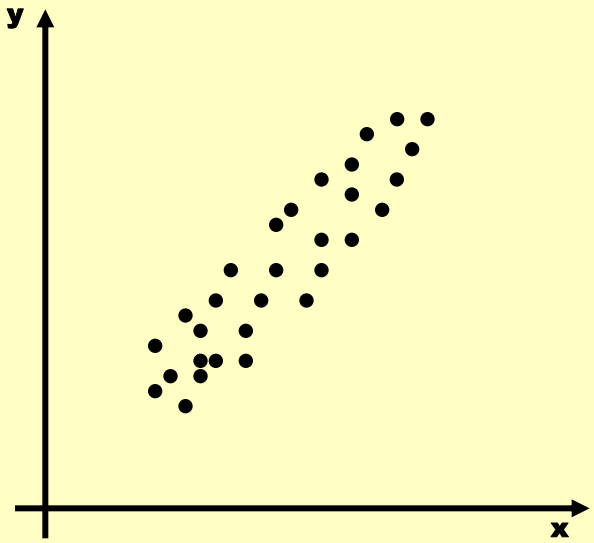
La distribution des points peut aussi montrer des formes rappelant certaines relations ou fonctions.





Un type de forme retiendra notre attention dans cette présentation,

l'alignement linéaire.



Lorsque l'alignement des points est linéaire, deux outils permettent d'interpréter le nuage :

- le coefficient de corrélation linéaire : r

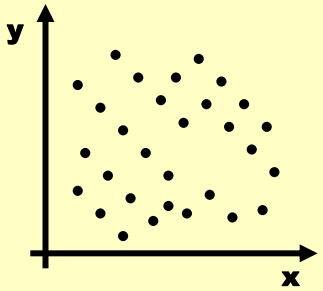
décrivant la densité de la corrélation;

- la droite de régression : $y = ax + b$

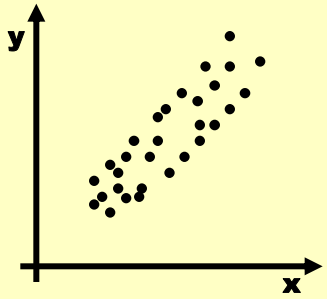
modélisant la corrélation; permettant ainsi de faire des prédictions.

Le coefficient de corrélation : méthode graphique

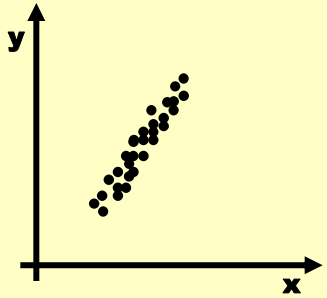
Certains nuages ne démontrent pas de corrélation.



D'autres montrent une corrélation faible.



Enfin, certains montrent une très forte corrélation.



Alors, comment faire pour quantifier cette corrélation ?

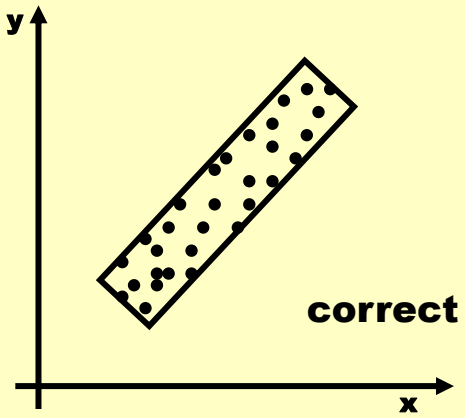
Il existe une méthode graphique dite « du rectangle ».

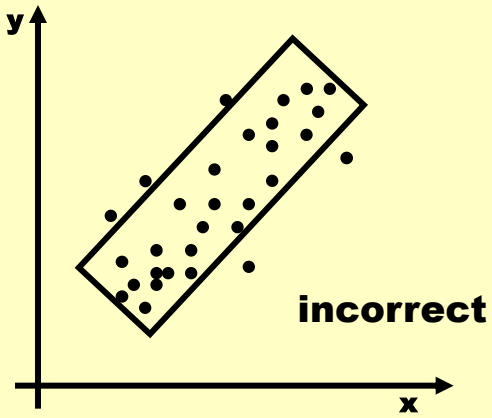
Elle est simple, mais comporte quelques règles.

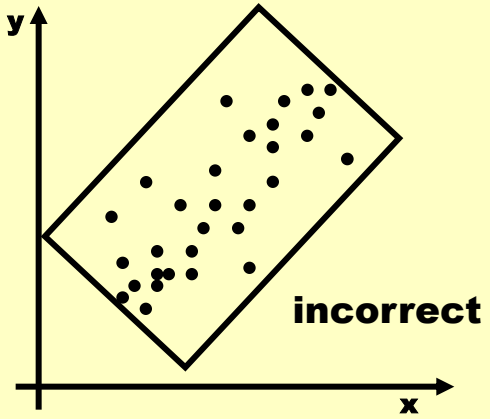
Prenons un exemple :

Étape 1 :

Dessiner un rectangle de plus petites dimensions qui contient tous les points.





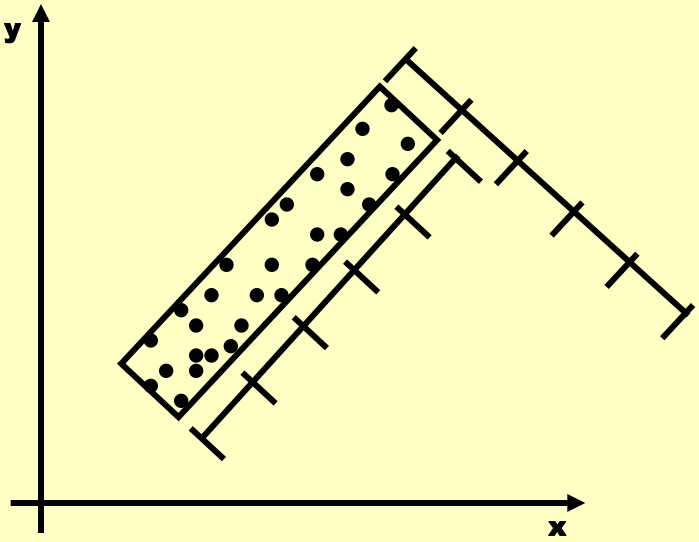


Étape 2 :

