FORMATIONS AUX STATISTIQUES : MANUEL DU FORMATEUR DE NIVEAU 1

Janvier 1991

SCÉNARIO 5

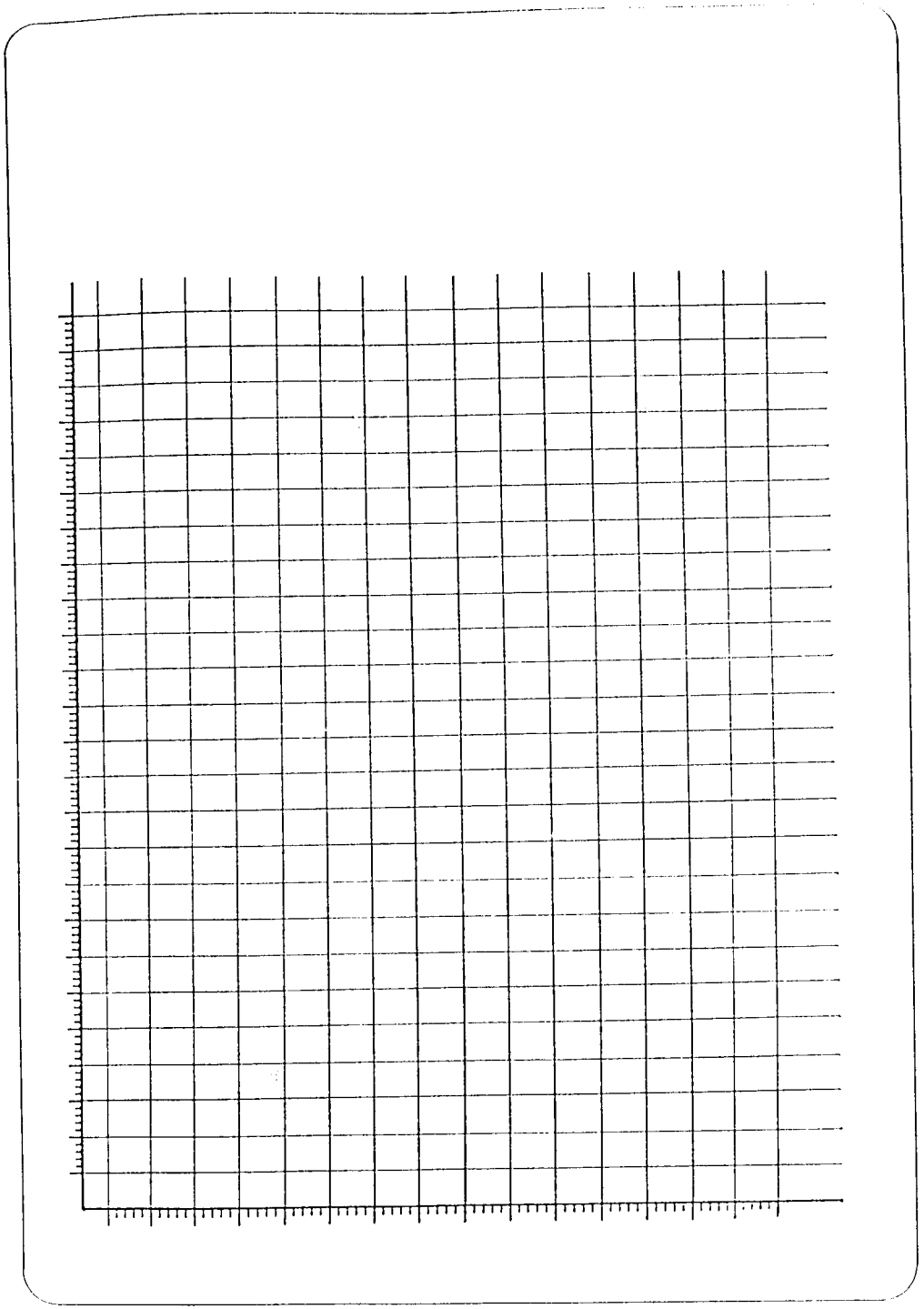
Liaison entre deux variables

INRA

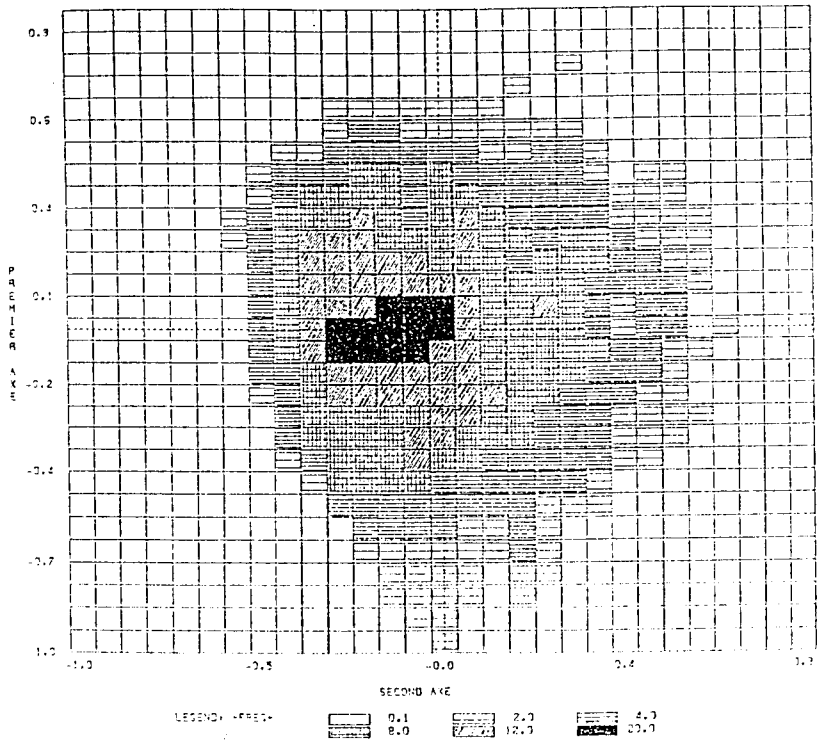
DOCUMENTS "FORMATEUR"

