Centre **N**ational d'**E**tudes **A**gronomiques des **R**égions **C**haudes

ANALYSE STATISTIQUE POUR LE TRAITEMENT D'ENQUÊTES

Mastère **D**éveloppement **A**gricole **T**ropical

Année 2004-2005, UV **DAT 104**

Stéphanie Laffont Vivien Rossi

UMR ENSAM-INRA
Analyse des Systèmes et Biométrie
rossiv@ensam.inra.fr

Table des matières

Première partie

Statistiques Descriptives pour le traitement d’enquêtes

Introduction

Le champs d'application des statistiques est très vaste. Il couvre la plupart des domaines scientifiques : économie, biologie, agronomie, psychologie, sociologie,... Tous ces domaines possèdent leurs problèmes spécifiques cependant les techniques statistiques employées sont souvent semblables. En effet, quelque soit le domaine considéré, lorsque l'on se pose une question sur un phénomène, on effectue des expériences puis on essaie de trouver une réponse en interprétant les résultats. Le rôle des statistiques est d'extraire les informations intéressantes des résultats afin de faciliter l'interprétation.
Il existe différentes façons d'obtenir des données et différentes façons de les traiter. L'étude statistique effectuée sur un jeu de données dépend essentiellement de la nature des données et de la question à laquelle on souhaite répondre.
L'objectif de ce cours est de présenter des outils permettant de réaliser les traitements statistiques élémentaires d'une enquête. L'aspect théorique des techniques présentées est abordé superficiellement, pour plus de précision on peut se réporter aux ouvrages de la bibliographie, notamment Saporta ([10]) ou Snedecor & Cochran ([12]).
Le cours s'articule autour de deux grandes parties respectant la chronologie des différentes étapes intervenant lors de la réalisation d'une enquête. La première concerne les travaux préliminaires, c'est à dire l'élaboration du questionnaire, le choix de la population à intérroger et la collecte des données. Ces trois étapes, dont chacune relève d'une théorie à part entière, ne constituant pas le coeur du cours sont restreintes à leurs éléments essentiels. Toutefois, il est important de garder à l'esprit qu'elles conditionnent toutes les autres étapes de l'enquête et par conséquence la pertinance des conclusions finales. Plus précisément, si ces trois étapes sont correctement effectuées, la réalisation des autres étapes en sera bien facilitée.
La partie suivante présente quelques outils statistiques utiles pour analyser les questionnaires d'une enquête. Les techniques proposées sont uniquement de type descriptif, c'est à dire qu'elles permettent de synthétiser les informations contenues dans les questionnaires. Il est bien sûr possible d'effectuer des traitements statisitiques plus avancés sur des enquêtes, de type classification ou modélisation.

On peut se reporter à la bibliographie pour obtenir des références présentant ce type de traitements.

Tous les supports utilisés pour réaliser ce document sont référencés dans la bibliographie.

Chapitre 1

Elaboration d'une enquête

1.1 Le questionnaire

Avant la rédaction du questionnaire réalisée souvent trop rapidement, il faut effectuer quelques étapes nécessaires à son efficacité future ; en terme de recueil des données ou de traitements statistiques.

Le problème posé doit être clairement identifié, la nature de la population à étudier doit être clairement définie. Il est également important de connaître la nature exacte des informations à recueillir, et le degré de précision désiré. Sans oublier le nerf de la guerre : le budget. Les différentes phases sont présentées dans ce qui suit.

1.1.1 Etapes d'une enquête par sondage

Pour effectuer une enquête par sondage, il est indispensable de respecter les instructions suivantes :

- Etude du domaine
- Formulation du problème : dresser une liste claire des objectifs de l'enquête.
- Détermination de la population : bien préciser la population à échantillonner.
- Détermination des objectifs : établir, lorsque c'est possible, le degré de précision désiré afin d'analyser le rapport des coûts et des avantages.
- Définition des informations à recueillir : établir une liste précise et courte des données à collecter.
- Choix de l'échantillon : déterminer l'unité de l'échantillonnage (personne physique, collectivité, ...). Etablir le plan de l'échantillonnage ou la méthode de sélection.
- Choix du mode de collecte : téléphone, convocations, visites à domicile, . . .
- Rédaction du projet de questionnaire et guide d'entretien
- Test du questionnaire : faire parfois une pré-enquête courte.
- Rédaction du questionnaire final et du guide d'entretien final

L'étude du domaine permet le recueil d'informations sur le domaine à étudier, elle peut se faire de diverses façons : recherches documentaires, réflexion individuelle ou collective, étude sur le même domaine ...(les options possible des décisions à prendre, les informations à connaître pour éclairer la décision, les hypothèses a priori).
Formuler correctement le problème posé permet de déterminer plus clairement les objectifs de l'étude et donc de définir les informations à recueillir.

La collecte des données

Nous présentons sommairement ci-dessous différents modes de collecte sont possibles.

• Entretien face à face

L 'enquêteur propose des thèmes de discussion, et enregistre le discours. Il en ressort une précision sémantique mais un coût énorme, donc un échantillon très limité.

• Questionnaire par enquêteur

L'enquêteur pose les questions et note les réponses, il s'agit d'un face à face au domicile de l'enquêté, dans la rue, dans un lieu spécialisé, par téléphone. L'enquêteur peut ainsi vérifier la bonne compréhension des questions et le taux de réponses peut être élevé mais l'approche de l'enquêté est parfois délicate, et l'influence des conditions d'enquête est importante (lieu et heure de l'entrevue, réaction de la personne, suggestion des réponses lorsque l'enquêteur explique une question mal comprise par l'enquêté). Ce mode de collecte reste cher.

• Dépouillement de document

L'enquêteur répond aux questions en fonction d'informations disponibles dans un dossier représentant l'enquêté. L'enquêteur a une bonne compréhension des questions et le taux de réponses est élevé, mais le problème de la fiabilité des informations contenues dans le dossier est indéniable, ainsi que l'interprétation des informations par l'enquêteur. Les informations sont limitées au contenu du dossier.

Rédaction du questionnaire

La rédaction du questionnaire peut être alors faite, le questionnaire ne doit pas être trop long pour ne pas lasser l'enquêté, les questions doivent être claires, le vocabulaire utilisé simple. C'est dans cette optique qu'il n'est pas inutile de tester ce questionnaire sur un petit échantillon, afin d'en faire ressortir les lacunes. Un manuel expliquant l'utilisation du questionnaire et le type de réponses attendues et sa finalité, permet à l'enquêteur de limiter le taux d'erreurs et de non réponses. L'analyse statistique des questions ne doit pas être perdue de vue, réponse quantitative pour des moyennes, analyse linéaire, des ACP, tableau de fréquence, AFC pour des variables qualitatives.

1.2 Méthodes d'échantillonnage

L'étude exhaustive d'un caractère donné dans une population est un recensement. Elle se heurte souvent à une impossibilité matérielle : coût trop élevé, ou destruction des individus étudiés. Les méthodes d'analyse quantitative ont alors recours à la théorie des sondages, qui consiste à étudier un sous-ensemble de la population qu'on appelle un échantillon. La théorie des sondages pose deux types de problèmes :

- L'échantillon doit être représentatif de la population : c'est la théorie de l'échantillonnage.
- Les techniques numériques utilisées sur les observations expérimentales doivent conduire à des résultats fiables, c'est-à-dire donnant une bonne représentation des paramètres inconnus de la population : c'est la théorie de l'estimation et des tests.

Les deux problèmes sont liés : la méthode d'échantillonnage utilisée a une influence sur les estimations obtenues.

1.2.1 Avantages de la méthode d'enquêtes par sondages

La méthode d'enquêtes par sondages présente sur le recensement (lorsqu'il est possible) les avantages suivants :

1. Coût plus réduit.
2. Plus grande vitesse d'exécution (notamment pour les sondages d'opinions).
3. Plus grande fiabilité des résultats : le personnel étant plus réduit, il peut être plus qualifié.
4. Moins de risque d'erreur : le volume des données à traiter est plus faible.
5. Plus grand champ d'application, notamment dans le cas de destruction des unités testées.

1.2.2 Divers types de sondages

Pour effectuer un sondage dans une population, c'est-à-dire pour en extraire un échantillon, deux types de méthodes sont employées : méthodes empiriques et méthodes aléatoires. Seules les méthodes aléatoires permettent d'utiliser la théorie de l'estimation.

Méthodes empiriques : sondages raisonnés

Ce sont les plus connues du grand public et les plus utilisées par les instituts de sondage d'opinion. La précision de ces méthodes ne peut être calculée et leur réussite n'est que le résultat d'une longue pratique et de l'habileté professionnelle. Les éléments sondés sont choisis dans la population suivant des critères fixés a priori.

Méthode des unités types : Elle repose sur l'idée suivante, les différentes variables attachées à un individu de la population n'étant pas indépendantes, un individu qui se trouve dans la moyenne de la population pour un certain nombre de caractères importants, sera également peu différent de la moyenne pour les autres caractères.
La méthode consiste donc à diviser la population en un certain nombre de sous-ensembles relativement homogènes et à représenter chacun d'eux par une unité-type.
On choisit donc des unités d'individus que l'on considère comme fortement représentatives de certaines catégories de population : cantons-types, bureau de vote pilotes, dont les résultats observés sur de longues périodes figurent les résultats définitifs d'une région ou d'une ville, etc.

Exemple 1.1 *L'INSEE décomposa en 1942 la France en 600 régions agricoles et, dans chaque région, désigna un canton-type.*
Comme il y a en France environ 3000 cantons, la désignation de 600 cantons-types permettait de réduire d'un facteur 5 l'ampleur d'une étude des cantons.

