

•
•
•

Analyse Statistique pour Le Traitement d'Enquêtes

Mastère Développement Agricole Tropical

Stéphanie Laffont & Vivien ROSSI

UMR ENSAM-INRA

Analyse des systèmes et Biométrie

rossiv@ensam.inra.fr

Plan du cours

- Introduction

Plan du cours

- Introduction
- Traitements préliminaires
 - Le questionnaire
 - L'échantillonnage
 - La collecte des données

Plan du cours

- Introduction
- Traitements préliminaires
- Traitements statistiques
 - Analyse uni-variée des variables : tris à plats, répartition, histogrammes,...
 - Analyse bi-variée des variables : tris croisés, corrélations, test du χ^2 ,...
 - Analyse multi-variée : ACP, AFC

Plan du cours

- Introduction
- Traitements préliminaires
- Traitements statistiques
- Conclusion

Plan du cours

- Introduction
- Traitements préliminaires
- Traitements statistiques
- Conclusion

Détails des phases d'une enquête (I)

I. L'idée

Détails des phases d'une enquête (I)

I. L'idée

1- Le commanditaire

Détails des phases d'une enquête (I)

I. L'idée

1- Le commanditaire

2- Le but

Détails des phases d'une enquête (I)

I. L'idée

- 1- Le commanditaire
- 2- Le but
- 3- La population

Détails des phases d'une enquête (I)

I. L'idée

- 1- Le commanditaire
- 2- Le but
- 3- La population
- 4- Les types d'informations à collecter

Détails des phases d'une enquête (I)

I. L'idée

- 1- Le commanditaire
- 2- Le but
- 3- La population
- 4- Les types d'informations à collecter
- 5- Le budget

Détails des phases d'une enquête (II)

II. La préparation

Détails des phases d'une enquête (II)

II. La préparation

1- Etude du domaine

Détails des phases d'une enquête (II)

II. La préparation

- 1- Etude du domaine
- 2- Formulation du problème posé

Détails des phases d'une enquête (II)

II. La préparation

- 1- Etude du domaine
- 2- Formulation du problème posé
- 3- Détermination de la population

Détails des phases d'une enquête (II)

II. La préparation

- 1- Etude du domaine
- 2- Formulation du problème posé
- 3- Détermination de la population
- 4- Détermination des objectifs

Détails des phases d'une enquête (II)

II. La préparation

- 1- Etude du domaine
- 2- Formulation du problème posé
- 3- Détermination de la population
- 4- Détermination des objectifs
- 5- Définition des informations à recueillir

Détails des phases d'une enquête (II)

II. La préparation

6- Choix de l'échantillon

Détails des phases d'une enquête (II)

II. La préparation

6- Choix de l'échantillon

7- Choix du mode de collecte

Détails des phases d'une enquête (II)

II. La préparation

6- Choix de l'échantillon

7- Choix du mode de collecte

8- Rédaction du projet de questionnaire, du guide d'entretien

Détails des phases d'une enquête (II)

II. La préparation

- 6- Choix de l'échantillon
- 7- Choix du mode de collecte
- 8- Rédaction du projet de questionnaire, du guide d'entretien
- 9- Test du projet de questionnaire et guide d'entretien

Détails des phases d'une enquête (II)

II. La préparation

- 6- Choix de l'échantillon
- 7- Choix du mode de collecte
- 8- Rédaction du projet de questionnaire, du guide d'entretien
- 9- Test du projet de questionnaire et guide d'entretien
- 10- Rédaction du questionnaire définitif, du guide d'entretien définitif

Etude du domaine

- Identifier les éléments pertinents

Etude du domaine

- Identifier les éléments pertinents
- Recueillir un maximum d'idées de personnes :
 - compétentes sur le domaine
 - concernées par le problème

Formulation du problème posé

- Les options possibles des décisions à prendre
- Les hypothèses a priori

Détermination de la population

- Qui en fait parti ?
- Qui en exclu ?

-
-
-

Détermination des objectifs

Les objectifs sont souvent limités par :

Détermination des objectifs

Les objectifs sont souvent limités par :

- Le budget disponible

Détermination des objectifs

Les objectifs sont souvent limités par :

- Le budget disponible
- La longueur du questionnaire acceptable pour l'enquêté → **Aller à l'essentiel**

Détermination des objectifs

Les objectifs sont souvent limités par :

- Le budget disponible
- La longueur du questionnaire acceptable pour l'enquêté → **Aller à l'essentiel**
- Les outils de traitements

Choix du mode de collecte

- Entretien face à face

Choix du mode de collecte

- Entretien face à face
- Questionnaire par enquêteur

Choix du mode de collecte

- Entretien face à face
- Questionnaire par enquêteur
- Dépouillement de documents

Choix du mode de collecte

- Entretien face à face
- Questionnaire par enquêteur
- Dépouillement de documents
- Questionnaire auto administré

Choix du mode de collecte

- Entretien face à face
- Questionnaire par enquêteur
- Dépouillement de documents
- Questionnaire auto administré
- Extraction de fichiers

Choix de l'échantillonnage

- **Définition** : Sous ensemble de la population censé la représenter dans son ensemble

Choix de l'échantillonnage

- **Définition** : Sous ensemble de la population censé la représenter dans son ensemble
- **Utilité** : Limiter le coût de l'enquête

Choix de l'échantillonnage

- **Définition** : Sous ensemble de la population censé la représenter dans son ensemble
 - **Utilité** : Limiter le coût de l'enquête
- Relève de la théorie des sondages

Détails des phases d'une enquête (III)

III. Recueil des données

Détails des phases d'une enquête (III)

III. Recueil des données

1- Approche de l'enquêté

Détails des phases d'une enquête (III)

III. Recueil des données

- 1- Approche de l'enquêté
- 2- Soumission des questions

Détails des phases d'une enquête (III)

III. Recueil des données

- 1- Approche de l'enquêté
- 2- Soumission des questions
- 3- Enregistrement des réponses

Détails des phases d'une enquête (IV)

IV. Analyse

Détails des phases d'une enquête (IV)

IV. Analyse

1- Codage et transformation des données

Détails des phases d'une enquête (IV)

IV. Analyse

- 1- Codage et transformation des données
- 2- Analyse univariée

Détails des phases d'une enquête (IV)

IV. Analyse

- 1- Codage et transformation des données
- 2- Analyse univariée
- 3- Analyse bivariée

Détails des phases d'une enquête (IV)

IV. Analyse

- 1- Codage et transformation des données
- 2- Analyse univariée
- 3- Analyse bivariée
- 4- Analyse multivariée

Détails des phases d'une enquête (V)

V. Rapport et conclusion

Détails des phases d'une enquête (V)

V. Rapport et conclusion

1- Structure

Détails des phases d'une enquête (V)

V. Rapport et conclusion

1- Structure

2- Contenu

Détails des phases d'une enquête (V)

V. Rapport et conclusion

- 1- Structure
- 2- Contenu
- 3- Présentation

Les Méthodes d'échantillonnages

- Méthodes empiriques :
 - Méthode des unités types
 - Méthode des quotas

Les Méthodes d'échantillonnages

- Méthodes empiriques :
 - Méthode des unités types
 - Méthode des quotas
- Méthodes probabilistes
 - Méthodes aléatoires
 - Sondage élémentaire
 - Sondage stratifié
 - Echantillonnage systématique
 - Echantillonnage à plusieurs degrés

Plan du cours

- Introduction
- Traitements préliminaires
- Traitements statistiques
- Conclusion

Introduction du cadre statistique

▶ Présentation sous forme d'un tableau :

$$\begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,n} \\ r_{2,1} & r_{2,2} & \cdots & r_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{q,1} & r_{q,2} & \cdots & r_{q,n} \end{bmatrix}$$

Introduction du cadre statistique

▶ Présentation sous forme d'un tableau :

$$\begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,n} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{q,1} & r_{q,2} & \cdots & r_{q,n} \end{bmatrix}$$

- $r_{i,j}$ réponse à la $i^{\text{ème}}$ question sur le $j^{\text{ème}}$ questionnaire, avec $i = 1, \dots, q$ et $j = 1, \dots, n$.

Introduction du cadre statistique

▶ Présentation sous forme d'un tableau :

$$\begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,n} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{q,1} & r_{q,2} & \cdots & r_{q,n} \end{bmatrix}$$

- La $j^{\text{ème}}$ colonne regroupe toutes les réponses du $j^{\text{ème}}$ questionnaire.

Introduction du cadre statistique

▶ Présentation sous forme d'un tableau :

$$\begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,n} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{q,1} & r_{q,2} & \cdots & r_{q,n} \end{bmatrix}$$

- La $i^{\text{ème}}$ ligne regroupe toutes les réponses à la $i^{\text{ème}}$ question.

Introduction du cadre statistique

▶ Présentation sous forme d'un tableau :

$$\begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,n} \\ r_{2,1} & r_{2,2} & \cdots & r_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{q,1} & r_{q,2} & \cdots & r_{q,n} \end{bmatrix}$$

▶ Toutes les enquêtes peuvent s'exprimer sous la forme du tableau ci-dessus

Formalisation des données

- La réponse R_i à la question Q_i , $i = 1, \dots, q$, du questionnaire est appelée **variable**.

Formalisation des données

- La réponse R_i à la question Q_i , $i = 1, \dots, q$, du questionnaire est appelée **variable**.
- La $i^{\text{ème}}$ ligne du tableau précédent rassemble n observations de la variable R_i .

Formalisation des données

- La réponse R_i à la question Q_i , $i = 1, \dots, q$, du questionnaire est appelée **variable**.
- La $i^{\text{ème}}$ ligne du tableau précédent rassemble n observations de la variable R_i .
- Deux types de variables : **quantitatives** et **qualitatives**
→ Deux traitements différents

Les types de variables

Les variables quantitatives

- Données numériques
- Exemples : taille, poids, concentrations, pH, ...

Les types de variables

Les variables qualitatives

- Données non numériques

Les types de variables

Les variables qualitatives

- Données non numériques
- Exemples : couleur des yeux, lieu de naissance, . . .

Les types de variables

Les variables qualitatives

- Données non numériques
- Exemples : couleur des yeux, lieu de naissance,...
- Les **modalités** sont les valeurs possibles d'une variable qualitatives :
 - Modalités ordonnées
Ex : Faible, Moyen, Bon, Très Bon
 - Modalités quelconques
Ex : Bleu, Vert, Marron

Plan du cours

- Introduction
- Traitements préliminaires
- Traitements statistiques
 - Analyse uni-variée des variables : tris à plats, répartition, histogrammes,...
 - Analyse bi-variée des variables : tris croisés, corrélations, test du χ^2 ,...
 - Analyse multi-variée : ACP, AFC
- Conclusion

Analyse univariée ou Tris à plats

- **Présentation générale** : Soient x_1, \dots, x_n des observations de la variable X
 - Exemple : les n réponses r_{i1}, \dots, r_{in} de la question Q_i .

Analyse univariée ou Tris à plats

- **Présentation générale** : Soient x_1, \dots, x_n des observations de la variable X
 - Exemple : les n réponses r_{i1}, \dots, r_{in} de la question Q_i .
- **Objectif** : Résumer l'information contenue dans x_1, \dots, x_n

Analyse univariée ou Tris à plats

- **Présentation générale** : Soient x_1, \dots, x_n des observations de la variable X
 - Exemple : les n réponses r_{i1}, \dots, r_{in} de la question Q_i .
- **Objectif** : Résumer l'information contenue dans x_1, \dots, x_n
- **Moyens** :
 - Approches numériques
 - Approches graphiques

Analyse univariée ou Tris à plats

- **Présentation générale** : Soient x_1, \dots, x_n des observations de la variable X
 - Exemple : les n réponses r_{i1}, \dots, r_{in} de la question Q_i .
- **Objectif** : Résumer l'information contenue dans x_1, \dots, x_n
- **Moyens** :
 - Approches numériques
 - Approches graphiques
- **Outils différents** suivant que la variables soit qualitative ou quantitative

Traitements numériques (1 va quant.)

- Estimation de la valeur centrale de X
 - La moyenne des x_1, \dots, x_n : $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
 - La médiane des x_1, \dots, x_n : $x_{(n/2)}$ “le x_i du milieu”.

Traitements numériques (1 va quant.)

- Estimation de la valeur centrale de X
 - La moyenne des x_1, \dots, x_n : $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
 - La médiane des x_1, \dots, x_n : $x_{(n/2)}$ “le x_i du milieu”.
- La dispersion de X
 - L'écart-type à la moyenne :
$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$
 - Les quartiles, le minimum, le maximum

Traitements numériques (1 va quant.)

