PARTIE IV-A

1 INTRODUCTION

2 PRESENTATION DES DONNEES IRIS

3 PRESENTATION GRAPHIQUE

- Variance
- Covariance
- Coefficient de corrélation

4 PRESENTATION MATRICIELLE

- Variance-covariance
- Inertie

5 REPRESENTATION GEOMETRIQUE DE LA DISPERSION

- Une variable
- Deux variables
- Exemples

1 INTRODUCTION

A

Données iris

Variabilité

↓

Variance et Covariance

↓

Corrélation

↓

Inertie

↓

Ellipse d'inertie → Valeur et vecteur propre

B

C

Différenciation, maximisation

D

Direction de variabilité maximum

1 INTRODUCTION

Les notions développées ici sont illustrées par l'exemple de *R. A. Fisher* limité aux deux populations iris *setosa* et iris *versicolor*. Les points suivants sont exposés :

- Présentation des données *iris* et présentation graphique :
  - de la variance,
  - de la covariance,
  - du coefficient de corrélation.

  Notion d'inertie et d'ellipse d'inertie.
- Valeurs et vecteurs propres.
- Différenciation et maximisation.
- Recherche des directions de variabilité maximum.

2 PRESENTATION DES DONNEES IRIS

Exemple de R. A. Fisher (1936)
Kendall et Stuart, vol 3, pp : 317-322

On veut décrire le tableau :

- Des longueurs des sépales notées X
- Des largeurs des sépales notées Y

mesurées sur des iris appartenant aux variétés :

Versicolor	: $n_1 = 50$
Setosa	: $n_2 = 50$
Nb. observations	: $n = 100$

2 PRESENTATION DES DONNEES IRIS

Exemple de R. A. Fisher (1936)
Kendall et Stuart, vol 3, pp : 317-322

L'exemple des *iris* de Fisher est utilisé pour illustrer l'idée que l'on peut, à partir des notions de moyenne, variance et covariance, décrire des observations en précisant des directions caractéristiques (*cf.* paragraphe 5.6).

2 PRESENTATION DES DONNEES IRIS

iris Versicolor				*iris Setosa*			
X	Y	X	Y	X	Y	X	Y
7.0	3.2	6.6	3.0	5.1	3.5	5.0	3.0
6.4	3.2	6.8	2.8	4.9	3.0	5.0	3.4
6.9	3.1	6.7	3.0	4.7	3.2	5.2	3.5
5.5	2.3	6.0	2.9	4.6	3.1	5.2	3.4
6.5	2.8	5.7	2.6	5.0	3.6	4.7	3.2
5.7	2.8	5.5	2.4	5.4	3.9	4.8	3.1
6.3	3.3	5.5	2.4	4.6	3.4	5.4	3.4
4.9	2.4	5.8	2.7	5.0	3.4	5.2	4.1
6.6	2.9	6.0	2.7	4.4	2.9	5.5	4.2
5.2	2.7	5.4	3.0	4.9	3.1	4.9	3.1
5.0	2.0	6.0	3.4	5.4	3.7	5.0	3.2
5.9	3.0	6.7	3.1	4.8	3.4	5.5	3.5
6.0	2.2	6.3	2.3	4.8	3.0	4.9	3.6

2 PRESENTATION DES DONNEES IRIS

- On a mesuré :
  - les longueurs des sépales notées X,
  - les largeurs des sépales notées Y,