Méthode des quotas : L'enquêteur prélève librement son échantillon, à condition de respecter une composition donnée à l'avance (pourcentage fixé d'agriculteurs, d'ouvriers, de cadres, etc., par exemple).
Cette méthode est facile, mais aucun intervalle de confiance ne peut être donné. Elle suppose implicitement que les catégories retenues pour la détermination des quotas sont pertinentes quant à l'objet de l'étude, ce qui est bien difficile à établir. Pour diminuer l'arbitraire du choix, on impose à l'enquêteur des normes de déplacement géographique : c'est la méthode de Politz. On utilise souvent des "panels", qui sont des échantillons permanents dont on étudie l'évolution.

Exemple 1.2 *Quelques panels souvent utilisés :*

- *Panel d'audience à la télévision : médiamétrie, centres d'études d'opinion,. . .*
- *Panel de consommateurs (SECODIF : 4 500 ménages).*
- *Panel de détaillants (SOFRES).*

Ces panels sont utilisés en marketing (lancement d'un produit, transfert de marques, etc.).

Méthodes Probabilistes

Ces méthodes autorisent l'emploi des techniques statistiques pour analyser les résultats.

Méthodes aléatoires : Les éléments sondés sont extraits au hasard d'une liste connue a priori de la population, appelée base de sondage.

Exemple 1.3 *Quelques bases de sondages :*

1. *Liste d'immatriculation des véhicules automobiles en France.*
   C'est une très bonne base car elle est mise à jour régulièrement (cartes grises neuves, cartes grises à détruire).
2. *Répertoire des entreprises (SIREN).*
   Chaque entreprise possède un numéro d'immatriculation à neuf chiffres, un nom ou raison sociale, une adresse exacte.
3. *L'annuaire téléphonique est une mauvaise base de sondage car d'une part, tout individu ne possède pas obligatoirement un téléphone et, d'autre part, un individu peut posséder un téléphone et ne pas figurer sur l'annuaire (la liste rouge représente environ 8 % des abonnés et l'annuaire ne recense pas les téléphones portables, soit environ 40 % des téléphones).*

Les bases de sondages sont en général établies à partir des résultats d'un recensement et elles sont corrigées périodiquement entre deux recensements.
Le tirage de l'échantillon est effectué dans la base de sondage selon des critères spécifiques à chaque méthode (plan de sondage).
Cette méthode de travail ne laisse aucune initiative aux enquêteurs : il est très simple de contrôler leur travail.

Sondage élémentaire : échantillon aléatoire simple. Dans un échantillon aléatoire simple, les éléments constituant l'échantillon sont extraits au hasard (à l'aide d'une table de nombres au hasard, par exemple) d'une liste de la population.
On extrait ainsi n individus d'une population de taille N.
Le tirage peut s'effectuer avec ou sans remise, renvoyant ainsi généralement à un modèle de loi binomiale (avec remise), ou hypergéométrique (sans remise).
Si le tirage s'effectue avec remise, l'échantillon aléatoire simple est dit indépendant (EASI = Echantillon Aléatoire Simple et Indépendant).
La méthode permet notamment de calculer des intervalles de confiance, comme nous le verrons plus loin.

Définition 1.1 *Le taux de sondage f est défini par le rapport*

$$f = \frac{n}{N}$$

Par exemple, l'INSEE utilise des taux de sondage de l'ordre de $1/1500$ pour les enquêtes sur les conditions de vie des ménages.

Exemple 1.4 *Nous voulons extraire un échantillon de 8 individus dans une population formée de 437 individus.*
Nous numérotons les individus de la population de 1 à 437.
Nous considérons trois colonnes consécutives d'une page de nombres au hasard : ils forment des nombres au hasard à trois chiffres.
Nous lisons ces nombres de trois chiffres en ne retenant que ceux qui sont compris entre 001 et 437.
Lorsque nous avons retenus 8 nombres, notre échantillon est constitué des 8 individus désignés dans la population par ces huit nombres.
Selon que nous effectuons un tirage avec ou sans remise, nous garderons ou écarterons un individu déjà tiré.

L'inconvénient majeur de la méthode élémentaire est son coût : les individus tirés peuvent être très éloignés géographiquement.

Sondage stratifié : La population étudiée W est partitionnée en q sous-populations $W_1, W_2, \ldots, W_q$, appelées "strates".

L'échantillon est constitué de la réunion de q échantillons choisis au hasard, un par strate : nous effectuons dans chaque strate un échantillonnage simple.

Exemple 1.5 $W = \{1, 2, 3, 4, 5\}$, $W_1 = \{1, 2\}$, $W_2 = \{3, 4, 5\}$.
Nous sélectionnons trois individus, dont un dans W_1 et deux dans W_2.
Nous obtenons l'un des six échantillons possibles.

Cette méthode se justifie par deux raisons essentielles :

$1-$ L'existence d'une stratification de fait, soit pour des raisons géographiques, soit pour des raisons administratives.

Exemple 1.6 *Enquête sur les conditions de vie pénitentiaire en France.*
La population est celle des détenus en France
Les strates sont les populations de détenus dans les divers établissements pénitentiaires.

Enquête sur la consommation par un organisme disposant de bureaux départementaux.
La population est celle des consommateurs français.
Les strates sont les consommateurs de chaque département.

$2-$ Un caractère étudié dans la population peut varier sous l'influence d'un certain nombre de facteurs.
Pour éliminer au mieux les risques de biais, nous créons des strates homogènes et, dans chacune d'elle, nous extrayons un échantillon aléatoire simple.

Exemple 1.7 *Pour étudier la consommation de tabac, si nous estimons que l'âge et le sexe sont des facteurs très influents, nous partageons la population en strates du type :*

- Hommes de moins de 20 ans,
- Hommes de 20 à 30 ans,
- etc.
- Femmes de moins de 20 ans,
- Femmes de 20 à 30 ans,
- etc.

De chaque strate, nous extrayons un échantillon aléatoire simple.

Echantillonnage systématique : les individus de la population W sont numérotés de 1 à N.

Pour sélectionner n individus, nous partageons la population en k groupes : 1, ... , k , 1 + k, ... , 2 k , ... , 1 + (n - 1) k, ... , N .
Nous choisissons au hasard l'individu i par les individus numérotés de 1 à k. Nous constituons notre échantillon des individus i, i + k, i + 2 k, ... , i + (n - 1) k .
Le choix de l'individu i détermine entièrement la constitution de l'échantillon.

Exemple 1.8 $W = \{1, ..., 20\}$, $k = 4$.
Les échantillons possibles sont : $\{1, 5, 9, 13, 17\}$, $\{2, 6, 10, 14, 18\}$, $\{3, 7, 11, 15, 19\}$, $\{4, 8, 12, 16, 20\}$.

Cette méthode est bien adaptée à la sélection de cartes dans un fichier, ou au prélèvement de pièces dans une fabrication pour un contrôle de qualité.
Elle présente une certaine analogie avec la méthode précédente d'échantillonnage stratifié.

Echantillonnage à plusieurs degrés : la population W est divisée en sous-populations appelées unités primaires. Chaque unité primaire est divisée en unités secondaires, etc.
Nous effectuons des tirages au hasard en cascade : nous tirons des unités primaires ; dans chaque unité primaire, nous tirons une unité secondaire, etc.

Exemple 1.9 *L'INSEE effectue des échantillonnages à quatre niveaux : départements, cantons, communes, ménages.*

Cette méthode permet une exécution rapide. Elle est économique, car elle focalise les tirages.
La méthode de tirage au hasard à chaque niveau peut varier suivant le cas, par exemple tirage proportionnel aux unités qu'il contient, ou tirage équiprobable. Nous disons alors que nous pouvons avoir des tirages avec probabilités inégales.
Cas particulier : tirage par grappes. Nous choisissons des grappes pour lesquelles nous gardons tous les "grains", ou individus. Une "grappe" est un groupe d'individus de même nature.

Exemple 1.10 *ménages d'un même immeuble.*

Chapitre 2

Traitements Statistiques d'une enquête

Dans ce chapitre, nous supposons être en possession de questionnaires d'une enquête. Nous notons n le nombre de questionnaires remplis et q le nombre de questions sur le questionnaire. Les données peuvent donc se présenter sous la forme du tableau suivant :

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

où $r_{i,j}$ désigne la réponse à la $i^{\text{ème}}$ question sur le $j^{\text{ème}}$ questionnaire, avec $i = 1, \ldots, q$ et $j = 1, \ldots, n$.
La $j^{\text{ème}}$ colonne de ce tableau regroupe ainsi toute les réponses du $j^{\text{ème}}$ questionnaire.
La $i^{\text{ème}}$ ligne de ce tableau regroupe toutes les réponses à la $i^{\text{ème}}$ question.

A chacune des questions Q_i, $i = 1, \ldots, q$, du questionnaire, on associe la réponse R_i qui est appelée variable. Par exemple, si Q_5= Quelle est la taille de la personne ?, alors la variable associée est R_5= La taille de la personne.
Ainsi la ligne i du tableau rassemble n observations de la variable R_i pour $i = 1, \ldots, q$.

Tout d'abord différencions les deux types de variables les quantitatives et les qualitatives :

Définition 2.1 *X est une variable quantitative (ou variable continue) si ses valeurs observées, $x_1, \cdots, x_n$ appartiennent* $\mathbb{R}$.