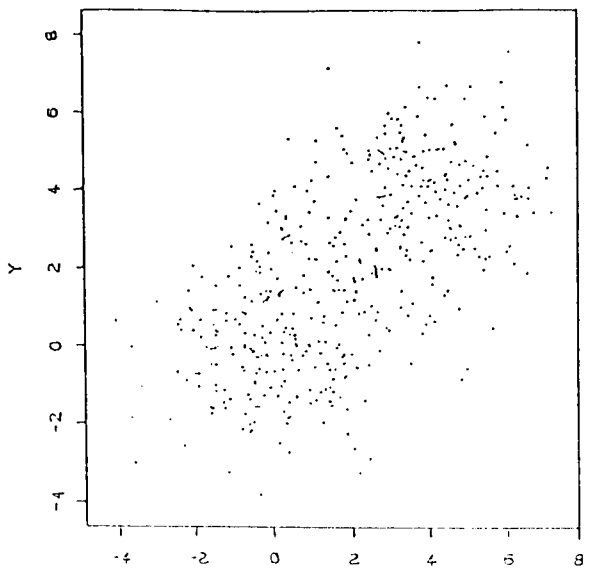
DÉPARTEMENT DE BIOMÉTRIE

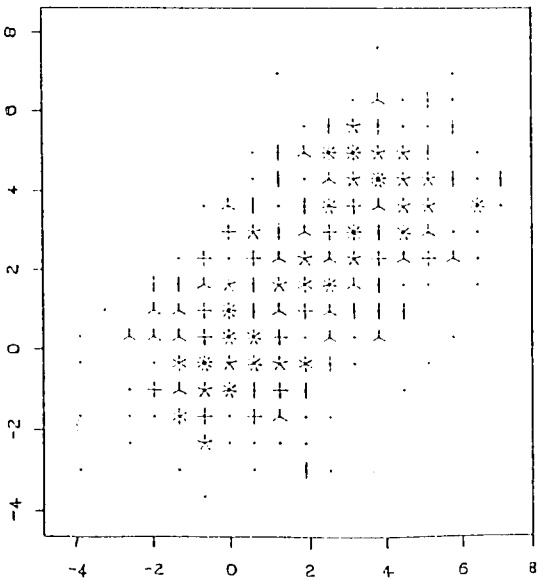
formation permanente fp



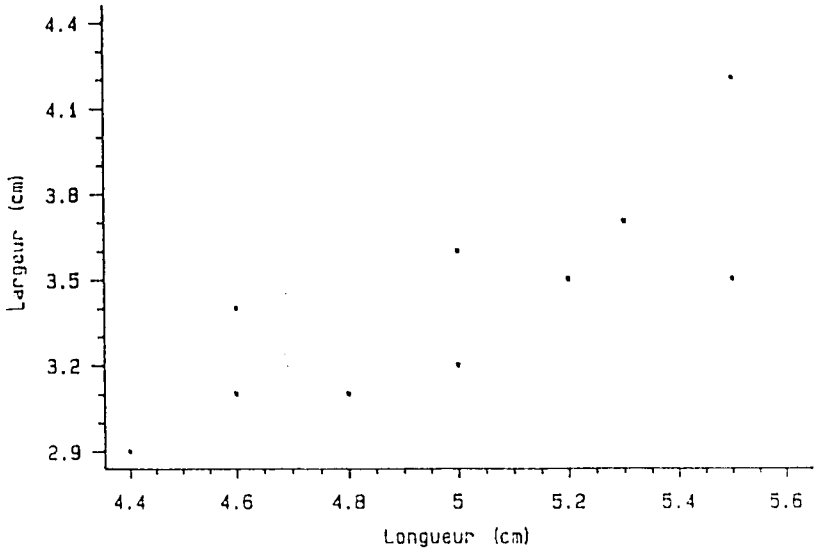
PLAGES D'EGALE DENSITE
NUAGE GLOBAL



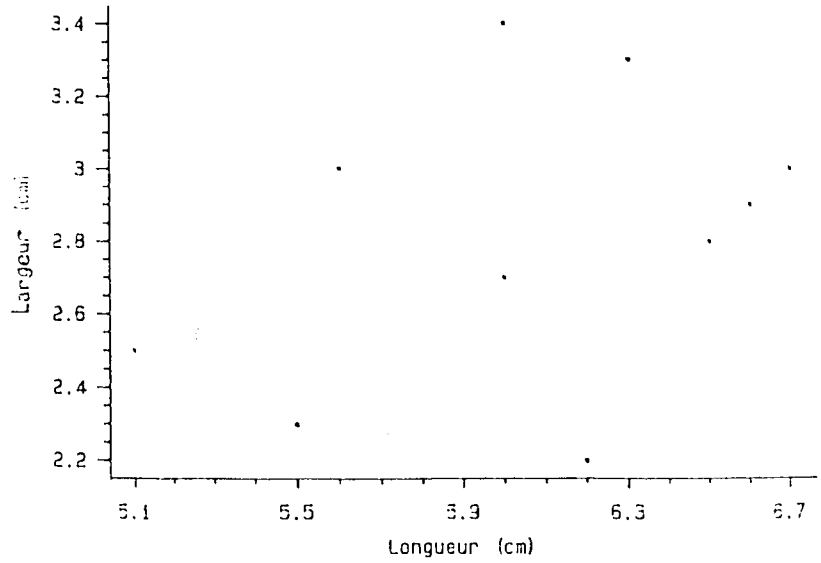

Y
8
6
4
2
0
-2
-4
-4
-2
0
2
4
6
8


8
6
4
2
0
-2
-4
-4
-2
0
2
4
6
8

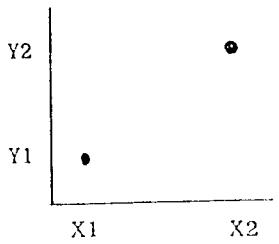
RELATION ENTRE LONGUEUR ET LARGEUR DES SEPALES CHEZ L'ESPECE Iris setosa



RELATION ENTRE LONGUEUR ET LARGEUR DES SEPALES CHEZ L'ESPECE Iris versicolor

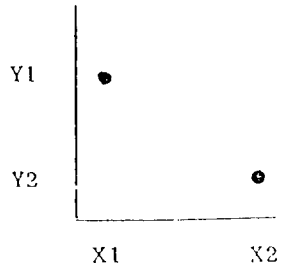


CONCORDANCE – DISCORDANCE



paire CONCORDANTE:

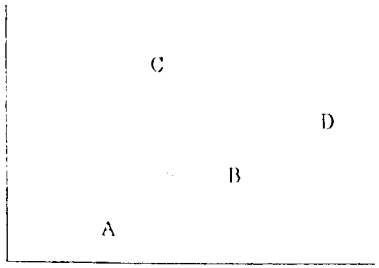
quand X augmente, alors Y augmente



paire DISCORDANTE:

quand X augmente, alors Y diminue

EXEMPLE



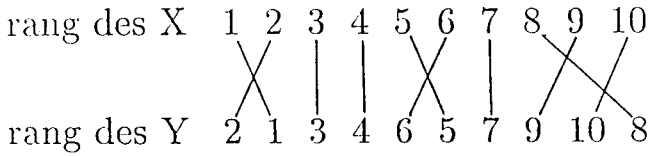
paires concordantes:
A/B
A/C
A/D
B/D

paires discordantes:
B/C
C/D

MÉTHODE GRAPHIQUE : POUR COMPTER LES DISCORDANCES

Les X sont rangées dans l'ordre croissant, on reporte en face les rangs des valeurs de Y correspondantes.

Exemple :



les rangs de X et de Y de même ordre sont reliés entre eux.

Nombre d'intersections = nombre de paires discordantes

COEFFICIENT DE CORRÉLATION τ DE KENDALL

calcul :

S = (nb de paires concordantes) − (nb de paires discordantes)

S est maximum si toutes les paires sont concordantes ; et avec n points : $S_{max} = n(n-1)/2$

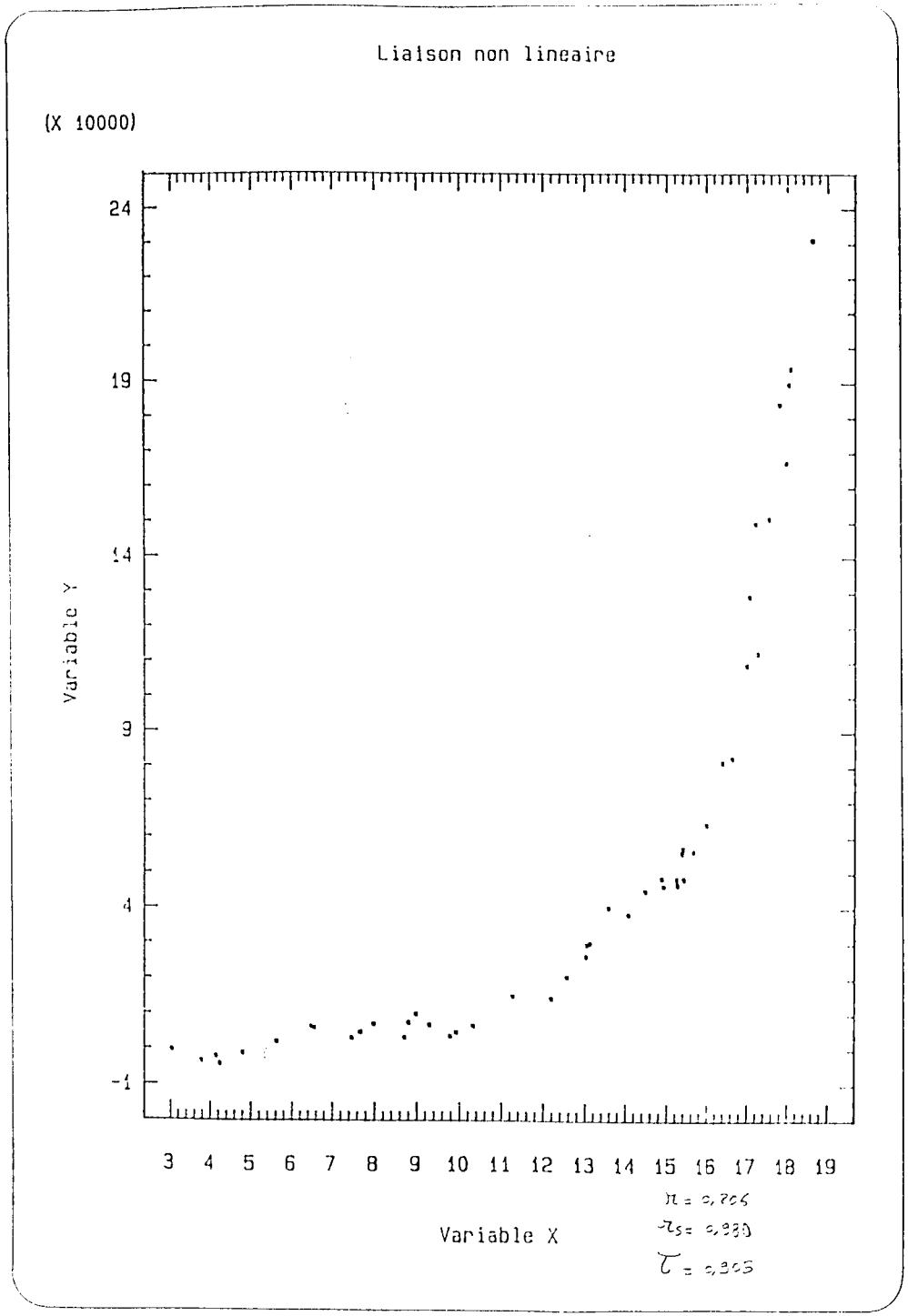
Le coefficient τ est donné par : $\tau = S/S_{max}$