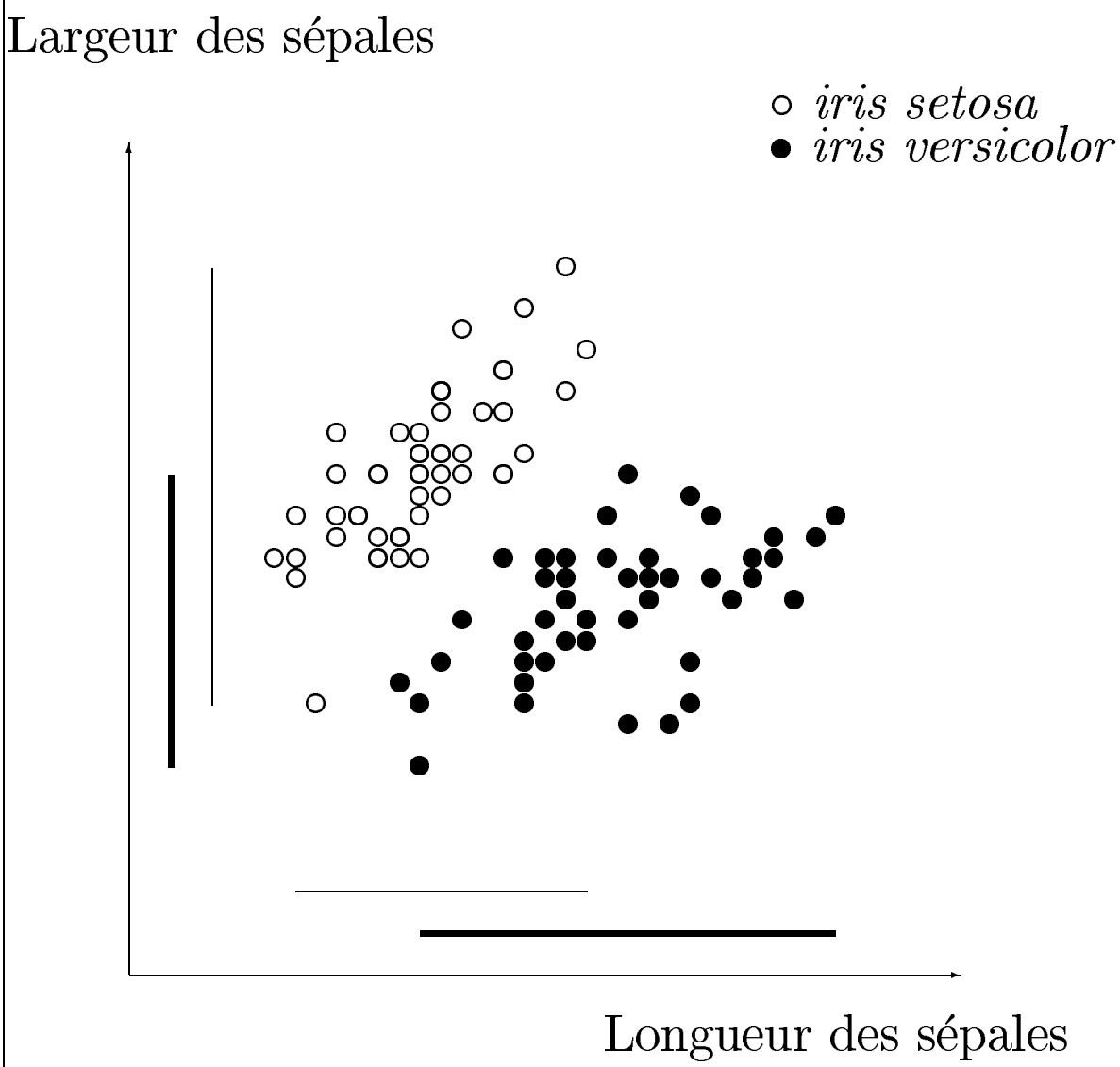
  sur $n_1 = 50$ iris *versicolor* et $n_2 = 50$ iris *setosa*.
- On veut :
  - Décrire ces deux variétés d'iris en estimant des paramètres de position et de dispersion.
  - Trouver une direction de l'espace de représentation qui les différencie au mieux.
- On représente les données dans le plan des X et des Y.

2 PRESENTATION DES DONNEES IRIS
(suite)

iris Versicolor				*iris Setosa*			
X	Y	X	Y	X	Y	X	Y
6.1	2.9	5.6	3.0	4.3	3.0	4.4	3.0
5.6	2.9	5.5	2.5	5.8	4.0	5.1	3.4
6.7	3.1	5.5	2.6	5.7	4.4	5.0	3.5
5.6	3.0	6.1	3.0	5.4	3.9	4.5	2.3
5.8	2.7	5.8	2.6	5.1	3.5	4.4	3.2
6.2	2.2	5.0	2.3	5.7	3.8	5.0	3.5
5.6	2.5	5.6	2.7	5.1	3.8	5.1	3.8
5.9	3.2	5.7	3.0	5.4	3.4	4.8	3.0
6.1	2.8	5.7	2.9	5.1	3.7	5.1	3.8
6.3	2.5	6.2	2.9	4.6	3.6	4.6	3.2
6.1	2.8	5.1	2.5	5.1	3.3	5.3	3.7
6.4	2.9	5.7	2.8	4.8	3.4	5.0	3.3

2 PRESENTATION DES DONNEES IRIS

(représentation du nuage de points)



2 PRESENTATION DES DONNEES IRIS

(représentation du nuage de points)

- Dans le plan, les deux populations constituent deux ensembles distincts.
- En projection sur les axes X et Y les deux populations se recouvrent largement.

On se pose les questions :

- **Existe-t-il un axe du plan tel que, en projection sur cet axe, les deux populations sont séparées au mieux ?**
- **Si oui, comment atteindre cet objectif ?**

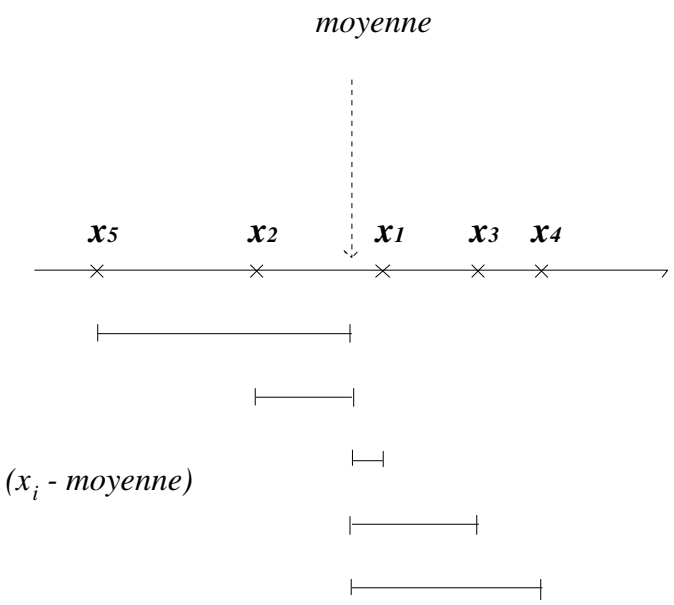
On résume chaque population par :

- Un paramètre de position.
- Des paramètres de dispersion.