Exemple 2.1 *La variable X = la superfie d'un terrain ou la taille d'une personne ou le poids d'une récolte,... est quantitative.*

Définition 2.2 *X est une variable qualitative (ou variable discrète) si ses valeurs observées, $x_1, \cdots, x_n$ appartiennent à $\{m_1, \ldots, m_k\}$ où m_i, $i = 1, \ldots, k$ est une modalité.*

Exemple 2.2 *Q= Quelle est la couleur de votre voiture ?*

- ☐ *Bleu*
- ☐ *Grise*
- ☐ *Rouge*
- ☐ *Verte*
- ☐ *Autres*

La variable de réponse R peut prendre seulement cinq valeurs différentes, c'est donc une variable qualitative avec les cinq modalités : Bleu, Grise, Rouge, Verte, Autres.

Les traitements statistiques vont portés sur les lignes de ce tableau, donc sur les variables $R_1, \ldots, R_q$. Dans un premier temps, on effectue des tris à plats, c'est à dire que l'on étudie la répartion des valeurs pour chacune des variables. Plus concrètement, on étudie la répartion des réponses à chacune des questions. Bien entendu, la nature de l'étude change suivant que les réponses soient quantitatives ou qualitatives (voir def 2.1 et 2.2).
Dans un deuxième temps, on effecute des tris croisés, c'est à dire que l'on étudie la répartion conjointe des valeurs de variables. Cette étude est réalisée afin de caractériser des liens éventuels entre les réponses de différentes questions. Toujours dans cet optique, nous présentons le test du χ^2 qui permet de tester l'indépendance entre les variables (i.e. entre les réponses de différentes questions). Comme pour les tris à plats, les techniques utilisées varient suivant la nature quantitative ou qualitative des réponses.
La dernière phase de traitement consiste à étudier globalement toutes les variables simultanément (i.e. toutes les colonnes simultanément) afin de dégager des tendances générales. A cette fin, on utilise les techniques statistiques d'analyse de données. Tout d'abord, on effectue une Analyse en Composante Principale (ACP) sur les variables qunatitatives (i.e. les lignes des réponses quantitatives). Ensuite, on pratique une Analyse Factorielle des Correspondances (AFC) sur l'ensemble des variables (i.e. des réponses quantitatives et qualitatives).

2.1 Analyse univariée ou Tris à plats des variables

Afin de simplifier les notations, nous travaillons avec n observations $x_1, \ldots, x_n$ d'une variable X, qui correspondent aux n réponses $r_{i1}, \ldots, r_{in}$ d'une question Q_i.
Nous présentons les différentes techniques couramment utilisées pour résumer l'information contenue dans l'échantillon $x_1, \ldots, x_n$ de n observations d'une variable X.

2.1.1 Cas d'une variable quantitative

Afin de fournir un bon aperçu de l'information contenue dans des données, il faut donner des indicateurs numériques et des représentations graphiques.

Traitements numériques

La première information importante à fournir est une idée de la valeur centrale des données de l'échantillon. On peut calculer plusieurs quantités à cet effet, nous rappelons les définitions des deux qui sont le plus utilisées.

Définition 2.3 *La Moyenne $\bar{x}$ de $x_1, \ldots, x_n$ est définie par*

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Exemple 2.3 *Soit $x_1 = 12,\ x_2 = 14,\ x_3 = 8,\ x_4 = 10,\ x_5 = 13$ les notes d'un examen. La moyenne est*

$$\bar{x} = \frac{1}{5}(12 + 14 + 8 + 10 + 13) = 11.4$$

Un des défauts de la moyenne est sa sensibilité aux valeurs extrêmes ou aberrantes. Il suffit par exemple d'une seule valeur anormalement élevée pour que la moyenne ne soit plus représentative des valeurs généralement observées. Une autre quantité lui est donc souvent préférée, la médiane. C'est la valeur de l'échantillon $x_1, \ldots, x_n$ telle qu'il y a autant de valeurs plus grandes et autant de valeurs plus petites au sein de l'échantillon.

Définition 2.4 *Soit $x_{(1)}, \ldots, x_{(n)}$ l'échantillon $x_1, \ldots, x_n$ ordonné.*
La médiane M_e de $x_1, \ldots, x_n$ est

$$M_e = \begin{cases} \dfrac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{si } n \text{ est pair} \\ \\ x_{(E[n/2]+1)} & \text{si } n \text{ est impair} \end{cases}$$

où $E[n/2]$ désigne la partie entière de $n/2$.

Exemple 2.4 *Soit $x_1 = 12,\ x_2 = 14,\ x_3 = 8,\ x_4 = 10,\ x_5 = 13$ les notes d'un examen. L'échantillon ordonné est $x_{(1)} = 8,\ x_{(2)} = 10,\ x_{(3)} = 12,\ x_{(4)} = 13,\ x_{(5)} = 14$ et la médiane est*

$$M_e = x_{(3)} = 12$$

Si on rajoute une note, $x_1 = 12,\ x_2 = 14,\ x_3 = 8,\ x_4 = 10,\ x_5 = 13,\ x_6 = 11$. L'échantillon ordonné devient $x_{(1)} = 8,\ x_{(2)} = 10,\ x_{(3)} = 11,\ x_{(4)} = 12,\ x_{(5)} = 13,\ x_{(6)} = 14$ et la médiane devient

$$M_e = \frac{x_{(3)} + x_{(4)}}{2} = \frac{11 + 12}{2} = 11.5$$

Une fois qu'on a un indicateur de la valeur centrale de l'échantillon, il faut préciser la dispersion des observations de l'échantillon autour de cette valeur. Lorsque l'indicateur de la valeur centrale est la moyenne, l'indicateur de dispersion utilisé est l'écart-type. Mais avant de pouvoir calculer l'écart-type, il faut calculer la variance de l'échantillon.

Définition 2.5 *La variance de l'échantillon $x_1, \ldots, x_n$ est*

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

La variance donne une idée de la dispersion des valeurs de l'échantillon puisque c'est la moyenne des carrés des écarts à la moyenne. Cependant la variance n'est pas à l'échelle des observations car on considère le carré des écarts. Pour se ramener à l'échelle des observations on calcule donc la racine carrée de la variance ce qui nous conduit à l'écart-type.

Définition 2.6 *L'écart-type de l'échantillon $x_1, \ldots, x_n$ est*

$$\sigma = \sqrt{V}$$

Exemple 2.5 *Soit $x_1 = 12,\ x_2 = 14,\ x_3 = 8,\ x_4 = 10,\ x_5 = 13$ les notes d'un examen. La moyenne est $\bar{x} = 11.4$, la variance est*

$$\begin{aligned} V &= \tfrac{1}{5}\Big[(12-11.4)^2 + (14-11.4)^2 + (8-11.4)^2 + (10-11.4)^2 \\ &\quad +(13-11.4)^2\Big] \\ &= 4.64 \end{aligned}$$

et l'écart-type est

$$\sigma = \sqrt{4.64} \approx 2.15$$

Comme l'écart-type dépend de la moyenne, il peut pâtir de ses faiblesses. Les autres indicateurs généralement employés complètent les informations apportées par la médiane. On considère les valeurs minimales et maximales de l'échantillon ainsi que les quartiles.

Définition 2.7 *Le premier quartile Q_1 de $x_1, \ldots, x_n$ est la valeur de $x_1, \ldots, x_n$ telle qu'un quart des observations de $x_1, \ldots, x_n$ lui soient inférieures et trois quarts lui soient supérieures.*
Le deuxièmee quartile de $x_1, \ldots, x_n$ est la médiane.
Le troisème quartile Q_3 de $x_1, \ldots, x_n$ est la valeur de $x_1, \ldots, x_n$ telle que trois quarts des observations de $x_1, \ldots, x_n$ lui soient inférieures et un quart lui soient supérieures.

Exemple 2.6 *Soient 10, 12, 7, 14, 11, 8, 9, 15, 5, 12, 10.5, 11, 14, 8, 16 les notes d'un examen. Dans la plupart des logiciels statistiques, tous les indicateurs évoqués ci-dessus sont présentés sous la forme du tableau suivants :*

Min	Q_1	*Médiane*	Q_3	*Max*
5	*8.5*	*11*	*13*	*16*

Le premier quart des notes est donc compris entre 5 et 8.5, le second quart est compris entre 8.5 et 11, le troisème quart est compris entre 11 et 13, le quatrième quart est compris entre 13 et 16.

Traitements graphiques

Les représentations graphiques d'un échantillon sont complémentaires aux indicateurs numériques car elles permettent une vision globale mais imprécise des données.
La représentation graphique généralement utilisée pour des données quantitatives est l'histogramme. Il faut au préalable définir des classes $C_1, \ldots, C_k$ partitionnant le domaine d'existance de $x_1, \ldots, x_n$. Ensuite, il suffit de compter le nombre d'élément de $x_1, \ldots, x_n$ apartenant à chacune des classes.

