Statistiques appliquées à la gestion
Cours d'analyse de donnés
Master 1

F. SEYTE : Maître de conférences HDR en sciences économiques – Université de Montpellier I

M. TERRAZA : Professeur de sciences économiques – Université de Montpellier I

Présentation de l'analyse de données

« L'analyse des données a pour but de fournir grâce à l'ordinateur un outil permettant d'appréhender le contenu de tableaux de taille importante à l'aide de représentations accessibles par l'utilisateur », Edwin Diday.

L'analyse des données c'est aujourd'hui l'expression consacrée pour désigner les analyses statistiques descriptives multidimensionnelles

L'analyse des données rassemble un groupe de techniques aux fondements mathématiques qui permet d'appréhender la structure de l'information contenue dans un espace à plusieurs dimensions.

L'information, c'est la position relative des points dans l'espace multidimensionnel.

L'analyse des données est réalisable lorsqu'il est possible de réduire l'espace multidimensionnel (où l'information n'est pas lisible) en un espace à deux ou trois dimensions (où l'information est lisible), de telle sorte que cet espace réduit conserve une part importante de l'information qui était contenue dans l'espace multidimensionnel d'origine.

Les espaces multidimensionnels ont pour origine des tableaux statistiques de données de toute nature mais où les dimensions des lignes et des colonnes sont importantes. Ce sont ces lignes et ces colonnes qui constituent les dimensions des espaces et les points qui forment les nuages informationnels.

L'analyse des données est utilisée par la plupart des sciences appliquées : les psychologues, les juristes, les historiens, les économistes, les gestionnaires…

L'analyse des données a ses premiers développements mathématiques au début du siècle précédent (1905). Elle a cependant connu un essor sans précédent dans les années 70 et 80, grâce à l'amélioration des instruments de calcul et au développement de la micro-informatique.

Sous l'expression générique de l'analyse des données, on rassemble deux grandes techniques :

- **les analyses factorielles** : ces méthodes doivent leur nom aux nouveaux axes de l'espace que l'on peut réduire, qui portent le nom d'axes principaux, mais aussi de facteurs.
- Les techniques de **classification automatique** : ce sont des algorithmes informatiques automatiques capables de dresser des typologies, des regroupement de points, bref d'effectuer des classifications.

Ce sont les analyses factorielles qui font l'objet de ce cours.

I du tableau de données à l'analyse des données

Les analyses de données ont pour matière principale le **tableau de données**. De la nature de ce tableau dépend la nature des variables qui le composent. L'individu est un élément d'un ensemble fini que l'on appelle l'ensemble des individus. Ils sont portés en ligne du tableau. La description de ces individus est réalisée par des variables. Les variables sont portées en colonne du tableau.

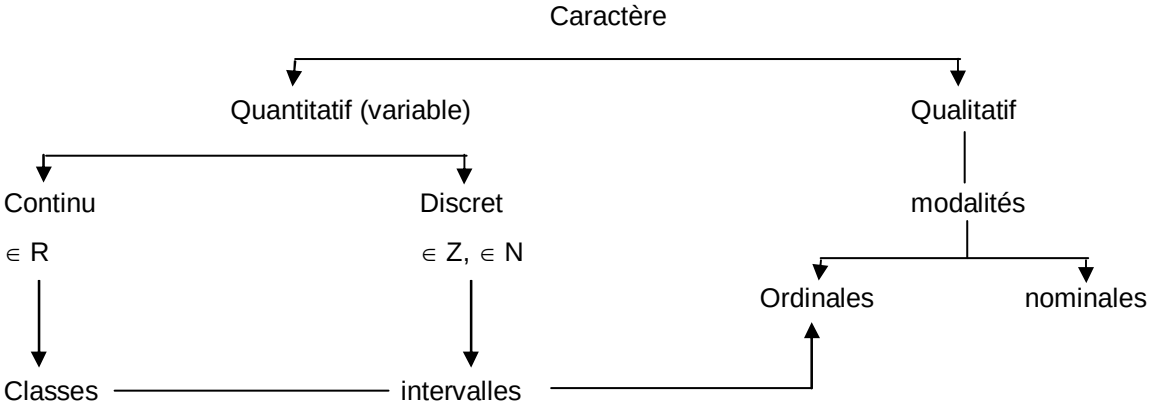
Une variable est définie par un ensemble que l'on appelle l'ensemble des observations (qui sont portées à l'intérieur du tableau) par une structure algébrique sur cet ensemble et par une application de l'ensemble des individus sur l'ensemble des observations.