On cherche des directions particulières :

- Pour une seule population afin de montrer sur un cas simple la démarche suivie.
- Lorsque l'on considère les deux populations conjointement.

Présentation graphique de la variance



Présentation graphique de la variance

- On dispose de 5 observations de la variable X.
- On représente les 5 valeurs des observations par 5 points sur un axe.
- On résume ces valeurs par un paramètre de position : **la moyenne.**

  La moyenne ne suffit pas pour résumer efficacement les observations, il faut en plus un paramètre de dispersion, mais pas n'importe lequel.

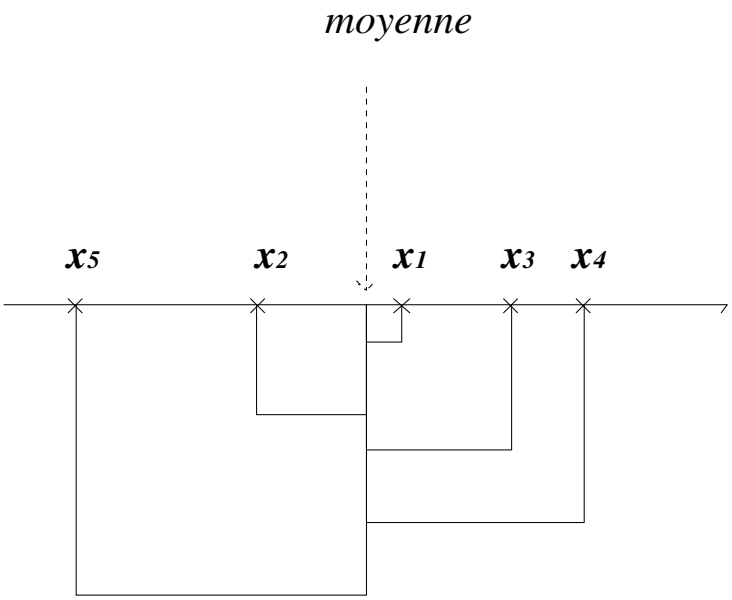
  - **Idée 1 :** utiliser les segments $(x_i - \bar{x})$.
    Ce n'est pas une bonne idée car :

$$\sum_{i=1}^{5} (x_i - \bar{x}) = 0$$

  - **Idée 2 :** utiliser les carrés des segments $(x_i - \bar{x})^2$. C'est une meilleure idée car :