Définition 2.8 *L'histogramme associé à l'échantillon $x_1, \ldots, x_n$ et aux classes $C_1, \ldots, C_k$ est la fonction h définie par*

$$h(x) = \begin{cases} \frac{n_i}{mes(C_i)} & \text{si } x \in C_i, i=1,\ldots,k \\ 0 & \text{sinon} \end{cases}$$

où $mes(C_i)$ est la longueur de la classe C_i et n_i est le nombre d'élément de $x_1, \ldots, x_n$ appartenants à la classe C_i

Exemple 2.7 *Soient 10, 12, 7, 14, 11, 8, 9, 15, 5, 12, 10.5, 11, 14, 8, 16 les notes d'un examen. Si on prend les classes suivantes* $C_1 = [3\ 6[$, $C_2 = [6\ 9[$, $C_3 = [9\ 12[$, $C_4 = [12\ 15[$, $C_5 = [15\ 18[$, *alors on a* $mes(C_1) = mes(C_2) = \cdots = mes(C_5) = 3$, $n_1 = 1$, $n_2 = 3$, $n_3 = 5$, $n_4 = 4$ *et* $n_5 = 2$. *Comme toutes les classes ont la même taille, il n'est pas nécessaire de diviser les effectifs des classes par leur taille. On obtient ainsi le graphique (2.1).*

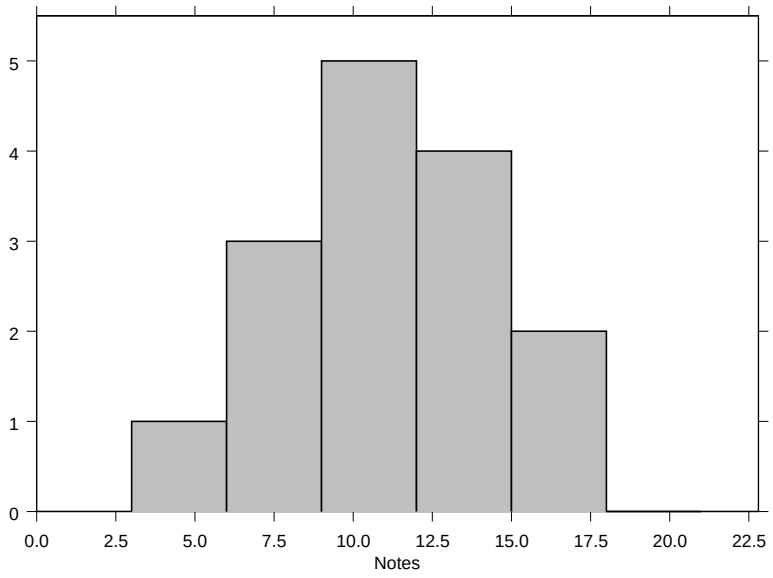

FIG. 2.1 – Histogramme des notes

2.1.2 Cas d'une variable qualitative

Modalités quelconques

Tout d'abord, considérons une variable qualitative x à k modalités $m_1, \ldots, m_k$ quelconques. Cette fois il n'est pas possible d'étudier une tendance centrale ou la dispersion car les réponses ne sont plus des nombres. On s'intéresse alors à la proportion de chaque modalité dans l'échantillon. Ainsi, on compte le nombre d'occurences de chaque modalité parmi l'échantillon $x_1, \ldots, x_n$. Puis on calcule la fréquence de chacune des modalités dans l'échantillon et on les regroupe dans un tableau appelé le tableau des fréquences.

Définition 2.9 *La fréquence f_i de la modalité m_i dans l'échantillon $x_1, \ldots, x_n$ est*

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

où n_i est le nombre d'occurence de la modalité m_i dans $x_1, \ldots, x_n$

Définition 2.10 *Le tableau des fréquences des modalités $m_1, \ldots, m_k$ dans $x_1, \ldots, x_n$ est*

Modalités	m_1	m_2	$\cdots$	m_k
Fréquences	f_1	f_2	$\cdots$	f_k

Les représentations graphiques servent à illustrer visuellement les proportions de chaque modalité. On utilise généralement le diagramme en batons ou le graphique en secteurs (camembert). Pour le diagramme en batons, un baton (ou barre) est associé à chaque modalité et sa hauteur correspond à l'effectif de la modalité. Pour le graphique en secteurs, on divise un disque en autant de secteurs qu'il y a de modalités et l'aire des secteurs est proportionnelle aux effectifs des modalités. Ces deux graphiques sont effectués automatiquement dans tous les logiciels de statistiques et même certains tableurs, l'exemple suivant en donne une illustration.

Exemple 2.8 *Dans un questionnaire sousmis à des agriculteurs, se trouvait la question suivante : Quel type d'engrais utilisez-vous ?*
a : chimiques
b : biologiques
c : mélange des deux
d : aucun
Les réponses obtenues : b, b, a, a, c, d, c, b, c, a, d, c, b, a, c, a, b, c, c,b.
Le tableau des fréquences associées est alors

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

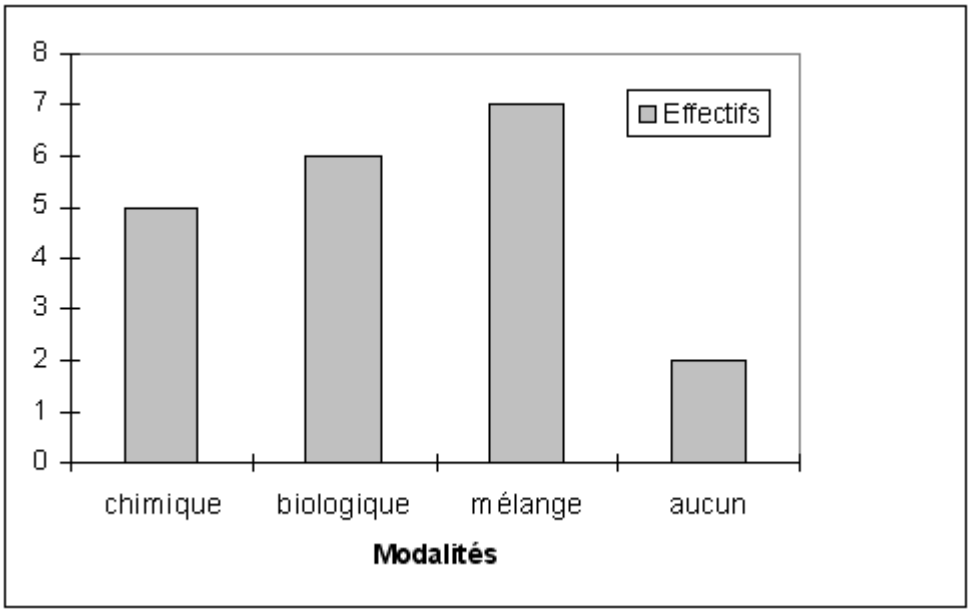

FIG. 2.2 – Histogramme de répartition des engrais

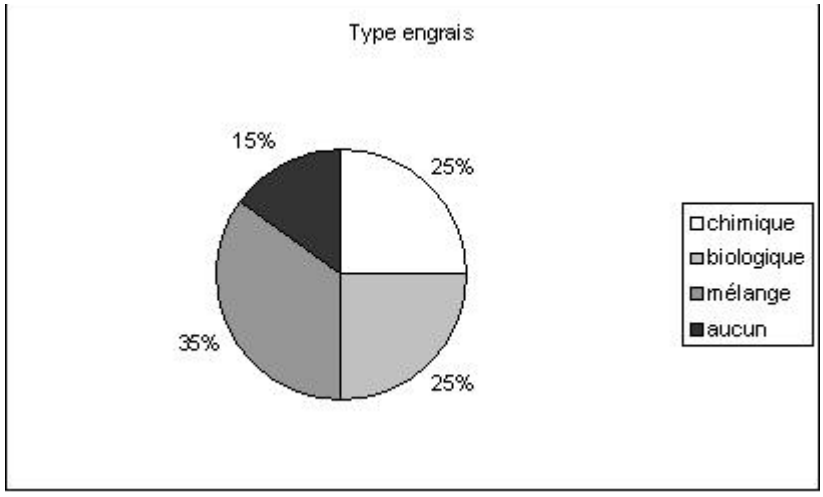

FIG. 2.3 – Graphique en secteurs de la répartition des engrais

Modalités ordonnées

Considérons une variable qualitative x à k modalités $m_1, \ldots, m_k$ ordonnées. Cette fois, il faut respecter l'ordre des modalités dans le tableau des fréquences et ajouter la ligne des fréquences cumulées car il peut être intéressant de faire des commentaires sur un groupe de modalités.

Définition 2.11 *Soit n_i est le nombre d'occurences de la modalité m_i dans $x_1, \ldots, x_n$ pour $i = 1, \ldots, k$.*
Le tableau des fréquences des modalités $m_1, \ldots, m_k$ dans $x_1, \ldots, x_n$ est :

Modalités	m_1	m_2	$\cdots$	m_k
Fréquences	$\frac{n_1}{n}$	$\frac{n_2}{n}$	$\cdots$	$\frac{n_k}{n}$
Fréquences cumulées	$\frac{n_1}{n}$	$\frac{n_1+n_2}{n}$	$\cdots$	$\frac{n_1+n_2+\cdots+n_k}{n} = 1$

Les représentations graphiques utilisées pour les modalités ordonnées sont les mêmes que pour les modalités quelconques. Cependant, il ne faut pas oublier de respecter l'ordre de modalités afin de faciliter les interprétations.

Exemple 2.9 *Comment trouvez-vous le café ?*

TB : Très bon
B : Bon
A : Acceptable
M : Mauvais

Les réponses obtenues : A, B, B, TB, M, A, B, A, TB, M, B, TB, A, A, B, M, M, TB, A, B
Le tableau des fréquences associées est alors

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

En regroupant les modalités TB et B, on peut dire que 50% des clients apprécient ce café. Si on regroupe les modalités TB,B et A on peut dire que 80% des clients sont satisfaits du café.

