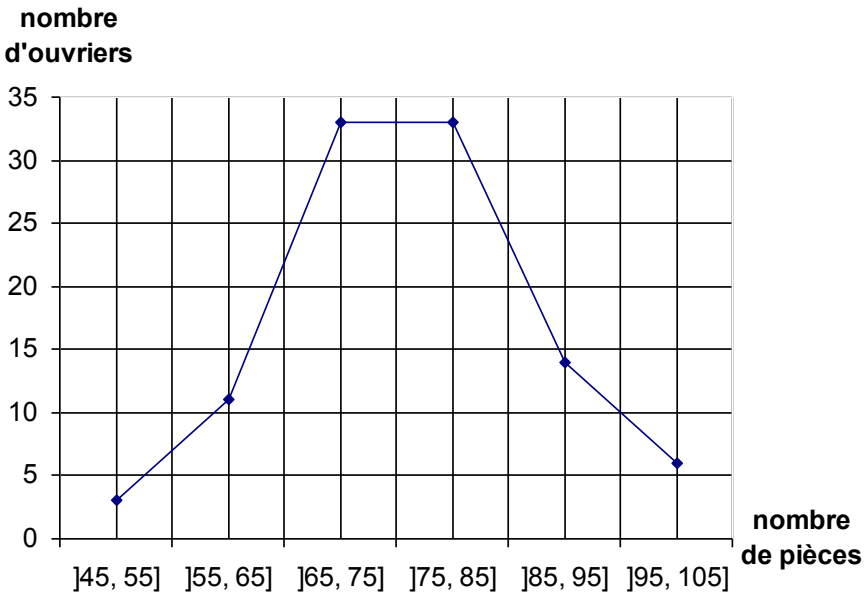
STATISTIQUE A DEUX VARIABLES

Ajustement : Position du problème

Considérons la série statistique suivante :

Dans une usine, 100 ouvriers fabriquent le même type de pièces ; en fin de journée, on a relevé le nombre de pièces fabriquées par chacun d'eux et obtenu le tableau ci-dessous à gauche. Ci-dessous, à droite, on trouve le polygone des effectifs de cette série.

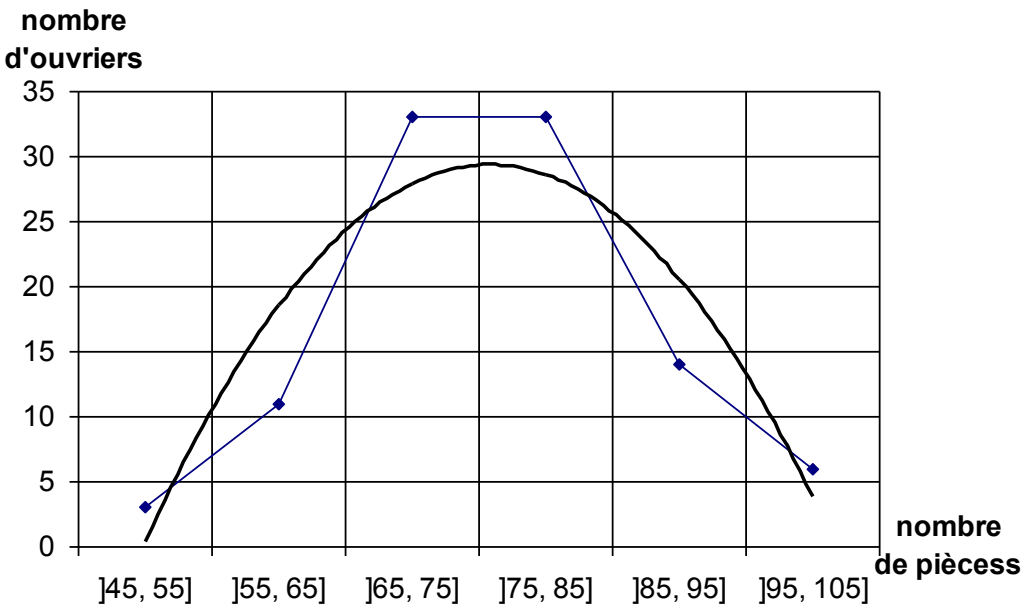
X	f
]45, 55]	3
]55, 65]	11
]65, 75]	33
]75, 85]	33
]85,95]	14
]95, 105]	6