$$\sum_{i=1}^{5} (x_i - \bar{x})^2 \neq 0$$

Présentation graphique de la variance

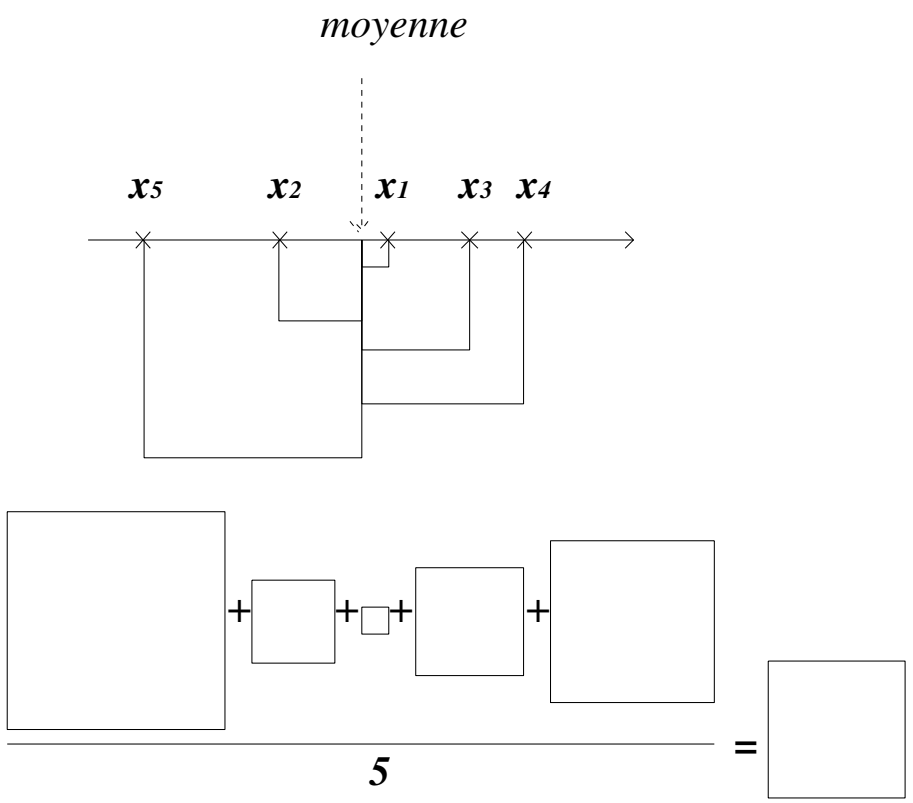


$(x_i - \text{moyenne})^2$

Présentation graphique de la variance

On peut facilement représenter graphiquement les carrés des segments $(x_i - \bar{x})^2$ par 5 carrés dont les côtés sont respectivement égaux à $x_i - \bar{x}$ pour $i = 1, \cdots, 5$.

Présentation graphique de la variance



Présentation graphique de la variance

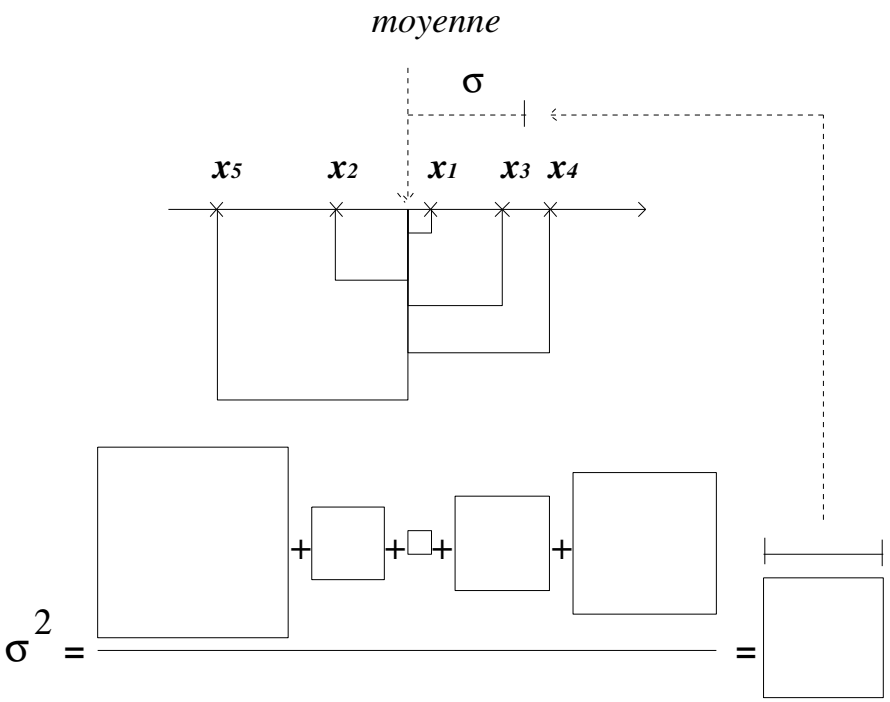
- **Idée 3 :** faire la somme des surfaces des 5 carrés puis, calculer une surface moyenne en divisant par le nombre de carrés.

- Cette surface moyenne est la **variance.**

- Cette surface n'a pas la même dimension que les données initiales (on ne peut pas la superposer sur l'axe des valeurs x_i).

  Comme la surface moyenne est représentée par un carré, il est bien évident que le côté de ce carré peut facilement être superposé sur l'axe des valeurs x_i d'où une représentation de la dispersion.

Présentation graphique de la variance



$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

Présentation graphique de la variance

- Le segment "côté du carré de la surface moyenne" est appelé l'**écart-type.**

- L'écart-type est donc la racine carrée de la variance.

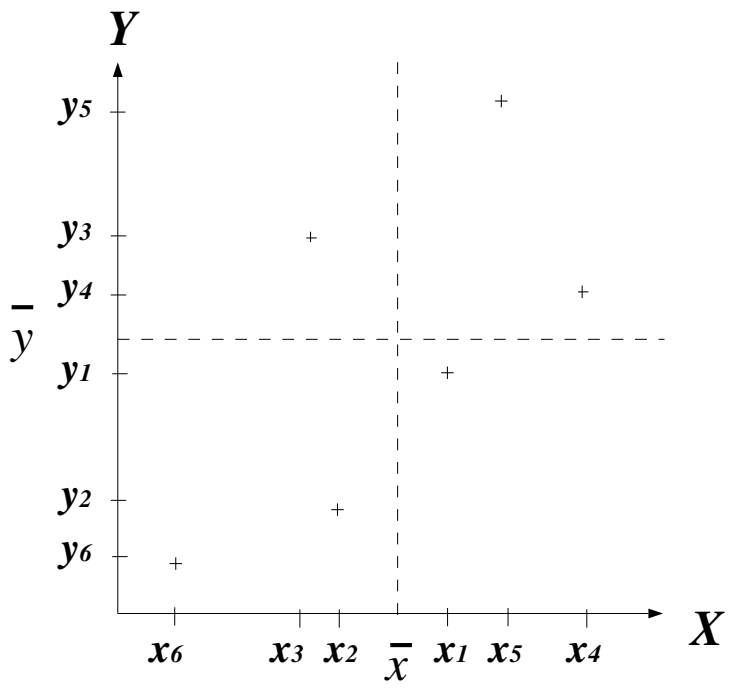
- L'écart-type est de la même dimension que les données initiales.