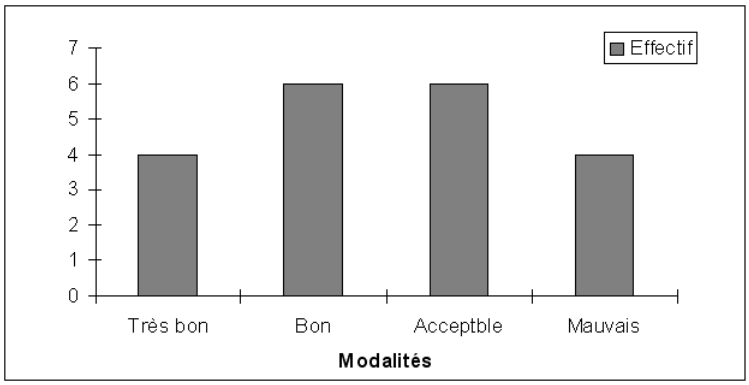

FIG. 2.4 – Histogramme de l'opinion des clients sur le café

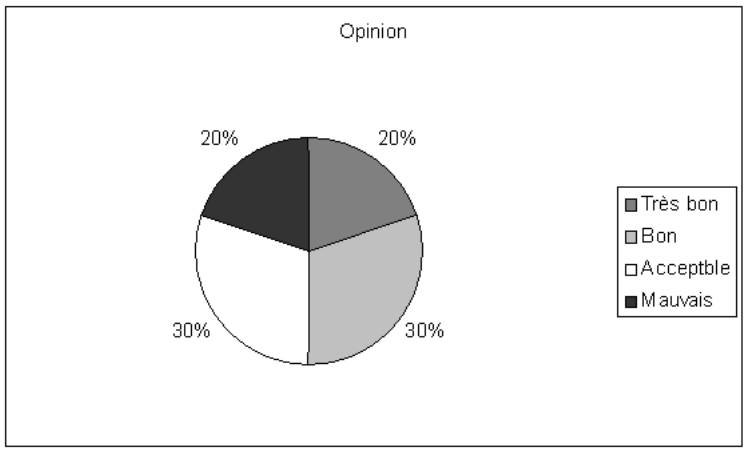

FIG. 2.5 – Graphique en secteurs de l'opinion des clients sur le café

2.2 Analyse bi-variée ou Tris croisés des variables

Dans cette section, nous présentons les traitements que l'on effectue pour étudier simultanément les réponses à deux questions du questionnaire.
Comme dans la section précédente, pour simplifier les notations, nous notons $x_1, \ldots, x_n$, les réponses des n questionnaires à "une question donnée" et $y_1, \ldots, y_n$ les réponses des n questionnaires à "une aute question".

Nous présentons les différentes techniques couramment utilisées pour étudier le lien entre les deux échantillons $x_1, \ldots, x_n$ et $y_1, \cdots, y_n$ des variables X et Y.

2.2.1 Cas de deux Variables Quantitatives

Il existe un indicateur simple à calculer qui permet de dire s'il existe une rélation affine entre X et Y. Il s'agit du coefficient de corrélation linéaire.

Définition 2.12 *Soit deux échantillons $x_1, \ldots, x_n$ et $y_1, \cdots, y_n$, le coefficient de corrélation linéaire entre X et Y, $\rho_{x,y}$ est*

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

Si $|\rho_{x,y}|$ est proche de 1 alors il existe deux réels $a, b \in \mathbb{R}$ tel que $y_i \approx ax_i + b$ pour tout $i = 1, \ldots, n$. Si $|\rho_{x,y}|$ n'est pas proche de 1, on peut seulement conclure qu'il n'existe pas de lien linéaire entre X et Y.

Dans tous les cas, il est toujours intéressant de représenter le nuage de points des couples $\{(x_i, y_i),\ i = 1, \ldots, n\}$ afin d'avoir une idée de la nature du lien entre X et Y. Les deux exemples suivants illustrent l'intérêt de telles représentations.

Exemple 2.10 *Soient les observtions suivantes des variables $X = \{$4.59, 3.37, 9.33, 4.85, 9.64, 3.68, 6.19, 5.39, 2.43, 2.64, 2.91, 3.39, 7.59, 6.79, 9.07$\}$ et $Y = \{$-0.99, -0.70, -3.31, -1.39, -4.35, 0.84, -1.73 -1.90 2.39, -0.38, 0.88, -1.11, -3.47, -2.07, -4.71 $\}$.*
Leur coefficient de corrélation linéaire est égal à $\rho_{XY} = -0.91$, sa valeur absolue est proche de 1, il existe donc un lien linéaire entre X et Y. La figure 2.6 donne une confirmation visuelle de ce résultat.

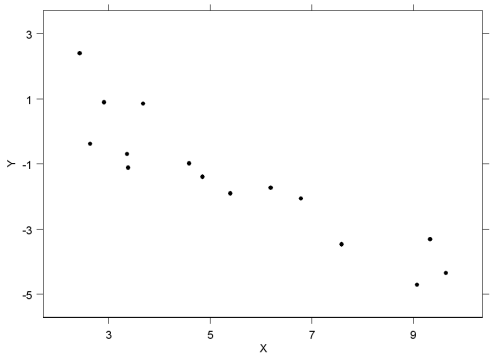

FIG. 2.6 – Représentation graphique du nuage de points

Exemple 2.11 *Soient les observations suivantes des variables* $X = \{$ *-3.03, -4.44, 1.45, -1.83, 0.66, 1.31, -3.69, -0.19, 1.69, 4.29, 3.69, -2.96, 4.61, -1.78, 3.19, 3.97, -4.72, -3.42, -3.45, -2.51*$\}$ *et* $Y = \{$ *774.08, 7676.3, 9.57, 37.51, 0.00864, 6.05, 2544.9, 1.70, 22.32, 6314.5, 2542.3, 679.51, 9634.13, 32.8, 1070.5, 3922.9, 11125.99, 1608.6 1699.0, 255.39* $\}$.
Leur coefficient de corrélation linéaire est égal à $\rho_{XY} = 0.025$*, sa valeur absolue est proche de 0, il n'existe donc pas de lien linéaire entre* X *et* Y*. Cependant, comme le montre la figure 2.7, il semble exister une relation de type quadratique.*

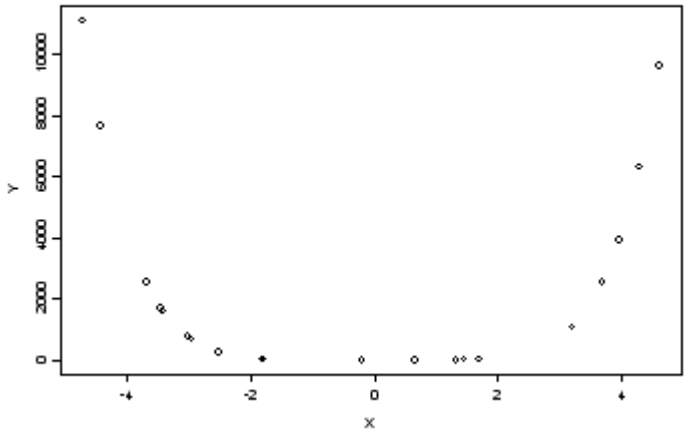

FIG. 2.7 – Représentation graphique du nuage de points

2.2.2 Cas de deux Variables Qualitatives

Soient toujours deux échantillons $x_1, \ldots, x_n$ et $y_1, \ldots, y_n$ mais cette fois de deux variables X, Y qualitatives. Notons respetivement $mx_1, \ldots, mx_k$ et $my_1, \ldots, my_l$ les modalités de respectivement x et y. Afin d'étudier le lien entre X et Y, on regarde les effectifs de modalités croisées de X et Y. Tous ces effectifs sont ensuite regroupés dans le tableau de contingence.

Définition 2.13 *Le Tableau de contingence de x et y est défini comme suit*

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

où
n_{ij} *est l'effectif de l'intersection des deux modalités* mx_i *et* my_j.
$n_{i\cdot} = \sum_{j=1}^{l} n_{ij}$ *(i.e. l'effectif de la modalité* mx_i*) et* $n_{\cdot j} = \sum_{i=1}^{k} n_{ij}$ *(i.e. l'effectif de la modalité* my_j*).*

Les $n_{i\cdot}$ et les $n_{\cdot j}$ s'appellent respectivement marges en lignes et marges en colonnes. La constitution de ce tableau est l'opération appelée "tri croisé".

Test d'indépendance de deux variables qualitatives

Caractériser l'indépendance entre deux variables X et Y est très utile dans une étude et en particulier pour une enquête. Théoriquement cela se caractérise par le fait que la distribution des valeurs de X ne change pas en fonction des valeurs de Y, le test d'indépendance évalue la probabilité de cette indépendance.
Nous ne rentrons pas dans les détails techniques de la construction des tests statistiques d'indépendance, nous présentons seulement le plus répandu (dans les logiciels), il s'agit du test du χ^2. Le χ^2 est la loi de probabilité qui sert à prendre la décision sur l'acceptation ou le rejet de l'hypothèse d'indépendance des deux variables.

Définition 2.14 *La mesure de liaison d^2 entre X et Y est*

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

Si les variables X et Y sont indépendantes d^2 suit approximativement une loi de $\chi^2_{(l-1)(k-1)}$. C'est à dire, on connait les valeurs vraisemblables que peut prendre d^2 sous l'hypothèse que X et Y sont indépendants. Ainsi à un degré de précision fixé α (en général 5%), on connait le seuil que d^2 ne doit pas dépasser pour accepter l'hypothèse d'indépendance.

Définition 2.15 *Si d^2 est supérieur à la valeur critique vc qu'une variable $\chi^2_{(l-1)(k-1)}$ a une probabilité α de dépasser alors on rejettera l'hypothèse d'indépendance de X et Y.*

La valeur critique vc est définie par $P(\chi^2_{(l-1)(k-1)} > vc) = \alpha$, pour trouver vc on doit utiliser des tables de probabilité. Mais heureusement, tous les logiciels calculent directement la valeur critique vc et d^2 ainsi on a plus qu'à les comparer. Donc si $d^2 < vc$, on accepte l'hypthotèse d'indépendance de X et Y au seuil α, sinon on la rejette. Bien entendu, si d^2 et vc sont proches, il est préférable de mitiger la conclusion.