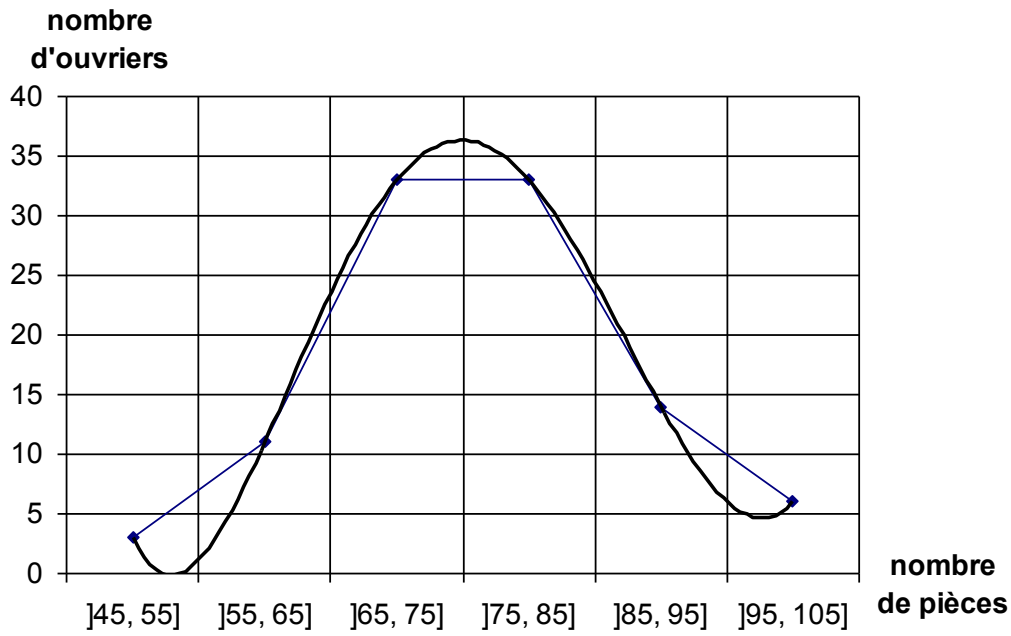


Le problème qui nous préoccupe est de voir si les sommets de ce polygone se trouvent **approximativement** sur une courbe d'équation **simple** et facilement utilisable pour des développements ultérieurs. C'est le problème de l'**ajustement.**

Dans les exemples étudiés en statistique, le polygone des effectifs peut généralement être approché par une courbe « en cloche » ou courbe de Gauss. L'équation de cette courbe n'est pas très simple et nous ne l'étudierons pas. Cependant, il peut aussi arriver que la courbe sur laquelle se situent approximativement les sommets du polygone des effectifs soit une droite (on parlera alors d'**ajustement linéaire**), une parabole, une hyperbole, une cubique, une courbe logarithmique, exponentielle, …

Dans le cas précédent, on peut déterminer une « courbe d'ajustement ». Au moyen d'un tableur, on trouve facilement une des courbes suivantes (trendline en anglais) :





Dans ce qui suit, nous nous limiterons à l'étude de l'ajustement linéaire et nous étudierons deux méthodes qui permettent d'y arriver : la méthode des **moyennes discontinues** et la **méthode des moindres carrés.**

A. Application à la physique

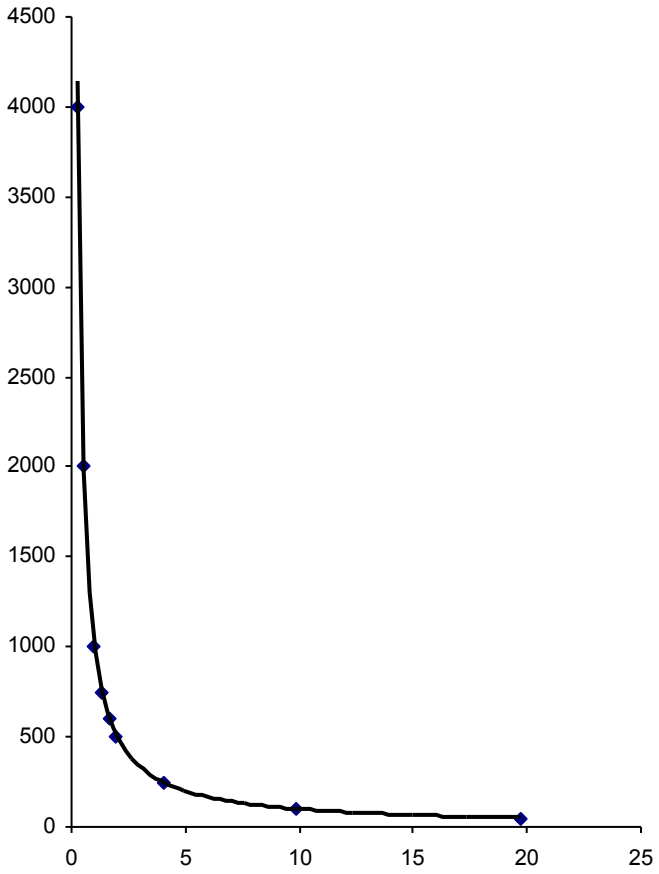
Le même problème se pose lorsque l'on étudie deux (ou plusieurs, mais nous nous limiterons à deux) caractères d'une même population et que l'on cherche à trouver une relation entre ces deux caractères. C'est souvent le cas en physique expérimentale.