- **Résumé**

  – Variance : $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

  – Ecart-type : $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$

Présentation graphique de la covariance



Présentation graphique de la covariance

- On dispose de 6 observations des variables X et Y. On dit aussi 6 couples de points (x_i, y_i) $i = 1, \cdots, 6$.

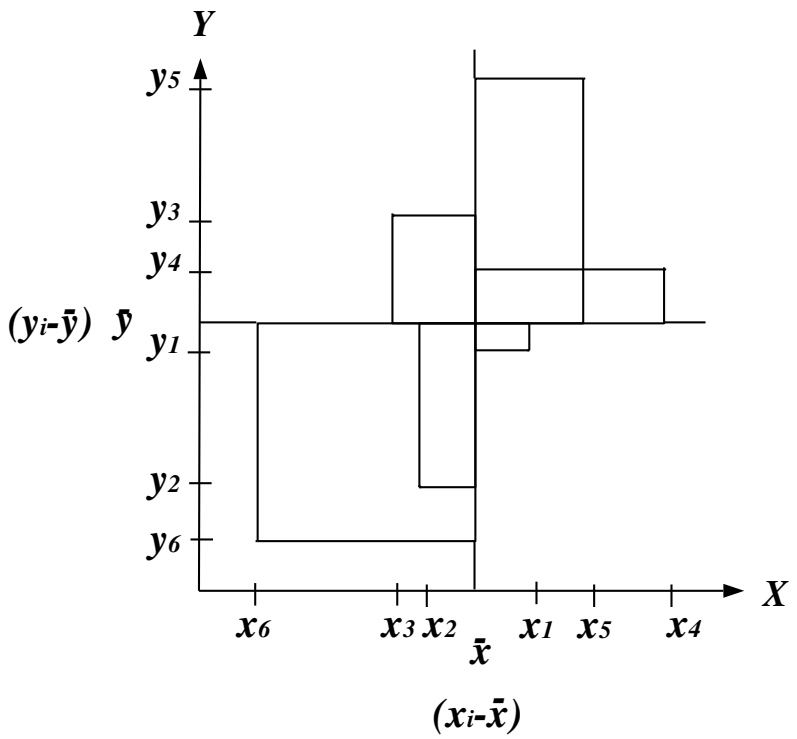
- On représente les 6 couples de points dans un système d'axes (X , Y). Une observation est alors représentée par un point dans le plan.

- L'ensemble des points forme un **nuage de points** dont on peut calculer le "centre de gravité". Le centre de gravité a pour coordonnées le couple : (moyenne des x_i , moyenne des y_i) = ($\bar{x}$, $\bar{y}$).

  On cherche une valeur qui résume la relation qui existe entre X et Y. Ce qui semble important, c'est la position de chaque point par rapport au centre de gravité.

  – **Idée 4 :** utiliser les valeurs $(x_i - \bar{x})(y_i - \bar{y})$

Présentation graphique de la covariance



$$\text{surface } i = (x_i - \overline{x}) \times (y_i - \overline{y})$$

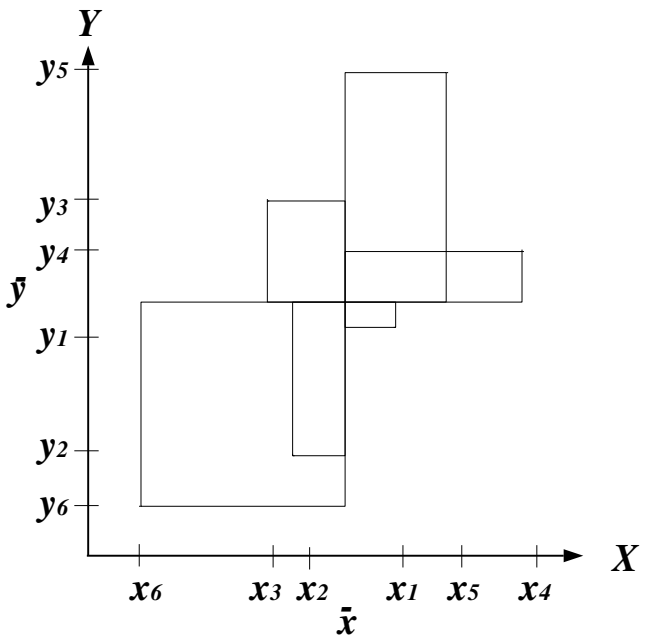
Présentation graphique de la covariance

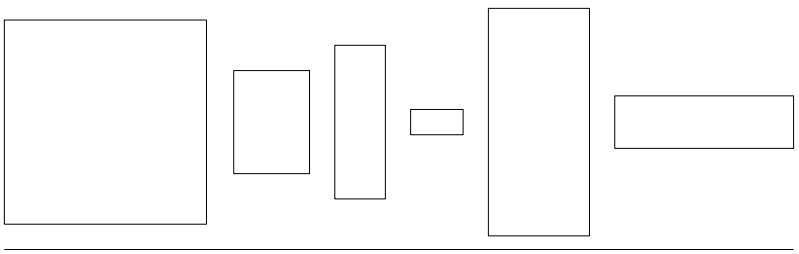
A chaque couple (x_i, y_i) on peut associer une surface i.

$$(x_i - \bar{x})(y_i - \bar{y})$$

On représente donc 6 surfaces sur le graphe.