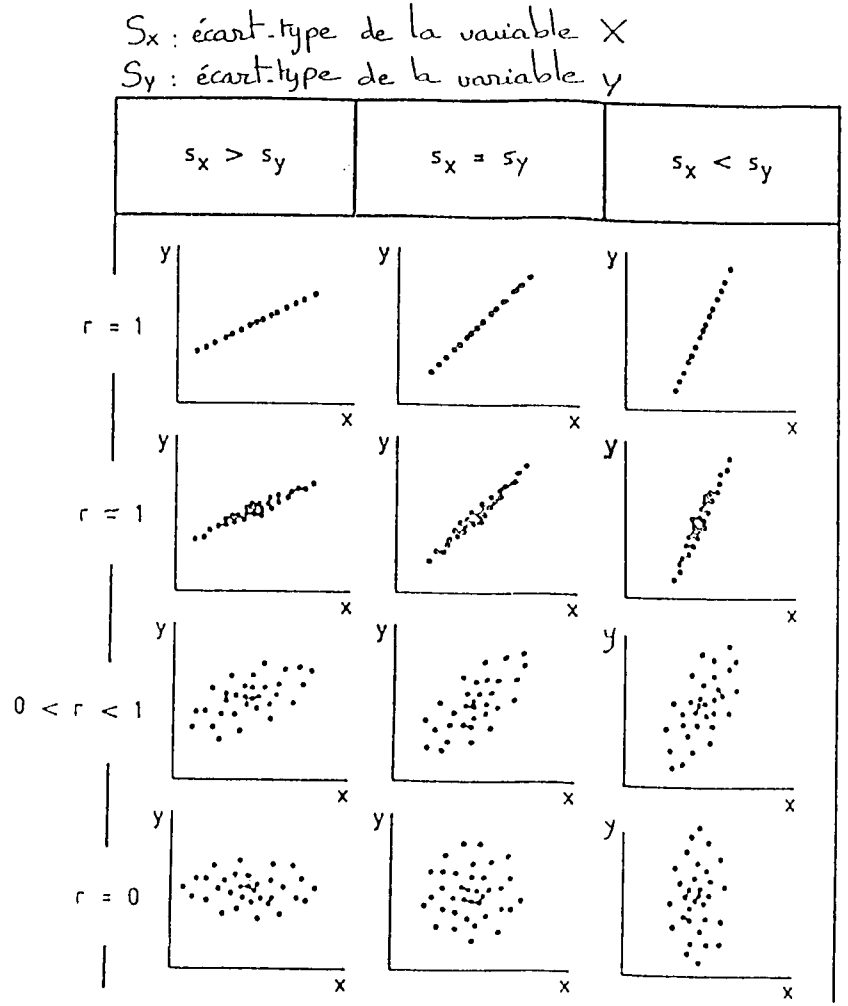
En récapitulant :

nb total de paires = $n(n-1)/2$
nb de paires discordantes = nb d'intersections
nb de paires concordantes = nb total de paires - nb paires discordantes

Sur l'exemple :

nb total de paires = (10x9)/2 = 45
nb paires discordantes = 4
nb paires concordantes = 45 - 4 = 41
$S = 41 - 4 = 37$
$S_{max} = 45$
donc $\tau = S/S_{max} = 37/45 = 0.822$

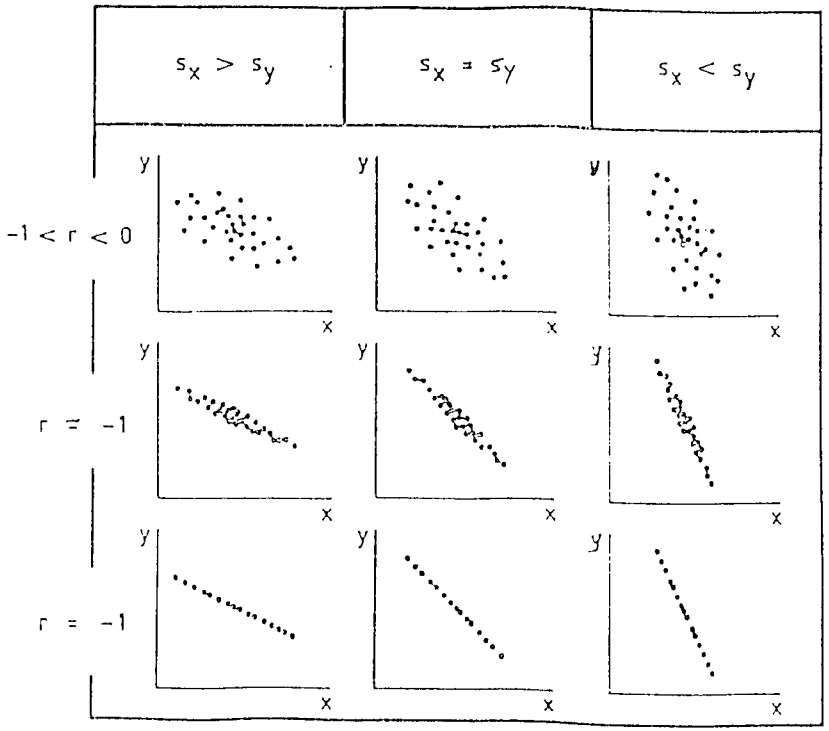

Liaison non lineaire
(X 10000)
24
19
14
9
4
-1
Variable Y
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Variable X
r = 0,706
rs = 0,980
τ = 0,905



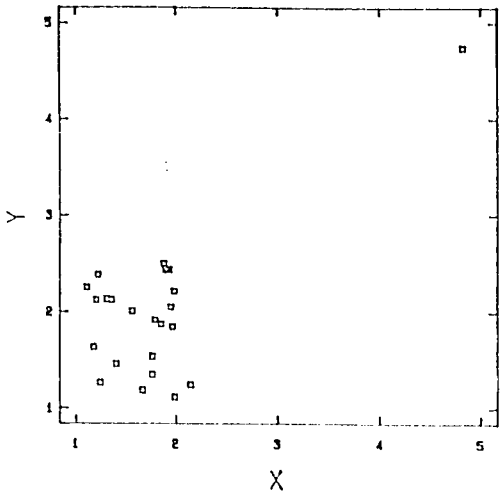
transparent 8 :

Quelques formes typiques de diagrammes de dispersion en relation avec les valeurs des coefficients de corrélation et des écarts-types (S_x, S_y)

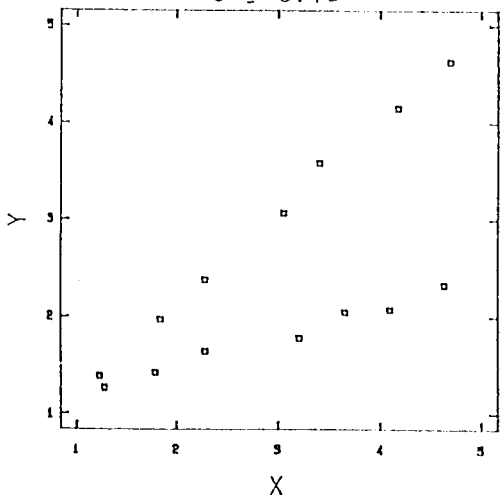
S_x : écart-type de la variable X
S_y : écart-type de la variable y