Considérons, par exemple, une masse de gaz à une température déterminée. Cette masse de gaz occupe un certain volume *V* et se trouve à une pression *P*. Réduisons le volume et mesurons la nouvelle pression et le nouveau volume. Recommençons l'opération plusieurs fois. Nous obtenons le tableau suivant :

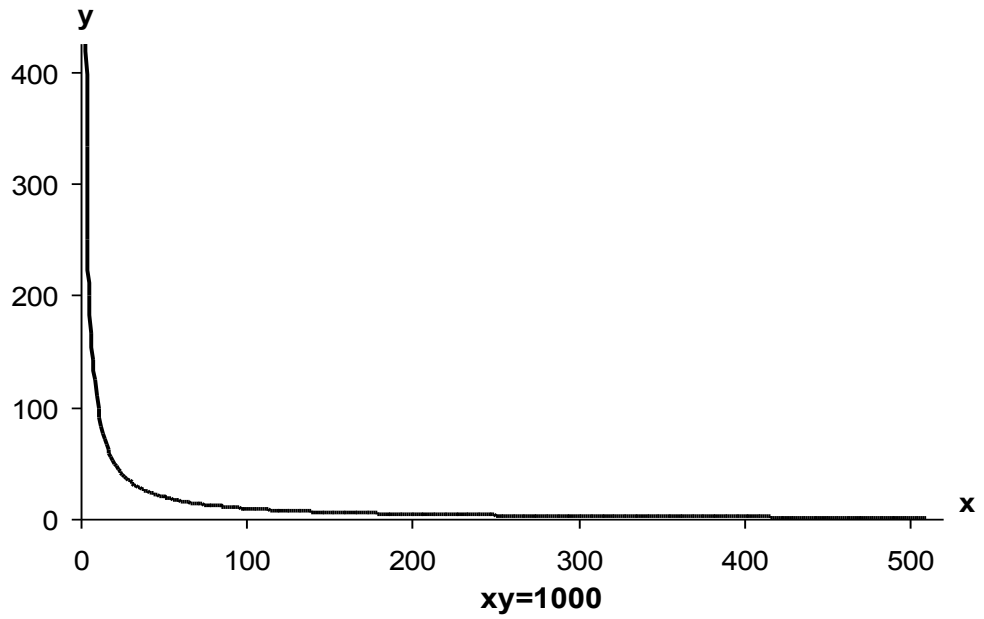
P	**4000**	**2000**	**1000**	**750**	**600**	**500**	**250**	**100**	**50**
V	**0,24**	**0,51**	**1**	**1,35**	**1,67**	**1,98**	**4,02**	**9,9**	**19,7**
PV	**960**	**1020**	**1000**	**1012,5**	**1002**	**990**	**1005**	**990**	**985**

Le problème est : *P* et *V* sont-ils liés par une relation ? Nous le savons c'est la loi de Boyle-Mariotte : $PV = c^{te}$. Représentons graphiquement les résultats expérimentaux. Ici aussi, un tableur permet d'obtenir rapidement une courbe ajustée. Elle est dessinée en gras sur la figure et semble être très proche d'une branche d'hyperbole équilatère. Nous avons **ajusté** cet ensemble de points suivant une hyperbole. En fait, s'il n'y avait pas eu d'erreur de mesure, ces points seraient situés sur la courbe d'équation $xy = 1000$, dont le graphique figure à la page suivante.

N.B. : Il peut, bien sûr, arriver que les caractères étudiés soient totalement indépendants l'un de l'autre.



Loi de Boyle-Mariotte



B. Table de corrélation

Définition : La **corrélation** entre deux variables est le lien (dépendance statistique) qui unit ces deux variables. C'est ce que nous recherchons. La corrélation peut s'étudier dans un domaine beaucoup plus vaste que la statistique, mais les méthodes sont les mêmes.

La **table de corrélation** est un tableau à double entrée donnant les effectifs des deux caractères.