Présentation graphique de la covariance



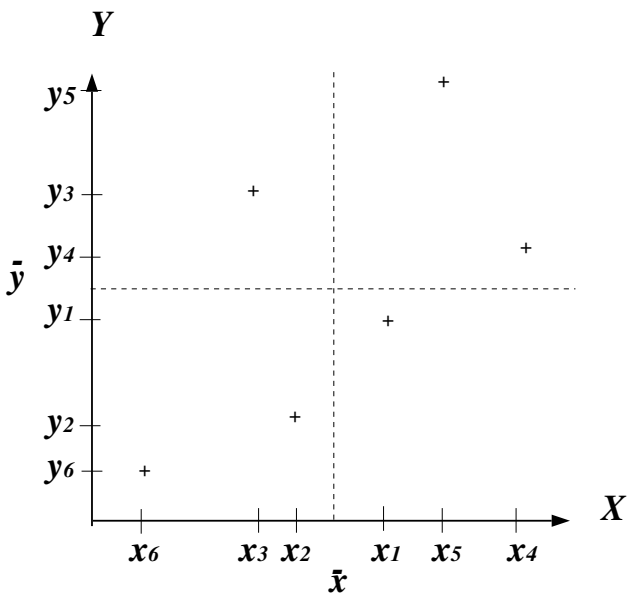


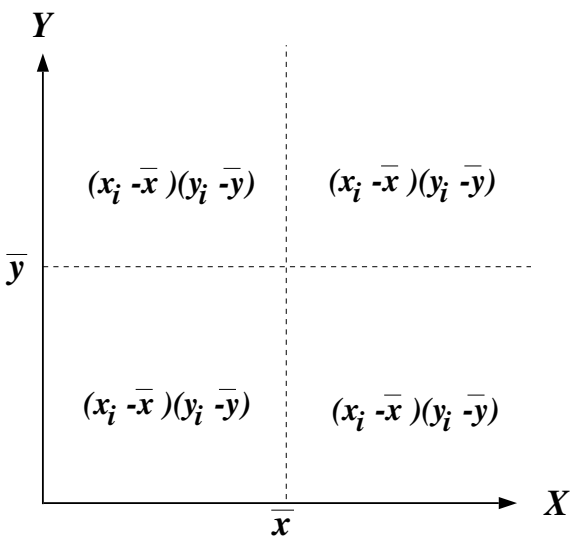
6

– **Idée 5 :** obtenir une surface moyenne c'est-à-dire faire une "somme" des surfaces que l'on divise par le nombre de surfaces

Peut-on additionner directement toutes ces surfaces ?

Présentation graphique de la covariance





Présentation graphique de la covariance

On considère le nuage de points initial et on calcule les produits indiqués dans les différents quadrants.

Ces produits ont-ils les mêmes signes ?

On détermine avec les stagiaires les signes de chacun des éléments du produit. Par exemple :

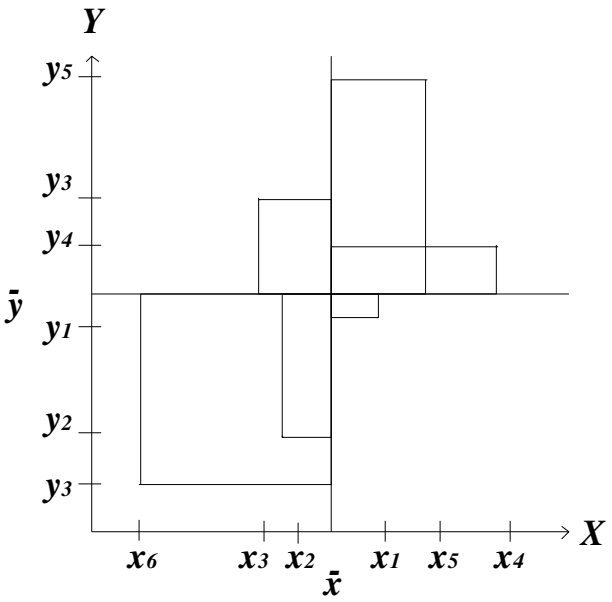
$x_i < \bar{x} \Rightarrow x_i - \bar{x} < 0,\ x_i > \bar{x} \Rightarrow x_i - \bar{x} > 0$