Plus généralement, une variable est un caractère statistique particulier. On dénombre deux types de caractères : quantitatif et qualitatif.

Le caractère quantitatif est mesurable, c'est-à-dire qu'il prend ses valeurs dans des ensembles mathématiques comme par exemple l'ensemble des entiers naturels relatifs, réels… C'est ce caractère qu'on appelle variable.

Le caractère qualitatif est non mesurable. Il est qualifié par des modalités. On considère qu'il existe deux types de modalités :

- des modalités qu'on peut classer (ex : petit, moyen, grand),
- des modalités où le classement est indifférent (ex yeux bleus, verts…).

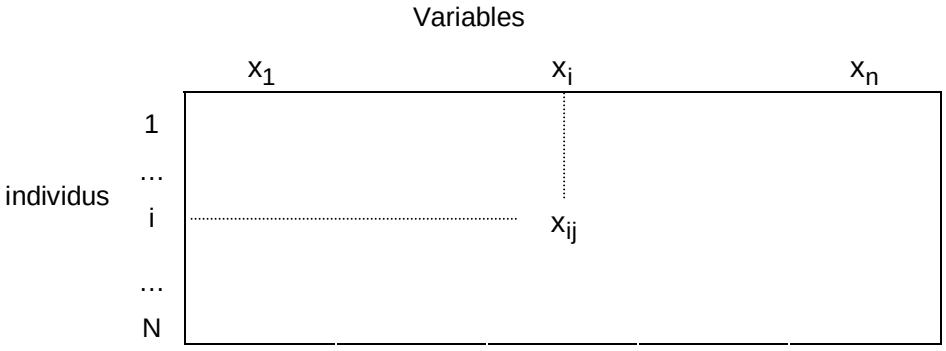


Fréquemment, les variables quantitatives sont transformées en classes (pour le cas continu) ou en intervalles (pour le cas discret). On considère alors que ces classes ou intervalles sont les modalités d'une variable qualitative ordinale. On constate alors que dans la plupart des tableaux, on ne dispose que d'un seul type de caractère : le caractère qualitatif, nominal ou ordinal. Dans la suite du cours, l'appellation caractère ne sera pas retenue. Comme dans la plupart des manuels, on retiendra le terme générique de variables.

Les différents caractères (variables) que l'on vient de définir permettent d'élaborer des tableaux différents et c'est cette différence qui, à son tour, définit les méthodes d'analyse de données.

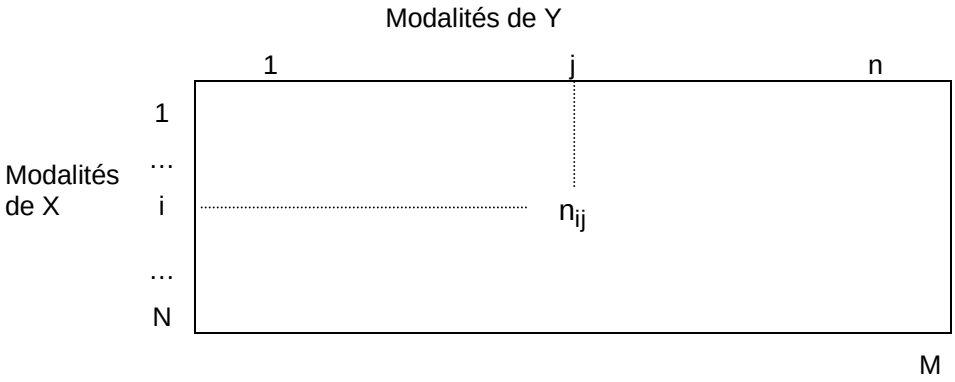
On considère dans la pratique quatre tableaux de données sur lesquels s'appliquent des méthodes d'analyses factorielles différentes.