Exemple : Le tableau ci-dessous donne la taille et le poids des élèves de l'école. Il s'agit d'une table de corrélation. Le problème est de savoir s'il existe une corrélation entre ces deux variables.

	j		1	2	3	4	5	6	7	8	9	10	11	12	13	
i		Y_j	$45 \leq Y<49$	$49 \leq Y<53$	$53 \leq Y<57$	$57 \leq Y<61$	$61 \leq Y<65$	$65 \leq Y<69$	$69 \leq Y<73$	$73 \leq Y<77$	$77 \leq Y<81$	$81 \leq Y<85$	$85 \leq Y<89$	$89 \leq Y<93$	$93 \leq Y<97$	Totaux
	X_i		47	51	55	59	63	67	71	75	79	83	87	91	95	
1	$125 \leq X<131$	128	1	3	5	4										13
2	$131 \leq X<137$	134		2	5	8	1									16
3	$137 \leq X<143$	140		1	3	6	2	1								13
4	$143 \leq X<149$	146			2	4	8	12								26
5	$149 \leq X<155$	152			30	40	20	12	2							104
6	$155 \leq X<161$	158		11	12	60	38	15								136
7	$161 \leq X167$	164				72	190	185	48	22	5					522
8	$167 \leq X<173$	170					60	170	180	82	22					514
9	$173 \leq X<179$	176						10	25	82	20	1				138
10	$179 \leq X<185$	182								1	5	12	14	2	1	35
	Totaux		1	17	57	194	319	405	255	187	52	13	14	2	1	1517

Comme d'habitude, chaque classe est caractérisée par son centre. Nous nous trouvons ici **devant une série statistique double.** Chaque colonne, chaque ligne (et en particulier la colonne et la ligne des totaux) donne lieu à une série statistique simple et peut être étudiée comme telle. Le polygone des effectifs aura la forme d'une courbe en cloche.

C. Définitions et notations

1. Les **centres des classes** sont notés X_i et Y_j .

2. L'**effectif** (ou répétition) du couple (X_i, Y_j) est le nombre d'éléments de la population dont les valeurs du caractère sont respectivement X_i et Y_j .
Nous la noterons n_{ij} (i^e ligne, j^e colonne).

Exemple : L'effectif du couple (X_7, Y_6) est $n_{76} = 185$.

3. L'**effectif marginal** n_i correspondant à l'élément X_i est la somme des effectifs de tous les couples (X_i, Y_j) quel que soit Y_j Y_j.
On a donc : $n_i = n_{i1} + n_{i2} + n_{i3} + \ldots$

De même, l'**effectif marginal** n'_j correspondant à l'élément Y_j est la somme des effectifs de tous les couples (X_i, Y_j) quel que soit X_i.
On a donc : $n'_j = n_{1j} + n_{2j} + n_{3j} + \ldots$

Exemples : $n_3 = 13$; $n'_4 = 194$

4. L'**effectif total *N*** de la série est le nombre des observations effectuées.
Il correspond à : la somme des effectifs n_{ij} ;
la somme des effectifs marginaux n_i correspondant aux valeurs X_i;
la somme des effectifs marginaux n'_j correspondant aux valeurs Y_j.

Donc $N = n_{11} + n_{12} + n_{13} + \ldots + n_{21} + n_{22} + n_{23} + \ldots + n_{31} + n_{32} + n_{33} + \ldots + n_{41} + n_{42} + n_{43} + \ldots$
$= n_1 + n_2 + n_3 + \ldots$
$= n'_1 + n'_2 + n'_3 + \ldots$

L'étude statistique du seul caractère *X* peut se faire au départ des valeurs X_i et des effectifs marginaux n_i correspondants. Cette étude se fera comme on l'a vu en quatrième et permettra de construire les diagrammes que nous connaissons bien et de déterminer, pour cette série, le mode, la médiane, les quartiles, la moyenne arithmétique $\bar{x}$, la variance $\sigma^2(X)$ et l'écart-type $\sigma(X)$. La notation $\sigma(X)$ et non plus σ tout seul, est imposée par le fait que l'étude statistique du seul caractère *Y* peut se faire au départ des valeurs Y_j et des effectifs marginaux correspondants n'_j et donnera lieu à la détermination du mode, de la médiane, des quartiles, de la moyenne arithmétique $\bar{y}$, de la variance $\sigma^2(Y)$ et de l'écart-type $\sigma(Y)$ de cette série.