Mesurer la longueur et la largeur du rectangle :

- longueur : 5 unités

- largeur : 1 unité



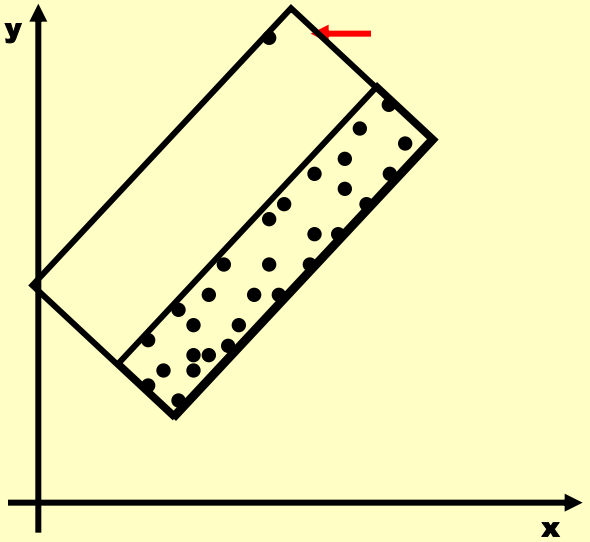
Étape 3 :

Utiliser la formule suivante en remplaçant la largeur et la longueur par les mesures effectuées.

$$r \approx \pm\left(1 - \frac{\text{largeur}}{\text{longueur}}\right) \rightarrow r \approx \pm\left(1 - \frac{1}{5}\right) \rightarrow r \approx 0{,}8$$

Remarque : Il peut arriver qu'un nuage de points contienne une (des) donnée(s) éloignée(s) des autres points.

Exemple :



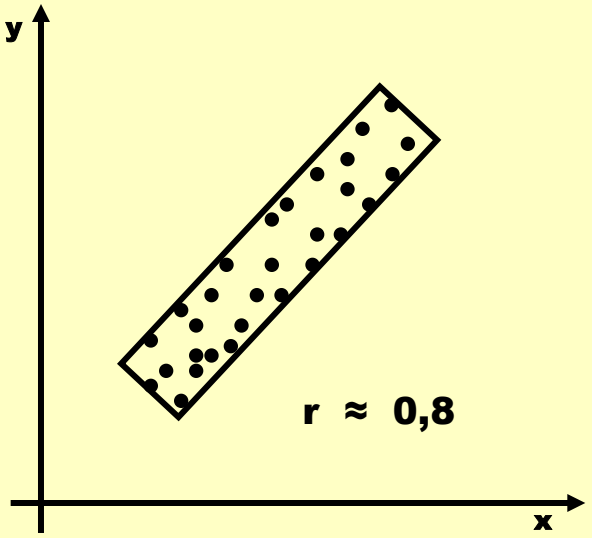
On ignore alors cette donnée, car elle modifierait beaucoup le rectangle.

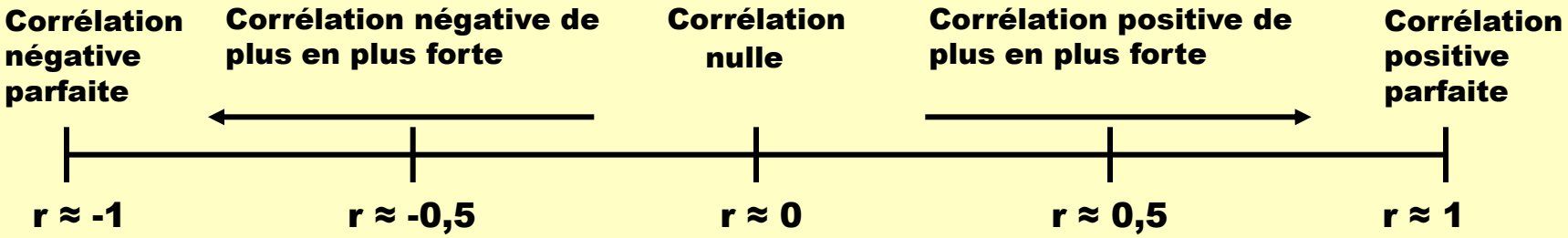
Cette donnée très éloignée des autres s'appelle « donnée aberrante ».

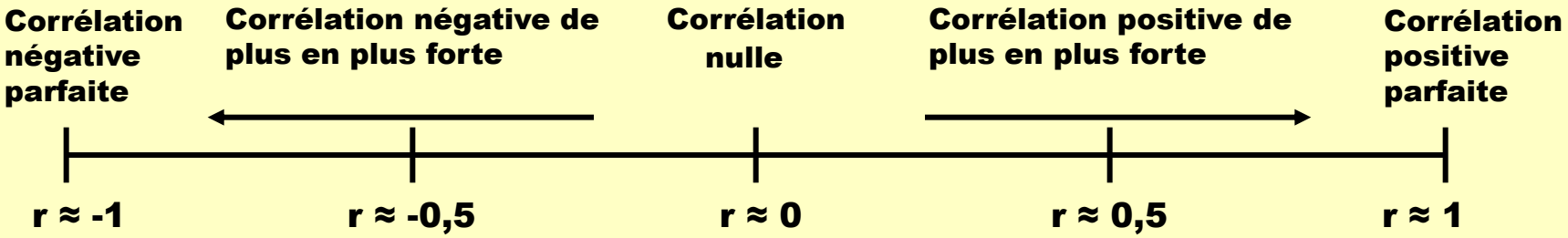
Il est préférable de l'ignorer.

Étape 4 :

On interprète le coefficient à l'aide de la règle suivante :







Remarque : La formule r ≈ ± (1 – largeur/longueur) et cette règle de correspondance ont été obtenues suite à des calculs assez complexes.

