CORRÉLATIONS ET REGRESSIONS

Régis Schlagdenhauffen

CORRÉLATION ET RÉGRESSION

- Il est rare de pouvoir mener une enquête statistique sur toute une population (P de taille N) dont chaque individu serait repéré au moyen d'un couple de variables aléatoires X et Y quantitatives. Si nous interrogions toute la population, nous disposerions de la série statistique bivariée suivante : $\{(x_i, y_i) ; i=1,..., (x_n, y_n) ; i= N\}$

- Dans la mesure où, nous ne pouvons pas interroger la population entière, nous nous limitons à un **échantillon de population** (de taille n) nous offrant la série statistique suivante : {(xi ,yi); i=1,...,n}

- Dans un premier temps, nous allons décrire ces variables et essayer de voir si elles ont un **lien entre elles**, c'est-à-dire s'il y a **corrélation entre les deux variables étudiées** X et Y.

- Si nous **voulons étendre cette notion** de corrélation à l'ensemble des individus de la population, c'est-à-**dire procéder par inférence**, nous devrons utiliser des tests, afin de nous assurer que l'échantillon est représentatif.

De la pratique à la théorie

- **Un exemple : le poids et la taille**

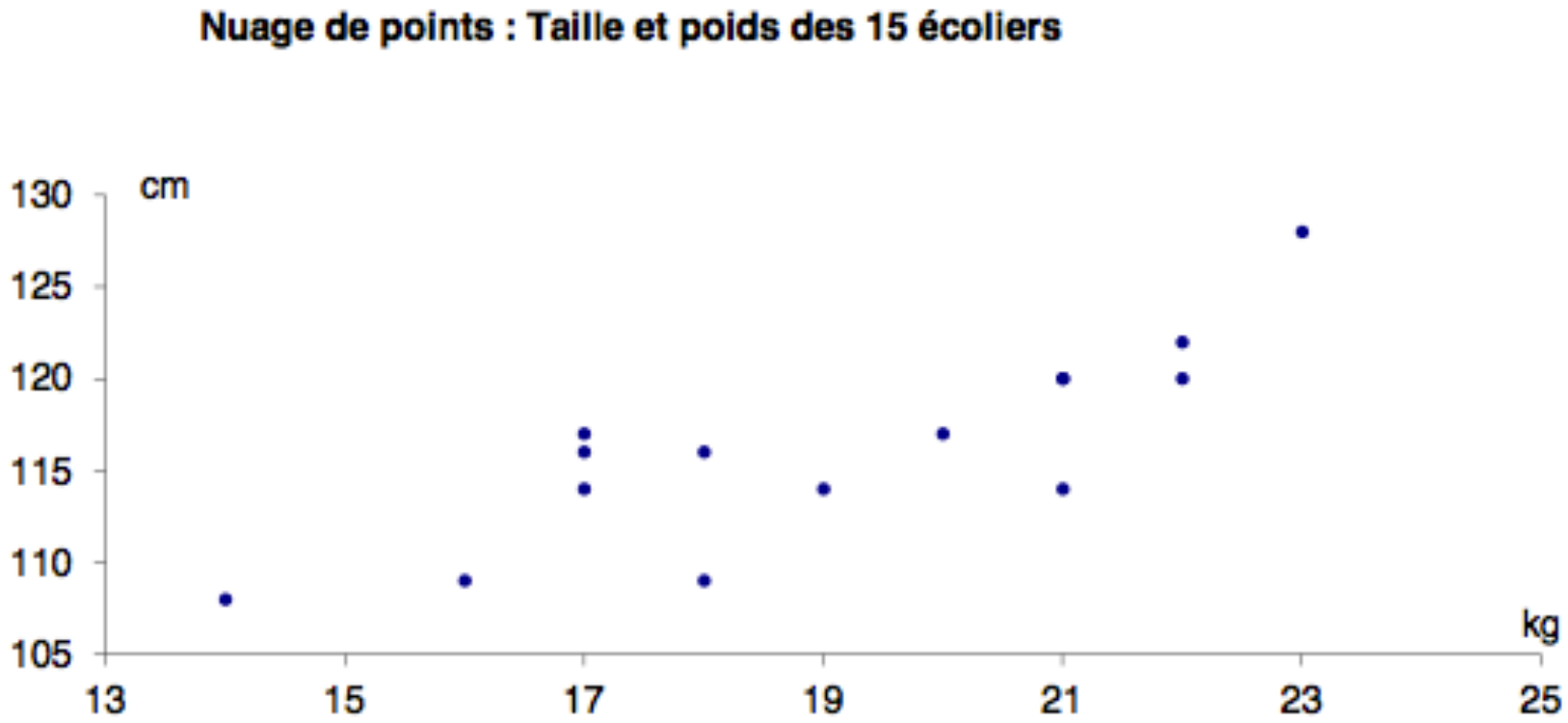
- Des collègues se sont intéressés au poids et à la taille des enfants en fin de dernière année de maternelle, des enfants d'environ 6 ans. 2 169 enfants ont fait partie de cette enquête (N), mais pour l'exemple qui nous intéresse ici, nous avons tiré un très faible échantillon de 15 individus (n=15).
- Chaque enfant a été pesé (X) et mesuré (Y), ce qui a donné 15 couples de mesures. Nous disposons dès lors de la série bivariée:$\{(x_i, y_i) ; i=1,...,15\}$
- Nous pouvons **consigner les données** de chaque individu ***i*** dans le tableau suivant; elles nous permettront de les visualiser au moyen d'une **représentation graphique** par un point défini dans un **système d'axes** au moyen de ses coordonnées :
  x_i et y_i.

Lecture : le premier écolier (x_1, y_1) pèse 17 kg et mesure 116cm, soit 1m16, etc.