Dans le cas qui nous préoccupe, on a :

$$\bar{x} = 165{,}04 \ ; \ \sigma(X) = 8{,}71 \ ; \ \bar{y} = 66{,}93 \ ; \ \sigma(Y) = 6{,}54$$

5. La **covariance** $\mathbf{cov(X,Y)}$ est la moyenne arithmétique des produits $(X_i-\overline{x})(Y_i-\overline{y})$. Donc

$$\begin{aligned}\operatorname{cov}(X,Y)=\frac{1}{N}[&n_{11}(X_1-\overline{x})(Y_1-\overline{y})+n_{12}(X_1-\overline{x})(Y_2-\overline{y})+n_{13}(X_1-\overline{x})(Y_3-\overline{y})+\ldots\\&+n_{21}(X_2-\overline{x})(Y_1-\overline{y})+n_{22}(X_2-\overline{x})(Y_2-\overline{y})+n_{23}(X_2-\overline{x})(Y_3-\overline{y})+\ldots\\&+n_{31}(X_3-\overline{x})(Y_1-\overline{y})+n_{32}(X_3-\overline{x})(Y_2-\overline{y})+n_{33}(X_3-\overline{x})(Y_3-\overline{y})+\ldots\\&+n_{41}(X_4-\overline{x})(Y_1-\overline{y})+n_{42}(X_4-\overline{x})(Y_2-\overline{y})+n_{43}(X_4-\overline{x})(Y_3-\overline{y})+\ldots\\&+\ldots]\end{aligned}$$

La recherche de la covariance peut se faire grâce à la calculatrice et au moyen d'un tableau du genre de celui de la page 7. Il est nécessaire cependant que la calculatrice possède au moins trois mémoires. Les résultats intermédiaires seront néanmoins compliqués et la précision risque de s'en ressentir. C'est pourquoi, il sera, à notre avis, plus simple d'utiliser la formule :

$$\begin{aligned}\operatorname{cov}(X,Y)=\frac{1}{N}[&n_{11}X_1Y_1+n_{12}X_1Y_2+n_{13}X_1Y_3+\ldots\\&+n_{21}X_2Y_1+n_{22}X_2Y_2+n_{23}X_2Y_3+\ldots\\&+n_{31}X_3Y_1+n_{32}X_3Y_2+n_{33}X_3Y_3+\ldots\\&+n_{41}X_4Y_1+n_{42}X_4Y_2+n_{43}X_4Y_3+\ldots\\&+\ldots]-\overline{x}\,\overline{y}\end{aligned}$$

que nous accepterons sans démonstration.

Nous allons illustrer ce procédé par la recherche de la covariance de la série donnée page 4. Le tableau qui suit donne les valeurs des différents termes $n_{ij}X_iY_j$ compris dans les crochets de la formule ci-dessus. Dans la ligne et la colonne « totaux », on trouve des sommes partielles, tandis, qu'à l'intersection de ces deux rangées, on trouve la valeur de la somme comprise entre ces crochets. On obtient donc :

$$cov(X,Y)=\frac{16.820.642}{1.517}-165{,}04\,x\,66{,}93=42$$

6. Le **coefficient de corrélation linéaire est le nombre**

$$r = \frac{\text{cov}(X,Y)}{\sigma(X)\sigma(Y)}$$

Dans l'exercice qui précède, on obtient :