Représentations graphiques

En général, on se contente du tableau de contingence lorsque l'on étudie deux variables qualitatives. Toutefois, il est possible de représenter un histogramme en trois dimensions du tableau de contingence mais ce n'est pas systématique.

Exemple 2.12 *Le questionnaire d'enquête de satisfaction des clients d'un bar comportait les deux questions suivantes :*
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

Voici les réponses recueillies pour 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 ; et pour 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.

Tout d'abord, on effectue un tri croisé des variables café et service en calculant le tableau de contingence (voir fig 2.8).

	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

FIG. 2.8 – Tableau de contingence

On peut aussi représentater graphiquement ce tableau, voir figure 2.9.

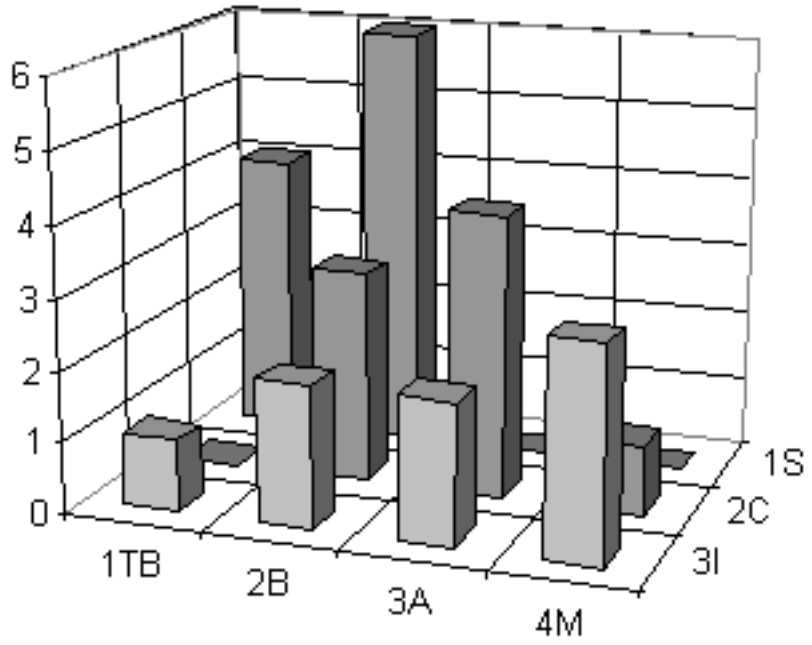

FIG. 2.9 – Histogramme en 3D du tableau de contingence

Il ressort des figures 2.8 et 2.9, que les clients semblent avoir la même opinion concernant le café et le service. Il y aurait donc une dépendance entre les deux variable. Effectuons un test statistique afin d'approfondir la question.

Voici la sortie de test d'indépendance du χ^2 réalisé avec StatBox :

Variable en lignes : Café
Variable en colonnes : Service

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

Valeur observée du khi^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 khi^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

Les observations faites sur les figures sont donc confirmées par le test.

2.3 Analyse multivariée de variables quantitatives ou Analyse en Composantes Principales

L'analyse en Composantes Principales (ACP) est une technique qui permet de faire la synthèse d'information contenue dans un grand nombre de variables quantitatives. Les " composantes principales " sont de nouvelles variables, indépendantes, combinaisons linéaires des variables initiales, possédant une variance maximum. Ces nouvelles variables permettent parfois d'éclairer des relations entre les données. Les composantes principales autorisent en outre la représentation graphique de grands tableaux de données trop complexes à décrire par les méthodes graphiques habituelles. C'est incontestablement cette dernière propriété qui est à l'origine de leur large utilisation. En toute rigueur, une analyse en composantes principales ne nécessite aucune condition de validité. L'usage de variables qualitatives (ordonnées ou pas) est à éviter, des techniques alternatives étant plus performantes.

2.3.1 Réduire la dimension

A la base de tout problème multivarié, se trouve un tableau de données, correspondant par exemple à n sujets chez lesquels sont mesurées p variables. Ce tableau peut être décrit en termes géométriques : n sujets correspondent alors à n points d'un espace de dimension p, les coordonnées de ces " points sujets " étant :

$$\begin{array}{lcccc} 1^{\text{er}} \text{ sujet} & (x_{11} & x_{12} & \cdots & x_{1p}) \\ 2^{\text{ème}} \text{ sujet} & (x_{21} & x_{22} & \cdots & x_{2p}) \\ \vdots & & & \vdots & \\ n^{\text{ème}} \text{ sujet} & (x_{n1} & x_{n2} & \cdots & x_{np}) \end{array}$$

De façon symétrique les p variables correspondent à p points d'un espace de dimension n, les coordonnées de ces "points variables" étant :

$$\begin{matrix} V_1 \\ \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} \end{matrix} \quad \begin{matrix} V_2 \\ \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix} \end{matrix} \quad \cdots \quad \begin{matrix} V_p \\ \begin{pmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{pmatrix} \end{matrix}$$

En pratique, n et p sont souvent supérieurs à 10 ou 20, les points sujets et les points variables sont donc plongés dans un espace de dimension élevée, inaccessible à notre intuition. L'analyse en composantes principales va extraire de cet espace de nouvelles dimensions plus parlantes. Si l'on ne retient parmi ces dernières que les deux ou trois qui contiennent le plus d'information, il est possible de représenter facilement les données sans perdre trop d'information.

Voyons maintenant les principes géométriques mise en oeuvre dans l'analyse en composantes principales.

2.3.2 Des composantes principales

Considérons maintenant, pour commencer, n points sujets inscrits dans un espace de dimension p. Pour simplifier, nous prendrons $p = 3$, les variables seront notées x, y, z. Une telle situation correspond à la figure 2.10 où chaque point est un sujet, et les trois variables les trois dimensions de l'espace.

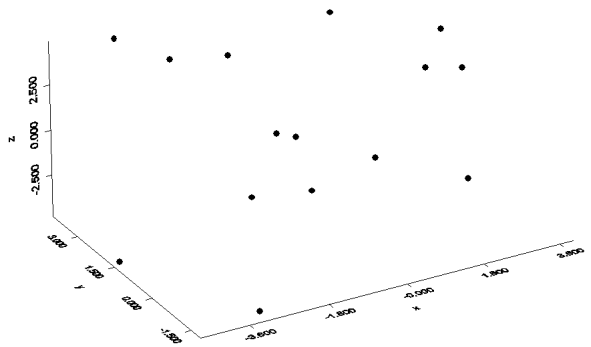

FIG. 2.10 – Représentation graphique d'un tableau de données : les sujets sont symbolisés par les points noirs, les variables par les dimensions de l'espace

Cherchons maintenant la direction suivant laquelle le nuage de points s'étire au maximum, voir figure 2.11. A cette direction correspond une nouvelle variable, combinaison linéaire de x, y, z pour laquelle la variance est maximale : cette direction est dénommée "première composante principale".

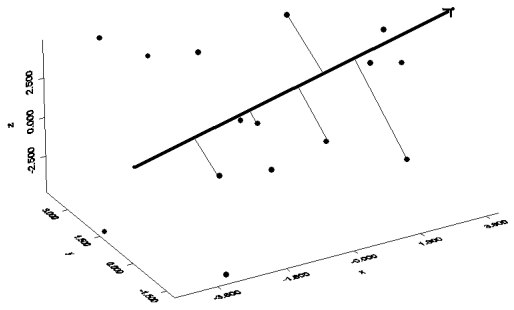

FIG. 2.11 – Recherche de la direction suivant l'amplitude du nuage est maximale

Pour obtenir la deuxième composante principale, projetons nos points sujets sur le plan perpendiculaire à la première composante (voir figure 2.12).

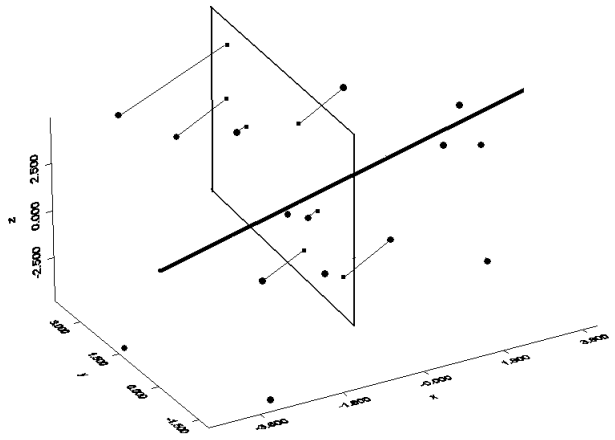

FIG. 2.12 – Projection des points sur un plan perpendiculaire à la première composante

Si l'on considère maintenant ces projections, il est possible de rechercher, comme précédemment, la direction suivant laquelle elles s'étirent au maximum (voir figure 2.13). Cette direction correspond la deuxième composante principale.

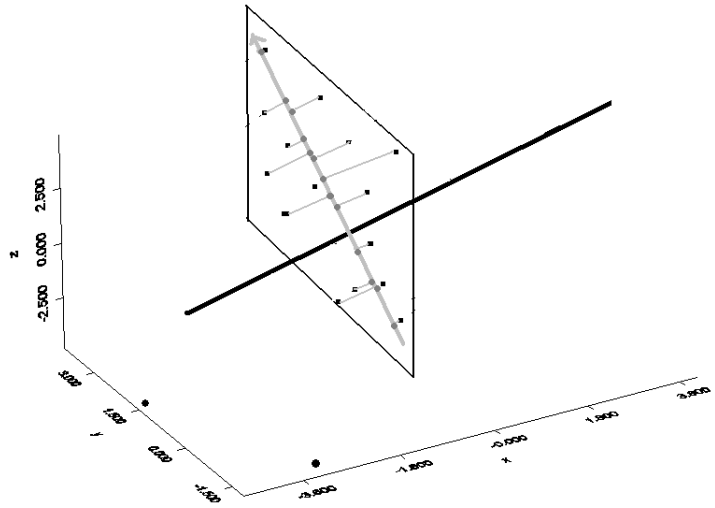

FIG. 2.13 – Recherche de la direction suivant laquelle les projections s'étirent au maximum. Cette direction correspond la seconde composante principale.