- le tableau de variables (caractères) quantitatives :



La méthode d'analyse factorielle qui permet de traiter ce tableau porte le nom d'analyse en composantes principales : ACP.

- le tableau de contingence :

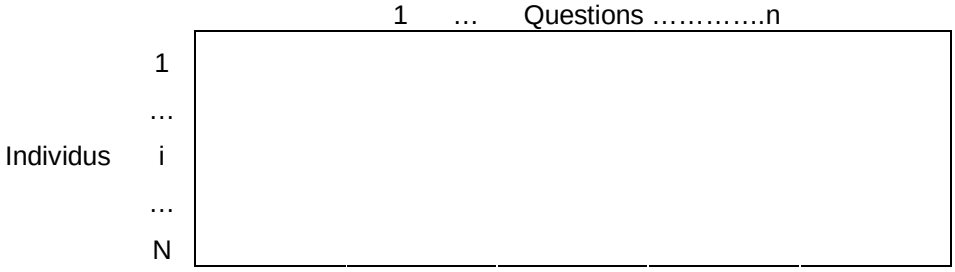


Les modalités doivent être exhaustives. (tous les M individus sont renseignés par les caractères X et Y) et exclusives les unes par rapport aux autres.

C'est la répartition de M individus selon les modalités des caractères X et Y.

La méthode d'analyse factorielle qui permet de traiter ce tableau porte le nom d'analyse factorielle des correspondances (AFC).

- les tableaux d'enquêtes ou de sondages :



Les réponses des N individus aux questions sont codées en affectant un chiffre aux diverses possibilités de réponses. Ces codes constituent pour une question le nombre de ses modalités. Le tableau codé est alors transformé en un tableau disjonctif complet. C'est-à-dire un tableau qui ne présente que des 0 ou des 1. Le chiffre 1 est donné à la modalité possédée par l'individu.

On applique à ce tableau une AFC et la méthode s'appelle analyse factorielle des correspondances multiples (AFCM)

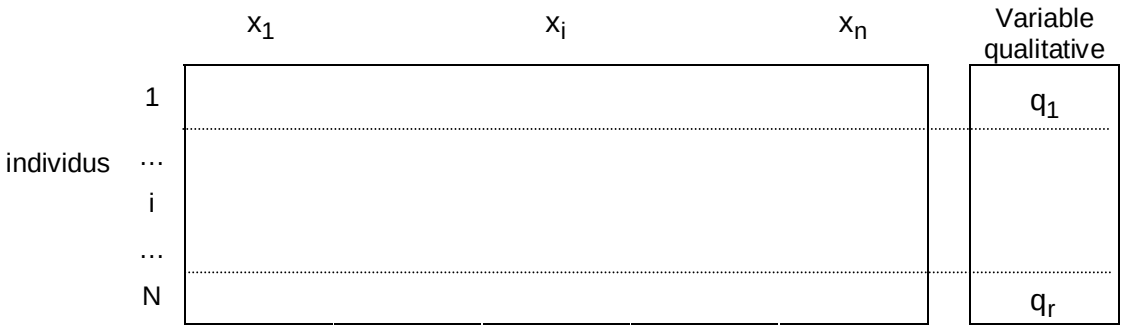
Exemple :

		Sexe	Nationnalité	Couleur Yeux
Individus	1	homme	Français	Bleu
	2	femme	Etranger	Marron
	3	femme	Etranger	Noir
	4	homme	Etranger	Bleu
	5	femme	Français	Marron
	6	homme	Français	Noir
	N	femme	Français	Bleu

Codification (pour la saisie des réponses)		
Sexe	Nationnalité	Couleur Yeux
1	1	1
2	2	2
2	2	3
1	2	1
2	1	2
1	1	3
2	2	1