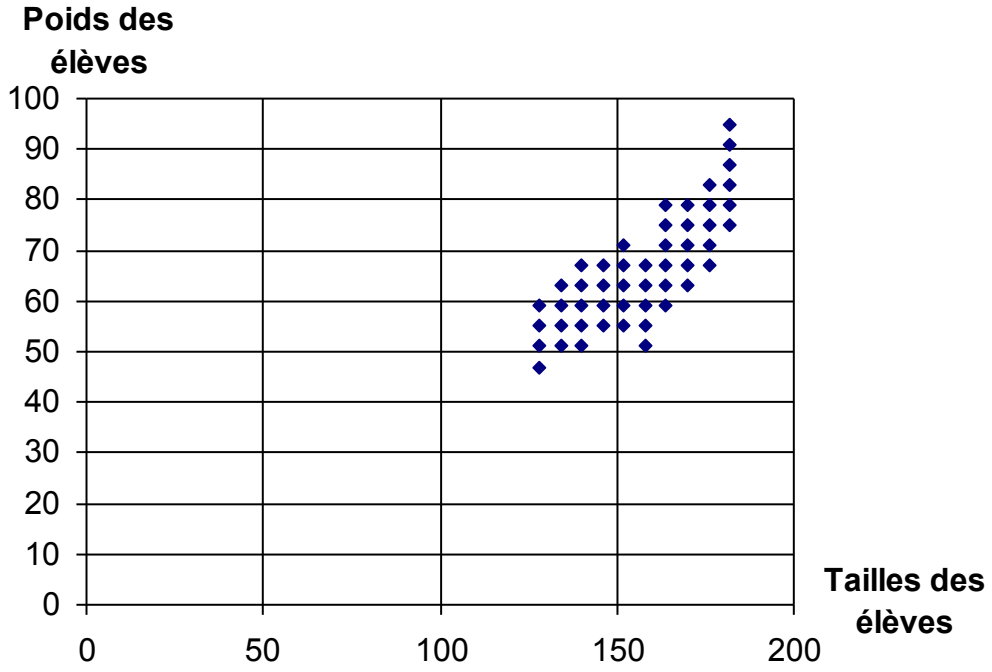
$$r = \frac{42}{8,71.6,54} = 0,74$$

N.B : On dit qu'il y a **forte corrélation si et seulement si** $0,87 \leq |r| \leq 1.$

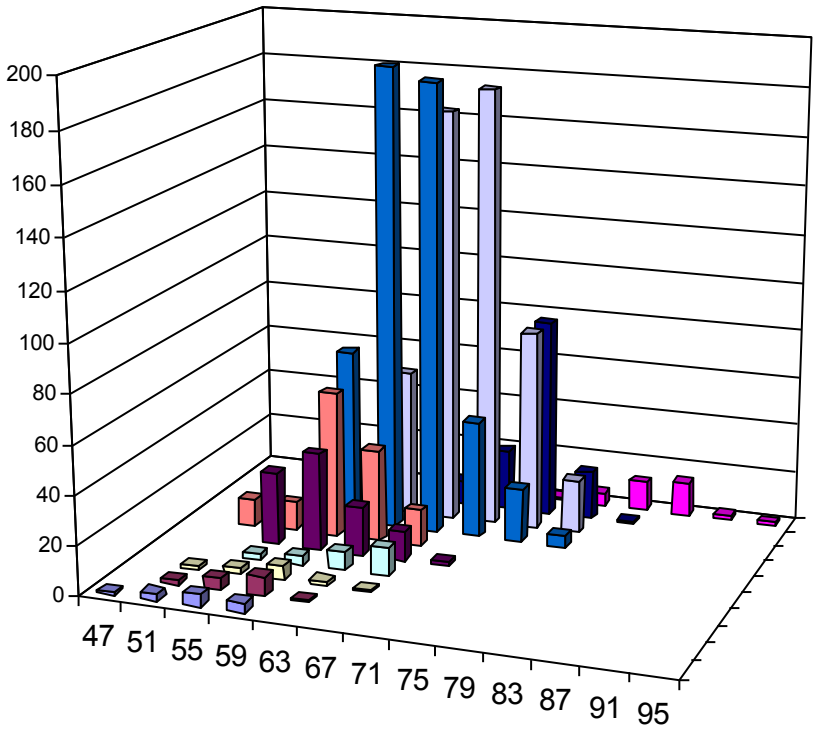
On démontre que $-1 \leq r \leq 1.$

D. Diagramme

Dans un système d'axes rectangulaires, portons en abscisse la taille des élèves et en ordonnée leurs poids. Nous obtenons ainsi **un nuage de points.**



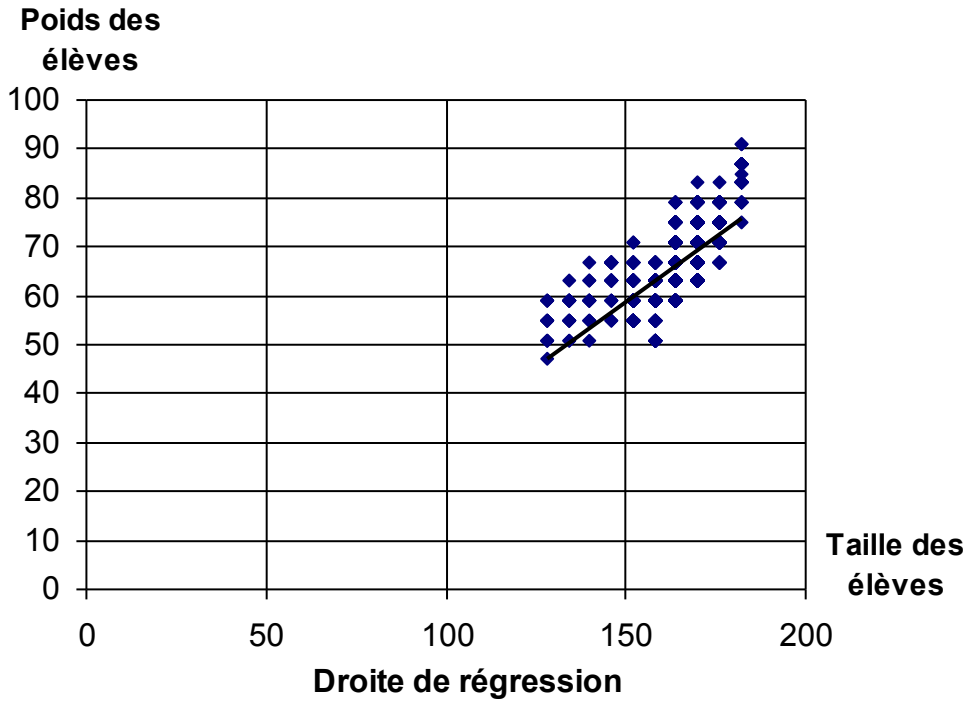
Remarque : Certains auteurs laissent le diagramme tel qu'il est présenté ci-dessus. D'autres préfèrent indiquer entre parenthèses, à côté de chaque point l'effectif correspondant. D'autres, enfin, suggèrent une représentation graphique, à trois dimensions, comme, par exemple, celle de la page suivante :