La formule permettant le calcul du coefficient de corrélation, noté r dans le cas d'un échantillon, est :

$$r = \frac{\sum x_i y_i - n\overline{xy}}{(n-1)s_x s_y}$$

où :

$\sum x_i y_i$ = la somme de tous les produits xy de chaque couple $(x\,;\,y)$;

n = la taille de l'échantillon ;

$\bar{x}$ = la moyenne de l'échantillon pour la variable X ;

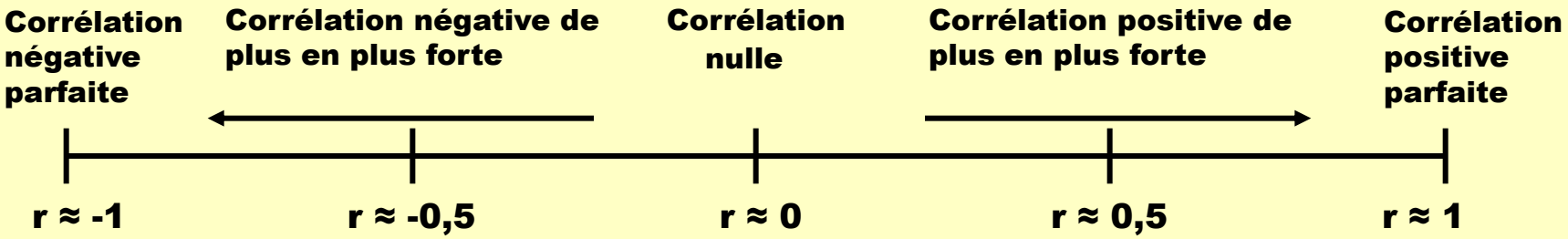
$\bar{y}$ = la moyenne de l'échantillon pour la variable Y ;

s_x = l'écart type de l'échantillon pour la variable X ;

s_y = l'écart type de l'échantillon pour la variable Y.

En introduisant dans l'équation les valeurs appropriées, on a alors ici :

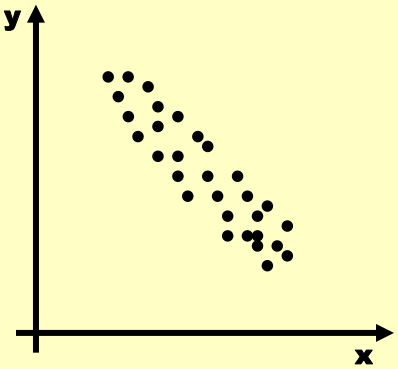
$$r \approx \frac{494{,}71 - 9 \times 6{,}9 \times 8{,}856}{(9-1) \times 2{,}134 \times 3{,}43} \approx -0{,}943.$$



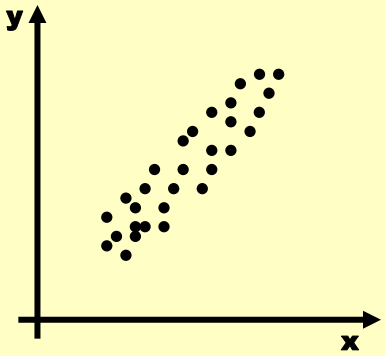
Suite à leurs calculs, les statisticiens ont remarqué que le coefficient de corrélation était un nombre qui variait toujours dans l'intervalle [-1 , 1].

Plus le coefficient se rapproche de 1 ou de -1 et plus la corrélation est forte.

Le coefficient variera de 0 à -1 si la corrélation est négative.

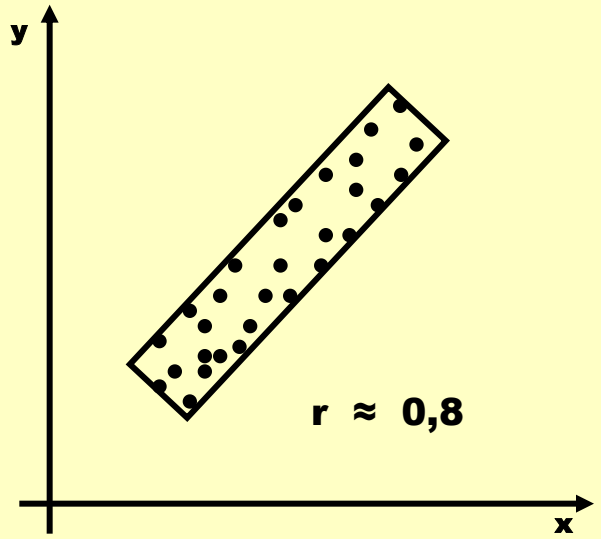


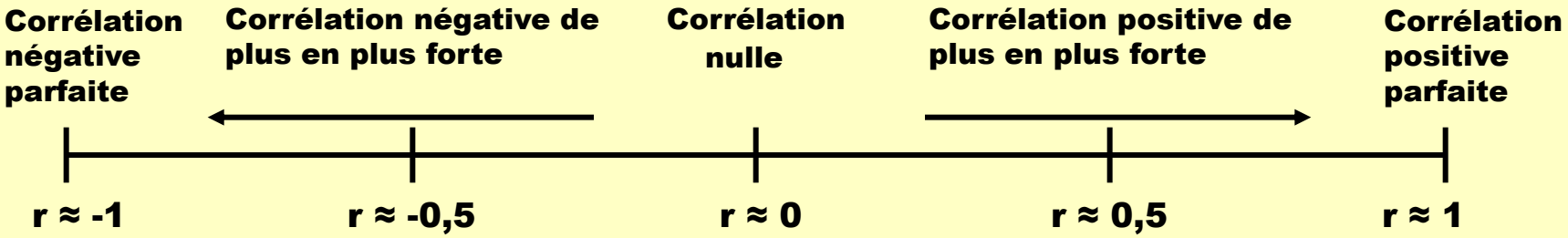
Le coefficient variera de 0 à 1 si la corrélation est positive.



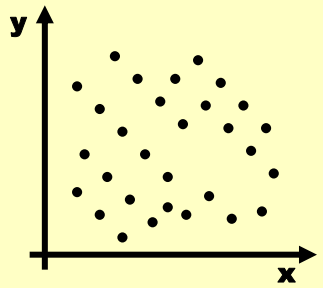
Le signe (+ ou -) du coefficient indique simplement l'orientation du nuage de points.

Dans cet exemple, la corrélation est relativement forte et positive.