Les composantes principales vérifient une propriété corollaire intéressante : les deux premières composantes principales forment un plan, nommé premier plan principal. Si l'on projette les points sur ce plan principal, la variance des projections est maximale. Le plan maximisant la variance est ainsi obtenu à partir des deux droites qui, successivement, extraient le maximum de variance.

Dans un espace de dimension trois, il n'y a que trois composantes principales. Une fois les deux premières déterminées, la dernière l'est aussi car elle est par construction perpendiculaire aux deux autres. Si chacune des trois composantes principales est représentée avec une longueur proportionnelle à la variance (ou, plus précisément, l'écart-type) qu'elle représente (ce qui était implicitement le cas dans les dessins précédents), il est possible d'obtenir une représentation " compacte " de nos données de départ. Non seulement nous avons une idée synthétique de leur situation générale, mais, de surcroît, nous connaissons les directions suivant lesquelles elles se dispersent préférentiellement. Les directions représentées sont désignées sous le terme de "vecteurs propres" (eigen vectors), leur longueur

respective prend le nom de “valeurs propres” (eigen values). La plupart des logiciels donnent systématiquement ces deux familles de paramètres.

2.3.3 Une représentation graphique optimisée

Dans une étude où un grand nombre de variables est à traiter, on souhaite souvent avoir, en début d'analyse, une vue d'ensemble de ces données : y-a-t-il une ou plusieurs sous populations de sujets ? comment sont reliées les différentes variables entre elles ? . . .
Dans le cas de deux variables cela se résume à regarder graphiquement, un simple coup d'oeil permet de détecter les sujets marginaux, dégager les tendances . . .
Avec trois variables cela devient plus compliqué mais faisable, au delà de trois variables le problème devient impossible.
Dans une telle situation, l'analyse en composantes principales est une technique de choix, car elle permet de réduire au mieux le nombre de dimensions afin de pouvoir disposer tout de même d'une représentation graphique.

Revenons au premier plan principal formé à partir des deux premières composantes principales. Nous avons vu que ce plan avait comme propriété de maximiser la variance des projections des points sujets. Formellement, cela revient à maximiser la somme des d_i^2 comme indiqué sur la figure 2.14.

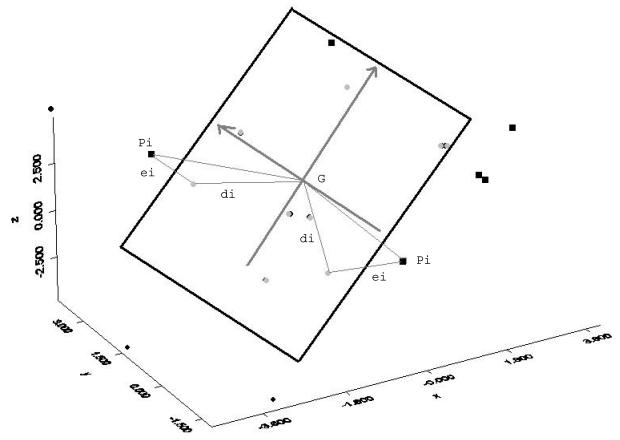

FIG. 2.14 – Le premier plan principal maximise la somme des d_i^2 et minimise la somme des e_i^2

Considérons un point sujet P_i quelconque. Si e_i représente la distance entre P_i

et sa projection, la distance GP_i, entre le centre du nuage et le point P_i vérifie la relation (théorème de Pythagore) : $GP_i^2 = d_i^2 + e_i^2$.
La distance GP_i étant fixe, si le premier plan principal est le plan qui maximise la somme des d_i^2, c'est aussi celui qui minimise la somme des e_i^2. Le premier plan principal est donc le plan qui passe " au plus près " des points sujets originaux.
Ce plan est donc susceptible de préserver au mieux la disposition des points originaux situés dans l'espace à trois dimensions, on dit par extension qu'il contient le maximum d'information.

2.3.4 L'intérêt de normaliser les données

• Si l'on s'intéresse aux relations qui associent plusieurs variables aléatoires exprimées dans des unités différentes, soit ayant un ordre de grandeur différent, il est fortement recommandé de les normaliser avant de procéder à leur analyse en composantes principales. On dit alors que l'analyse est réalisée à partir de la matrice des corrélations des variables aléatoires. Dans le cas contraire, on dit que l'analyse en composantes principales est réalisée à partir de la matrice de covariance des variables aléatoires.

• Si les variables sont mesurées dans les mêmes unités et si elles ont le même ordre de grandeur, il est préférable d'éviter de les normaliser car les propriétés statistiques des résultats d'une analyse en composantes principales sont bien moins complexes quand l'analyse est réalisée à partir de la matrice de covariance.

Rappel : nous rappelons qu'une variable X est normalisée si une mesure x_i est transformée en $x_i' = (x_i - \bar{x})/\sigma)$ où $\bar{x}$ et σ sont la moyenne et l'écart-type des x_i. Les x_i' sont ainsi de moyenne nulle et d'écart-type unité, on dit aussi qu'ils sont centrés réduits.

2.3.5 Les composantes principales en pratique

Considérons les notes obtenues par des collégiens :

	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	6,0 0	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 11	12,00	14,00	8,00	12,00	8,00	6,00
Elève 12	16,00	10,00	4,00	10,00	12,00	10,00
Elève 13	12,00	16,00	8,00	14,00	6,00	4,00
Elève 14	7,00	16,00	13,00	14,00	6,00	4,00
Elève 15	16,00	9,00	4,00	10,00	13,00	11,00
Elève 16	11,00	15,00	9,00	13,00	7,00	5,00
Elève 17	12,00	13,00	8,00	11,00	9,00	7,00
Elève 18	14,00	10,00	6,00	10,00	12,00	10,00

Traitements de base : les tris à plat, les tris croisés

Dans un premier temps, on effectue les traitements de base (cf. table 2.15) : moyenne, minimum, maximum, médiane, quartiles...

	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

FIG. 2.15 – Traitements de base du jeu de données Notes avec StatBox

Les box plots ou en français " boîte à moustache " (figure 2.16), donnent une idée d'ensemble des ces différentes valeurs, on y retrouve la moyenne, la médiane, les quartiles, les valeurs minimales et maximales ainsi que les valeurs extrêmes.

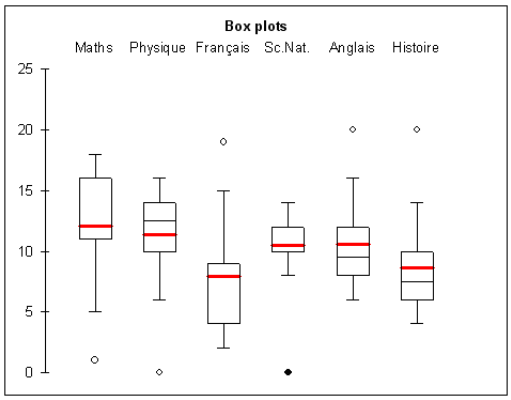

FIG. 2.16 – Box plots des variables de jeu de données Notes

Une autre méthode pour avoir une idée des caractéristiques d'une variable est de faire son histogramme, on a ainsi une idée de sa répartition. Les histogrammes obtenus pour la variable " Math " avec respectivement 10 et 5 classes sont réalisées sur la figure 2.17.

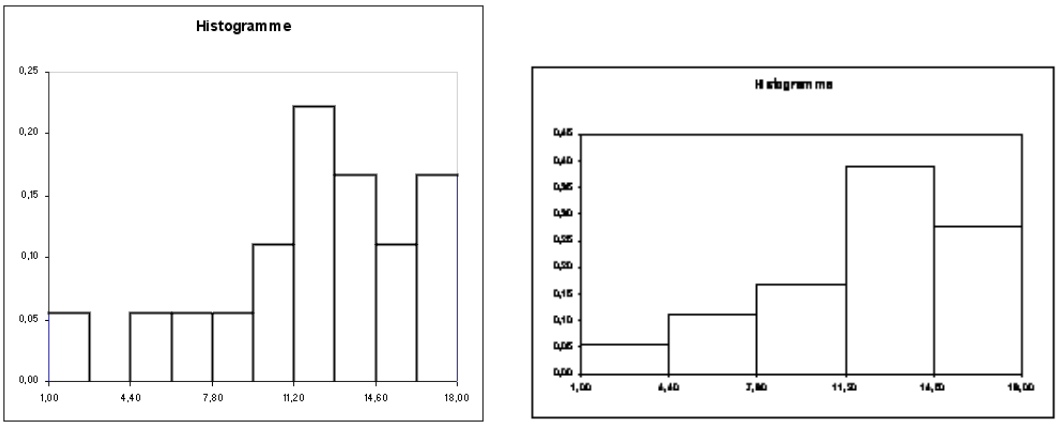

FIG. 2.17 – Histogramme des notes de Math avec dix et cinq classes

Tris croisés

Il est ensuite utile de regarder les relations que peuvent avoir les variables entre elles. Les nuages de points de la figure 2.18 permettent d'avoir une idée des relations entre les différentes variables.