- Exemple, la série de notes : 10, 12, 7, 14, 11, 8, 9, 15, 5, 12, 10.5, 11, 14, 8, 16

Min	Q_1	Médiane	Q_3	Max
5	8.5	11	13	16

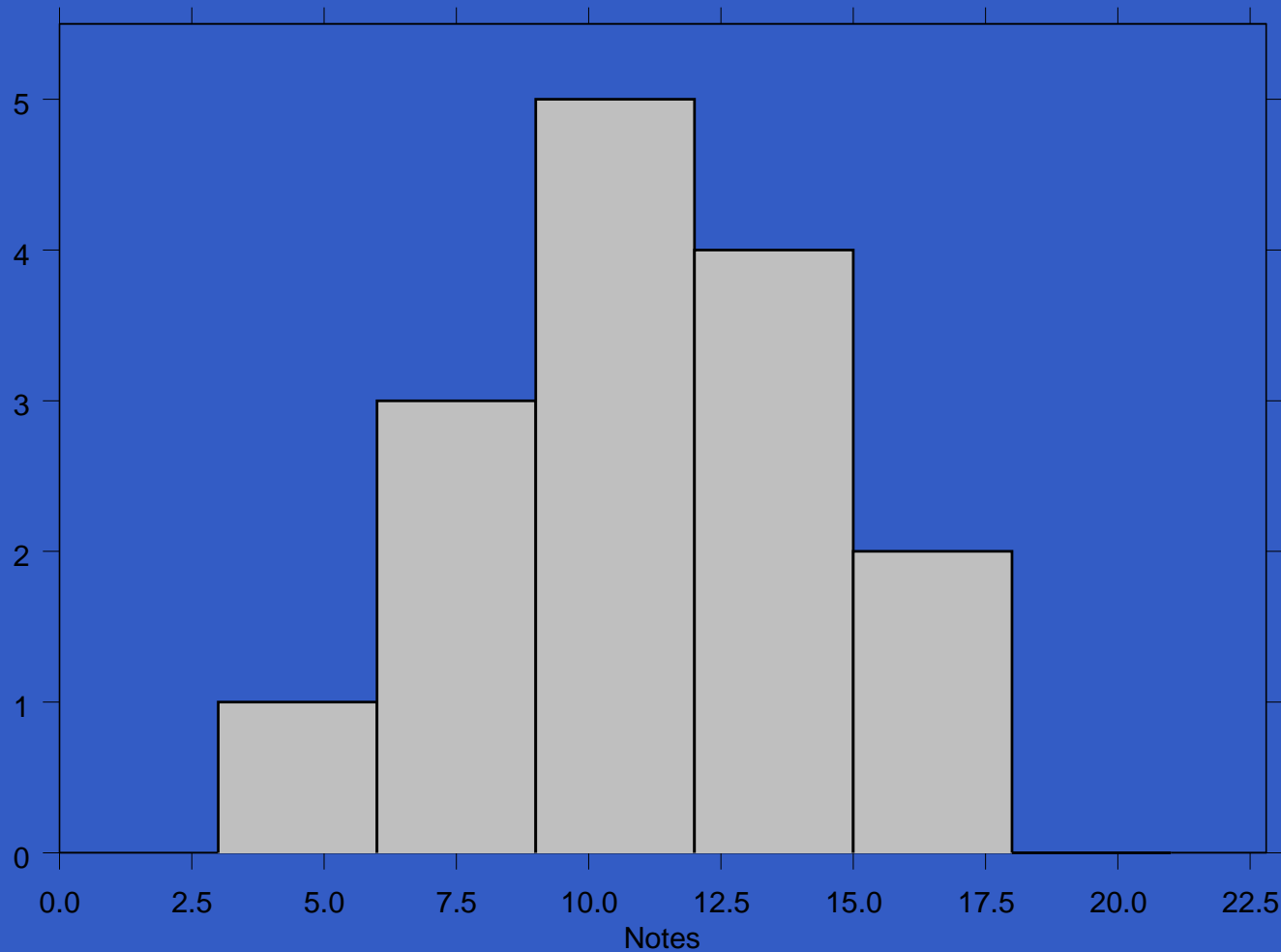
Traitement graphique (1 va quant.)

→ Histogramme

- Représentation des effectifs par classe
- Dépendant des classes : nombre et taille
- Exemple : Les mêmes notes 10, 12, 7, 14, 11, 8, 9, 15, 5, 12, 10.5, 11, 14, 8, 16.

Les classes : $C_1 = [3\ 6[$, $C_2 = [6\ 9[$,
 $C_3 = [9\ 12[$, $C_4 = [12\ 15[$, $C_5 = [15\ 18[$

Traitement graphique (1 va quant.)



Traitements numériques (1 va qualit.)

- Soient m_1, \dots, m_k les k modalités de X

Traitements numériques (1 va qualit.)

- Soient m_1, \dots, m_k les k modalités de X
- La **fréquence** de la modalité m_i dans l'échantillon x_1, \dots, x_n :

$$f_i = \frac{n_i}{n}$$

où n_i est le nombre d'occurrences de la modalité m_i dans x_1, \dots, x_n

Traitements numériques (1 va qualit.)

- Modalités quelconques

Le tableau des fréquences des modalités m_1, \dots, m_k dans x_1, \dots, x_n est

Modalités	m_1	m_2	\dots	m_k
Fréquences	f_1	f_2	\dots	f_k

Traitements numériques (1 va qualit.)

- Modalités ordonnées

Le tableau des fréquences des modalités

m_1, \dots, m_k dans x_1, \dots, x_n est :

Modalités	m_1	m_2	\dots	m_k
Fréquences	f_1	f_2	\dots	f_k
Fréquences cumulées	$\frac{n_1}{n}$	$\frac{n_1+n_2}{n}$	\dots	$\frac{n_1+\dots+n_k}{n} = 1$

Traitements graphiques (1 va qualit.)

- Modalités quelconques :
 - Illustration visuelle de la répartition dans les classes
 - Diagramme en batons, histogramme, graphique en secteurs, ...

Traitements graphiques (1 va qualit.)

- Modalités quelconques :
 - Illustration visuelle de la répartition dans les classes
 - Diagramme en batons, histogramme, graphique en secteurs, ...
- Modalités ordonnées → Idem mais il faut **respecter l'ordre des modalités**

Exemple : modalités quelconques

- Quel type d'engrais utilisez-vous ?
 - a : chimiques, b : biologiques
 - c : mélange des deux, d : aucun

Exemple : modalités quelconques

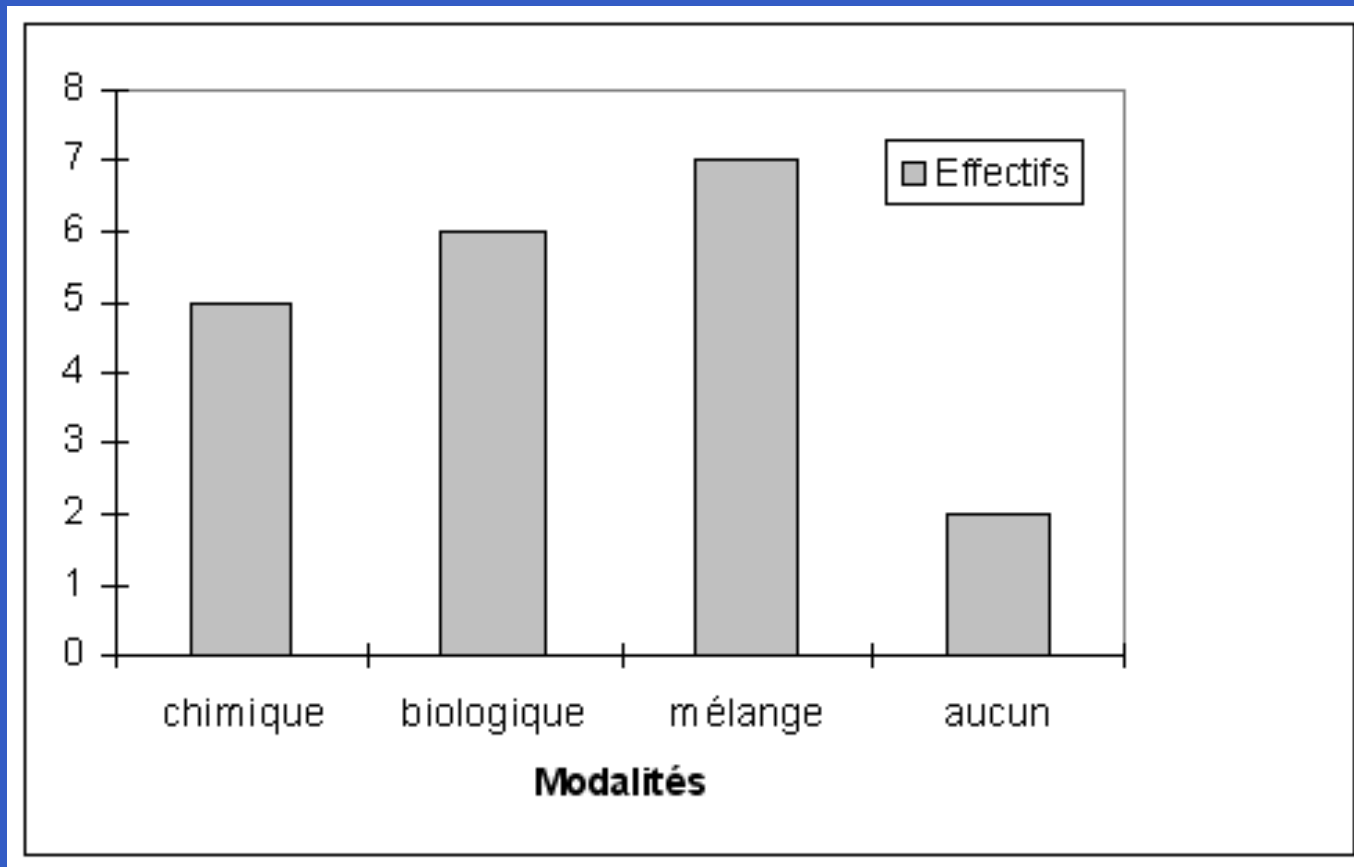
- Quel type d'engrais utilisez-vous ?
a : chimiques, b : biologiques
c : mélange des deux, d : aucun
- Réponses : b, b, a, a, c, d, c, b, c, a, d, c, b, a, c, a, b, c, c,b.

Exemple : modalités quelconques

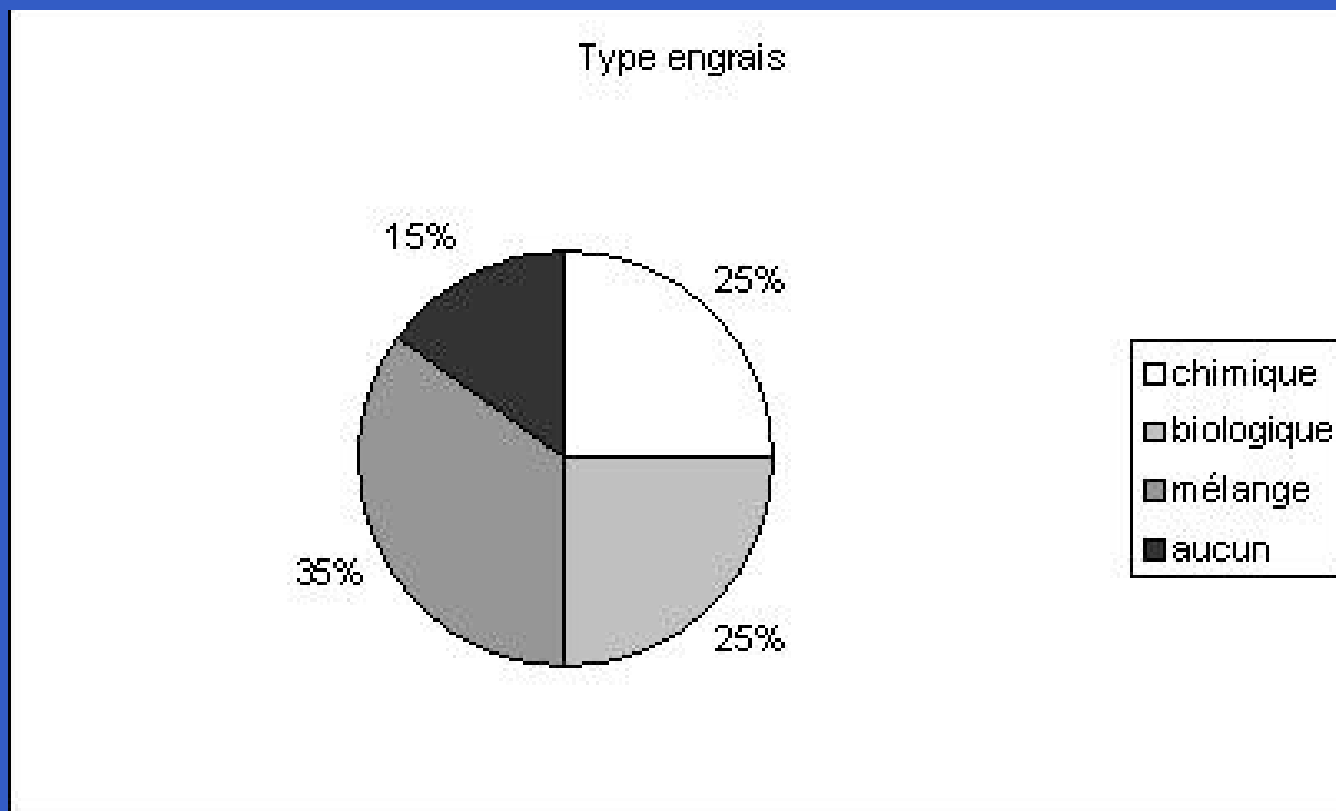
- Quel type d'engrais utilisez-vous ?
a : chimiques, b : biologiques
c : mélange des deux, d : aucun
- Réponses : b, b, a, a, c, d, c, b, c, a, d, c, b, a, c, a, b, c, c, b.
- Le tableau des fréquences :

a	b	c	d
5/20	6/20	7/20	2/20

Exemple : modalités quelconques



Exemple : modalités quelconques



Exemple : modalités ordonnées

- Comment trouvez-vous le café ?
TB : Très bon, B : Bon,
A : Acceptable, M : Mauvais

Exemple : modalités ordonnées

- Comment trouvez-vous le café ?