Tableau disjonctif						
Homme	Femme	Français	Etranger	Yeux bleus	Marron	Noir
1	0	1	0	1	0	0
0	1	0	1	0	1	0
0	1	0	1	0	0	1
1	0	0	1	1	0	0
0	1	1	0	0	1	0
1	0	1	0	0	0	1
0	1	1	0	1	0	0

- les tableaux quantitatifs où les individus sont regroupés par paquet en fonction d'une variable qualitative :



La méthode factorielle appliquée à ce tableau porte le nom d'analyse factorielle discriminante (notée AFD)

Les calculs de l'analyse de données ne se font jamais à la main. Les logiciels pour l'utiliser sont très nombreux et l'on peut les segmenter selon plusieurs types :

- les logiciels de traitement d'enquête (Le Sphinx, ethnos, Question, …). Bien que leur spécialité soit le traitement de questionnaires, ils intègrent quelques méthodes d'analyses factorielles. Les sorties sont relativement sommaires et les options disponibles sont limitées (pas de rotation des axes, …)
- les logiciels boîtes à outils (XLSTAT, Statbox). Ils permettent de réaliser diverses analyses factorielles (ACP, AFC, ACM), quelques techniques de classification (Classification hiérarchique, K moyennes) ainsi que les techniques de prévision classiques. Les données sont gérées à partir du logiciel Microsoft Excel et les sorties s'effectuent dans des feuilles de calculs. Globalement, ils offrent un bon rapport qualité/prix
- Les logiciels de statistique (SPSS, SPAD, SAS, …). Conçus pour manipuler et analyser de grands tableaux de données, ils sont très complets sur le plan des méthodes présentes et sur les options disponibles. L'utilisation est plus complexe et nécessite parfois plusieurs journées (voire plusieurs mois) de formation. Leur prix en fait un outil réservé aux cabinets statistiques ou aux directions statistiques de grandes entreprises.

Dans ce cours, nous utiliserons les sorties du logiciel Statbox.

Les bases de l'analyse de données

Après avoir introduit les principes généraux de l'analyse de données, nous rappellerons ici certaines statistiques élémentaires qui forment les fondations de l'analyse des données.

Présentation des données et types de variables

Généralement, le problème à résoudre se présente sous forme de table contenant les observations (ou individus ou exemples) en ligne et les variables (ou attributs) en colonne.

Les différents types de variables vont conditionner le choix des techniques utilisées.

On distingue généralement :

Type de variables		Caractéristiques
Qualitatives	Disjonctives (ou dichotomiques)	Elles peuvent prendre deux états (exemple vrai ou faux)
	Catégoriques non ordonnées ou qualitatives non ordonnées	Les différentes catégories ne contiennent pas de notions d'ordre (exemple : couleur des yeux)
Quantitatives	Catégoriques ordonnées ou qualitatives ordonnées	Les différentes catégories peuvent être classées (ex classes d'âges, échelles de Lickert)
	Continues	Elles peuvent prendre des valeurs numériques sur lesquelles des calculs, tels que la moyenne peuvent être effectués.

La notion d'association

Les associations sont des critères permettant de regrouper des variables. Elles se mesurent différemment selon que l'on s'intéresse à des variables quantitatives ou qualitatives.

L'association sur des variables quantitatives

La corrélation linéaire

Elle mesure la covariation qui existe entre deux variables X et Y. Le coefficient de corrélation indique si deux variables évoluent dans le même sens ou en sens contraire.

Il est compris entre -1 (corrélation négative) et +1 (corrélation positive). Lorsqu'il est nul on dit que les variables ne sont pas corrélées.

Le coefficient de corrélation s'écrit : $r_{xy} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y}$ avec :

$\text{cov}(x,y) = \sum_{i=1}^{n} p_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} p_i x_i y_i - \bar{x}\,\bar{y}$ avec p_i poids de l'individu i et $\sum_{i=1}^{n} p_i = 1$, n le nombre d'observations.

> *En général,* $\forall i \{1,....n\}\ p_i = \frac{1}{n}$. *C'est le cas le plus classique, tous les individus ont le même poids. La formule de la moyenne devient alors :* $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *et*
>
> $\text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ ....