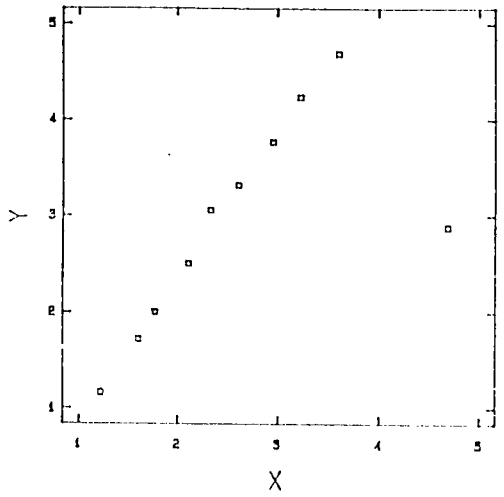
$\tau = 0.04$
$\tau_s = 0.02$
$r = 0.70$



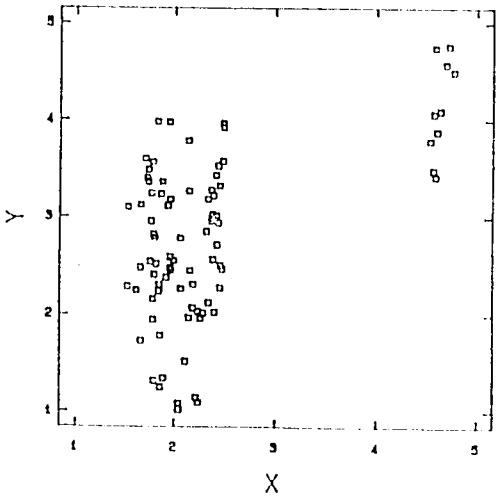
$\tau = 0.66$
$\tau_s = 0.78$
$r = 0.70$



$\tau = 0.78$
$\tau_S = 0.82$
$r = 0.70$



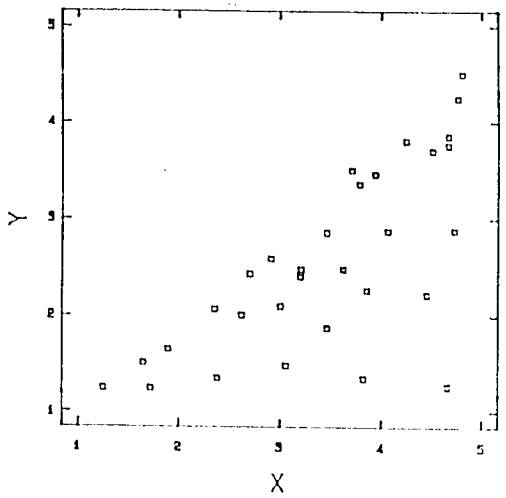
$\tau = 0.32$
$\tau_S = 0.45$
$r = 0.70$