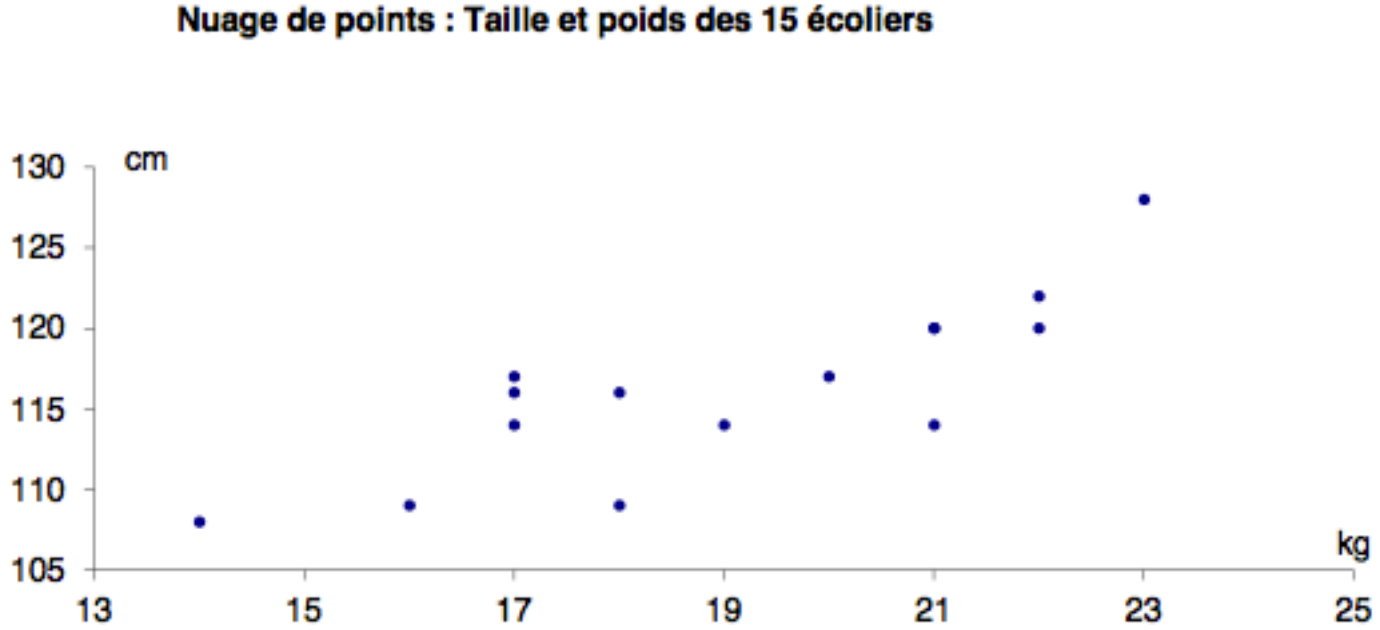
écolier	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	17	116	289	13 456	1 972
2	16	109	256	11 881	1 744
3	17	117	289	13 689	1 989
4	22	122	484	14 884	2 684
5	22	120	484	14 400	2 640
6	23	128	529	16 384	2 944
7	21	114	441	12 996	2 394
8	14	108	196	11 664	1 512
9	18	116	324	13 456	2 088
10	21	120	441	14 400	2 520
11	17	114	289	12 996	1 938
12	18	109	324	11 881	1 962
13	19	114	361	12 996	2 166
14	21	120	441	14 400	2 520
15	20	117	400	13 689	2 340
total	286	1 744	5 548	203 172	33 413

Les colonnes x_i^2 ; y_i^2 ; $x_i y_i$ serviront aux calculs que nous étudierons plus loin.

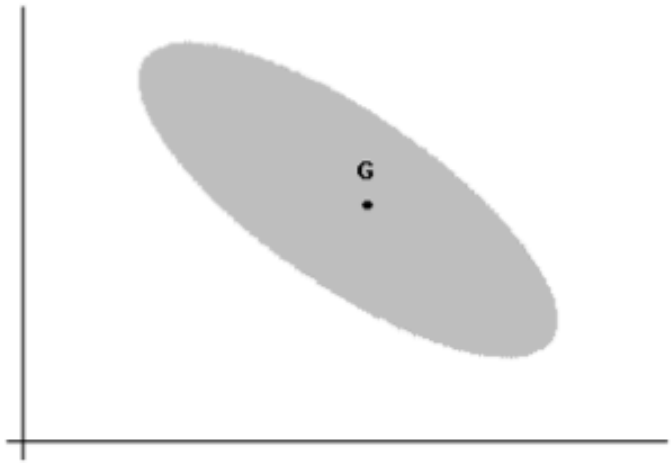
- Les données de ce tableau peuvent aussi être représentées graphiquement. On retrouve le premier écolier dont nous avons consigné le poids et la taille (17 kg et 116 cm de hauteur) à l'intersection des deux coordonnées des deux valeurs en abscisse et ordonnée. Il en va de même pour les 14 autres. Ceci nous donne le **nuage de points** suivant:



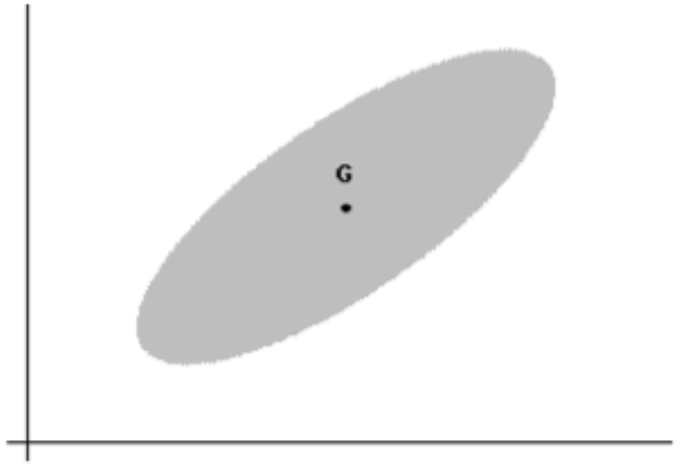
- À l'œil nu, nous observons que tous les points côte-à-côte forment une sorte de « nuage ». Ils sont plus ou moins rapprochés et croissent d'une manière qu'il reste à déterminer. La **forme du nuage** est une première indication du lien existant entre X et Y. Ici, plus la taille augmente, plus le poids augmente, ce qui laisse augurer une **corrélation linéaire positive** entre le poids et la taille.