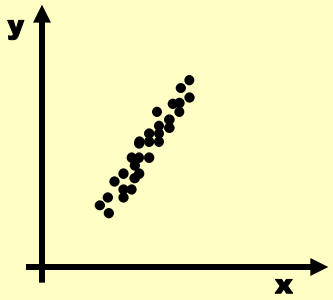


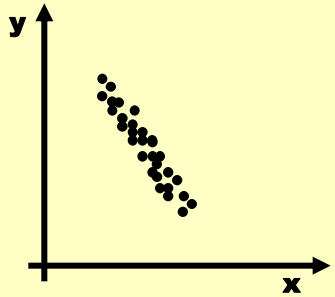


Un coefficient de 0 indique un lien linéaire nul entre les deux variables.



Un coefficient de 1 ou -1 indique un lien linéaire très fort entre les deux variables.





Bien entendu, il ne faut pas se fier qu'à ce coefficient.

Il faut connaître le sujet de l'étude, car le coefficient n'est qu'un indicateur.

Par exemple, un coefficient de 0,4 est considéré assez fort si on met en relation un nouveau médicament et la guérison d'une maladie.

La méthode graphique est approximative; elle permet quand même une estimation intéressante de la corrélation.

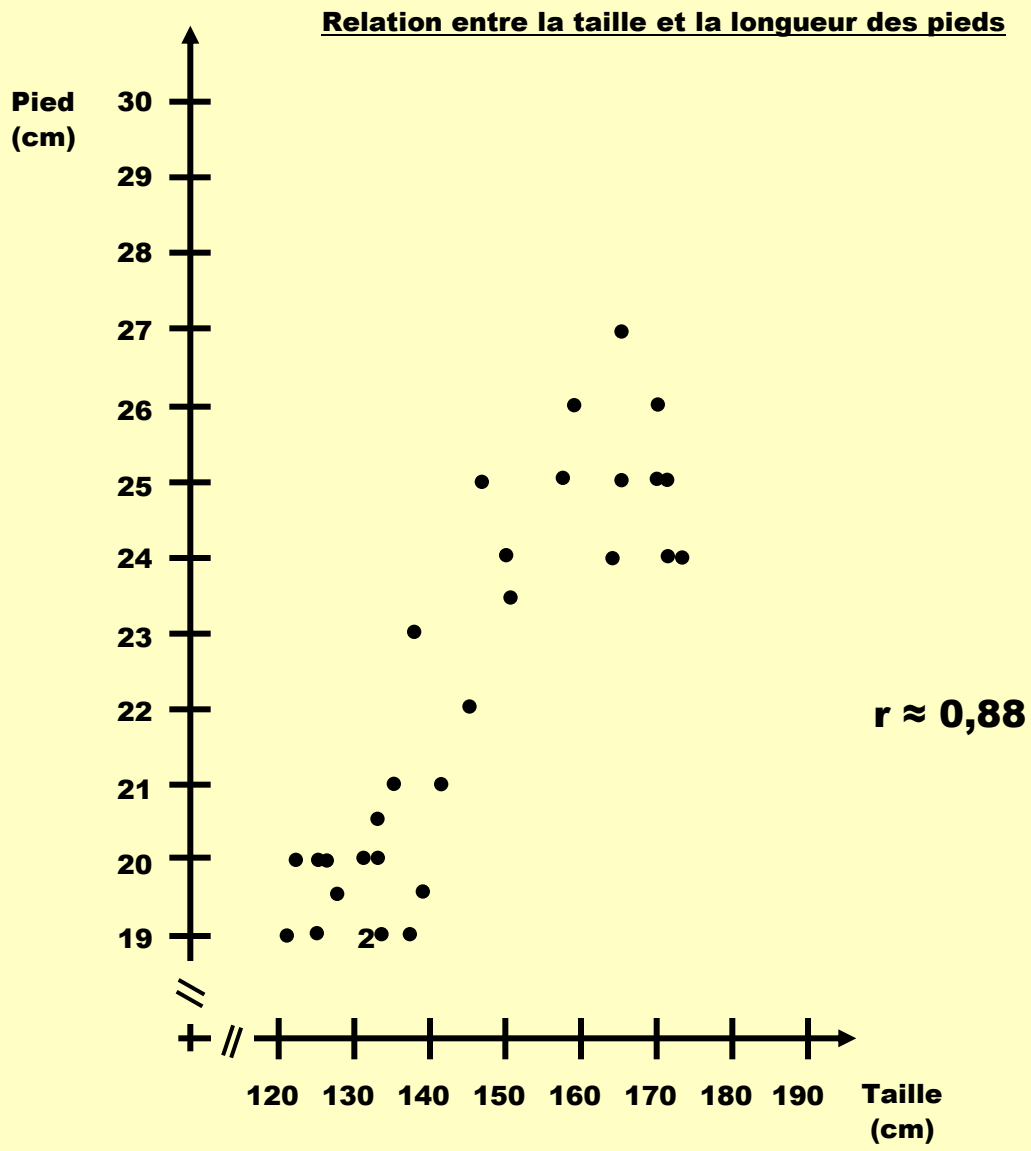
Lorsque le coefficient se rapproche de -1 ou 1, alors la relation entre les variables peut être qualifiée de linéaire. Nous pouvons alors utiliser la droite de régression.

$$y = ax + b$$

Voici une distribution mettant en relation la taille d'une personne et la longueur de ses pieds et le nuage de points représentant cette distribution.

Taille (cm)	Pied (cm)
122	19
123	20
125	20
125	19
126	20
127	19,5
131	20
133	20
134	19
134	19
134	20,5
135	21
138	19
138	23
139	19,5
141	21
145	22
147	25
150	24
151	23,5
158	25
159	26
164	24
165	27
165	25
170	26
170	25
171	25
172	24
173	24

n = 30 n = 30



Remarque :