TB : Très bon, B : Bon,

A : Acceptable, M : Mauvais

- Réponses :

A, B, B, TB, M, A, B, A, TB, M, B, TB, A, A, B,
M, M, TB, A, B

Exemple : modalités ordonnées

- Tableau des fréquences :

	TB	B	A	M
Fréquences	0.2	0.3	0.3	0.2
Fréq. cumulées	0.2	0.5	0.8	1

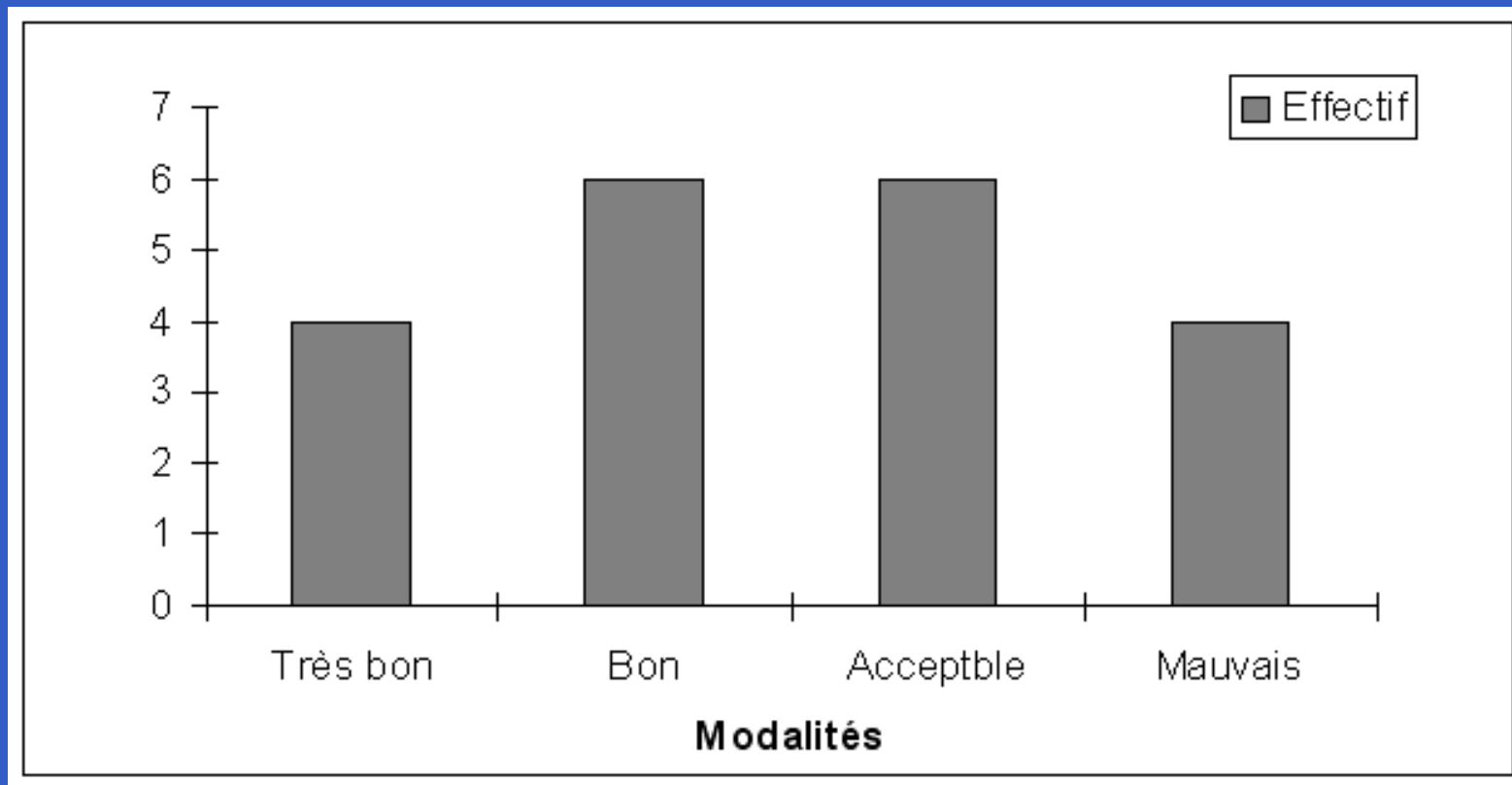
Exemple : modalités ordonnées

- Tableau des fréquences :

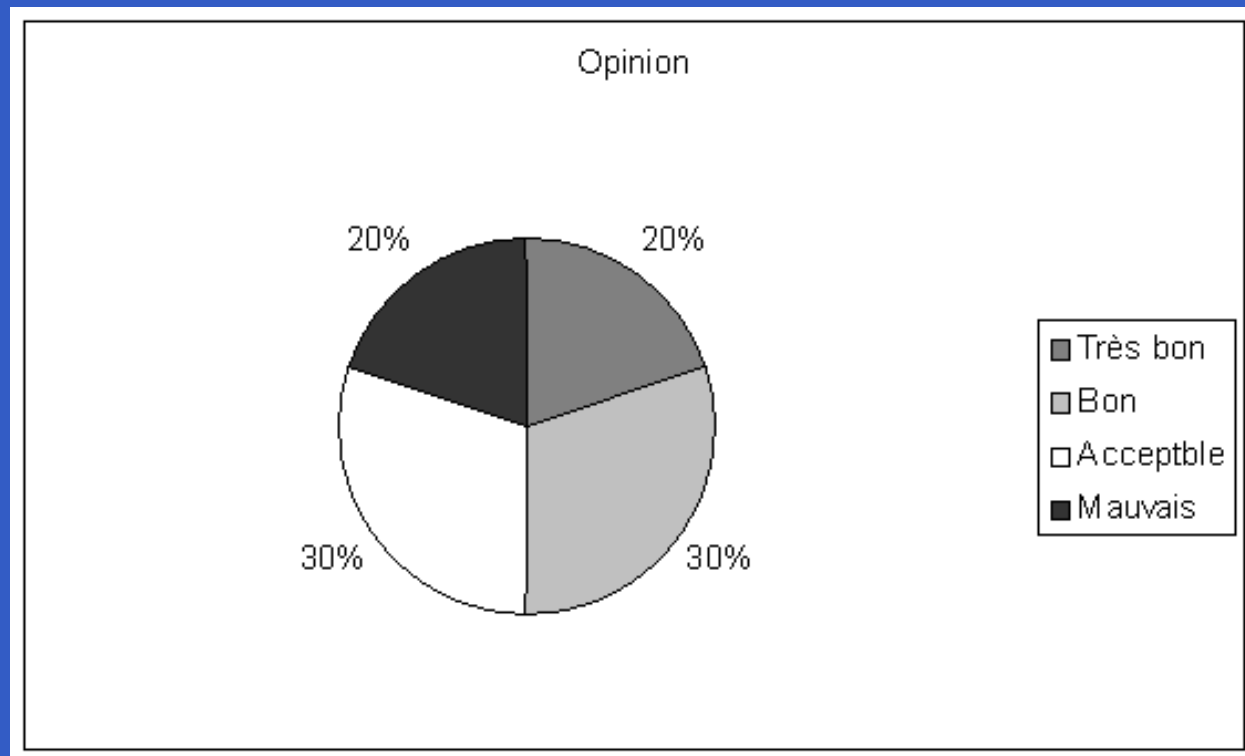
	TB	B	A	M
Fréquences	0.2	0.3	0.3	0.2
Fréq. cumulées	0.2	0.5	0.8	1

- Commentaires :
 - 50% des clients apprécient ce café (TB+B)
 - 80% des clients sont satisfaits du café (TB+B+A)

Exemple : modalités ordonnées



Exemple : modalités ordonnées



Plan du cours

- Introduction
- Traitements préliminaires
- Traitements statistiques
 - Analyse uni-variée des variables : tris à plats, répartition, histogrammes,...
 - Analyse bi-variée des variables : tris croisés, corrélations, test du χ^2 ,...
 - Analyse multi-variée : ACP, AFC
- Conclusion

Analyse bivariée ou Tris Croisés

- **Présentation générale** : Soient x_1, \dots, x_n des observations de la variable X et y_1, \dots, y_n des observations conjointes de la variable Y .
→ Exemple : les n réponses à deux questions différentes.

Analyse bivariée ou Tris Croisés

- **Présentation générale** : Soient x_1, \dots, x_n des observations de la variable X et y_1, \dots, y_n des observations conjointes de la variable Y .
→ Exemple : les n réponses à deux questions différentes.
- **Objectif** : Etudier le lien entre X et Y

Analyse bivariée ou Tris Croisés

- **Présentation générale** : Soient x_1, \dots, x_n des observations de la variable X et y_1, \dots, y_n des observations conjointes de la variable Y .
→ Exemple : les n réponses à deux questions différentes.
- **Objectif** : Etudier le lien entre X et Y
- **Moyens** :
 - Approches numériques
 - Approches graphiques

Analyse bivariée ou Tris Croisés

- **Présentation générale** : Soient x_1, \dots, x_n des observations de la variable X et y_1, \dots, y_n des observations conjointes de la variable Y .
→ Exemple : les n réponses à deux questions différentes.
- **Objectif** : Etudier le lien entre X et Y
- **Moyens** :
 - Approches numériques
 - Approches graphiques
- **Outils différents** suivant que les variables soient **qualitatives ou quantitatives**

Cas de deux variables quantitatives

- Recherche d'un lien linéaire entre X et Y
→ Coefficient de corrélation linéaire entre X et Y :

$$\rho_{x,y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$

Cas de deux variables quantitatives

- Recherche d'un lien linéaire entre X et Y
→ Coefficient de corrélation linéaire entre X et Y :

$$\rho_{x,y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$

- Interprétation de $\rho_{x,y}$
 - Si $|\rho_{x,y}|$ est proche de 1, le lien entre X et Y est linéaire
 - Sinon le lien n'est pas linéaire (on peut rien dire de plus)

Cas de deux variables quantitatives

- Etude graphique du lien entre X et Y
 - Représentation du nuage de points :
 x_1, \dots, x_n en abscisse et y_1, \dots, y_n
 - Si le nuage a une forme spécifique \implies il existe un lien
 - Si le nuage n'a pas de structure particulière \implies pas de lien ?

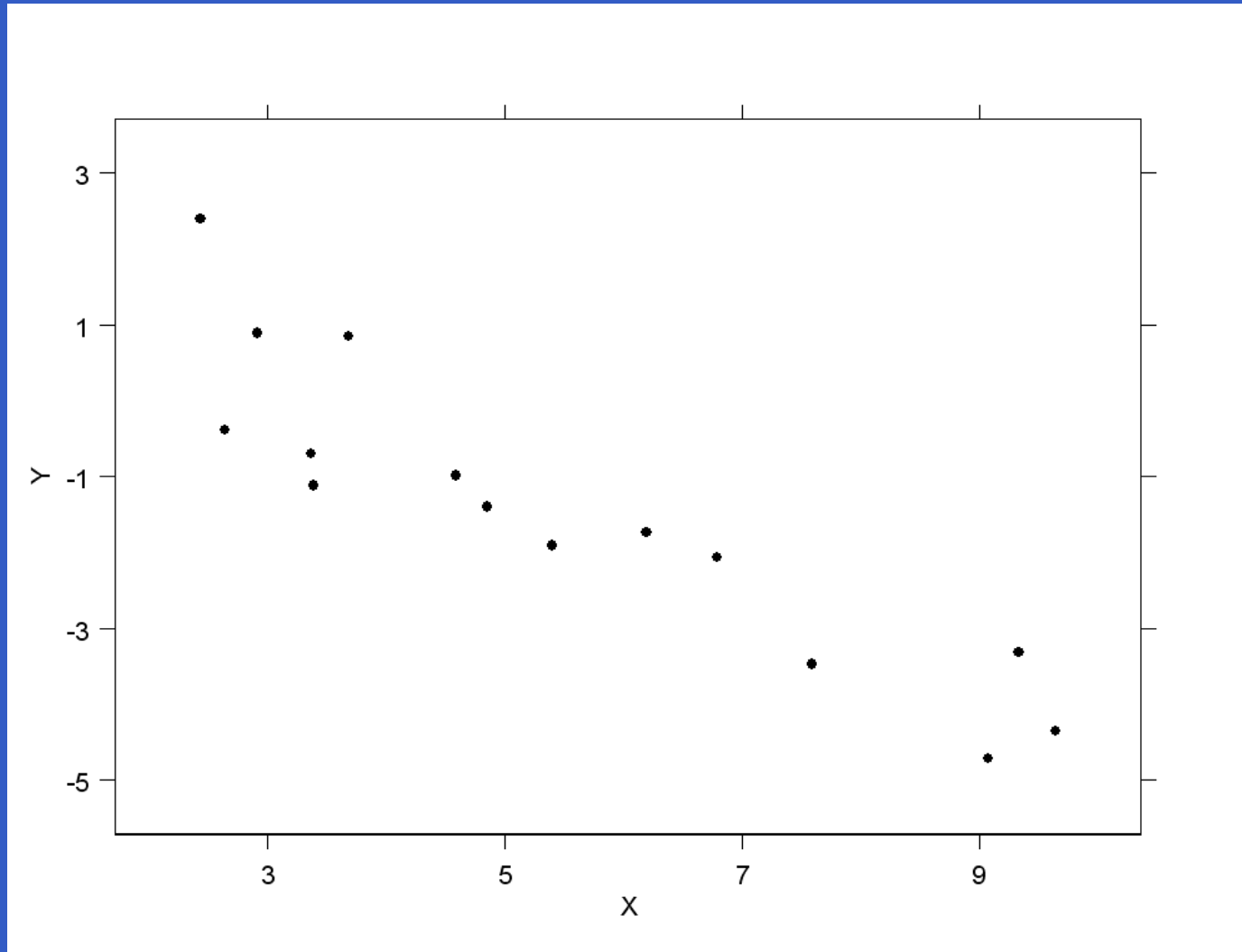
Exemple : deux variables quantitatives

$$X = \{4.59, 3.37, 9.33, 4.85, 9.64, 3.68, 6.19, 5.39, 2.43, \dots\}$$

$$Y = \{-0.99, -0.70, -3.31, -1.39, -4.35, 0.84, -1.73, -1.90, \dots\}.$$

$\rho_{XY} = -0.91$ proche de 1 \implies lien linéaire entre X et Y .

Exemple : deux variables quantitatives



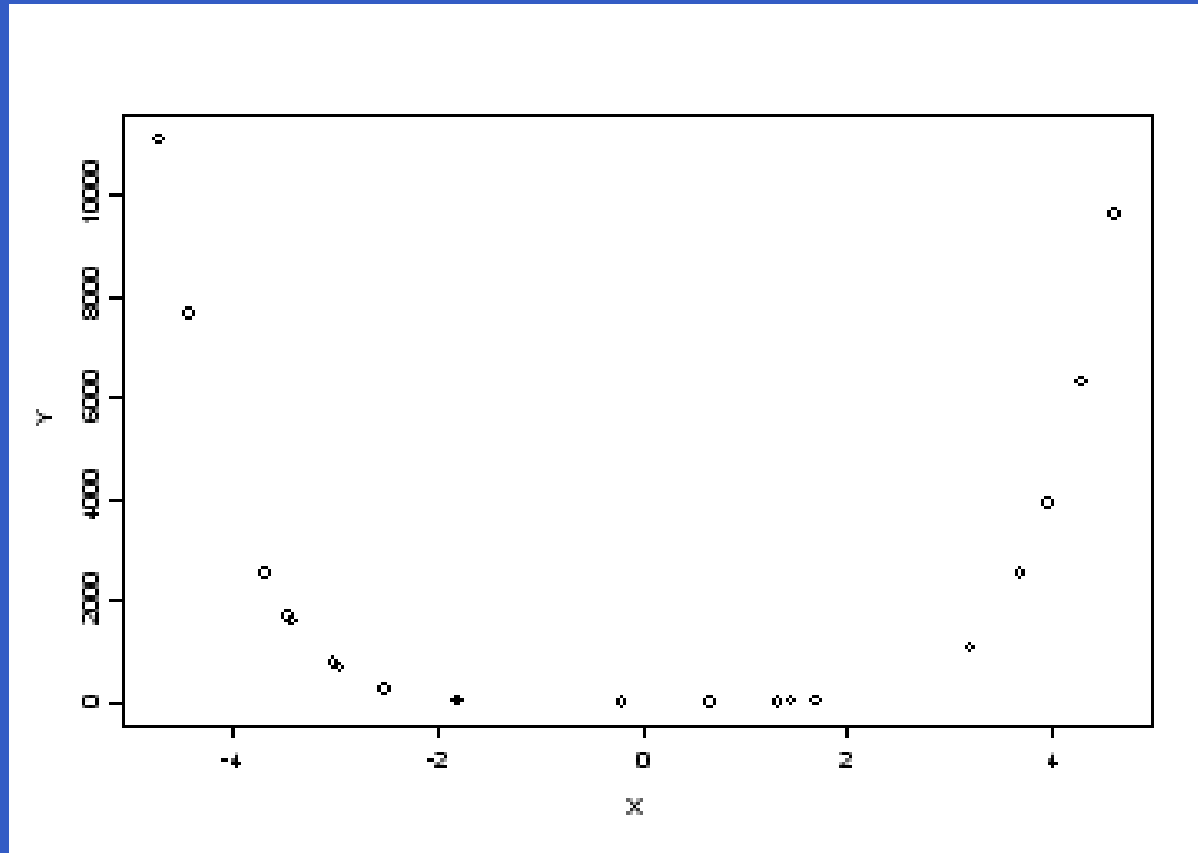
Exemple : deux variables quantitatives

$$X = \{ -3.03, -4.44, 1.45, -1.83, 0.66, 1.31, -3.69, -0.19, \dots \}$$

$$Y = \{ 774.08, 7676.3, 9.57, 37.51, 0.00864, 6.05, \dots \}.$$

$\rho_{XY} = 0.025$ pas proche de 1 \implies pas de lien linéaire.