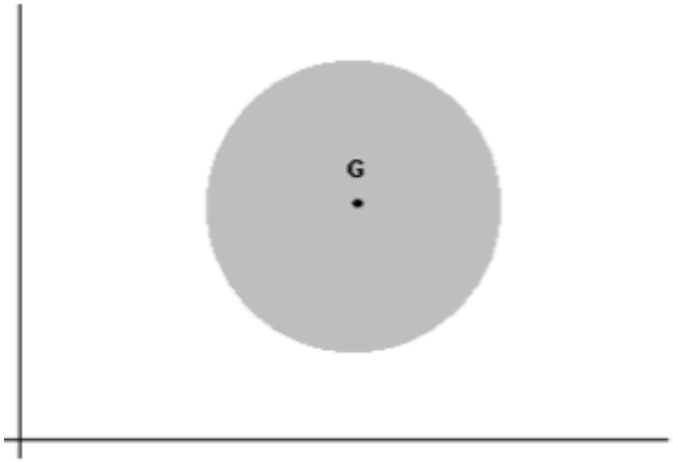
- Les formes de nuage les plus fréquentes sont les suivantes :



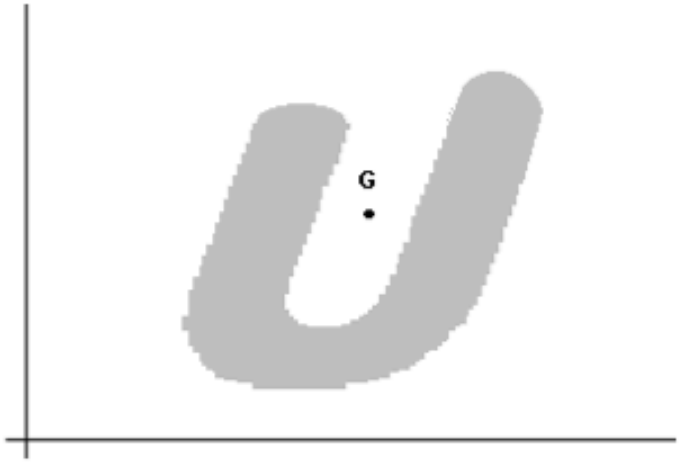
Quand X croît Y décroît ou inversement.
Corrélation négative



Quand X croît Y croît et inversement.
Corrélation positive



Indépendance linéaire



Indépendance linéaire

Démonstration de la corrélation linéaire

- Dans un premier temps nous allons observer le poids, puis la taille – autrement dit la **distribution de X et de Y**.
- Si nous étudions la distribution de X (indépendamment de celle de Y), nous pouvons calculer sa moyenne et son écart-type. Pour ce faire nous utilisons la colonne des x_i.
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{i=n} x_i$ donc $\bar{x} = \frac{286}{15} = 19{,}0667 \text{ kg}$

écolier	x_i
1	17
2	16
3	17
4	22
5	22
6	23
7	21
8	14
9	18
10	21
11	17
12	18
13	19
14	21
15	20
total	286

- Les enfants de notre échantillon pèsent donc en moyenne 19,1 kg. Maintenant, nous allons calculer la variance de X afin de pouvoir connaître l'écart-type, donc la dispersion des valeurs autour de la moyenne.
- Nous utiliserons la somme de la colonne x_i^2 (=5548) pour calculer la **variance** = $\mathbf{x_i^2/n - my}$ (avec x_i^2=5548; n=15 et my=19)

$$s_x^2 = \frac{5\,548}{15} - 19{,}0667^2 = 6{,}3289$$

- D'où **l'écart type** :

$$s_x = \sqrt{6{,}3289} = 2{,}604 \text{ kg}$$

écolier	x_i	x_i^2
1	17	289
2	16	256
3	17	289
4	22	484
5	22	484
6	23	529
7	21	441
8	14	196
9	18	324
10	21	441
11	17	289
12	18	324
13	19	361
14	21	441
15	20	400
total	286	5 548

- Nous procédons de même pour Y (la taille)
- Au terme des calculs nous pouvons affirmer que les 15 enfants mesurent en moyenne 116,2667 cm (soit 116,3 cm) avec un écart type de 5,3647 cm
- Désormais nous pouvons déterminer le **centre de gravité** (**G**) de notre nuage de points.
- Il se situe à l'intersection du poids moyen et de la taille moyenne.
- Donc G a pour coordonnées : Moyenne de X ; moyenne de Y
- G = (19,07 ; 116,27)
- Le **centre de gravité (G) va nous permettre de** lier les deux variables X et Y et d'étudier leur **covariance** : **Cov (X,Y)**.