Même s'il y a une relation assez forte entre les deux variables (r ≈ 0,88),

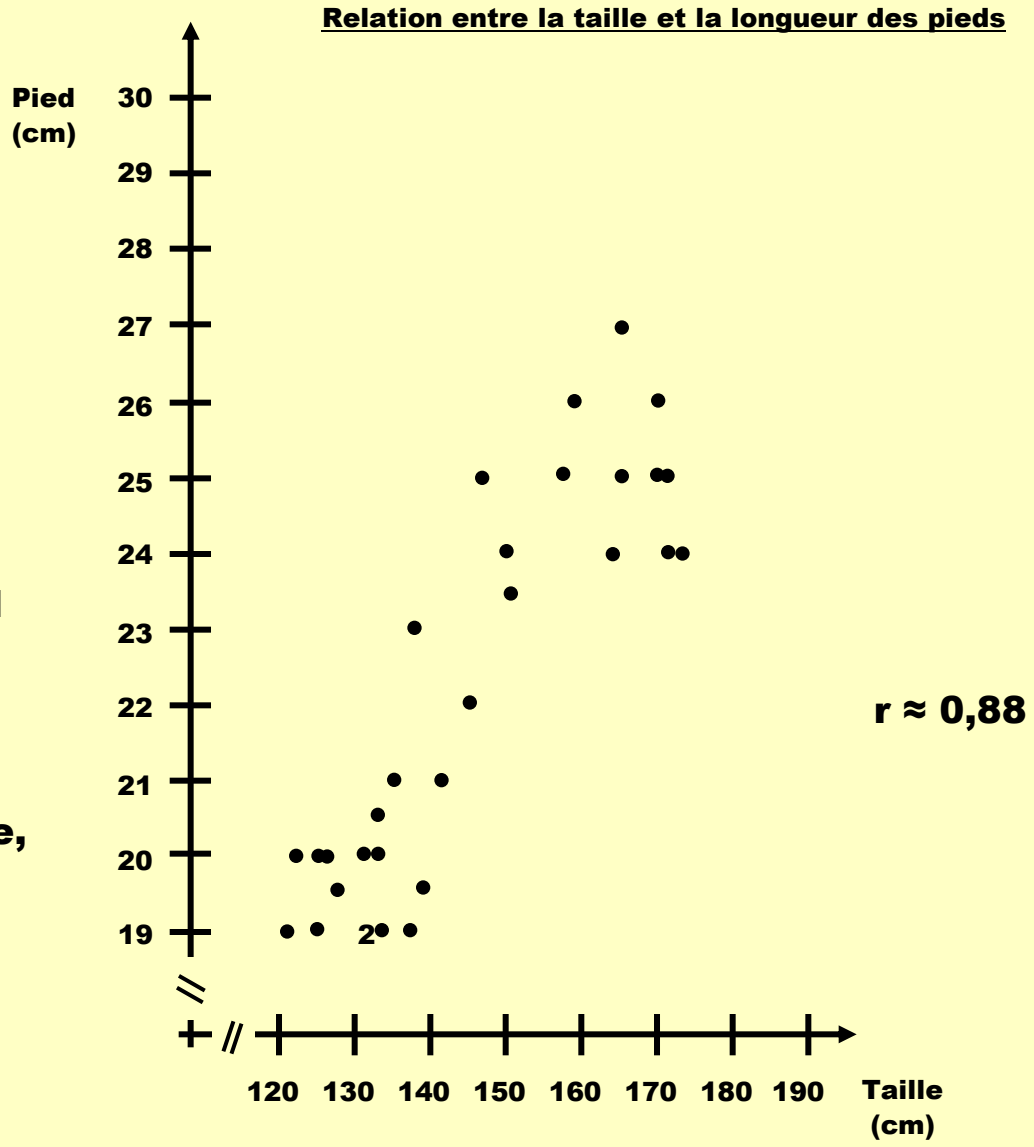
cette corrélation ne veut pas dire qu'il y ait une relation de cause à effet.

En effet, la taille ne fait pas pousser les pieds.

Par contre, une corrélation très forte peut laisser supposer qu'il y ait une relation de cause à effet.

Il faut alors poursuivre l'étude avec l'ensemble de la recherche, afin d'avoir plus d'informations.

Il ne faut jamais tirer des conclusions trop rapidement.



Le coefficient de corrélation est suffisamment grand. Nous pouvons alors tenter une régression linéaire. C'est-à-dire chercher la droite dont l'équation est $y = a x + b$ et qui passe le plus près de ces points.

Il existe plusieurs méthodes pour déterminer la droite de régression.

Nous allons en regarder 3.

Méthode 1 : Tracer manuellement la droite selon certains critères.

Méthode 2 : La droite de Mayer.

Méthode 3 : Les moindres carrés.

Méthode 1 : Tracer manuellement la droite selon certains critères.

Ces critères doivent être suivis dans l'ordre qu'ils sont énoncés :

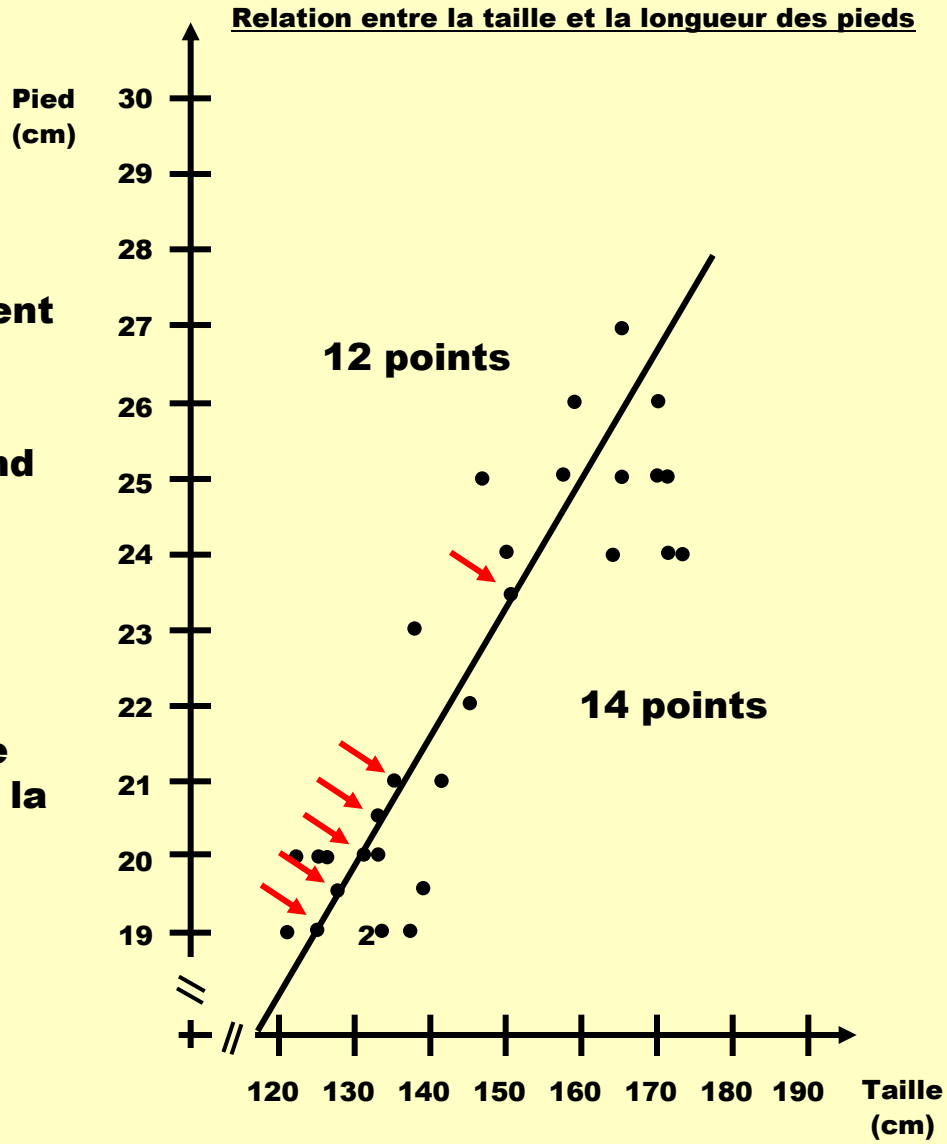
1) **La droite doit respecter la direction de l'ensemble des points.**