La régression

La régression permet d'analyser la manière dont une variable (dite expliquée) est affectée par les valeurs d'une ou plusieurs autres variables (dites explicatives). Exemple : $y = ax + b$

La méthode des MCO (Moindres Carrés Ordinaires), par exemple, permet de calculer les paramètres a et b en fonction des observations x_i et y_i :

$$\hat{a} = \frac{cov(x,y)}{V(x)} \text{ et } \hat{b} = \bar{x} - \hat{a}\bar{y}$$

L'association sur les variables qualitatives

Le test du χ^2

Principe du test : il permet de tester l'existence ou non d'une relation entre deux variables quelconques. Il repose sur une comparaison de la fréquence de distribution de ces variables à une distribution théorique. Il consiste à calculer (χ^2 calculé) la somme des écarts entre la distribution théorique et la distribution observée et à comparer ce résultat à une valeur prédéterminée (χ^2 lu dans une table ou χ^2 tabulé)

Si le χ^2 calculé est supérieur au χ^2 tabulé alors il existe une relation entre les deux variables.

(voir le rappel de cours de L3 dans le fichier MS1_M1M2Res : Module 2)

La notion de similarité

Similarité sur des variables dichotomiques

On dit que deux objets A et B, décrits par p attributs sont similaires, si le maximum d'attributs sur les p attributs sont identiques. Le nombre de points communs (ou coïncidences) permet de construire une mesure quantitative de la similarité entre des objets.

Il existe deux types de coïncidences :

Valeur de l'attribut A	Valeur de l'attribut A	Coïncidence
Oui	Oui	Positive
Oui	Non	Non coïncidence
Non	Oui	Non coïncidence
Non	Non	Négative

Selon la manière de prendre en compte les coïncidences négatives, on obtiendra différentes valeurs de similarité :

L'indice de Russel n'accorde aucun poids aux coïncidences négatives. C'est donc le nombre de coïncidences positives divisé par le nombre de comparaisons

L'indice de Jaccard consiste à donner un poids moins important aux coïncidences négatives qu'aux positives. C'est donc le nombre de coïncidences positives divisé par la différence entre le nombre de comparaisons et le nombre de coïncidences négatives.

L'indice de Sokal donne le même poids aux coïncidences négatives et positives. Nombre de coïncidence positives et négatives divisé par le nombre de comparaisons.

Le choix du bon indice de coïncidence ne peut s'effectuer qu'après une analyse des variables de comparaison et une étude de la distribution des valeurs.

Un exemple : comparons la composition de trois desserts selon leur composition

	Barre de céréales	Crème dessert	Gâteau de Riz
Chocolat	Oui	Non	Oui
Beurre	Non	Non	Oui
Liquide	Non	Oui	Non
Parfum mandarine	Non	Non	Oui
Emballage métal	Non	Oui	Oui
Mini-dose	Oui	Oui	Non
Sucre	Oui	Oui	Oui
Riz	Oui	Non	Oui
Edulcorant	Non	Non	Oui
Colorant	Non	Non	Oui

Matrice de Coïncidence

		Barre de céréales	
		Oui	Non
Crème dessert	Oui	2	2
	Non	2	4
Gâteau de Riz	Oui	3	5
	Non	2	0

Indices de similarité :

Indice	Formule	Barre de Céréale / Crème dessert	Barre de Céréales / Gâteau de Riz	Conclusion
Russel	Coïncidences positives / Nombre de comparaisons	20%	30%	Gâteau de riz proche de barre de céréales
Jaccard	Coïncidences positives / (Nombre de comparaisons-coïncidences négatives)	33%	30%	Crème dessert proche de barre de céréales
Sokal	Coïncidences positives et négatives / Nombre de comparaisons	60%	30%	Crème dessert proche de barre de céréales

Dans cet exemple, on voit que le choix de l'indice de similarité a une importance capitale car la conclusion dépendra de l'indice choisi.