$$\bar{y} = \frac{1744}{15} = 116{,}2667 \text{ cm}$$

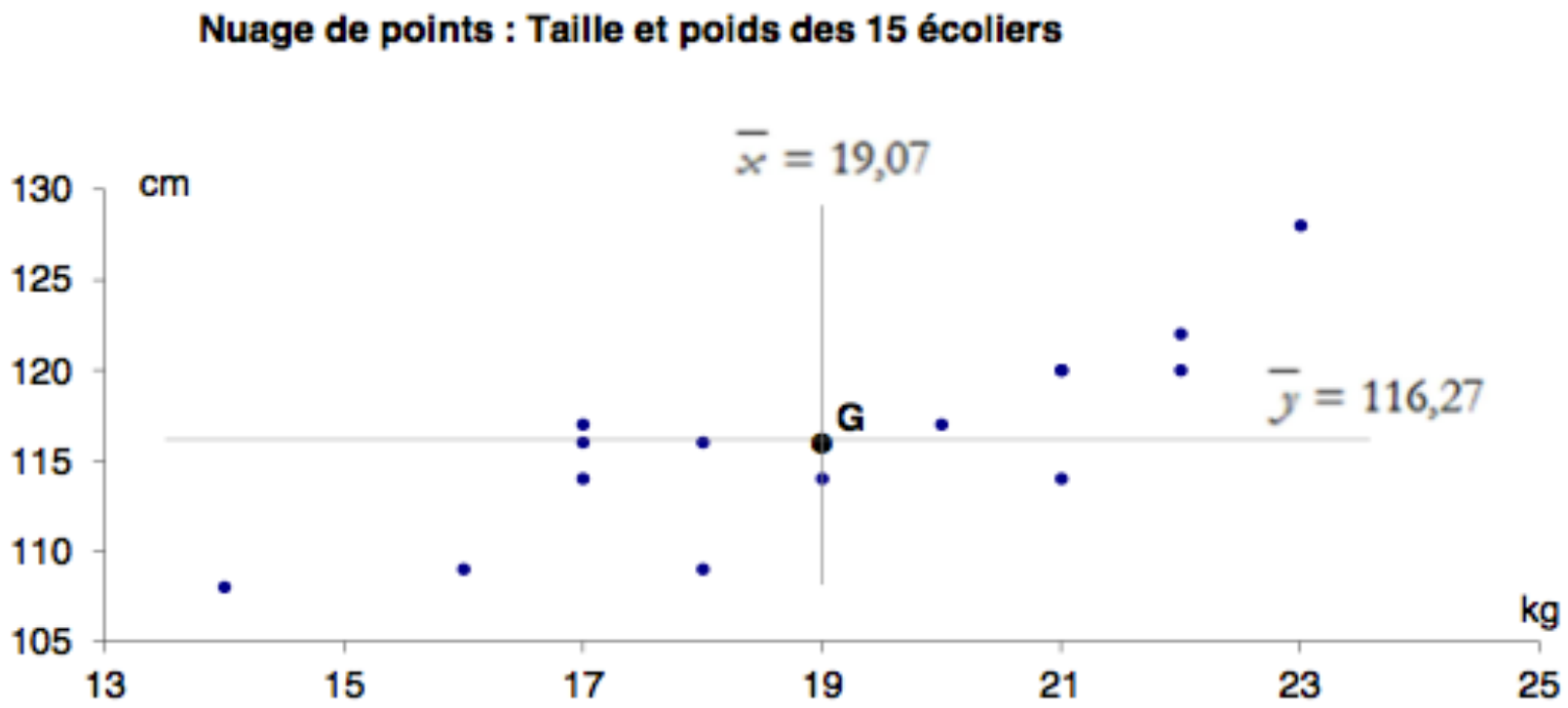
$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$$

$$= \frac{203\,172}{15} - 116{,}2667^2$$

$$= 26{,}8622$$

$$s_y = \sqrt{26{,}8622}$$

$$= 5{,}3647 \text{ cm}$$



En traçant 2 lignes perpendiculaires à l'intersection des coordonnées de G, nous faisons ainsi 4 quadrants (*que l'on numérote dans le sens des aiguilles d'une montre, par ex. en bas à gauche = cadran 4*), on observe que ce sont dans les cadrans 2 et 4 que se trouve le plus grand nombre de points, *i.e.* les deux cadrans positifs ↗. Cela indique que dans notre nuage de points X et Y augmentent en même temps.

- Le **calcul de la covariance** nous permet de prouver le lien positif qui unit X et Y (*N.B. - elle aurait été négative dans le cas où X augmente et Y diminue*).
- Pour calculer la covariance dans notre exemple, nous avons besoin de la **somme des x_i*y_i** = 33413 ; de **l'effectif** n=15 ; de la **moyenne des poids** (My_x) = 19,0667 et de la **moyenne des tailles** (My_y) = 116,2667.
- Nous appliquons la formule :

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\, y_i\right) - \bar{x}\,\bar{y}$$

$$\mathrm{Cov}(X,Y) = \frac{33\,413}{15} - 19{,}0667 \times 116{,}2667 = 10{,}7155$$

$x_i y_i$
1 972
1 744
1 989
2 684
2 640
2 944
2 394
1 512
2 088
2 520
1 938
1 962
2 166
2 520
2 340
33 413

Le coefficient de corrélation

- La **covariance** est un outil permettant de calculer le **coefficient de corrélation r**.
- Ce coefficient de corrélation est toujours compris entre –1 et +1. Plus exactement entre –1 et 0, lorsque la corrélation est négative (X croît Y décroît ou inversement) ; et entre 0 et +1 lorsque la corrélation est positive (X croît Y croît ou inversement) comme dans l'exemple précédent.
- La formule du coefficient de corrélation linéaire est :

$$r = \frac{Cov(X, Y)}{S_x \, S_y}$$

- Pour calculer $r = \frac{Cov(X, Y)}{s_x \, s_y}$

Nous avons donc besoin de connaître la Covariance de X et Y *Cov (X,Y)* ainsi que les écarts type de X (S_x) et de Y (S_y).

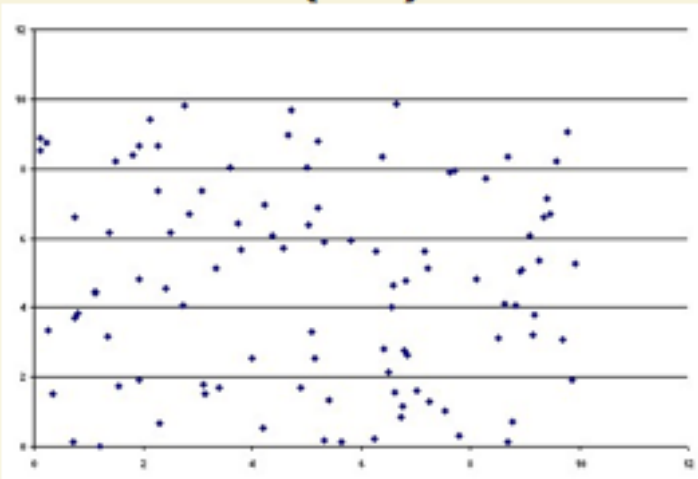
Dans le cas de notre exemple, nous avons déterminé précédemment Cov(X,Y) = 10,7155 ; S_x= 2,604 et Y_x=5,3647 d'où:

$$r = \frac{10,7155}{2,604 \text{ x } 5,3647} = 0,8218$$

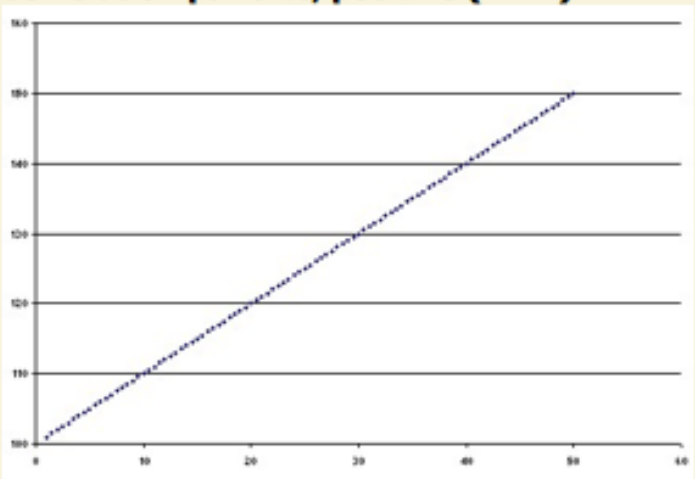
Conclusion: La corrélation est forte entre la taille et le poids. Il est évident que plus un enfant est grand, plus il va être lourd. Cependant, ce coefficient n'est pas égal à 1, ce qui aurait signifié une dépendance linéaire totale entre la taille et le poids, c'est-à-dire que le poids ne dépendrait exclusivement que de la taille.

Quelques exemples

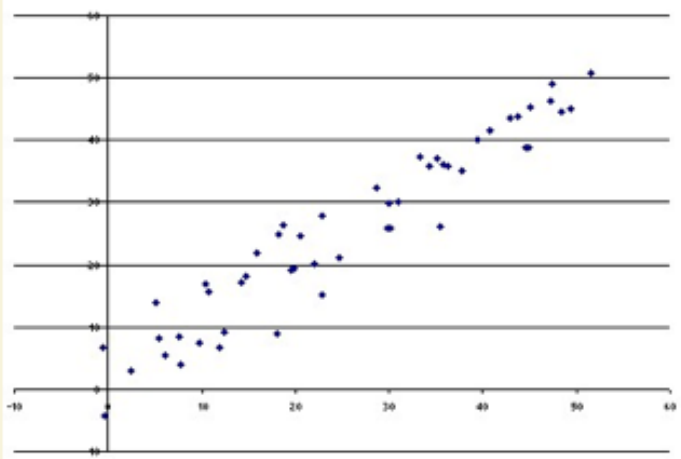
Corrélation nulle (r = 0)



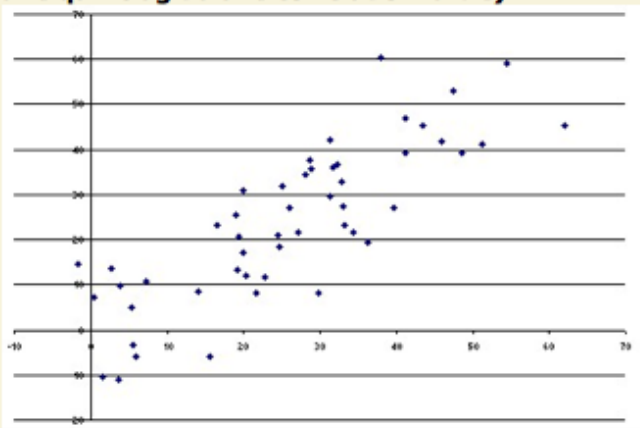
Corrélation parfaite, positive (r = 1)



Corrélation forte positive

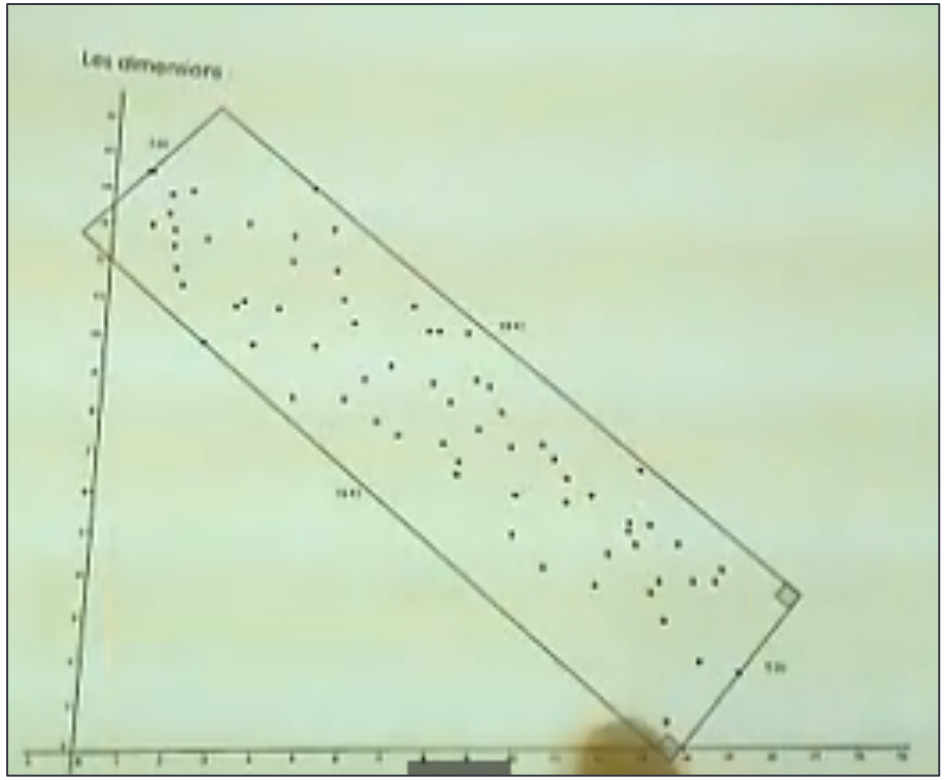


Corrélation un peu moins forte que l'exemple précédent, dire qu'il s'agit d'une corrélation faible)



Approximer la corrélation à l'œil nu

- Pour approximer la corrélation linéaire sans outil, il suffit de tracer un rectangle qui englobe tous les points du nuage et qui ait la plus petite surface (*donc le plus proche possible de l'ensemble du nuage de points*).