2) **La droite doit diviser le plus également possible l'ensemble des points.**

3) **La droite doit passer par le plus grand nombre de points possibles.**

Remarque :

Cette méthode est approximative. Cependant, elle peut être intéressante surtout lorsque le nuage de points est la seule information disponible pour représenter les données.



Il faut déterminer deux couples de coordonnées situés sur la droite.

Point 1 ≈ (125, 19)

Point 2 ≈ (151, 23,5)

Il faut déterminer le taux de variation :

$$a = \frac{y_2 - y_1}{x_2 - x_1} = \frac{23,5 - 19}{151 - 125} \approx 0,17$$

L'équation débute donc par :

$$y = 0,17x + b$$

En utilisant le début de l'équation, il faut déterminer la valeur du paramètre b en utilisant un des couples :

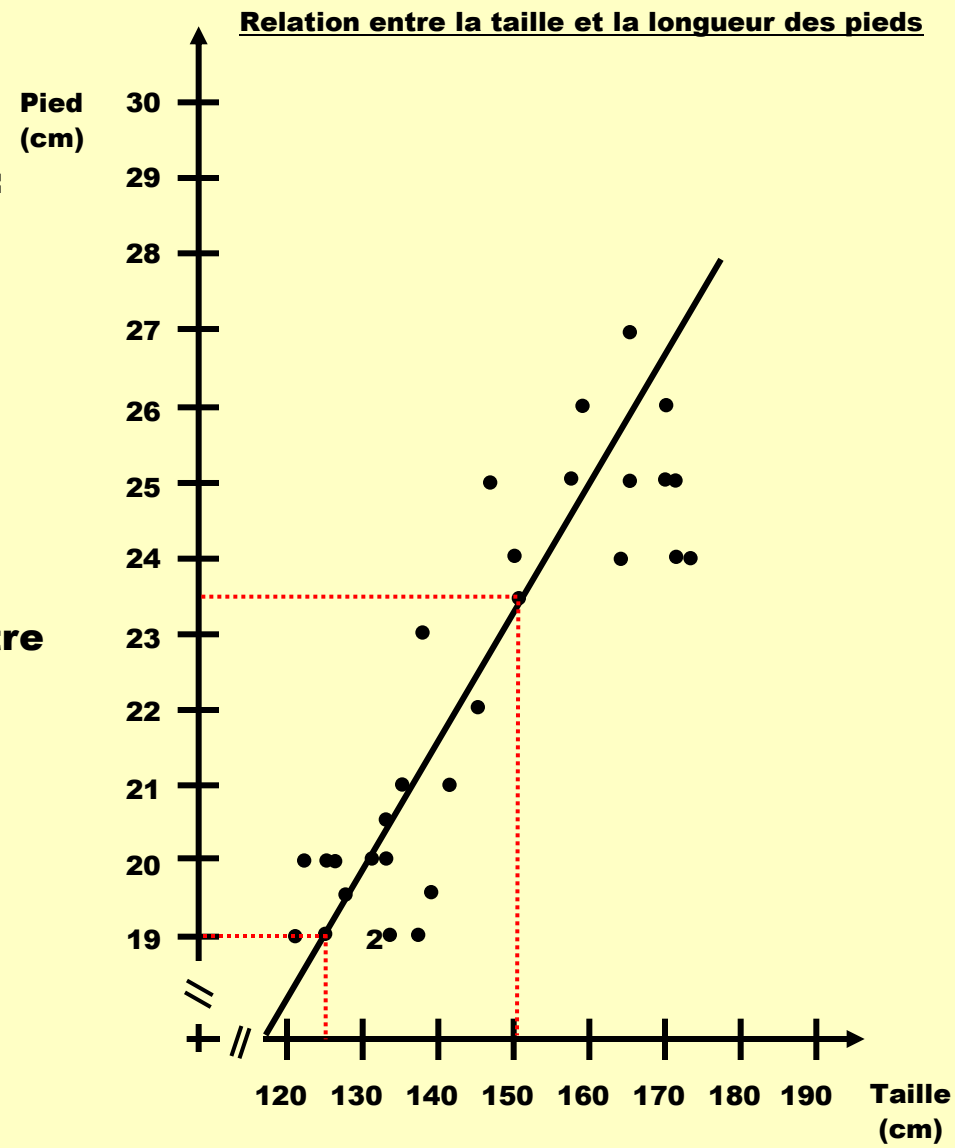
$y = 0,17x + b$ avec (125, 19)

$$19 = 0,17 \times 125 + b$$

$$19 = 21,25 + b$$

$$-2,25 = b$$

L'équation est $y = 0,17x - 2,25$



Méthode 2 : La droite de Mayer

La droite de Mayer est utilisée avec un tableau de distribution.