Exemple : deux variables quantitatives



→ il semble exister un lien quadratique entre X et Y

Cas de deux Variables Qualitatives

- m_{x_1}, \dots, m_{x_k} les modalités de X .

Cas de deux Variables Qualitatives

- m_{x_1}, \dots, m_{x_k} les modalités de X .
- m_{y_1}, \dots, m_{y_l} les modalités de Y .

Cas de deux Variables Qualitatives

- mx_1, \dots, mx_k les modalités de X .
- my_1, \dots, my_l les modalités de Y .
- n_{ij} est l'effectif de l'intersection des modalités mx_i et my_j .

Cas de deux Variables Qualitatives

- mx_1, \dots, mx_k les modalités de X .
- my_1, \dots, my_l les modalités de Y .
- n_{ij} est l'effectif de l'intersection des modalités mx_i et my_j .
- $n_{i.} = \sum_{j=1}^l n_{ij}$ (i.e. l'effectif de la modalité mx_i).

Cas de deux Variables Qualitatives

- mx_1, \dots, mx_k les modalités de X .
- my_1, \dots, my_l les modalités de Y .
- n_{ij} est l'effectif de l'intersection des modalités mx_i et my_j .
- $n_{i.} = \sum_{j=1}^l n_{ij}$ (i.e. l'effectif de la modalité mx_i).
- $n_{.j} = \sum_{i=1}^k n_{ij}$ (i.e. l'effectif de la modalité my_j).

Le tableau de contingence

	my_1	\dots	my_j	\dots	my_l	
mx_1	n_{11}	\dots	n_{1j}	\dots	n_{1l}	$n_{1.}$
\vdots	\vdots		\vdots			\vdots
mx_i	n_{i1}	\dots	n_{ij}	\dots	n_{il}	$n_{i.}$
\vdots	\vdots		\vdots			\vdots
mx_k	n_{k1}	\dots	n_{kj}	\dots	n_{kl}	$n_{k.}$
	$n_{.1}$	\dots	$n_{.j}$	\dots	$n_{.l}$	n

→ La constitution de ce tableau est l'opération appelée "tri croisé".

Test d'indépendance du χ^2 (2 va qualit.)

- Caractériser l'indépendance entre deux variables X et Y est très utile dans une étude et en particulier pour une enquête.

Test d'indépendance du χ^2 (2 va qualit.)

- Caractériser l'indépendance entre deux variables X et Y est très utile dans une étude et en particulier pour une enquête.
- La mesure de liaison d^2 entre X et Y est

$$d^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

Test d'indépendance du χ^2 (2 va qualit.)

- Caractériser l'indépendance entre deux variables X et Y est très utile dans une étude et en particulier pour une enquête.
- La mesure de liaison d^2 entre X et Y est

$$d^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

- Si les variables X et Y sont indépendantes d^2 suit approximativement une loi de $\chi^2_{(l-1)(k-1)}$.

Test d'indépendance du χ^2 (2 va qualit.)

- Sous l'hypothèse que X et Y sont indépendantes, on connaît donc les valeurs vraisemblables que peut prendre d^2

Test d'indépendance du χ^2 (2 va qualit.)

- Sous l'hypothèse que X et Y sont indépendantes, on connaît donc les valeurs vraisemblables que peut prendre d^2
- Si d^2 est supérieur à la valeur critique vc qu'une variable $\chi^2_{(l-1)(k-1)}$ à une probabilité α de dépasser alors on rejettera l'hypothèse d'indépendance de X et Y .

Test d'indépendance du χ^2 (2 va qualit.)

- La valeur critique v_c est définie par

$$P(\chi_{(l-1)(k-1)}^2 > v_c) = \alpha$$

pour trouver v_c on doit utiliser des tables de probabilité.

Test d'indépendance du χ^2 (2 va qualit.)

- La valeur critique v_c est définie par

$$P(\chi_{(l-1)(k-1)}^2 > v_c) = \alpha$$

pour trouver v_c on doit utiliser des tables de probabilité.

- Si $d^2 < v_c$, on accepte l'hypothèse d'indépendance de X et Y au seuil α
→ Sinon on la rejette.

Test d'indépendance du χ^2 (2 va qualit.)

- La valeur critique v_c est définie par

$$P(\chi_{(l-1)(k-1)}^2 > v_c) = \alpha$$

pour trouver v_c on doit utiliser des tables de probabilité.

- Si $d^2 < v_c$, on accepte l'hypothèse d'indépendance de X et Y au seuil α
→ Sinon on la rejette.
- ▶ Bien entendu, si d^2 et v_c sont proches il est préférable de mitiger la conclusion.

Exemple : 2 va qualit.

Q_1 Comment trouvez-vous le café ?

R_1 1-TB très bon, 2-B bon,
3-A acceptable, 4-M mauvais

Exemple : 2 va qualit.

Q_1 Comment trouvez-vous le café ?

R_1 1-TB très bon, 2-B bon,
3-A acceptable, 4-M mauvais

Q_2 Comment jugez-vous la qualité du service ?

R_2 1-S satisfaisante, 2-C convenable, 3-Insuffisante

Exemple : 2 va qualit.

Q_1 Comment trouvez-vous le café ?

R_1 1-TB très bon, 2-B bon,
3-A acceptable, 4-M mauvais

Q_2 Comment jugez-vous la qualité du service ?

R_2 1-S satisfaisante, 2-C convenable, 3-Insuffisante

R_1 : 1TB, 2B, 3A, 2B, 2B, 3A, 4M, 2B, 1TB, 3A,
4M, 3A, 2B, 2B, 2B, 1TB, 2B, 3A, 4M, 2B, 2B,
1TB, 2B, 4M, 3A, 1TB

R_2 : 1S, 1S, 2C, 3I, 1S, 2C, 2C, 2C, 3I, 3I, 3I, 2C,
1S, 2C, 2C, 1S, 1S, 3I, 3I, 1S, 1S, 1S, 3I, 3I, 2C,
1S.

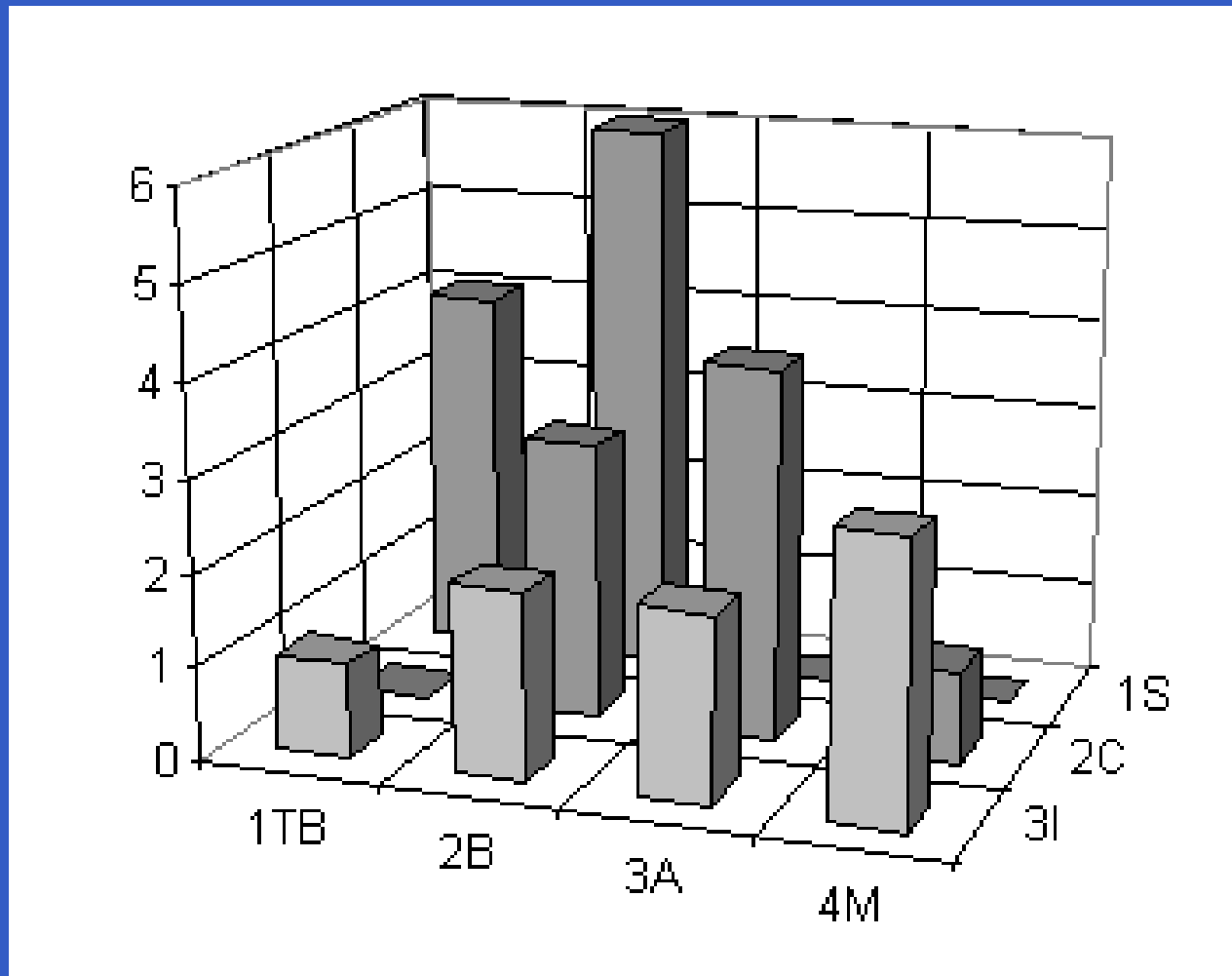
Exemple : 2 va qualit.

Tri croisé des variables café et service :

	1S	2C	3I	Total
1TB	4	0	1	5
2B	6	3	2	11
3A	0	4	2	6
4M	0	1	3	4
Total	10	8	8	26

Figure 1: Tableau de contingence

Exemple : 2 va qualit.



Exemple : 2 va qualit.

- Commentaires :
 - Les clients semblent avoir la même opinion concernant le café et le service.
 - Y aurait-il une dépendance entre les deux variables ?

Exemple : 2 va qualit.

- Commentaires :
 - Les clients semblent avoir la même opinion concernant le café et le service.
 - Y aurait-il une dépendance entre les deux variables ?

→ Effectuons un test statistique afin d'approfondir la question.

Test d'indépendance du χ^2 réalisé avec StatBox :

Exemple : 2 va qualit.

Variable en lignes : Café

Variable en colonnes : Service

Tests d'indépendance entre les lignes et les colonnes du tableau de contingence :

Valeur observée du χ^2 (ddl = 6) : 14,28

P-value associée : 0,03

Le test étant unilatéral, la p-value est comparée au seuil de signification : $\alpha = 0,05$

Valeur critique du χ^2 (ddl = 6) : 12,57

Conclusion : Au seuil de signification $\alpha = 0,05$ on peut rejeter l'hypothèse nulle d'indépendance entre les lignes et les colonnes.