$\tau = 0.57$

$r_S = 0.71$

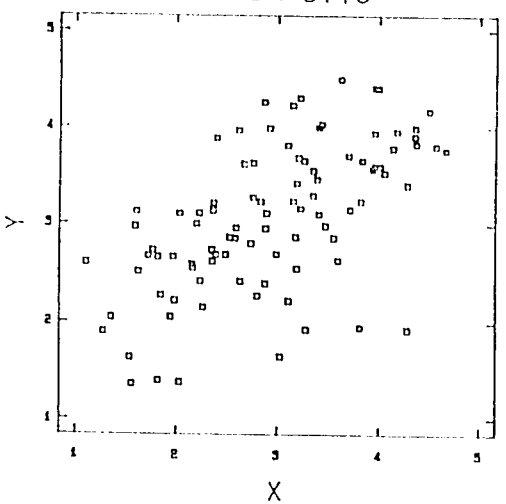
$r = 0.70$



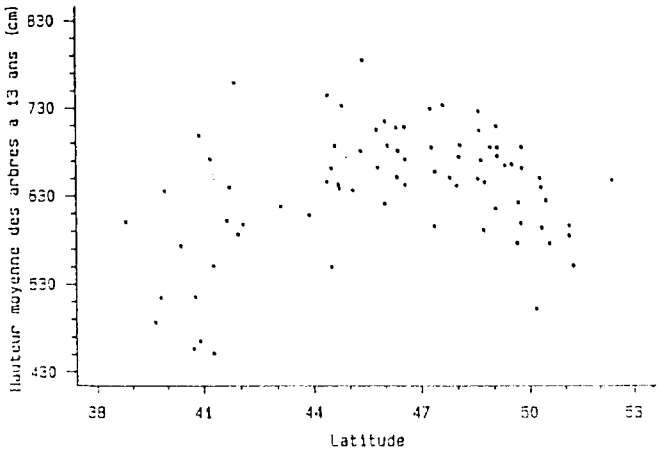
$\tau = 0.51$

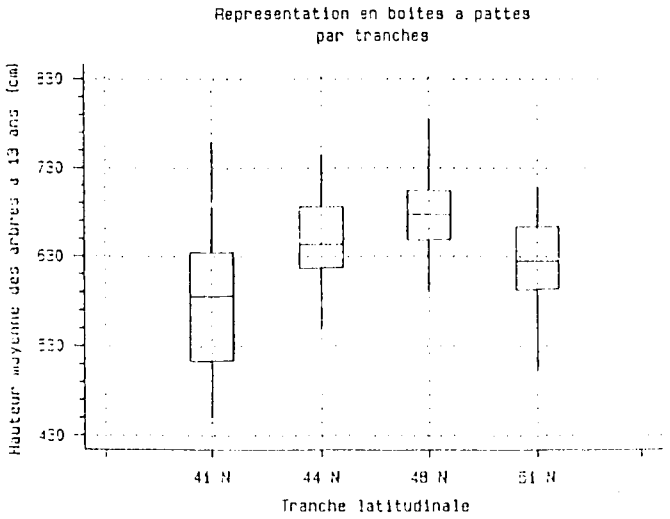
$r_S = 0.69$

$r = 0.70$

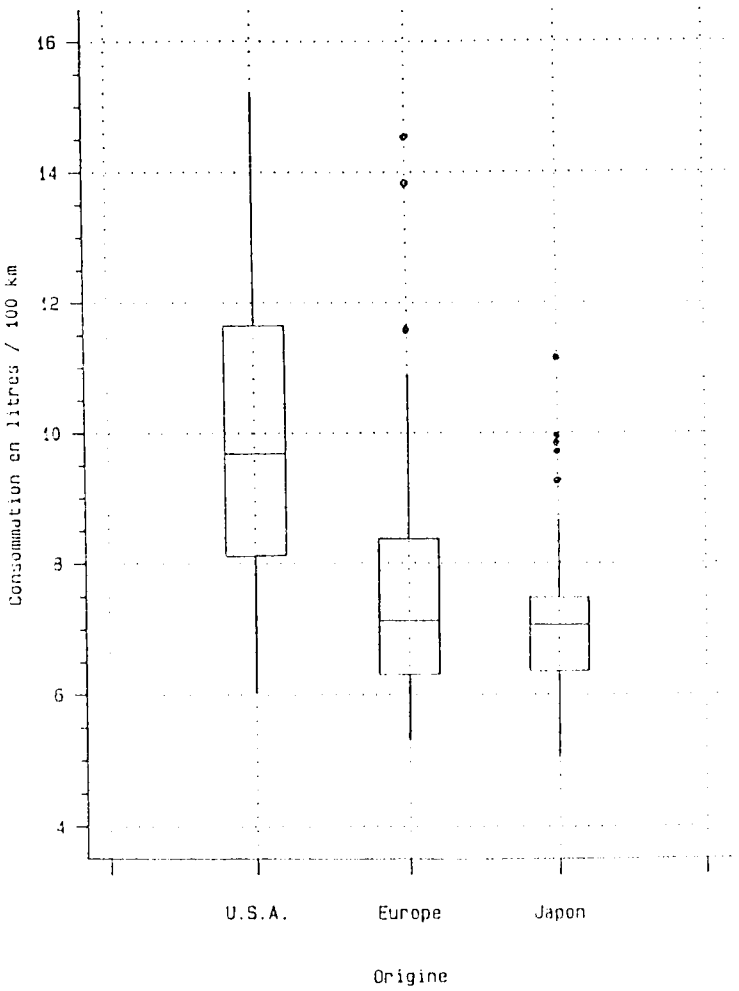


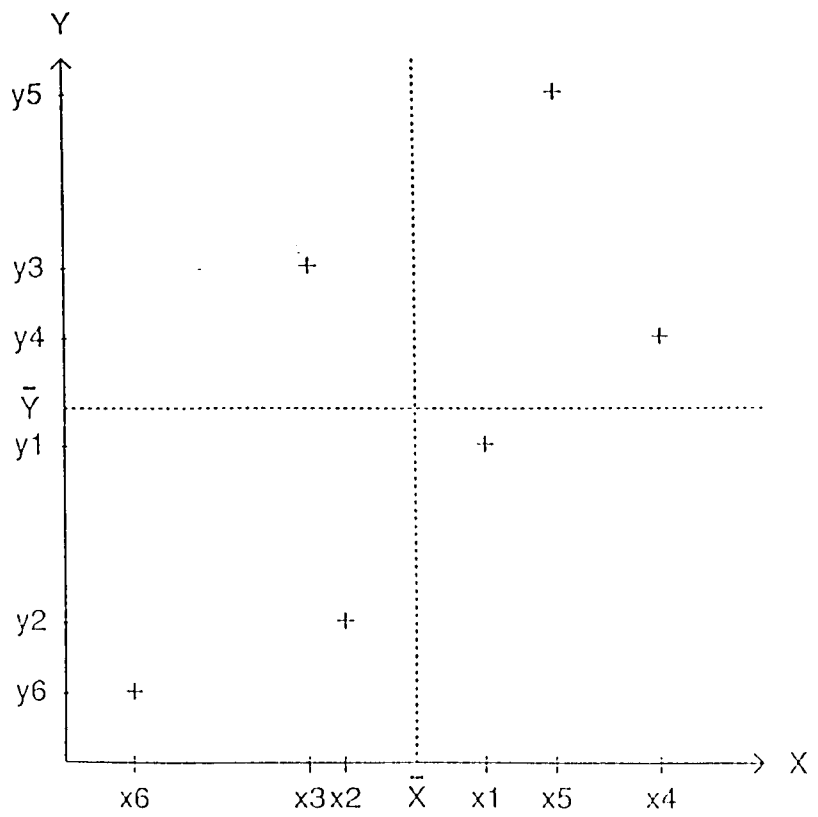
Liaison latitude d'origine - croissance
en hauteur de provenances de Douglas





Relation entre la consommation des voitures et leur pays d'origine





transparent 12 :

on dispose de 6 "individus", sur lesquels on a mesuré les valeurs de deux variables : X et Y ; on dispose donc de 6 couples de "points" :

(x_i, y_i) pour l'individu i, avec $1 \leq i \leq 6$

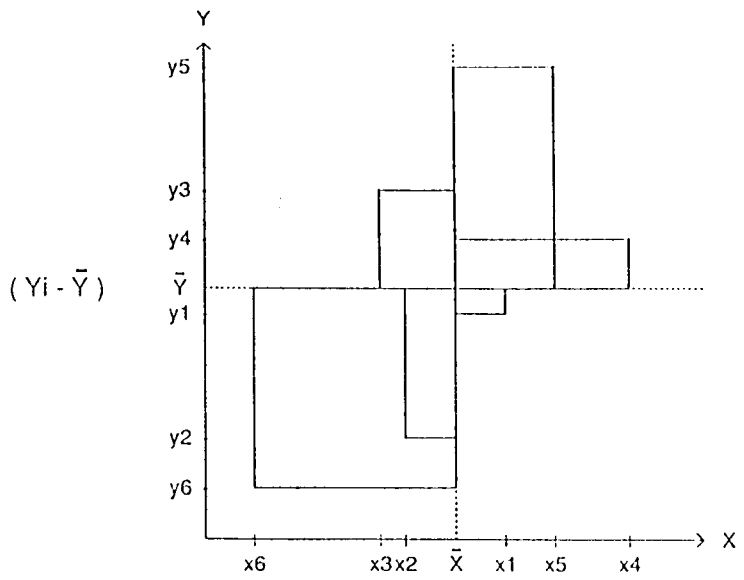
Une façon de les représenter est de le faire sur un graphe utilisant deux axes : l'un pour la variable X, et l'autre pour la variable Y. Un individu est alors représenté par un point.