Démarche :

Étape 1 : Mettre la liste en ordre croissant selon la première variable.

Étape 2 : Séparer la distribution en deux groupes égaux ou approximativement égaux.

Étape 3 : Dans le premier groupe, calculer la moyenne des x et des y.

Moyenne des x du premier groupe :

(122 + 123 + 125 + 125 + 126 + 127 + 131 + 133 + 134 + 134 + 134 + 135 + 138 + 138 + 139) ÷ 15

moyenne des x ≈ 131

Moyenne des y du premier groupe :

(19 + 20 + 20 + 19 + 20 + 19,5 + 20 + 20 + 19 + 19 + 20,5 + 21 + 19 + 23 + 19,5) ÷ 15

moyenne des y ≈ 20

x	y
Taille (cm)	Pied (cm)
122	19
123	20
125	20
125	19
126	20
127	19,5
131	20
133	20
134	19
134	19
134	20,5
135	21
138	19
138	23
139	19,5
141	21
145	22
147	25
150	24
151	23,5
158	25
159	26
164	24
165	27
165	25
170	26
170	25
171	25
172	24
173	24

n = 30 n = 30

Ces deux moyennes serviront de coordonnées pour le premier point.

Point 1 : (131, 20)

Par la suite, il faut déterminer les moyennes des x et des y du deuxième groupe.

Moyenne des x du deuxième groupe ≈ 160,1

Moyenne des y du deuxième groupe ≈ 24,4

Point 2 : (160,1 , 24,4)

On établit alors l'équation de la droite :

$$y = ax + b$$

$$y = 0{,}15x - 0{,}35$$

x	y
Taille (cm)	Pied (cm)
122	19
123	20
125	20
125	19
126	20
127	19,5
131	20
133	20
134	19
134	19
134	20,5
135	21
138	19
138	23
139	19,5
141	21
145	22
147	25
150	24
151	23,5
158	25
159	26
164	24
165	27
165	25
170	26
170	25
171	25
172	24
173	24
n = 30	n = 30

Les deux méthodes,

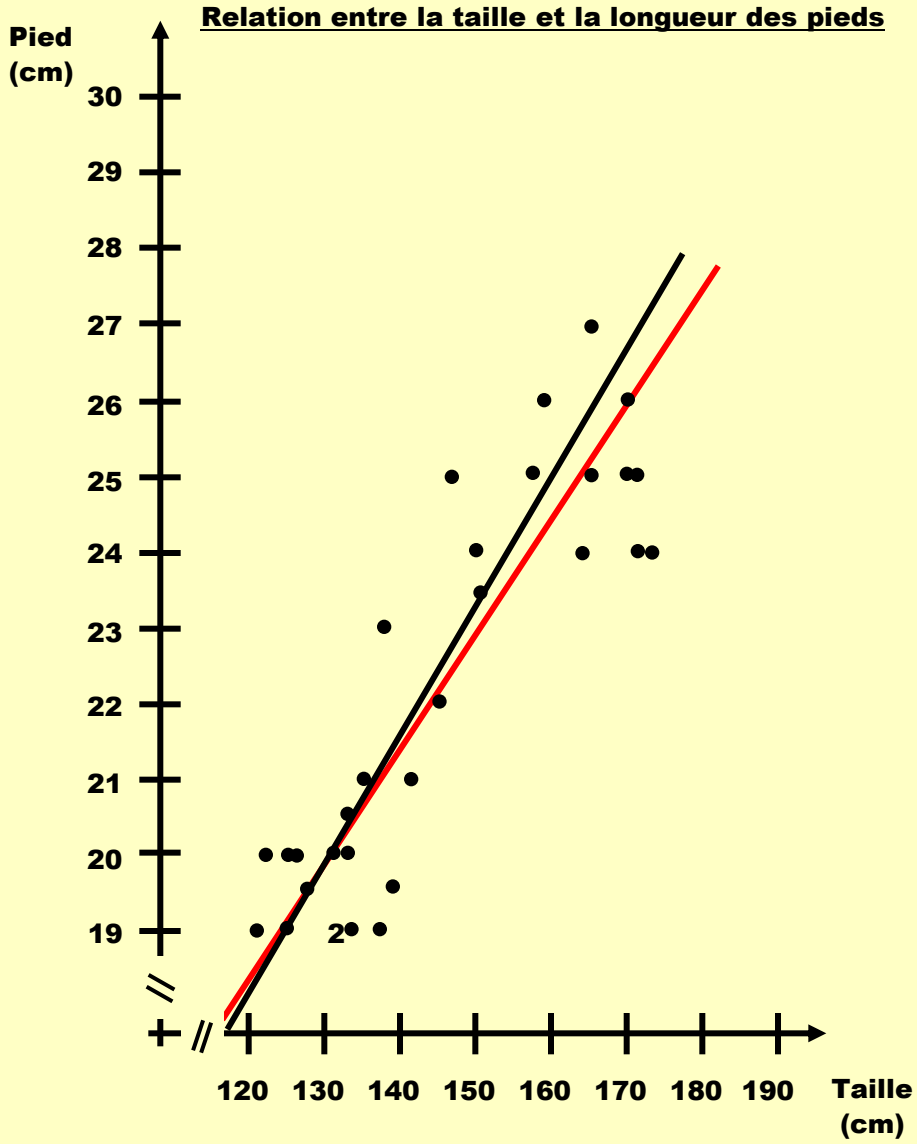
la droite tracée manuellement

et

la droite de Mayer

diffèrent quelque peu.

Cela est dû aux méthodes qui sont approximatives.



Il existe en statistique une formule pour établir l'équation d'une droite de régression; cette formule est beaucoup plus précise.

Il s'agit de « la droite de régression des moindres carrés ».

$$y = \frac{cov(x,y)}{V(x)}(x - \overline{x}) + \overline{y}$$

$$a = \frac{cov(x,y)}{V(x)} \qquad b = \overline{y} - \frac{\overline{x} \cdot cov(x,y)}{V(x)} = \overline{y} - a \cdot \overline{x}$$

Pour comprendre cette formule, il faut connaître passablement de notions statistiques.

Pour l'instant, nous pourrons calculer précisément une droite de régression avec la calculatrice, car c'est avec cette formule que la TI-80 détermine cette équation.