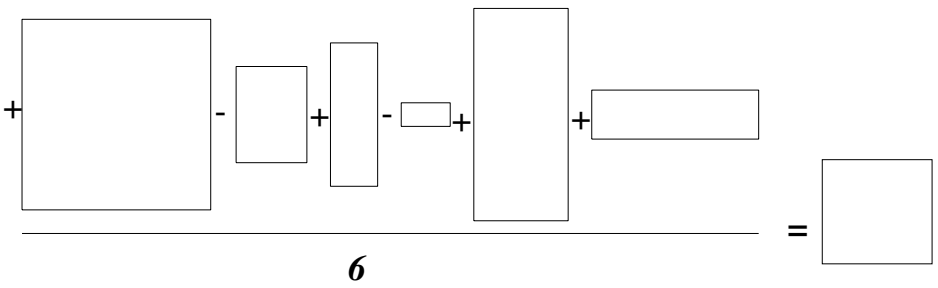
Pour chaque quadrant, on fait les produits et on obtient :

< 0	> 0
> 0	< 0

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

Présentation graphique de la covariance





Présentation graphique de la covariance

- On a obtenu les signes des surfaces et on calcule la surface moyenne en tenant compte du signe associé à chaque surface élémentaire.
- On a tracé la surface moyenne.

Quel est le signe de la surface moyenne ?

Les surfaces élémentaires positives sont plus importantes que les surfaces négatives $\Rightarrow$ la surface moyenne est positive, elle peut donc être reportée sur le graphe supérieur dans les quadrants positifs. Cette surface est appelée **covariance** de X et Y.

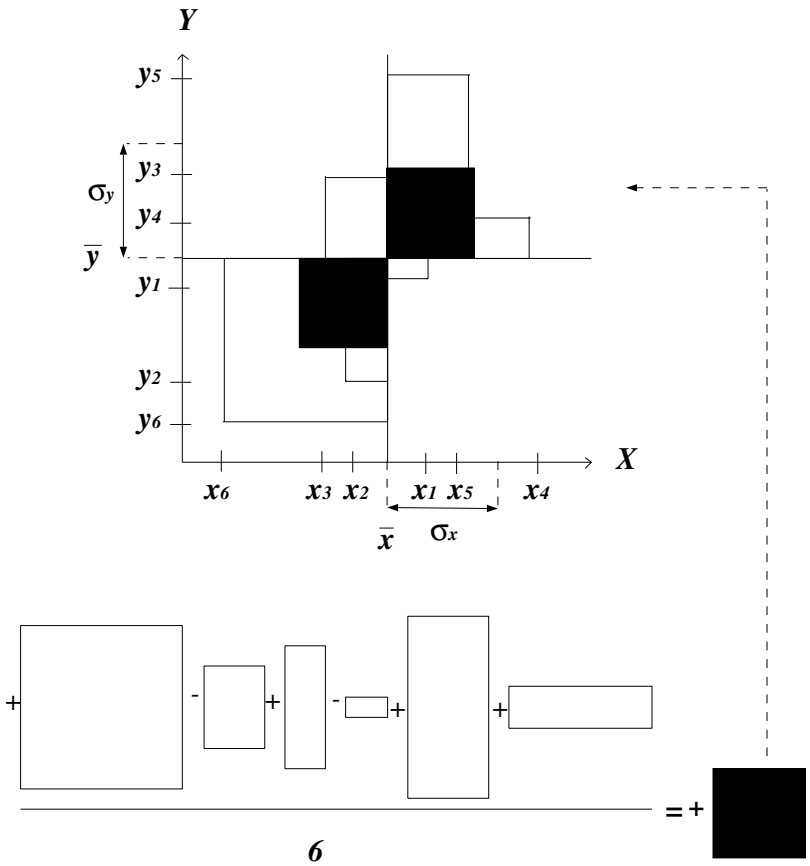
$$cov(X,Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Remarque : bien préciser qu'au produit :

$$(x_i - \bar{x})(y_i - \bar{y})$$

est associé une surface dont le signe est celui du produit.

Présentation graphique de la covariance



$$cov\,(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})$$

Présentation graphique de la covariance

- On reporte la covariance sur le graphe "en haut à droite" et "en bas à gauche" (de la même façon qu'on a reporté σ de part et d'autre de $\bar{x}$).