E. Détermination de la droite ajustée par la méthode graphique

Il suffit de tracer une droite qui traverse le nuage de points de telle manière que de chaque côté de la droite, il y ait, compte tenu des effectifs, à peu près le même nombre de points. La précision des résultats dépend de l'opérateur. Un peu d'habitude donne cependant des résultats très acceptables.

On peut obtenir, au moyen de l'ordinateur, une droite du genre de celle tracée ci-dessous.



F. Méthode des moyennes discontinues (Méthode de Mayer)

1. Définition

Le centre de gravité d'un ensemble de points est le point dont l'abscisse est la moyenne arithmétique des abscisses des points donnés et dont l'ordonnée est la moyenne arithmétique des ordonnées des points donnés, compte tenu des effectifs.

Exemple : Le centre de gravité de l'ensemble des points formant le nuage dessiné à la page précédente est le point de coordonnées $(165{,}04;66{,}93)$.
En effet, nous avons vu que $\bar{x} = 165{,}04$; $\bar{y} = 66{,}93$.

2. La méthode consiste à :

- diviser le nuage en deux ensembles de points d'effectifs à peu près égaux. Le plus simple est d'effectuer cette division en tenant compte soit des abscisses, soit des ordonnées. Dans l'exemple choisi, l'effectif total est 1517. Les deux ensembles devront comprendre environ 750 éléments. On pourra donc procéder d'une des deux manières suivantes :

  1er ensemble : les 7 premières classes (suivant les abscisses – taille des élèves).
  Effectif : 830.
  2ème ensemble : les autres classes. Effectif : 687.

  1er ensemble : les 6 premières classes (suivant les ordonnées – poids des élèves).
  Effectif : 524.
  2ème ensemble : le reste. Effectif : 993.

- On recherche les centres de gravité de chacun de ces ensembles :
  1er ensemble : (161,72 ; 63,16) ; 2ème ensemble : (171,34 ; 74,07).
  1er ensemble : (159,43 ; 63,23) ; 2ème ensemble : (171,82 ; 71,41).

- Il ne reste plus maintenant qu'à écrire l'équation de la droite joignant les deux centres de gravité : cette droite est la droite de régression cherchée. Normalement, elle doit passer par le centre de gravité du nuage de points.

Rappelons la formule générale de l'équation de la droite passant par les points (x_0,y_0) et (x_1,y_1):

$$y - y_0 = \frac{y_1 - y_0}{x_1 - x_0}(x - x_0).$$

On a donc :

- $y - 63{,}16 = \frac{74{,}07 - 63{,}16}{171{,}34 - 161{,}72}(x - 161{,}72)$

ou : $y = 1{,}13x - 120{,}25$

- $y - 63{,}23 = \frac{71{,}41 - 63{,}23}{171{,}82 - 159{,}43}(x - 159{,}43).$

ou : $y = 0{,}66x - 42{,}03$

On trouve donc deux droites de régression suivant que l'on a créé les deux ensembles de points au départ des abscisses ou au départ des ordonnées.

Remarques

Vérifions si le centre de gravité $(\bar{x}, \bar{y})$ du nuage appartient effectivement aux droites de régression. Nous avons calculé plus haut que $\bar{x} = 165{,}04$ et $\bar{y} = 66{,}93$.