Autrement dit, la dépendance entre les lignes et les colonnes est significative

Plan du cours

- Introduction
- Traitements préliminaires
- Traitements statistiques
 - Analyse uni-variée des variables : tris à plats, répartition, histogrammes,...
 - Analyse bi-variée des variables : tris croisés, corrélations, test du χ^2 ,...
 - Analyse multi-variée : ACP, AFC
- Conclusion

Analyse multi-variée

- Formalisations

→ Les p “points variables” : n individus par variable

$$\begin{array}{ccc} V_1 & V_2 & V_p \\ \left(\begin{array}{c} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{array} \right) & \left(\begin{array}{c} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{array} \right) & \dots \left(\begin{array}{c} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{array} \right) \end{array}$$

Analyse multi-variée

- Formalisations
- Objectif : Etudier globalement toutes les variables et tous les individus

Analyse multi-variée

- Formalisations
- Objectif : Etudier globalement toutes les variables et tous les individus
- Moyens :
 - Variables quantitatives : ACP
 - Variables qualitatives : AFC

Variables quantitatives : ACP

- Difficulté : n et p sont souvent supérieurs à 10 ou 20
 - Les points sujets ou variables sont dans un espace de dimension élevée

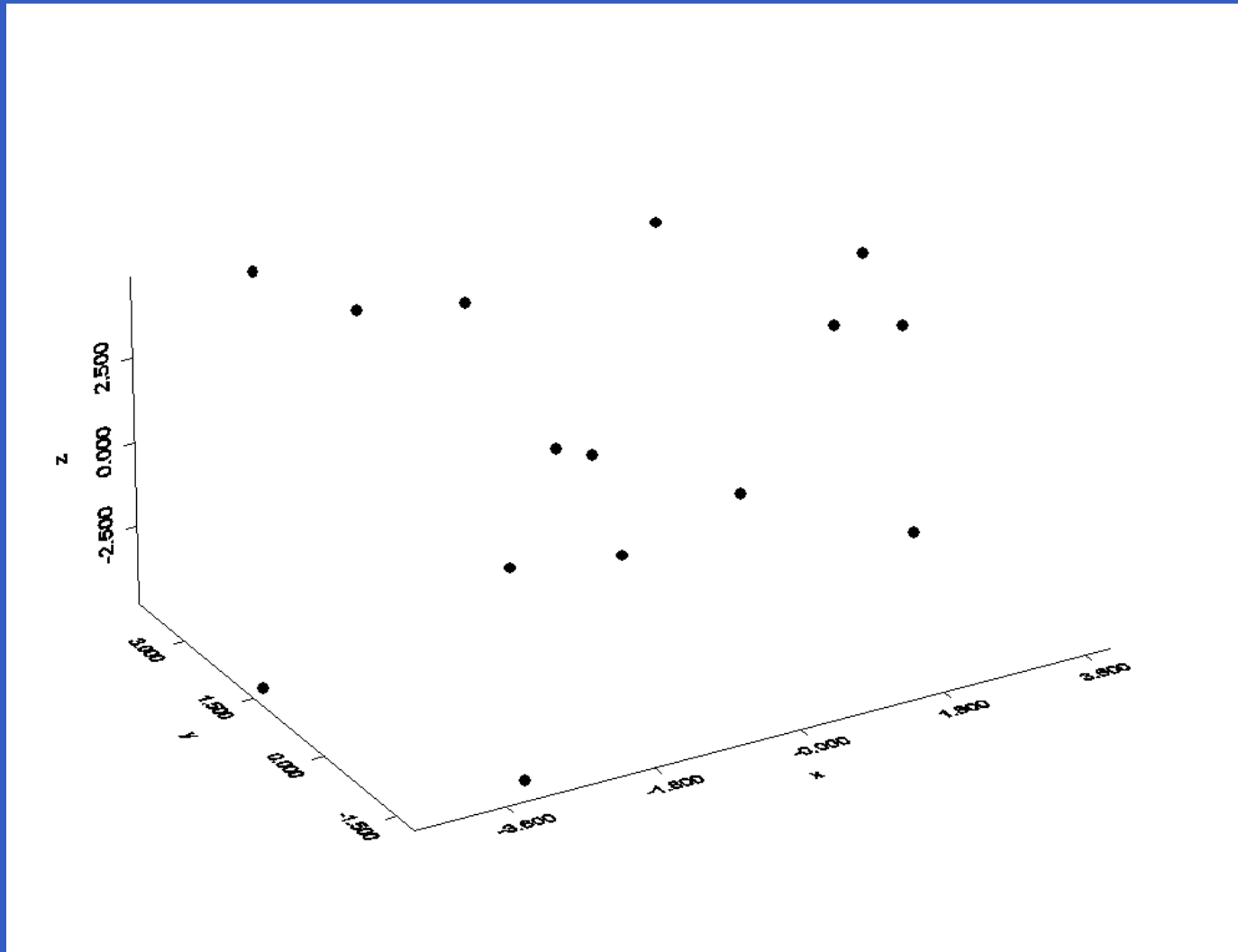
Variables quantitatives : ACP

- Difficulté : n et p sont souvent supérieurs à 10 ou 20
→ Les points sujets ou variables sont dans un espace de dimension élevée
- Extraction d'un espace de dimension plus petite contenant beaucoup d'information → l'ACP

Variables quantitatives : ACP

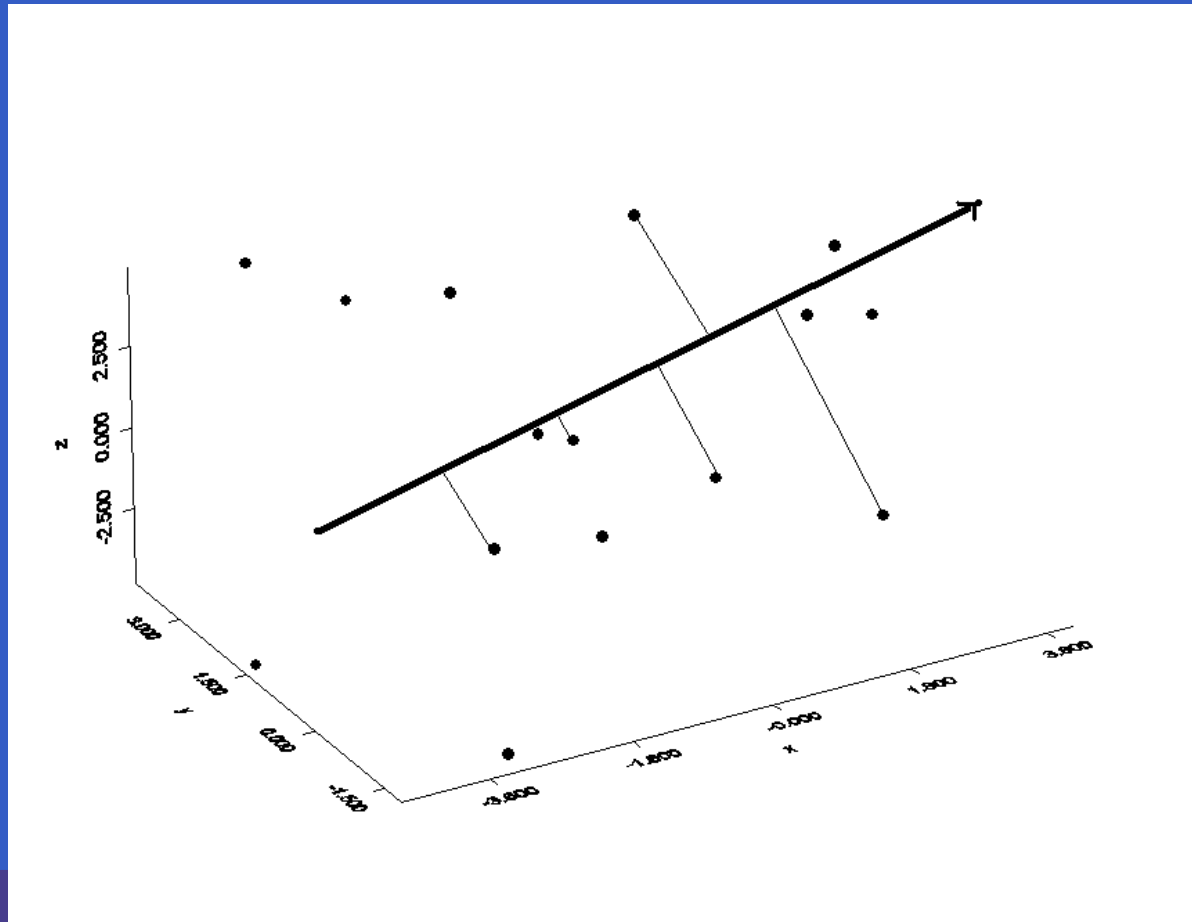
- Difficulté : n et p sont souvent supérieurs à 10 ou 20
→ Les points sujets ou variables sont dans un espace de dimension élevée
- Extraction d'un espace de dimension plus petite contenant beaucoup d'information → l'ACP
- Illustration en dimension 3 (3 variables)

Variables quantitatives : ACP



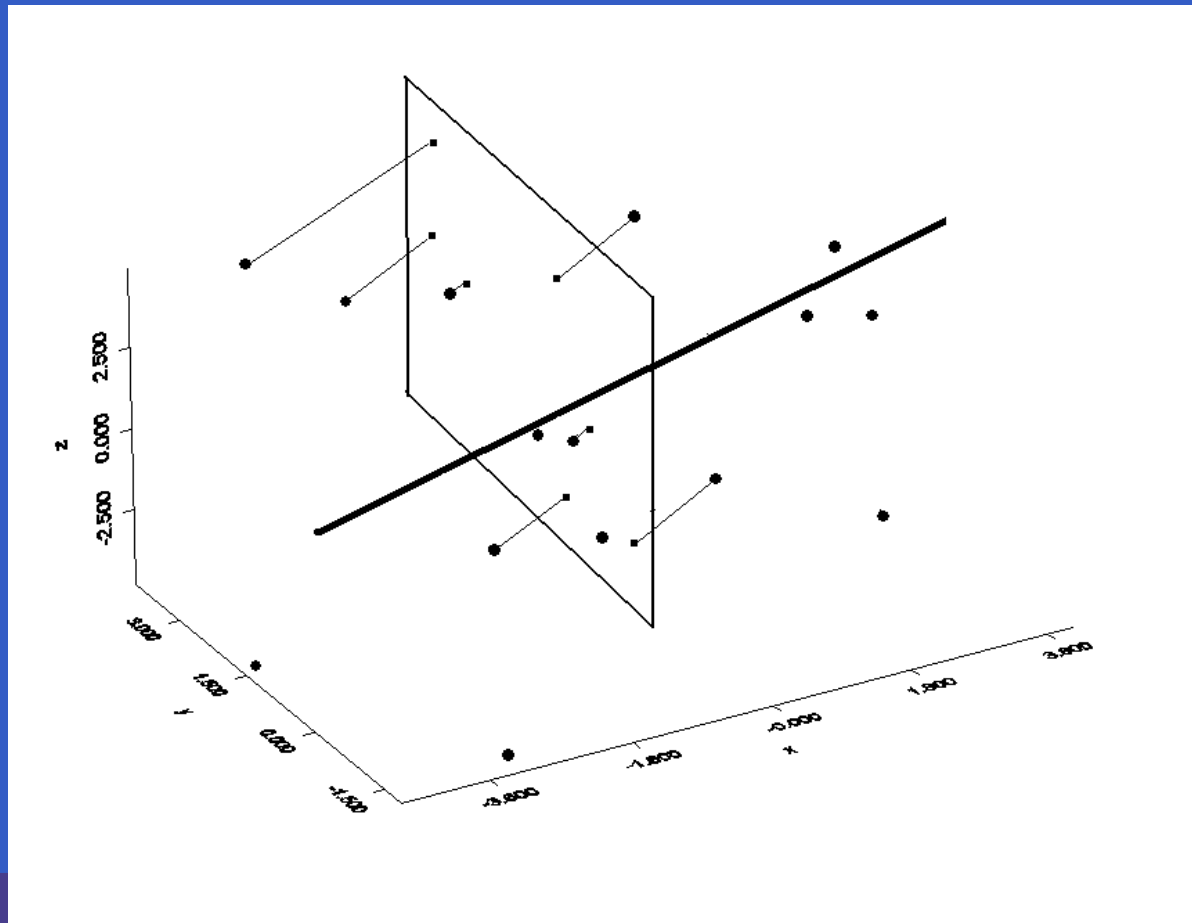
Première composante principale

Première composante : direction suivant laquelle le nuage est le plus étiré



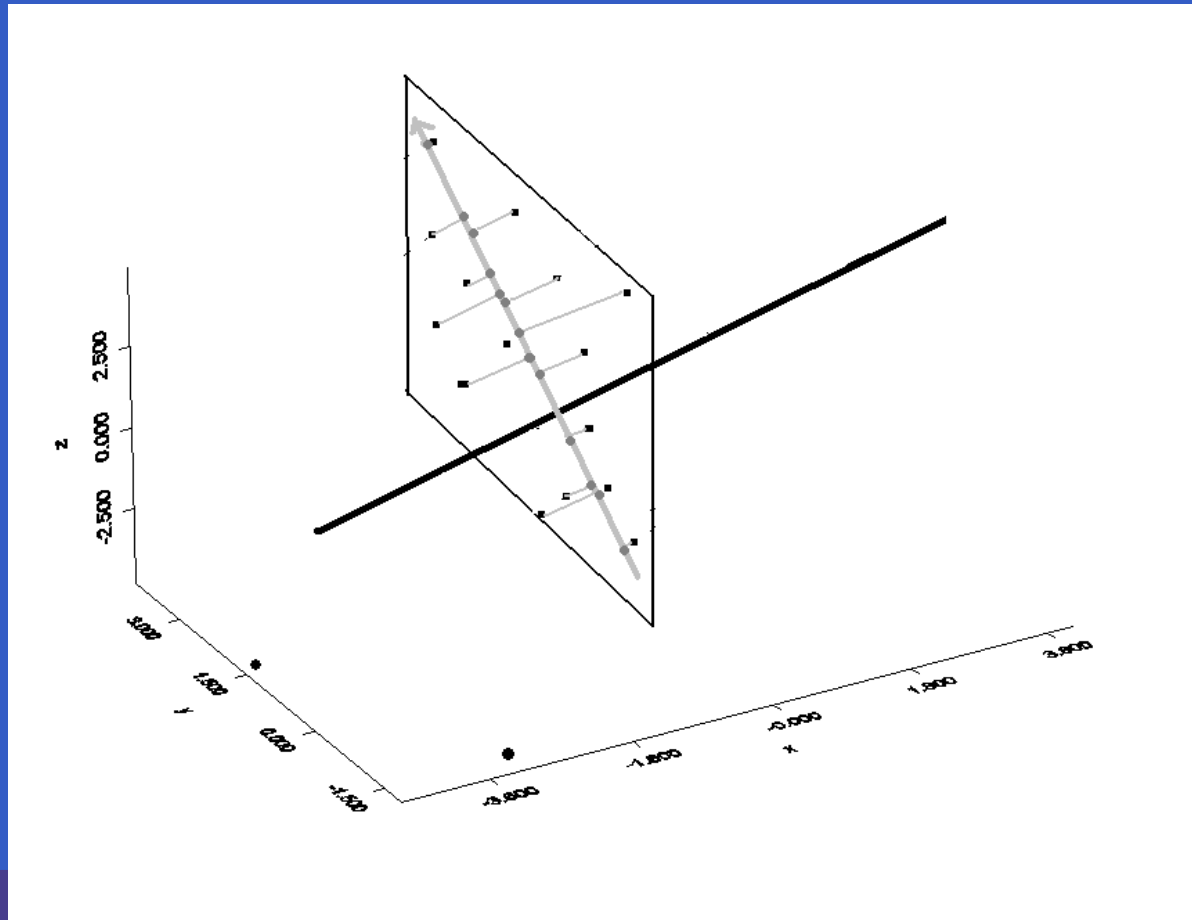
Seconde composante principale

Projection des individus sur le plan perpendiculaire à la première composante :



Seconde composante principale

Deuxième composante : direction selon laquelle le nuage des projections est le plus étendu



Remarques générales sur l'ACP

- Représentation graphique optimisée
 - Réduit la dimension en conservant un maximum d'information
 - Préserve au mieux la disposition originale des points

Remarques générales sur l'ACP

- Représentation graphique optimisée
 - Réduit la dimension en conservant un maximum d'information
 - Préserve au mieux la disposition originale des points
- ACP normalisée ?
 - Quand ? Si les données ont des échelles très différentes
 - Pourquoi ? Pour ne pas donner trop d'importance aux variables qui ont les plus grandes valeurs.