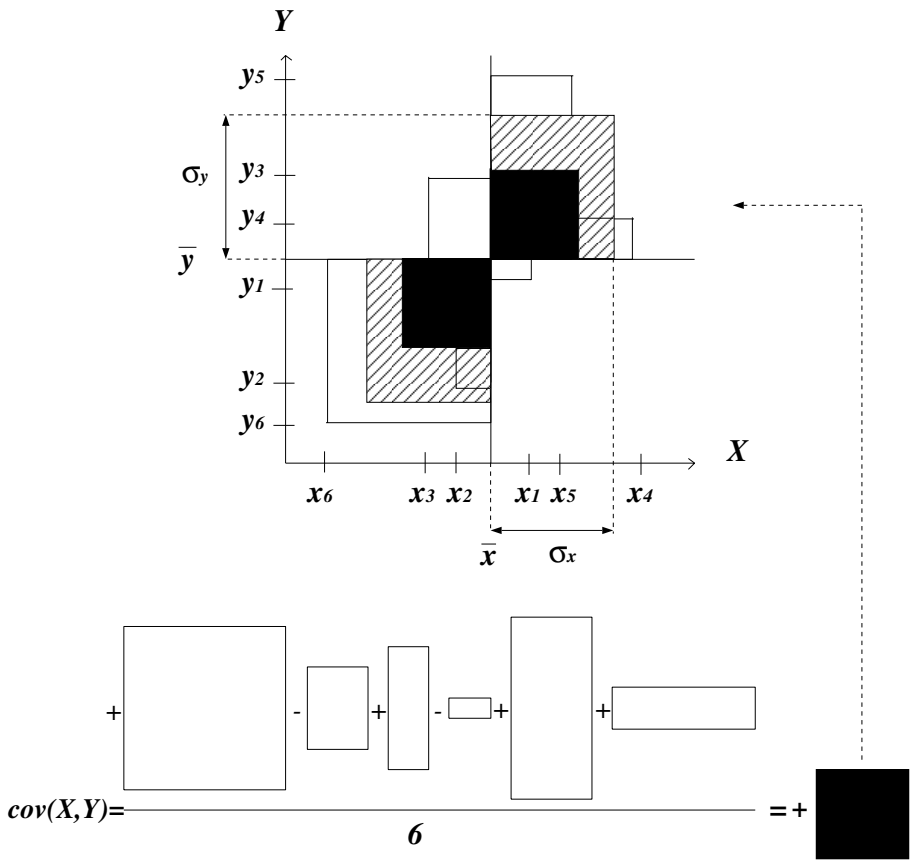
Qu'est-ce que la covariance de X et X ?

La formule de la covariance $\Rightarrow$ $cov(X, X) = var(X)$ = carré de l'écart type de X $\Rightarrow$ $cov(X, X)$ est la surface d'un carré de côté égal à l'écart-type de X.

- On trace sur les axes X et Y respectivement les segments de longueurs égales aux écart-types σ_X et σ_Y de X et Y. Ces segments ne sont pas de même grandeur, on n'a donc aucune raison de représenter la covariance par un carré.

  Par contre, ce qui peut être intéressant, c'est de représenter la surface associée à la covariance par rapport au produit $\sigma_X \sigma_Y$ considéré comme surface de référence.

Présentation graphique du coefficient de corrélation



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

Présentation graphique du coefficient de corrélation

La surface $\sigma_X \sigma_Y$ joue le rôle de surface de référence.

On veut comparer la surface associée à $cov(X, Y)$ à la surface $\sigma_X \sigma_Y$. Pour cela, on forme le rapport :

$$\rho = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

Ce rapport est appelé **coefficient de corrélation linéaire**.

- Il est positif si la covariance est positive.
- Il est négatif si la covariance est négative.

Présentation matricielle : variance-covariance

$$\begin{array}{ccc} & X & Y \\ X & cov(X,X) & cov(X,Y) \\ Y & cov(Y,X) & cov(Y,Y) \end{array}$$

$$\Updownarrow$$

$$\begin{array}{ccc} & X & Y \\ X & var(X) & cov(X,Y) \\ Y & cov(X,Y) & var(Y) \end{array}$$

$$\Sigma = \begin{bmatrix} var(X) & cov(X,Y) \\ cov(X,Y) & var(Y) \end{bmatrix}$$

Exemple : Population iris *setosa*
X : longueur des sépales
Y : largeur des sépales

$$\Sigma = \begin{bmatrix} 0.124 & 0.099 \\ 0.099 & 0.144 \end{bmatrix}$$

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

Présentation matricielle : variance-covariance

- $\Leftarrow$ Car :

  $cov(X, X) = var(X)$

  $cov(X, Y) = cov(Y, X)$

- $\Leftarrow$ Des "surfaces moyennes" ont été utilisées pour représenter graphiquement la notion de variance et de covariance. En fait, les estimateurs non biaisés de la variance et de la covariance sont :

  $$var(X) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

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

Ce sont ces estimateurs qui sont généralement utilisés.

Présentation matricielle : inertie

$$\Sigma = \begin{bmatrix} \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 & \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) \\ \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) & \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2 \end{bmatrix}$$

$$\Sigma = \frac{1}{n-1}\underbrace{\begin{bmatrix} \sum_{i=1}^{n}(x_i-\bar{x})^2 & \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\ \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) & \sum_{i=1}^{n}(y_i-\bar{y})^2 \end{bmatrix}}_{\text{matrice d'inertie}}$$

Exemple : Population d'iris *setosa*

$$\text{Matrice d'inertie} = \begin{bmatrix} 6.088 & 4.865 \\ 4.865 & 7.046 \end{bmatrix}$$

Présentation matricielle : inertie

Pourquoi utiliser la matrice d'inertie plutôt que la matrice de variance-covariance ?

Parce que les éléments de la matrice d'inertie sont des sommes de carrés qui s'additionnent simplement.