L'ensemble des points, forme un "nuage de points", dont on peut calculer le "centre de gravité", qui a pour coordonnées : ("moyenne des X_i", "moyenne des Y_i").

Or on cherche une valeur qui permette de résumer la relation qui relie plus ou moins bien X et Y.

On peut constater que ce qui semble important est, la position de chaque point en fonction de ce centre de gravité. On va donc faire intervenir les valeurs :

(X_i - moyenne des X_i) et (Y_i - moyenne des Y_i)



(Xi - X̄)

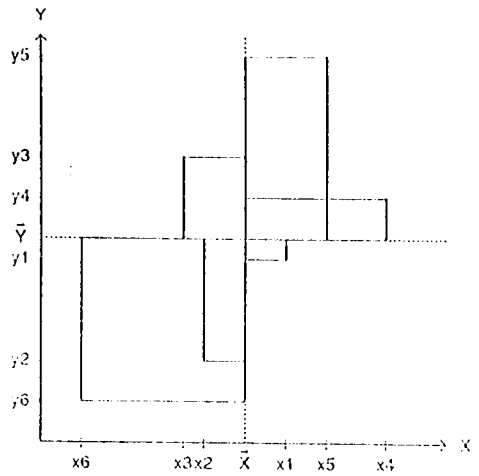
surface i = (Xi - X̄) . (Yi - Ȳ)

transparent 13 :

pour chaque point i, on peut donc définir une surface :

(X_i - moyenne des X_i) * (Y_i - moyenne des Y_i)

Sur le graphe, on obtient donc 6 surfaces.

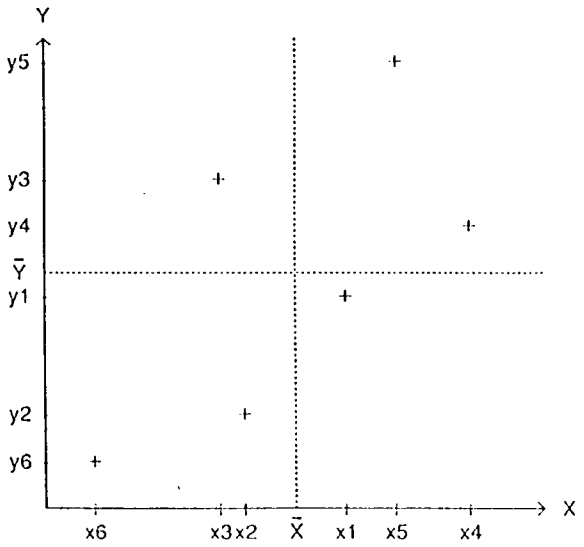


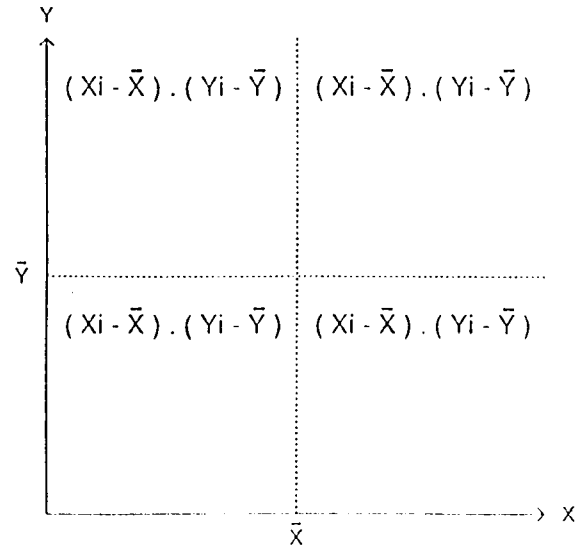
6

transparent 14 :

L'idée est ensuite d'obtenir une surface moyenne, soit donc de faire une "somme" de ces surfaces, divisée par le nombre de ces surfaces.

Peut-on additionner directement toutes ces surfaces?





transparent 15 :

On revient au nuage de points initial. On a besoin de calculer les produits indiqués dans les différents quadrants.
Ces produits vont-ils avoir des signes différents?
On remplit avec les stagiaires les signes de chacun des éléments des produits, qui interviennent, soient :

- pour les x_i tel que $x_i <$ moyenne des x_i, on a :

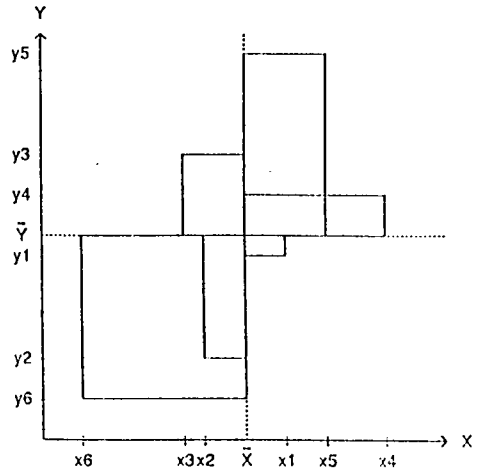
$$(x_i - \text{ moyenne des } X_i) < 0$$

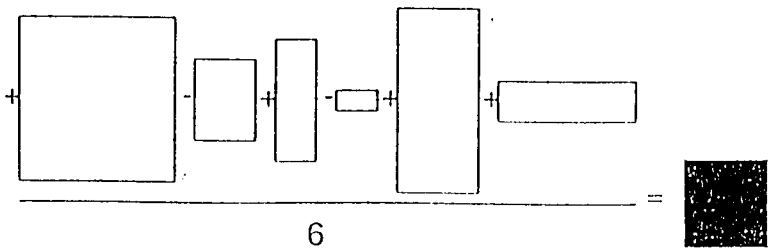
donc pour les x_i tel que $x_i >$ moyenne des X_i, ce produit est > 0;
On indique le signe de chaque (x_i moyenne des Xi) dans chaque quadrant.

- pour les y_i, on fait la même chose.