Exemple de réalisation d'une ACP

Les notes d'une classe de collège :

	Maths	Physique	Français	Sc.Nat.	Anglais	Histoire
Elève 1	18,00	13,00	2,00	11,00	9,00	7,00
Elève 2	18,00	14,00	2,00	12,00	8,00	6,00
Elève 3	14,00	11,00	6,00	10,00	11,00	9,00
Elève 4	5,00	8,00	15,00	10,00	14,00	12,00
Elève 5	14,00	14,00	6,00	12,00	8,00	6,00
Elève 6	1,00	0,00	19,00	0,00	20,00	20,00
Elève 7	8,00	6,00	12,00	8,00	16,00	14,00
Elève 8	12,00	10,00	8,00	10,00	12,00	10,00
Elève 9	17,00	13,00	3,00	11,00	9,00	7,00
Elève 10	11,00	12,00	9,00	10,00	10,00	8,00
⋮	⋮	⋮	⋮	...	⋮	⋮
Elève 18	14,00	10,00	6,00	10,00	12,00	10,00

Exemple de réalisation d'une ACP

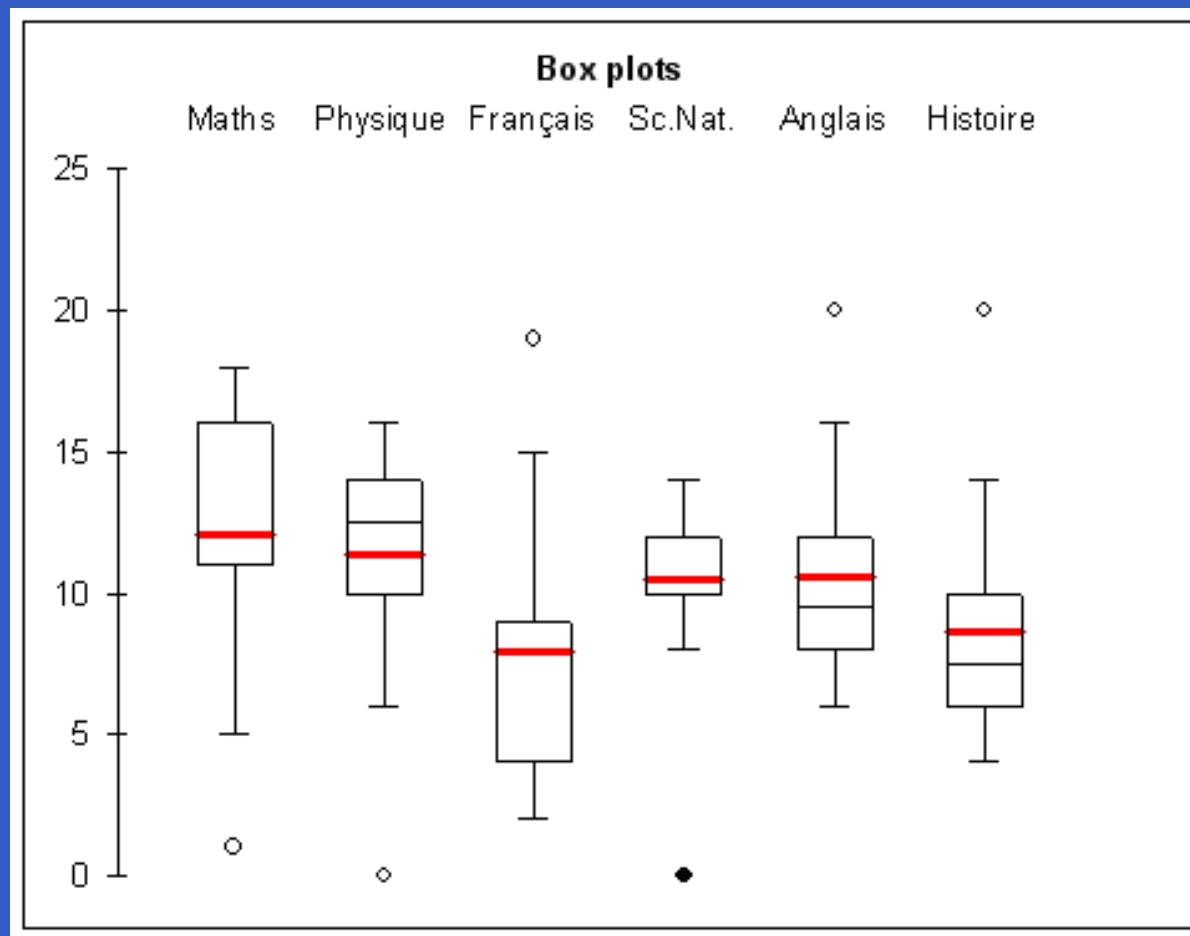
→ Première étape tris à plats :

	Maths	Physique	Français	Sc.Nat.	Anglais	Histoire
Nbr de valeurs utilisées	18	18	18	18	18	18
Nbr de valeurs ignorées	0	0	0	0	0	0
Nbr de val. min.	1	1	2	1	2	2
% de val. min.	5,56	5,56	11,11	5,56	11,11	11,11
Minimum	1,00	0,00	2,00	0,00	6,00	4,00
1er quartile	11,00	10,00	4,00	10,00	8,00	6,00
Médiane	12,00	12,50	8,00	10,50	9,50	7,50
3ème quartile	16,00	14,00	9,00	12,00	12,00	10,00
Maximum	18,00	16,00	19,00	14,00	20,00	20,00
Etendue	17,00	16,00	17,00	14,00	14,00	16,00
Total	218,00	204,00	142,00	188,00	190,00	156,00
Moyenne	12,11	11,33	7,89	10,44	10,56	8,67
Moyenne géométrique	10,57		6,63		10,02	7,94
Moyenne harmonique	7,22		5,40		9,54	7,32
Aplatissement (Pearson)	-0,15	1,26	-0,15	5,08	0,25	1,26
Asymétrie (Pearson)	-0,75	-1,20	0,75	-2,06	0,90	1,20
Aplatissement	0,69	2,82	0,69	8,58	1,29	2,82
Asymétrie	-0,89	-1,43	0,89	-2,46	1,07	1,43
CV (écart-type/moyenne)	0,38	0,35	0,58	0,29	0,34	0,46
Variance d'échantillon	19,65	14,78	19,65	8,69	12,47	14,78
Variance estimée	20,81	15,65	20,81	9,20	13,20	15,65
Ecart-type d'échantillon	4,43	3,84	4,43	2,95	3,53	3,84
Ecart-type estimé	4,56	3,96	4,56	3,03	3,63	3,96
Ecart absolu moyen	3,35	2,96	3,35	1,78	2,84	2,96
Ecart-type de la moy.	1,08	0,93	1,08	0,72	0,86	0,93

Traitements de base du jeu de données Notes

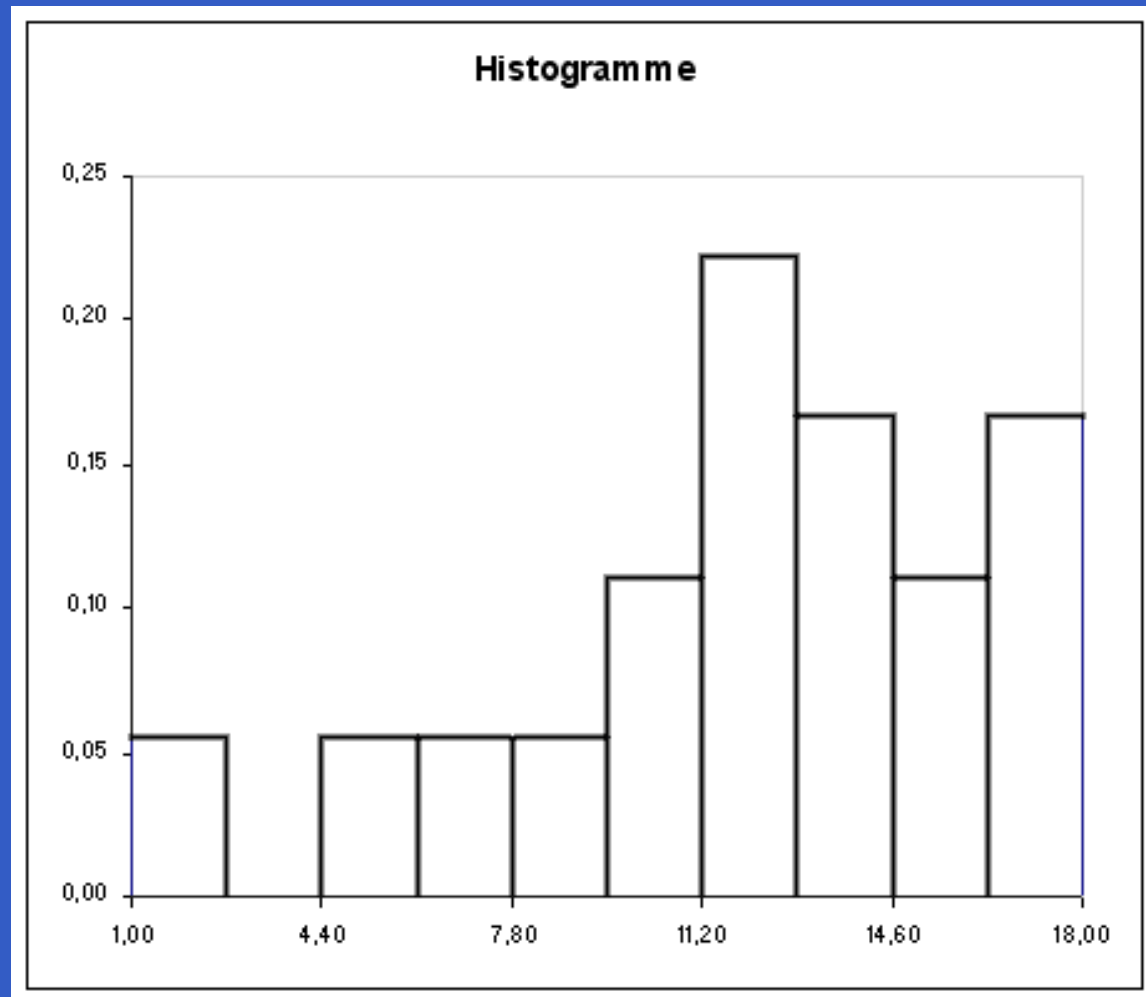
Exemple de réalisation d'une ACP

Représentation graphique globale : Box plots



Exemple de réalisation d'une ACP

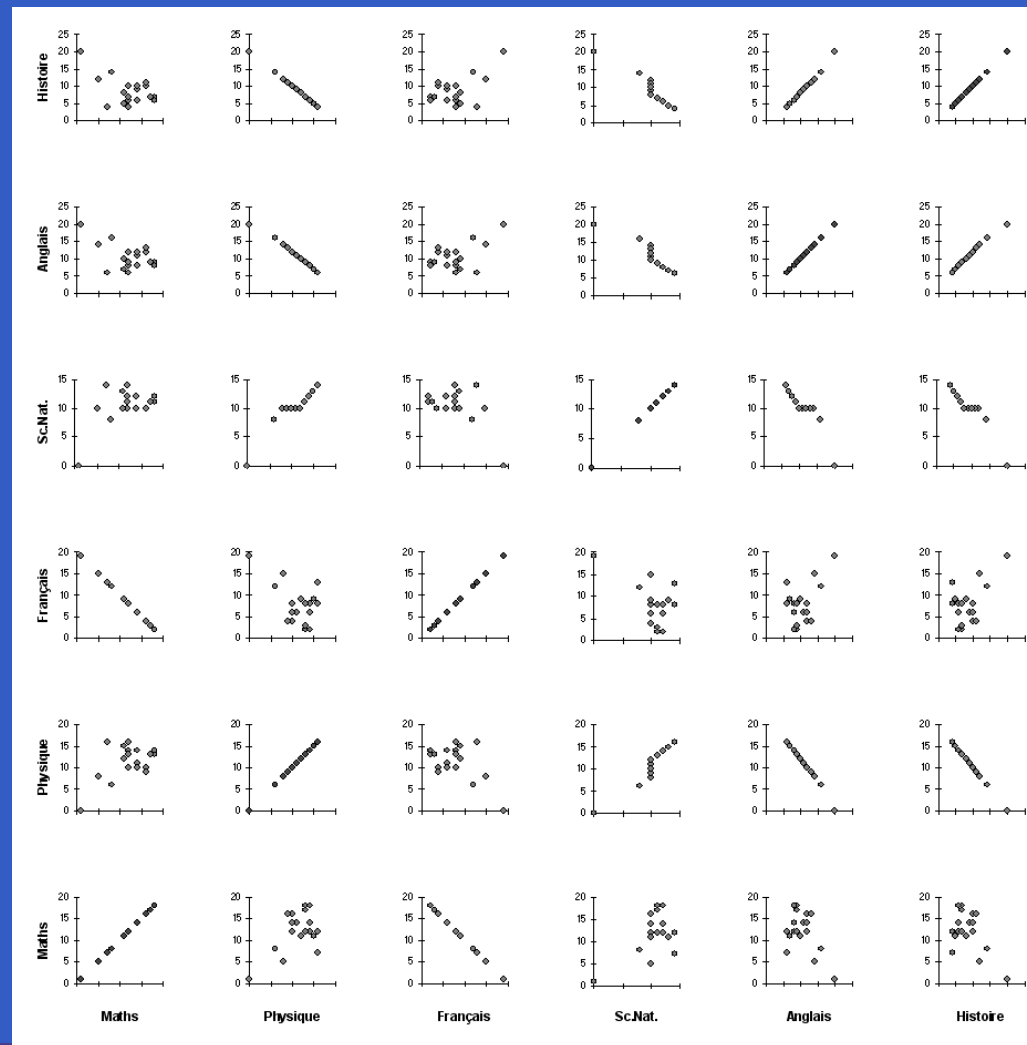
Représentation graphique par variable : histogrammes