Similarité sur variables quelconques

Il s'agit de construire un indice composite de toutes les similarités sur différents critères :

- la similarité sur variables dichotomique est égale à 1 si les deux objets présentent le même critère
- la similarité sur les variables qualitatives est égale à 1 si les objets présentent la même caractéristique
- la similarité sur les variables quantitatives mesure l'écart entre les deux objets de manière relative par rapport à l'étendue de la distribution de la variable.

Exemple de similarité sur variables quantitatives

	Produit A	Produit B	Produit C	Produit D
Prix	1300	1500	1800	1600

Etendue de la distribution : C'est l'écat entre la valeur maximale et la valeur minimale, donc ici elle est égale à (1800-1300)=500

La Similarité entre A et B sera égale au complément à 1 de la valeur absolue de l'écart entre A et B, divisé par l'étendue.

Soit ici : = 1 – (abs(1500-1300)/500) = 0,6

On voit aisément que deux produits qui ont un même prix auront une similarité=1 et les deux extrêmes auront une similarité=0.

La notion de distance

Cette notion est très utilisée dans les analyses multidimensionnelles et notamment dans les techniques de classification.

La notion de distance est le complément à la notion de similarité. Deux objets similaires ont en effet une distance nulle et une distance maximale sépare deux objets différents.

La notion de distance

S'il existe plusieurs façons de calculer des distances, l'une des plus utilisées est la distance euclidienne.

La distance euclidienne se définie dans R^n de la façon suivante :

$$d : R^n x R^n \rightarrow R^+$$

$$(x, y) \mapsto d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

> *Les différents types de distances :*
>
> *Distances euclidiennes générales : ce sont les distances les plus classiques, elles vérifient:*
>
> *$d_M^2(wi,w_{i'}) = {}^t(x_i - x_{i'})M (x_i - x_{i'})$ où M est une matrice symétrique définie positive. On les nomme également distances quadratiques ou métriques. Nous listons quelques cas particuliers :*
>
> *· distance euclidienne simple : c'est le cas où M=1 : $d^2(w_i,w_{i'}) = \sum^p_{j=1}(x_i^j - x_{i'}^j)^2$ (C'est le cas précédent)*
>
> *· distance de Mahalanobis : elle se rencontre fréquemment en analyse des données et surtout en analyse discriminante. Son expression analytique est la suivante : $d^2(w_i,w_{i'}) = {}^t(x_i - x_{i'})V^{-1} (x_i - x_{i'})$ où V est la matrice de variance-covariance.*
>
> *· distance du χ^2 : la distance du chi2 (lire « qui deux ») est importante en analyse des données. Elle est particulièrement bien adaptée aux tableaux de contingence. Elle est utilisée en analyse factorielle des correspondances. Rappelons qu'elle s'exprime ainsi : $d^2(w_i,w_{i'}) = \sum^p_{j=1} 1/x_.^j (x_i^j/x_i - x_{i'}^j/x_{i'})^2$ où $x_.^j = \sum^n_{i=1} x_i^j$ et $x_i = \sum^p_{j=1} x_i^j$.*

La notion de variance et les techniques de typologie

Pour mesurer le degré d'homogénéité d'une population, certaines techniques utilisent la notion de variance.

Considérons les notes en math et en français obtenues par des élèves d'une classe :

	Maths	Français
Elève 1	3	7
Elève 2	4	8
Elève 3	6	9
Elève 4	11	11
Elève 5	16	13
Elève 6	18	14
Elève 7	19	15
Moyenne	11	11

La variance des notes se calcule en calculant les écarts par rapport à la moyenne, en élevant ces écarts au carré et en divisant par le nombre d'observations.