- Ensuite, il convient de reporter la longueur du grand côté et celle du petit côté, ici 19,42 cm et 5,06 cm.

- Pour *approximer r* =

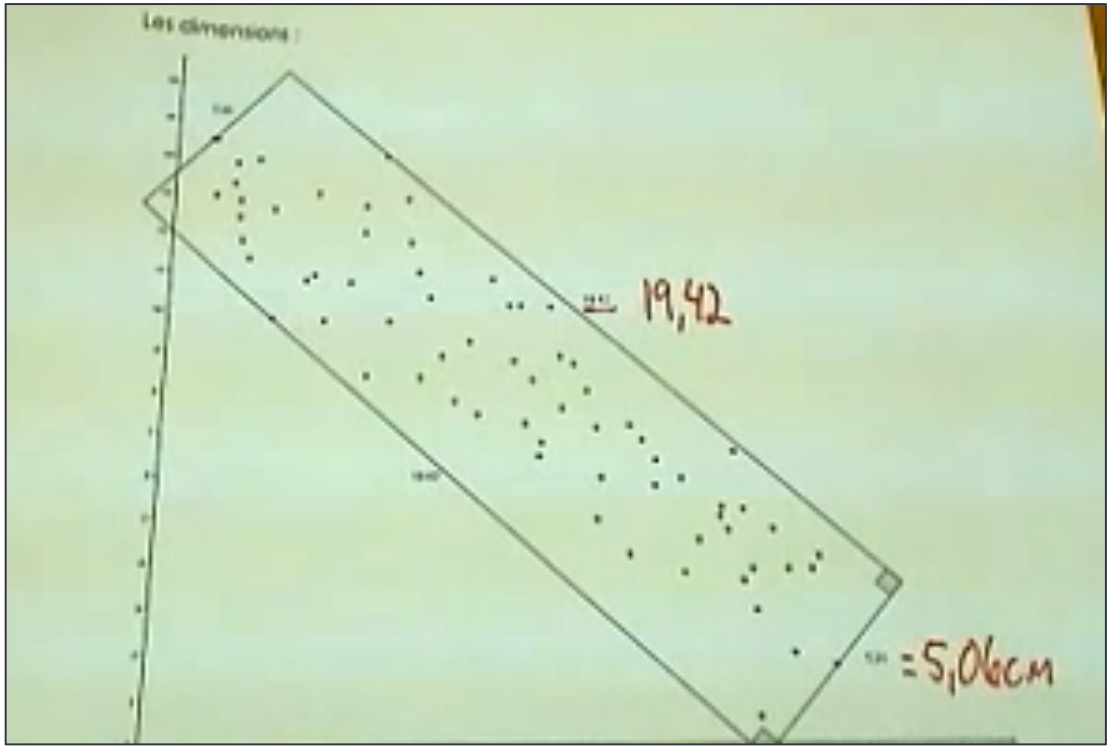
  ± 1 - (petit côté/
  grand côté)

Ici la pente décroît, donc r négatif, d'où

r = - (1 – (5,06/19,42))

r = - (1-0,26)

r = - 0,74



Conclusion : il y a une corrélation moyenne entre X et Y car <0,75. (*Contrairement à l'ex. poids taille où r ≥0,75 donc fort*).
NB : Si r <0,5, corrélation faible, si r <0,25 elle est faible à nulle.

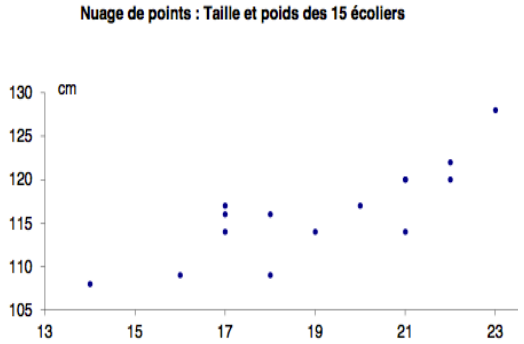
La droite de régression

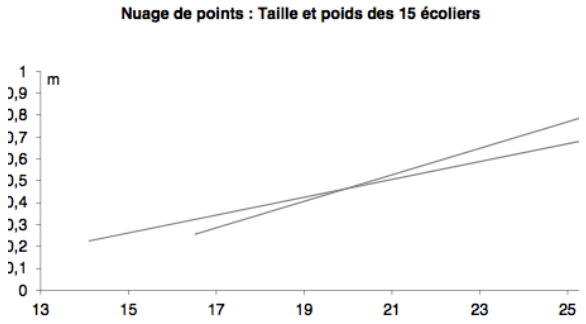
- Le coefficient de corrélation linéaire constitue un outil utile pour tracer une **droite de régression linéaire**. Cette droite synthétise de deux coups de crayon les données recueillies puis présentées sous forme d'un nuage de points.