Nous verrons comment utiliser la calculatrice TI-80 pour effectuer ce travail dans la prochaine présentation « Corrélation, TI-80 et interprétation.ppt ».

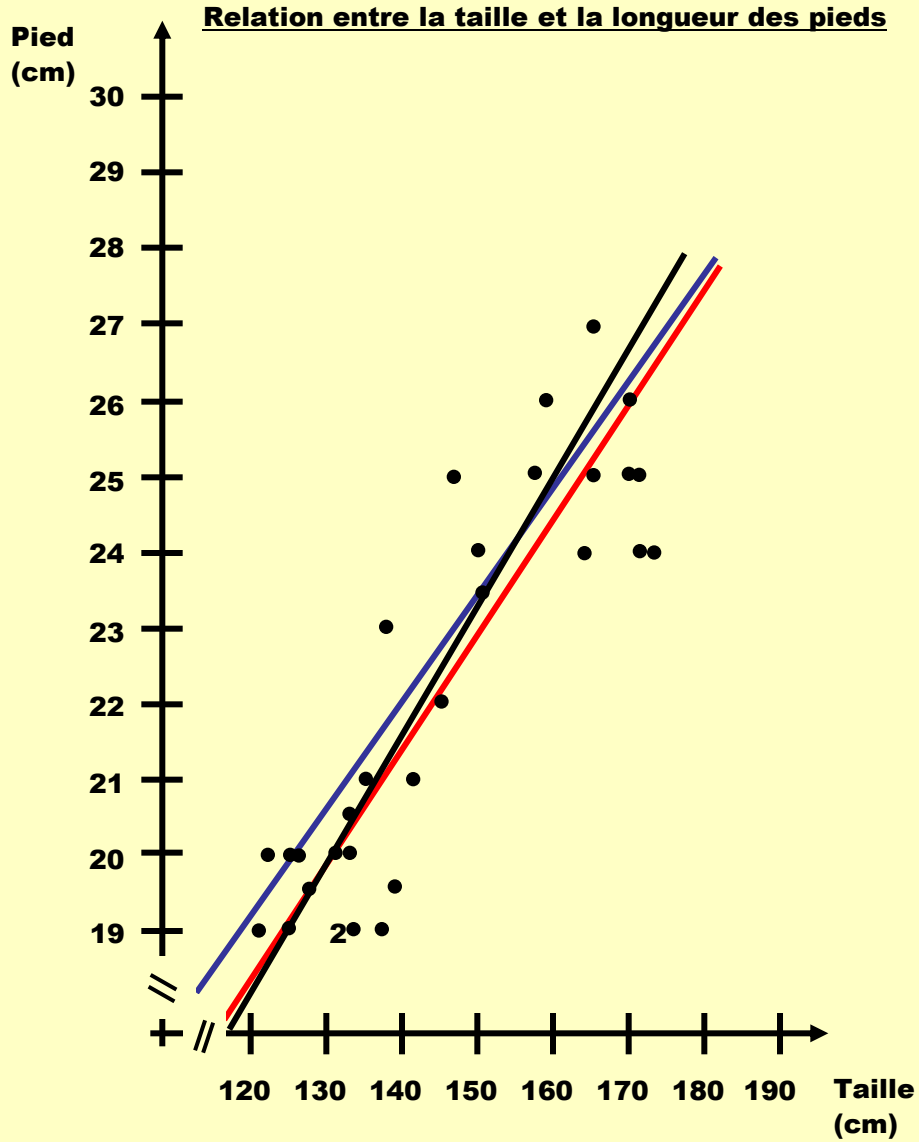
Droite tracée manuellement

Droite de Mayer

Droite de régression des moindres carrés

Remarque :

Même s'il y a de légères différences entre les droites, les trois méthodes traduisent approximativement la même réalité.



Lorsque la droite de régression est établie, elle peut servir à extrapoler, c'est-à-dire, à faire des prédictions.

Exemple

Nous pourrions nous demander quelle serait, théoriquement, la longueur des pieds d'un homme qui mesurerait 200 cm (soit 2 mètres).

x représente la taille, donc $y = 0{,}15x - 0{,}35$

$$y = 0{,}15 \times 200 - 0{,}35 = 29{,}65 \text{ cm}$$

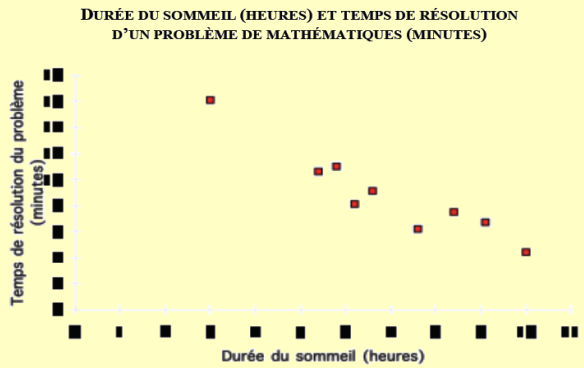
Cet exemple est un peu farfelu, mais il démontre l'utilité de la droite de régression.

Remarque

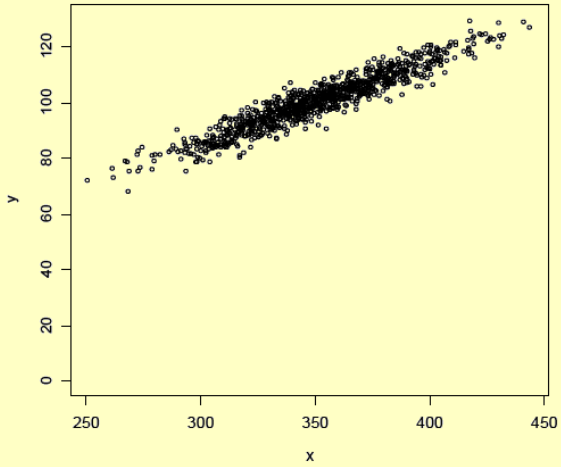
Les différents outils que nous venons de voir sont significatifs en autant que les échantillons étudiés soient assez nombreux.

Exemples

Diagramme de dispersion :



Très peu de données, donc étude peu significative.



Beaucoup de données, donc étude beaucoup plus significative.

En statistique, un échantillon trop petit est souvent une source de biais.