Histogramme des notes de Math

Exemple de réalisation d'une ACP

→ Deuxième étape : étude des corrélations



Exemple de réalisation d'une ACP

Etude numérique des corrélations : **Matrice des corrélations**

	Maths	Physique	Français	Sc.Nat.	Anglais	Histoire
Maths	1	0,53	-1,00	0,51	-0,50	-0,53
Physique	0,53	1	-0,53	0,95	-1,00	-1,00
Français	-1,00	-0,53	1	-0,51	0,50	0,53
Sc.Nat.	0,51	0,95	-0,51	1	-0,92	-0,95
Anglais	-0,50	-1,00	0,50	-0,92	1	1,00
Histoire	-0,53	-1,00	0,53	-0,95	1,00	1

→ Beaucoup de liens linéaires :
Entre Math et Français
Entre Physique Sc.Nat, Anglais et Histoire

Exemple de réalisation d'une ACP

Calcul des composantes principales :

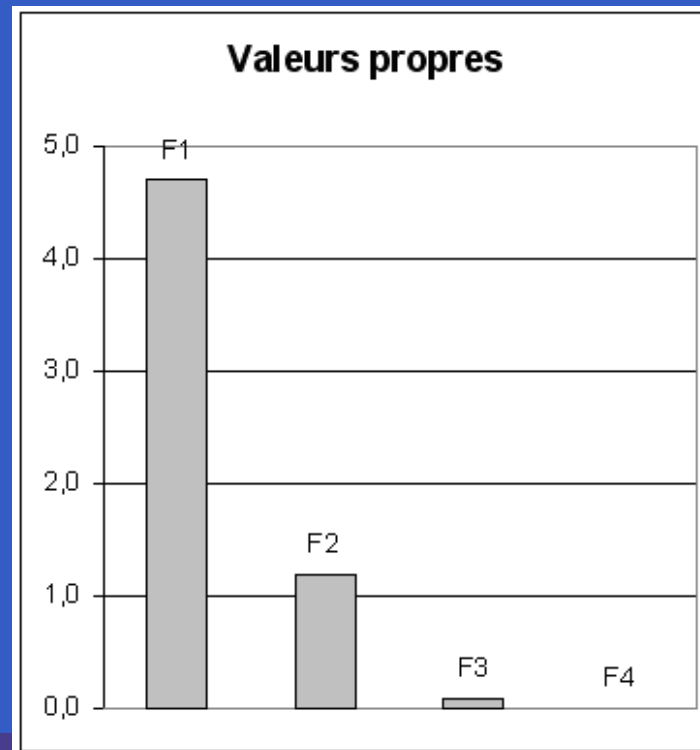
- Les valeurs propres :

	F1	F2	F3	F4
Valeur propre	4,70	1,20	0,09	0,00
% variance	78,41	20,01	1,56	0,02
% cumulé	78,41	98,42	99,98	100,00

Exemple de réalisation d'une ACP

Calcul des composantes principales :

- Les valeurs propres :



Exemple de réalisation d'une ACP

- Les vecteurs propres associés i.e. les composantes principales

	F1	F2	F3	F4
Maths	0,34	0,62	-0,01	-0,01
Physique	0,44	-0,23	-0,19	0,46
Français	-0,34	-0,62	0,01	0,01
Sc. Nat.	0,43	-0,23	0,86	-0,17

Exemple de réalisation d'une ACP

Les deux premiers axes principaux :

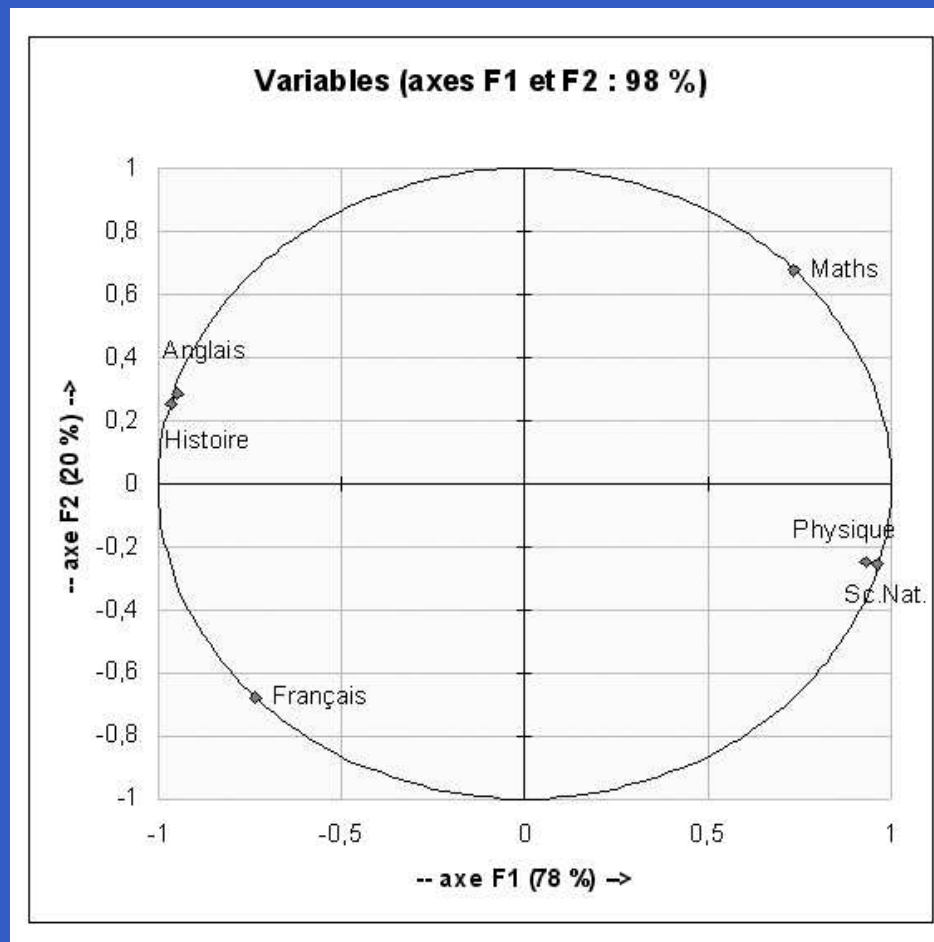
$$F1 = 0.34 * \text{Maths} + 0.44 * \text{Physique} - 0.34 * \text{Français} \\ + 0.43 * \text{Science Nat}$$

$$F2 = 0.62 * \text{Maths} - 0.23 * \text{Physique} - 0.62 * \text{Français} \\ - 0.23 * \text{Science Nat}$$

expliquent 98,42% de la variance : c'est
exceptionnel !

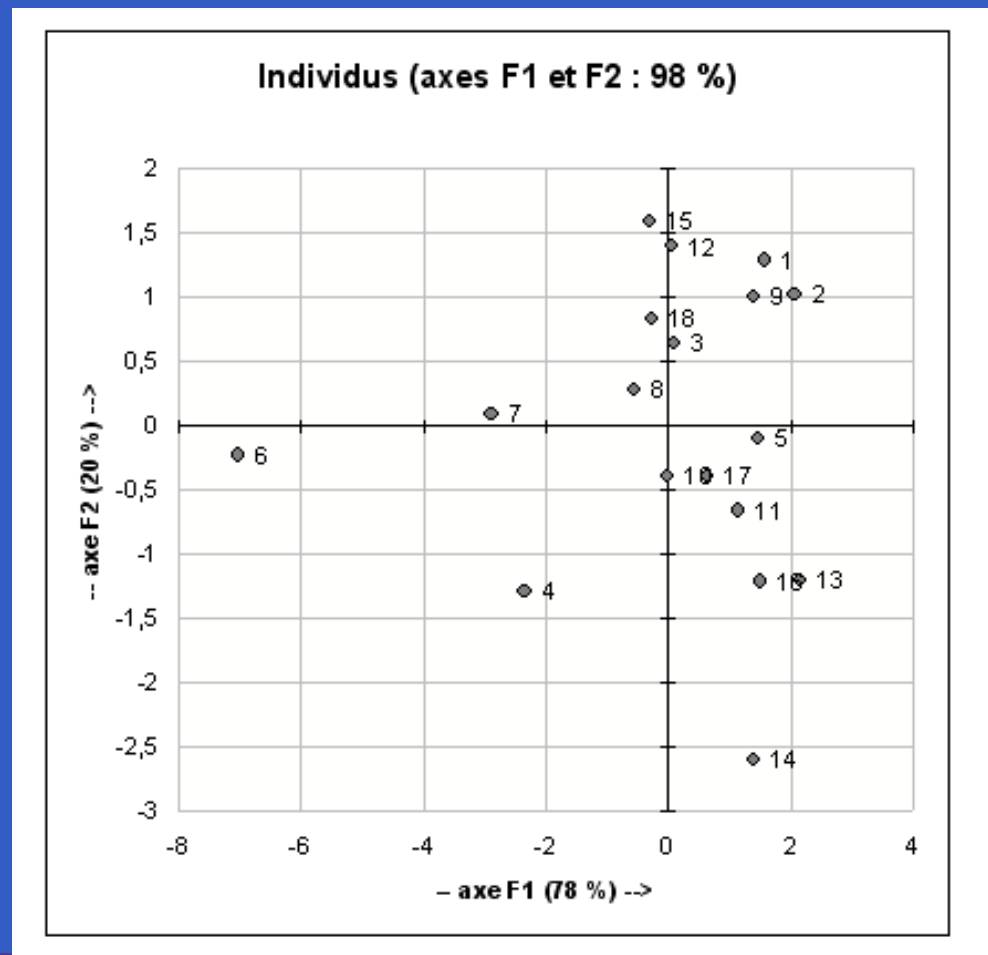
Exemple de réalisation d'une ACP

Représentation des variables dans le plan principal (F1,F2) :



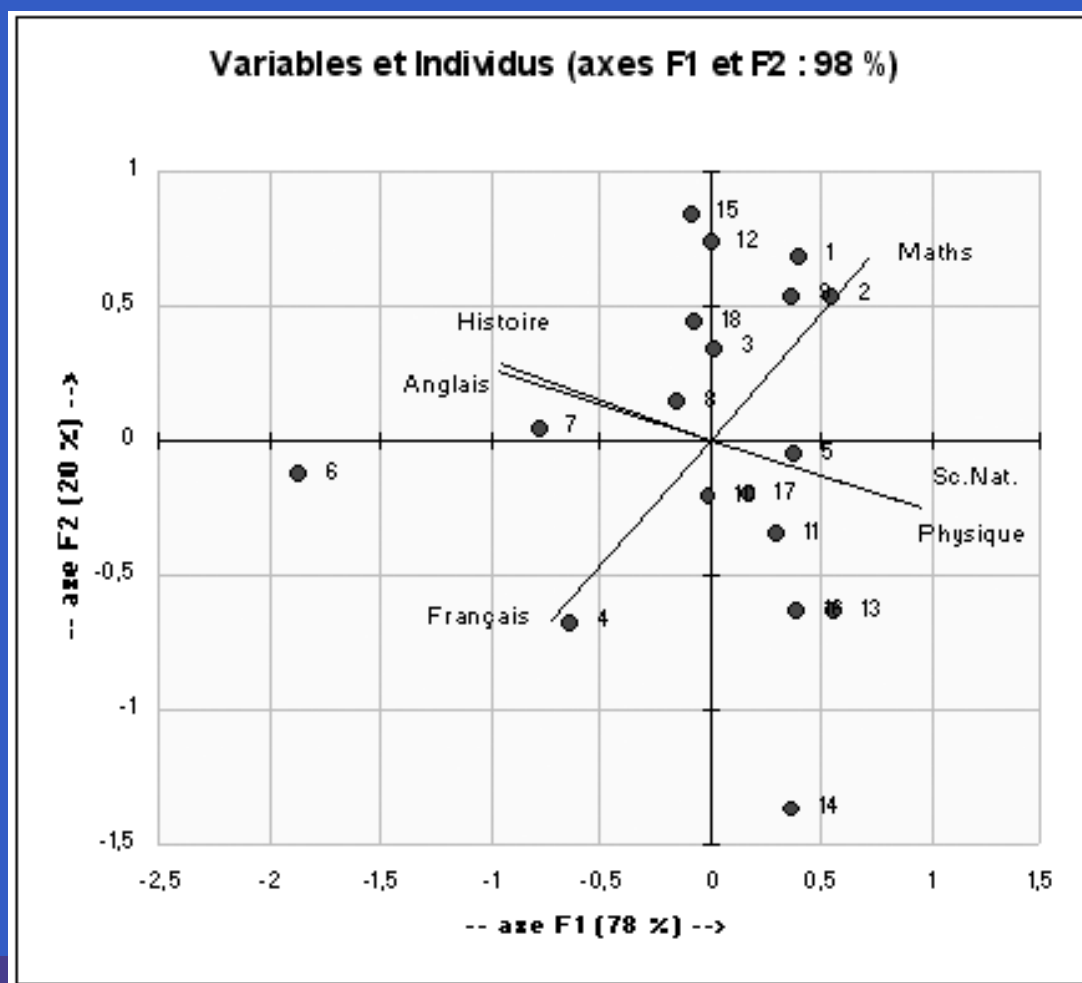
Exemple de réalisation d'une ACP

Représentation des individus sur le premier plan principal:



Exemple de réalisation d'une ACP

Représentation des individus et variables sur le premier plan principal:



Exemple de réalisation d'une ACP

- Graphique très utile pour faire les commentaires
- **Attention** : éviter l'interprétation simultanée variable-individu

Plan du cours

- Introduction
- Traitements préliminaires
- **Traitements statistiques**
 - Analyse uni-variée des variables : tris à plats, répartition, histogrammes,...
 - Analyse bi-variée des variables : tris croisés, corrélations, test du χ^2 ,...
 - Analyse multi-variée : ACP, **AFC**
- Conclusion

Variables qualitatives : AFC

- Adaptation de l'ACP au cas des variables qualitatives

Variables qualitatives : AFC

- Adaptation de l'ACP au cas des variables qualitatives
- Les principales différences entre AFC et ACP
 - AFC : Chaque modalité d'une variable est représentée par un point
 - ACP : Chaque variable est représentée par un point

Variables qualitatives : AFC

- Adaptation de l'ACP au cas des variables qualitatives
- Les principales différences entre AFC et ACP
 - ACP : sa qualité est évaluée sur le pourcentage de la variance restitué
 - AFC : plus délicat, car les variances restituées par les valeurs propres sont sous estimées