- **Tableau -> Nuage -> Droites**

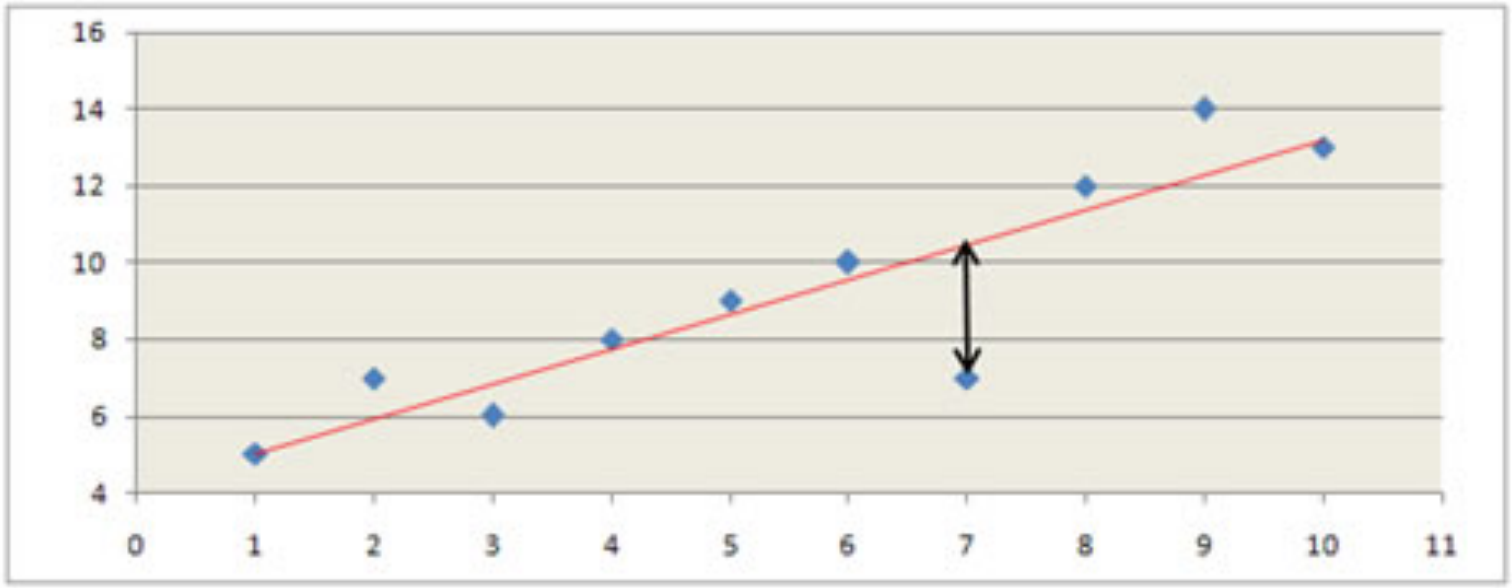
écolier	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	17	116	289	13 456	1 972
2	16	109	256	11 881	1 744
3	17	117	289	13 689	1 989
4	22	122	484	14 884	2 684
5	22	120	484	14 400	2 640
6	23	128	529	16 384	2 944
7	21	114	441	12 996	2 394
8	14	108	196	11 664	1 512
9	18	116	324	13 456	2 088
10	21	120	441	14 400	2 520
11	17	114	289	12 996	1 938
12	18	109	324	11 881	1 962
13	19	114	361	12 996	2 166
14	21	120	441	14 400	2 520
15	20	117	400	13 689	2 340
total	286	1 744	5 548	203 172	33 413

lonnes x_i^2 ; y_i^2 ; $x_i y_i$ serviront aux calculs que nous étudierons plu





- ✓Dès lors que les variables X et Y sont dépendantes linéairement, on peut exprimer X en fonction de Y, comme on peut exprimer Y en fonction de X, en les liant par une formule du type ***y* = a*x*+b** (avec a désignant le « taux de variation » ou « coefficient directeur » et b « l'ordonnée à l'origine »).
- ✓La droite y = ax+b, est dite **droite des moindres carrés ;** elle résume au mieux le nuage de points des observations, passe <u>toujours</u> par le centre de gravité (**G**) et minimise les « résidus » (i.e. les différences entre les valeurs observées et les valeurs prédites par le modèle)



- Par rapport au nuage de points, la droite de régression linéaire de y sur x minimise la somme des distances verticales des points à la droite. La droite de régression linéaire de x sur y minimise la somme des distances horizontales. Les deux droites se coupent au centre de du nuage de points. L'écart entre les deux est d'autant plus grand que la corrélation est faible.

- Dans notre exemple, il est évident que le poids dépend de la taille et non le contraire. Nous dirons alors que **le poids** est une **variable dépendante** ou expliquée et que **la taille** est une **variable explicative**.
- Il s'agit alors de trouver les équations des droites qui ajustent le mieux le nuage de point.
- Plus encore, la droite des moindres carrés permet de **prédire** des valeurs manquantes ou non-observées !

- Calcul de la **droite de régression de *Y* en *X***

- Il s'agit ici de déterminer les valeurs de a et b. Cette détermination s'effectue en utilisant la **méthode des moindres carrés**.
- Pour cela, calculons les coefficients de la droite r de forme **Y = a X + b**.

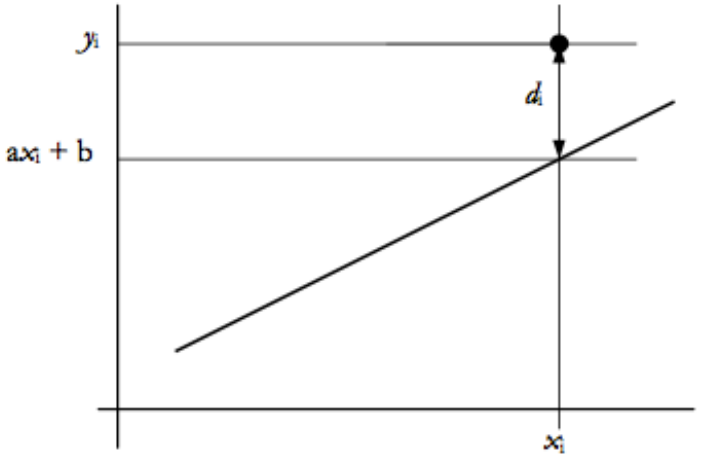
- **Calcul de a** : le coefficient directeur.
- Pour cela nous avons besoin de la Covariance de X et Y ainsi que de la variance de X (puisque nous déterminons Y en fonction de X).
- Or, Cov (X,Y)= 10,7155 et de $S_x^2 = 2{,}604^2 = 6{,}7809$