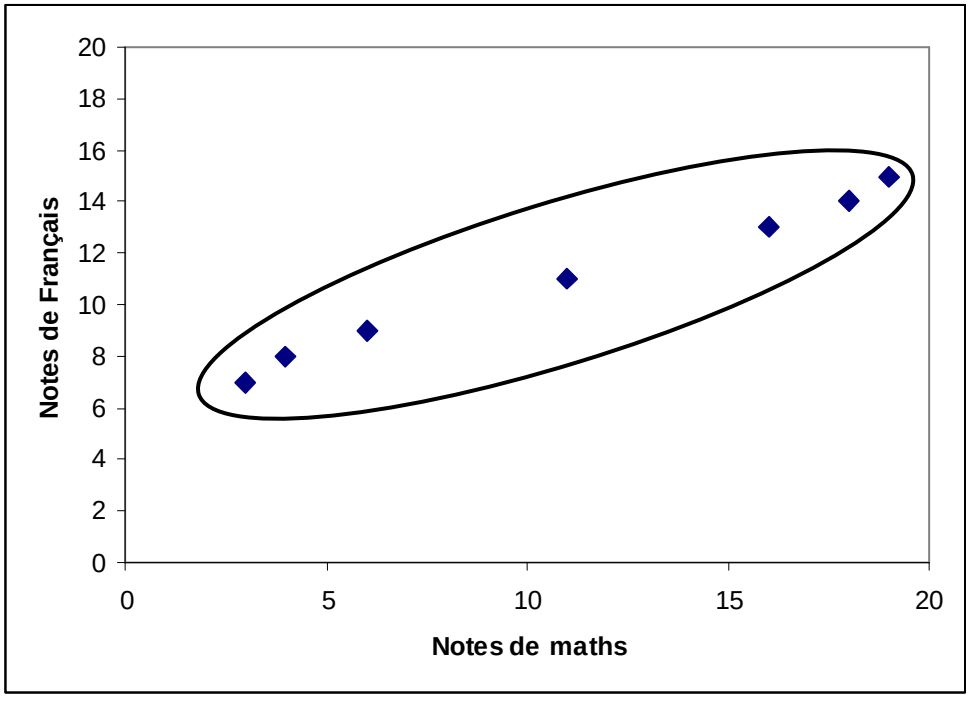
La formule de la variance est :

$$V = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

On peut appréhender la variance comme étant une surface. Plus elle est importante, plus la distribution s'éloigne de la moyenne. Si on considère cette surface comme étant un carré, la racine carrée de la variance représentera un coté de ce carré. Ce sera l'écart-type qui sera lui aussi une mesure de la dispersion autour de la moyenne.

Dans l'exemple, la variance des notes de maths est de 39,43, celles des notes de français de 8,28. De même, l'écart type des notes de maths est de 6,27 et celui des notes de français de 2,87. Ainsi, le professeur de math construit une échelle de différenciation plus importante que le professeur de français.

Comme la distance euclidienne, la variance permet de découper une population en sous ensembles homogènes.



On peut envisager l'algorithme suivant :

La variable maths possédant la variance la plus forte, on découpe la population selon la note de math. On crée les groupes suivants : Groupe 1 : élèves 1,2 et 3 , Groupe 2 : élèves 4,5,6 et 7

Le centre de gravité du nuage total est le point moyen (11,11)

Le centre de gravité du groupe 1 est égal aux moyennes en math et français des trois individus de ce groupe. Idem pour le groupe 2

La variance totale du nuage se calcule comme le carré de la distance entre l'ensemble des points et le centre de gravité. Ce qui donne (théorème de décomposition de la variance) :

La variance du groupe 1 correspond aux écarts entre les points du groupe 1 et le centre de gravité du groupe 1. De même, la variance du groupe 2 correspond aux écarts entre les points du groupe 2 et le centre de gravité du groupe 2

La variance intraclasse aussi appelée variance résiduelle est une moyenne des variances à l'intérieur des groupes

La variance interclasse correspond aux écarts entre les centres de gravité des groupes 1 et 2 et le centre de gravité de l'ensemble des points. On l'appelle également variance expliquée (par la répartition en groupe).

Une bonne typologie (ou segmentation) se juge sur la variance intraclasse (plus elle est faible, plus les points d'un groupe sont proches) et sur la variance interclasse (plus elle est forte, plus les groupes sont éloignés). Elle aura donc un ratio variance interclasse/ variance intraclasse maximal.