Variables qualitatives : AFC

- Adaptation de l'ACP au cas des variables qualitatives
- Les principales différences entre AFC et ACP
 - AFC : Ajout de variables illustratives n'intervenant pas dans le calcul des axes principaux → Aide à l'interprétation

Exemple de réalisation d'une AFC

On a demandé aux élèves de classer les matières par ordre de préférence :

C.Maths	C.Physique	C.Français	C.Sc. Nat.	C.Anglais	C.Histoire
a	b	f	c	d	e
a	b	f	c	d	e
a	b	f	d	c	e
f	e	a	d	b	c
a	b	f	c	d	e
d	e	c	f	a	b
e	f	c	d	a	b
a	c	f	d	b	e
a	b	f	c	d	e
b	a	e	c	d	f
b	a	d	c	e	f
a	c	f	d	b	e
c	a	d	b	e	f
d	a	c	b	e	f
a	e	f	d	b	c
c	a	d	b	e	f
c	a	d	b	e	f
a	e	f	c	b	d

Exemple de réalisation d'une AFC

- Toutes les variables sont qualitatives \implies
AFC

Exemple de réalisation d'une AFC

- Toutes les variables sont qualitatives \implies AFC
- On considère alors toutes les variables suivantes :
 - Math.a, Math.b, Math.c, Math.d, Math.e, Math.f,
 - Phy.a, Phy.b, Phy.c, Phy.e, Phy.f,
 - Fran.a, Fran.c, Fran.d, Fran.e, Fran.f,
 - ScNat.b, ScNat.c, ScNat.d, ScNat.f,
 - Ang.a, Ang.b, Ang.c, Ang.d, Ang.e,
 - Hist.b, Hist.c, Hist.d, Hist.e et Hist.f

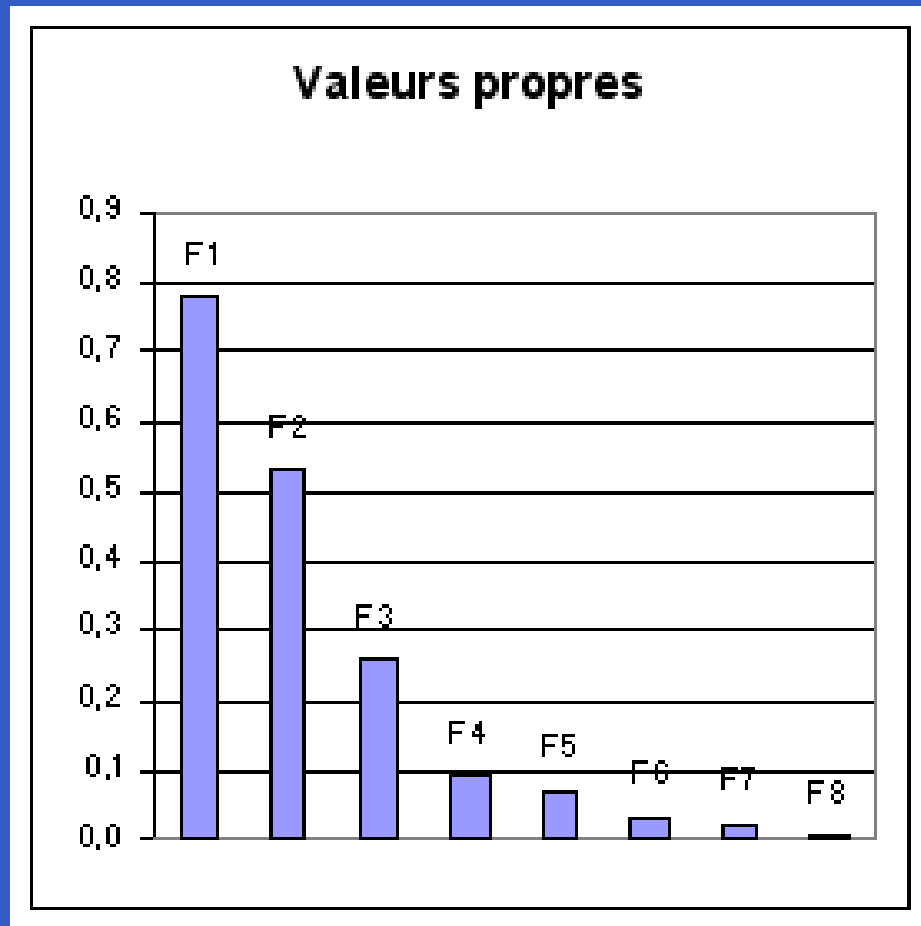
Exemple de réalisation d'une AFC

Etude des valeurs propres :

F1	F2	F3	F4	F5	F6	F7	F8
0,78	0,53	0,26	0,09	0,07	0,03	0,02	0,01
43,72	29,75	14,38	5,13	3,87	1,73	1,10	0,32
43,72	73,47	87,85	92,98	96,85	98,57	99,68	100,00

Exemple de réalisation d'une AFC

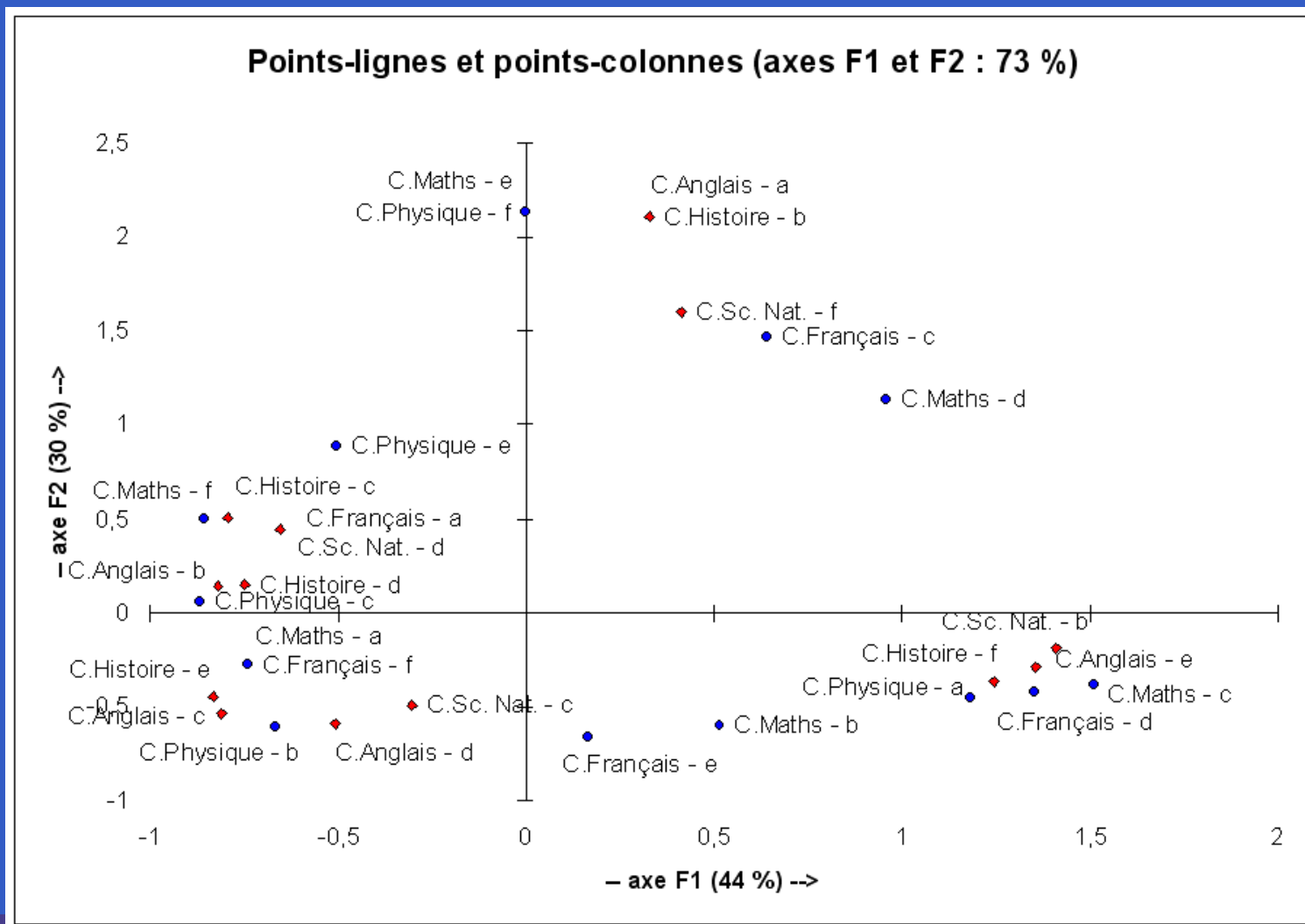
Histogramme des valeurs propres :



Le premier plan principal restitue 73.47% de la variance \implies Il restitue suffisamment d'information pour faire une interprétation des données.

Exemple de réalisation d'une AFC

Représentation des variables sur le plan principal :



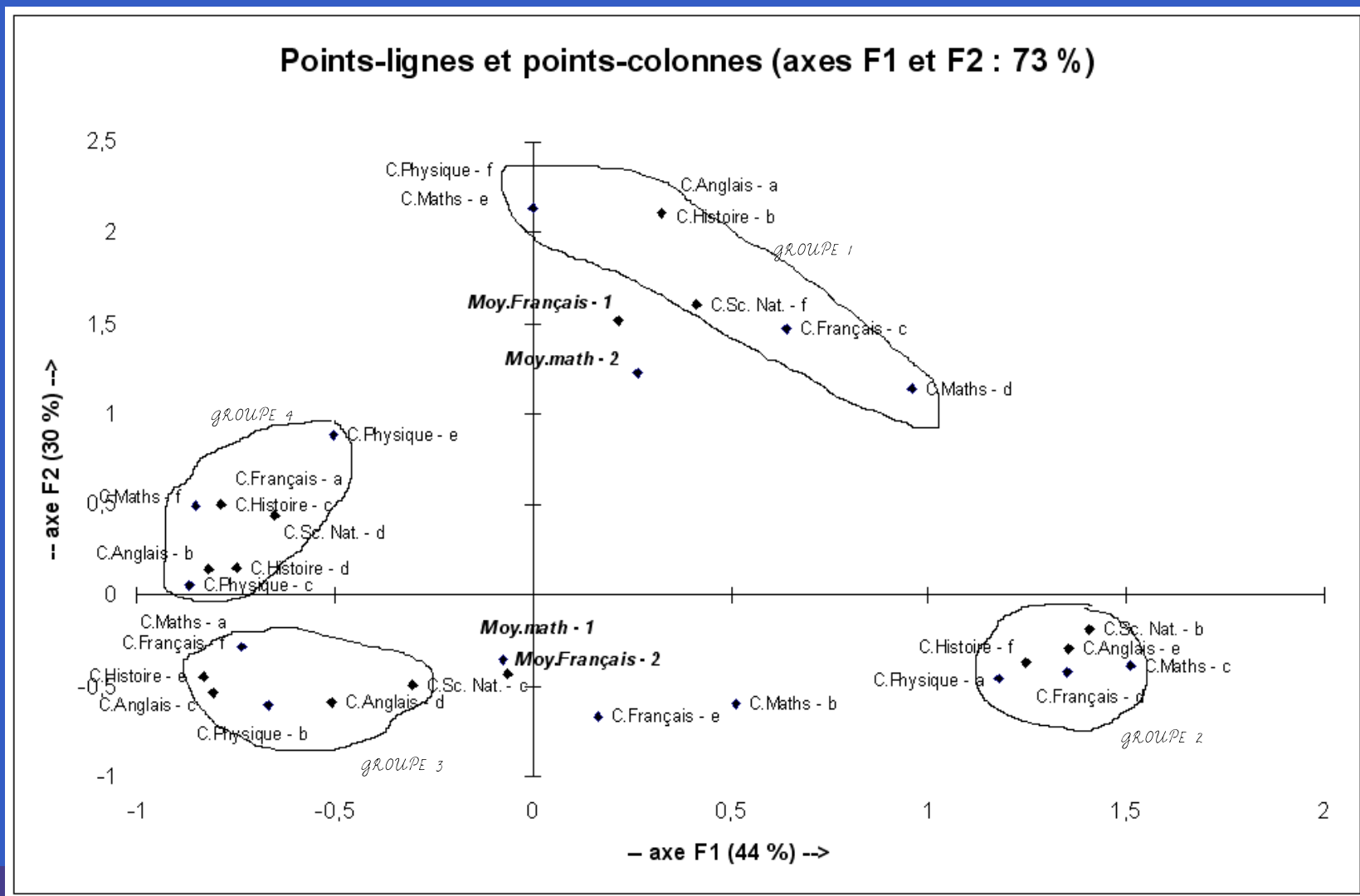
Exemple de réalisation d'une AFC

Ajoutons deux variables illustratives :

- Moy.Fran (1 ou 2)
- Moy.Math (1 ou 2)

Exemple de réalisation d'une AFC

Représentation de ces variables illustratives :



Exemple de réalisation d'une AFC

Interprétations

- Les élèves qui ont la moyenne en Français et pas la moyenne en Math préfèrent les matières littéraires

Exemple de réalisation d'une AFC

Interprétations

- Les élèves qui ont la moyenne en Français et pas la moyenne en Math préfèrent les matières littéraires
- Les élèves qui ont la moyenne en Math et pas la moyenne en Français préfèrent les matières scientifiques

Exemple de réalisation d'une AFC

Interprétations

- Il émerge quatre groupes dans la classe
 - Groupe 1 : Littéraire avec préférence Anglais-Histoire
 - Groupe 2 : Scientifique avec préférence Physique-Sc.Nat
 - Groupe 3 : Scientifique avec préférence Math
 - Groupe 4 : Littéraire avec préférence Français

Conclusion : Traitement d'enquête

- La création du questionnaire est une étape fondamentale
 - Bien adapter les questions au problème que l'on se pose et aux traitements qu'on souhaite leur faire subir.

Conclusion : Traitement d'enquête

- La création du questionnaire est une étape fondamentale
 - Bien adapter les questions au problème que l'on se pose et aux traitements qu'on souhaite leur faire subir.
- Utilisation des outils statistiques généraux
 - fonctionnement facilité pour des données de bonnes qualités

Conclusion : Outils statistiques

Nombre et Nature des variables	Méthodes
1 variable quantitative	histogramme, box plots
1 variable qualitative	diagramme à secteur
2 variables quantitatives	nuage de points
2 variables qualitatives	tableau de contingence
k variables quantitatives	ACP
k variables qualitatives	AFC