Il suffit enfin de faire, pour chaque quadrant le produit correspondant.
On obtient :

Les surfaces doivent donc être comptées positivement ou négativement, suivant le quadrant dans lequel elles se trouvent.





transparent 16 :

on obtient donc les signes des surfaces, quand on calcule cette surface moyenne. (Reprendre chaque signe en faisant remarquer de quel quadrant provient la surface).

On a tracé la surface moyenne obtenue en noire; quel est son signe?

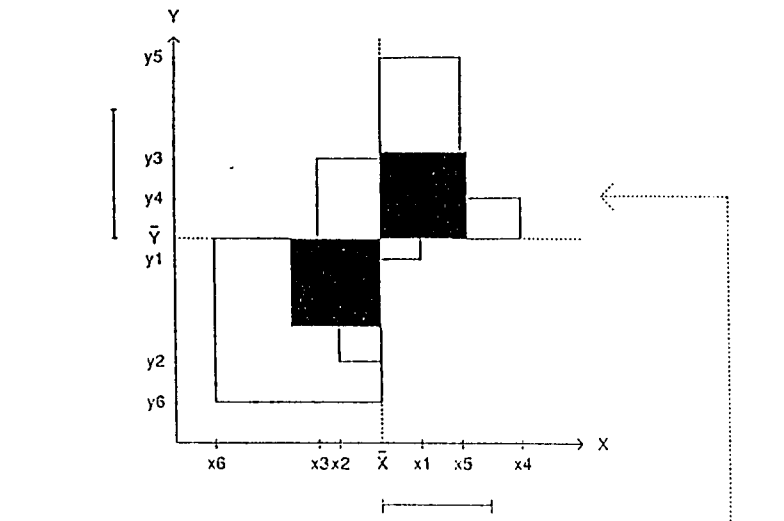
On voit que les surfaces moyennes positives sont plus importantes que les négatives; cette surface moyenne est donc positive; elle doit donc pouvoir être reportée sur le graphe supérieur, dans les quadrants positifs.

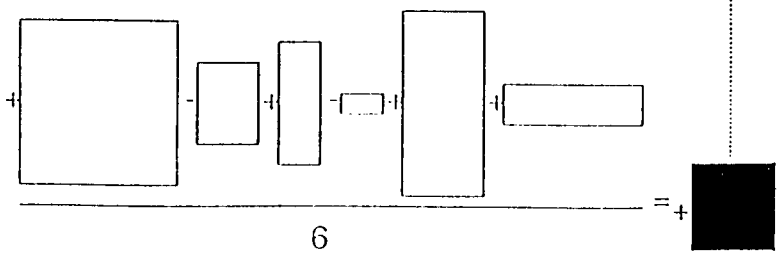
Cette surface moyenne est en faite appelée :

covariance de X et de Y ,

et nous avons vu sa formule :

$$cov(X,Y) = \frac{1}{n}.\sum_{i=1}^{n}(X_i - \overline{X}).(Y_i - \overline{Y})$$





6

transparent 17 :

On a donc reporté cette covariance sur le graphe de départ.

Et si on avait voulu calculer la covariance de X et de X?

Avec la formule, on arrive à :

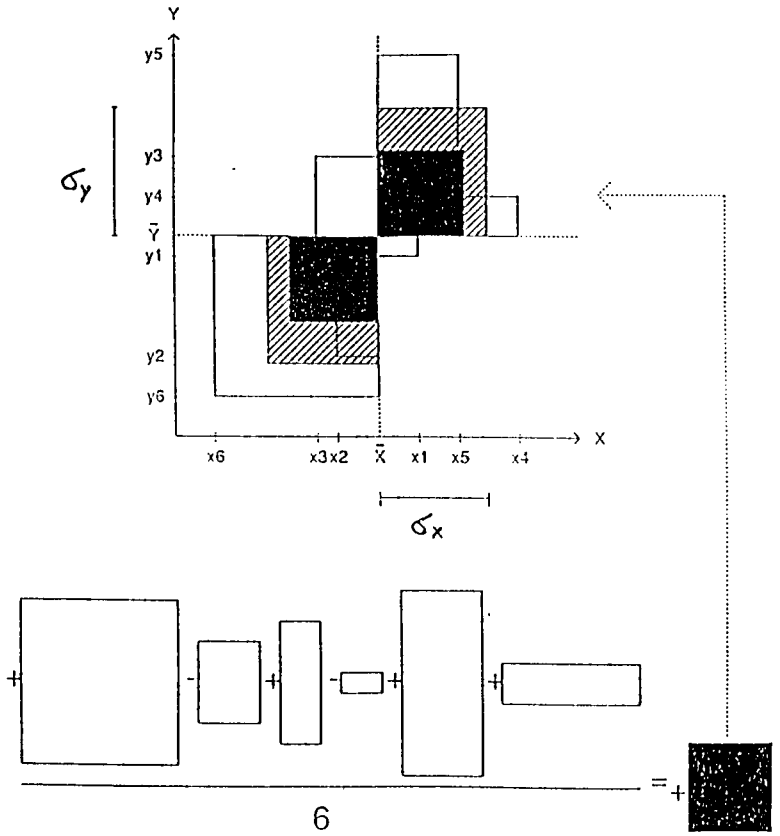
$$cov(X, Y) = var(X) = \text{(carré de l'écart-type de } X)$$

ce qui montre que la $cov(X, X)$ est une surface carrée, de côté égal à l'écart-type de X.

On peut donc tracer sur chaque axe X et Y, la valeur de leur écart-type; ce qui correspond aux deux segments présents près des axes. On constate alors que ceux-ci ne sont pas de la même grandeur; cela montre bien que la covariance que l'on a dessinée comme un carré, n'a pas de raison a priori de l'être (pas même échelle sur X et sur Y).

il serait par contre intéressant de considérer la surface obtenue avec :

(σ_x, σ_y) qui pourrait nous servir de surface de référence.



transparent 18 :

On a donc calculer cette nouvelle surface, et on l'a reportée sur la graphe. (surface hachurée)

En fait, elle joue un peu le rôle d'une surface de référence; il est donc peut être intéressant de comparer la covariance de X et de Y, à cette surface de référence?

On obtient alors le coefficient de corrélation linéaire, comme rapport des deux surfaces; avec les notions de corrélation positive si la cov. est positive, et de corrélation négative si la cov. est négative .

$$\rho = \frac{cov(X,Y)}{\sigma_x.\sigma_y}$$