On obtient $1{,}13 \times 165{,}04 - 120{,}25 = 66{,}24$ et le point de coordonnées $(\bar{x}, \bar{y})$ appartient effectivement à cette droite de régression (compte tenu des approximations).

De même : $0{,}66 \times 165{,}04 - 42{,}03 = 66{,}90$ et la même conclusion s'impose.

G. Méthode des moindres carrés

Comme dans la méthode précédente, nous trouverons deux droites des moindres carrés. Nous nous contenterons de donner leurs équations sans les démontrer.

On a

$$y - \bar{y} = \frac{\mathrm{cov}(X,Y)}{\sigma^2(X)}(x - \bar{x})$$

et

$$x - \bar{x} = \frac{\mathrm{cov}(X,Y)}{\sigma^2(Y)}(y - \bar{y})$$

1. Nous avons vu que $cov(X,Y) = 42; \sigma(X) = 8{,}71; \sigma(Y) = 6{,}54; \bar{x} = 165{,}04; \bar{y} = 66{,}93$

Il en résulte que : $\frac{\mathrm{cov}(X,Y)}{\sigma^2(X)} = 0{,}55$ et $\frac{\mathrm{cov}(X,Y)}{\sigma^2(Y)} = 0{,}98$

Les équations des droites des moindres carrés s'écrivent respectivement :

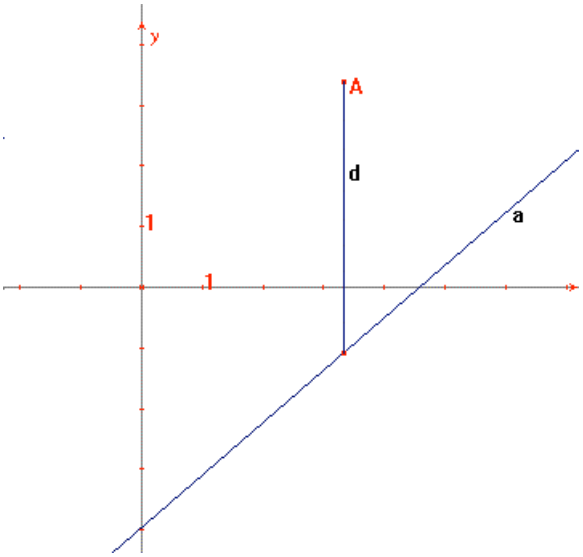
$$y = 0{,}55x - 23{,}84$$

et $x = 0{,}98y + 99{,}46$ ou $y = 1{,}02x - 101{,}49$

2. La **distance *d* d'un point *A* à une droite *a* parallèlement à** *OY* est la différence entre l'ordonnée de ce point et l'ordonnée du point d'intersection de cette droite et de la parallèle à *OY* passant par ce point.

Si :
- on recherche les distances parallèlement à *OY* de tous les points du nuage à une droite *d*,
- on élève chacune de ces distances au carré,
- on fait la somme de tous ces carrés, compte tenu des effectifs

on obtient un nombre égal à $1-r^2\sigma^2(Y)$.



Ce nombre est minimum quand *d* est la première droite de régression.

Telle est la signification de l'expression « moindres carrés ».

De même si l'on travaille parallèlement à *OX*, la somme obtenue est minimum quand *d* est la deuxième droite de régression.
Elle vaut alors $1-r^2\sigma^2(X)$.

H. Corrélation

1. Les coefficients $a=\dfrac{\text{cov}(X,Y)}{\sigma^2(X)}$ et $a'=\dfrac{\text{cov}(X,Y)}{\sigma^2(Y)}$ sont liés par la relation $aa'=r^2$.

Il en résulte que les droites des moindres carrés sont toutes deux croissantes ou toutes deux décroissantes. En effet *a* et *a'* sont leurs coefficients angulaires et sont de même signe.

Dans notre exemple, $aa'=0{,}55\times 1{,}02=0{,}56$ et $r=0{,}75$

2. Il est possible de démontrer que $-1\le r\le 1$. Dès lors, $0\le aa'\le 1$.

3. Si $r^2=1$, alors $aa'=1$ et les deux droites des moindres carrés ont même coefficient angulaire et sont confondues.
Les expressions $1-r^2\sigma^2(Y)$ et $1-r^2\sigma^2(X)$ sont nulles, ce qui a lieu sous la condition nécessaire et suffisante que toutes les distances considérées plus haut soient nulles, c'est-à-dire que tous les points soient alignés et appartiennent aux deux droites de régression confondues.

4. La corrélation est **forte** si et seulement si $0{,}87\le|r|\le 1$. Dans ce cas, le nuage est **très** allongé.