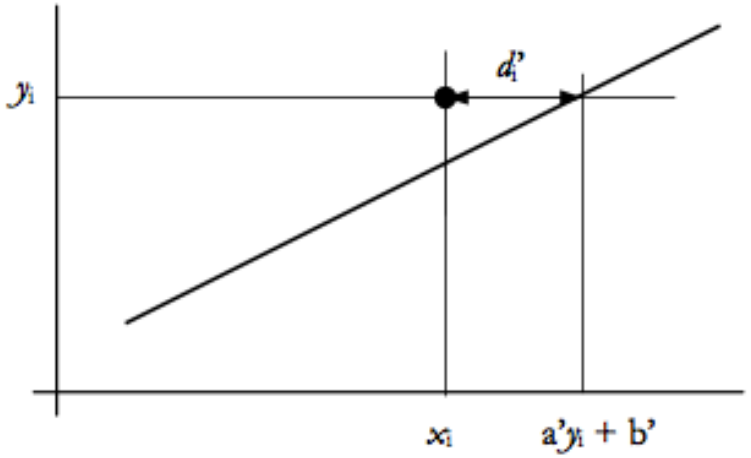
$$a = \frac{Cov(X,Y)}{s_x^2} \qquad a = \frac{10{,}7155}{6{,}7809} = 1{,}5802$$

- **Calcul de b:** Pour cela nous avons besoin de la moyenne de x, de y et de a. Or, $My_x = 19{,}067$; $My_y = 116{,}267$ et a = 1,5802.
-
- $$b = \bar{y} - a\bar{x}$$

Donc b = 116,267-1,5802*19,067
= 116,267-30,129 **= 86,138**

- On peut procéder de même pour calculer la **droite de régression de *X* en Y.** Car dans le premier cas on a cherché à minimiser les distances verticales (Y en X) tandis que maintenant on minimise les distances horizontales (X en Y).





- ***Droite de régression de Y en X*** ***Droite de régression de X en Y***

- Pour calculer la **droite de régression de *X* en Y** on procède de la même manière afin de calculer la droite d'équation : X = a'Y + b'
- Il s'agit ici de déterminer la valeur de a' et b'. Cette détermination s'effectue comme précédemment en utilisant la méthode des moindres carrés, d'où:

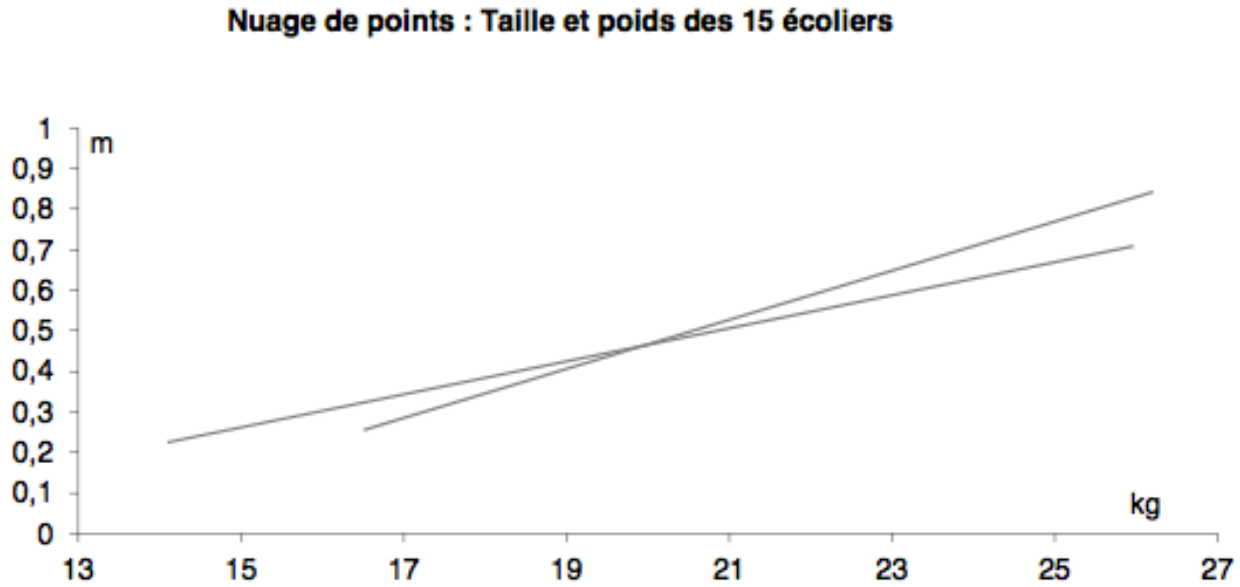
$$a' = \frac{Cov(X, Y)}{s_y^2} \qquad b' = \bar{x} - a'\,\bar{y}$$

$$a' = \frac{10{,}7155}{28{,}7809} = 0{,}3723 \quad \Rightarrow \quad b' = -24{,}2210$$

- L'équation du poids en fonction de la taille qui est la droite qu'il convient de calculer puisque le poids peut s'expliquer par la taille, s'écrit donc :
- **X = 0,3723Y – 24,2210**

Positionnement des droites de régression

- Maintenant que nous disposons des 2 coefficients directeurs des deux droites, à savoir a= 1,5802 et a'= 0,3723 nous pouvons tracer les deux droites de régression linéaire:



Formulaire : Corrélation et régression

Moyenne de X : $\bar{x} = \frac{\sum x_i}{n}$	Moyenne de Y : $\bar{y} = \frac{\sum y_i}{n}$
Écart-type de X : $s_x = \sqrt{\left(\frac{\sum x_i^2}{n}\right) - \bar{x}^2}$	Écart-type de Y : $s_y = \sqrt{\left(\frac{\sum y_i^2}{n}\right) - \bar{y}^2}$
Covariance de X et Y : $Cov(X,Y) = \left(\frac{\sum x_i y_i}{n}\right) - \bar{x}\bar{y}$	
Coefficient de corrélation de X et Y : $r = \frac{Cov(X,Y)}{s_x s_y}$	
Droite de régression de Y en fonction de X : $Y = aX + b$; $a = \frac{Cov(X,Y)}{{s_x}^2}$; $b = \bar{y} - a\bar{x}$	Droite de régression de X en fonction de Y : $X = a'Y + b'$; $a' = \frac{Cov(X,Y)}{{s_y}^2}$; $b' = \bar{x} - a'\bar{y}$