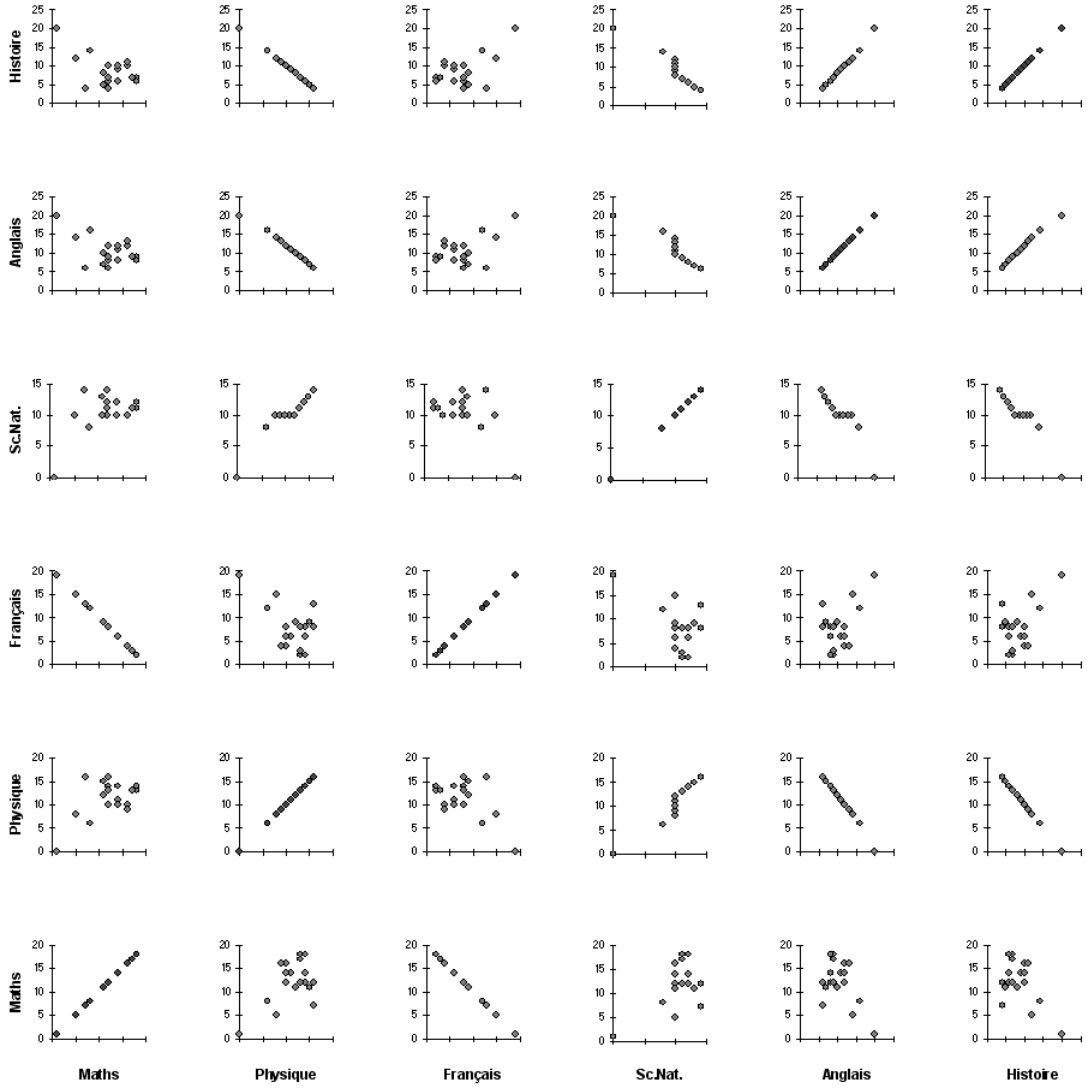

FIG. 2.18 – Nuages de points des variables croisées

On voit par exemple sur la figure 2.18 que les variables Science Naturelle et Physique sont fortement corrélées positivement, cela sous entend que les élèves

ayant des notes élevées dans une des deux matières ont également de bonnes notes dans l'autre, il en est de même pour les matières Histoire et Anglais. De la même façon nous voyons que les matières Anglais et Physique ont un coefficient de corrélation négatif, cela signifie que les notes qui seront élevées dans une des deux matières sera basse pour l'autre.
Ces relations sont également visibles sur le tableau des corrélations (fig 2.19).

	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

FIG. 2.19 – Matrice des corrélations

En effet, le coefficient de corrélation linéaire entre entre Histoire et Anglais est 0.99, il y a donc une relation linéaire positive entre les notes des deux matières. C'est à dire ceux qui ont une bonne note en Anglais ont une bonne note en Histoire et ceux qui ont une mauvaise note en Anglais ont une mauvaise note en Histoire.
De plus, le coefficient de corrélation linéaire entre entre Math et Français est de -1, il existe donc une relation linéaire négative entre les notes des deux matières. C'est à dire ceux qui ont une bonne note en Math ont une mauvaise note en Français et réciproquement.
. . .

Analyse multivariée : l'ACP

Dans un premier temps, on regarde le tableau des valeurs propres (voir fig 2.20), souvent on en fait aussi une représentation graphique (2.21). Les logiciels calculent aussi les vecteurs propres (fig 2.22) associés utilisés par la suite pour réaliser les représentations graphiques.

	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

FIG. 2.20 – Valeurs propres

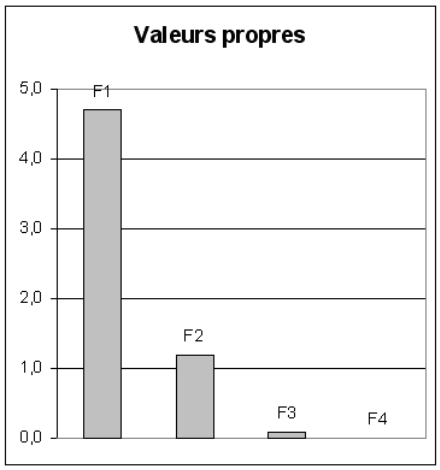

FIG. 2.21 – Histogramme des valeurs propres

	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

FIG. 2.22 – Vecteurs propres

Les figures 2.20 ou 2.21 nous permettent d'évaluer la part de variance que représente chaque composante principale. Les deux premières composantes principales restituent à elles deux 98,42% de la variance, on peut dire qu'elles ont récupéré la quasi totalité de l'information contenue dans les données. Dans la pratique, on a très rarement à faire à des situations si confortables. Nous rappelons que les deux premières composantes principales sont orthogonales et combinaisons linéaires des variables. Voici leurs formes :

F1 = 0.34*Maths + 0.44*Physique - 0.34*Français + 0.43 * Science Nat
F2 = 0.62*Maths - 0.23*Physique - 0.62*Français - 0.23*Science Nat

La première composante principale permet par exemple d'opposer les matières scientifiques au Français, de plus il est à noter que cette première composante principale restitue 78.41% de la variance ce qui est très important. C'est à dire qu'elle a un rôle discriminant très fort parmi les élèves.
On retrouve également les résultats obtenus lors des tris à plat, la Physique et les Sciences Nat. sont fortement corrélées, ainsi que l'Anglais et l'Histoire, les notes

bonnes dans une matière le sont aussi dans l'autre. Les Math et le Français eux sont corrélés négativement, la tendance est inverse, les élèves ayant de bonnes notes en Math auront tendance a en avoir de moins bonnes en Français.

Le premier plan principal, définit par les deux premières composantes principales F1 et F2 restitue 98% de la variance des données Notes. Donc en projetant ces données sur ce plan, on peut visualiser sur un graphique de deux dimensions la quasi-totalité de l'information contenue dans les données Notes qui sont de dimension six. Ainsi, l'analyse des données est beaucoup plus simple, mais il faut procéder avec méthode.

Tout d'abord, il faut regarder les coordonnées des différentes variables (ici les matières Anglais, Français,. . .) dans le plan principal (voir figure 2.23). Ensuite

	F1	F2	F3	F4
Maths	0,73	0,68	0,00	0,00
Physique	0,96	-0,26	-0,06	0,01
Français	-0,73	-0,68	0,00	0,00
Sc.Nat.	0,93	-0,25	0,26	-0,01
Anglais	-0,95	0,29	0,13	0,02

FIG. 2.23 – Axes Principaux

on réalise un graphique représentant les variables (les matières) sur le plan principal (voir fig 2.24).

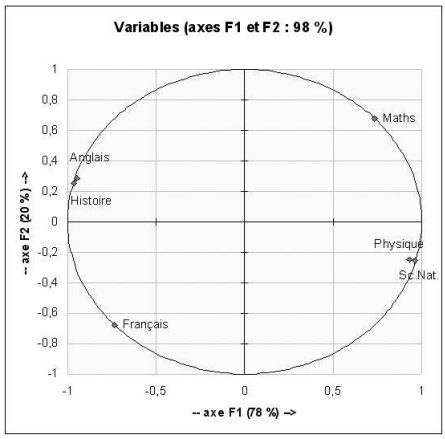

FIG. 2.24 – Représentation des variables

Bien entendu, les commentaires réalisés sur F1 et F2, sont visibles sur la figure 2.24. A savoir, F1 discrimine les matières littéraires et scientifques. F2 discrimine

essentiellement le Français et les Math.
Du point de vue pratique, si un élève se projette vers la gauche sur le plan principal, il a des bonnes notes dans les matières litéraires et mauvaise en sciences, s'il est situé vers le haut il a des meilleures notes en Anglais Histoires qu'en Français.
Pour avoir une vision globale des compétences de tous les élèves, on les projette tous sur le plan principal (figure 2.25

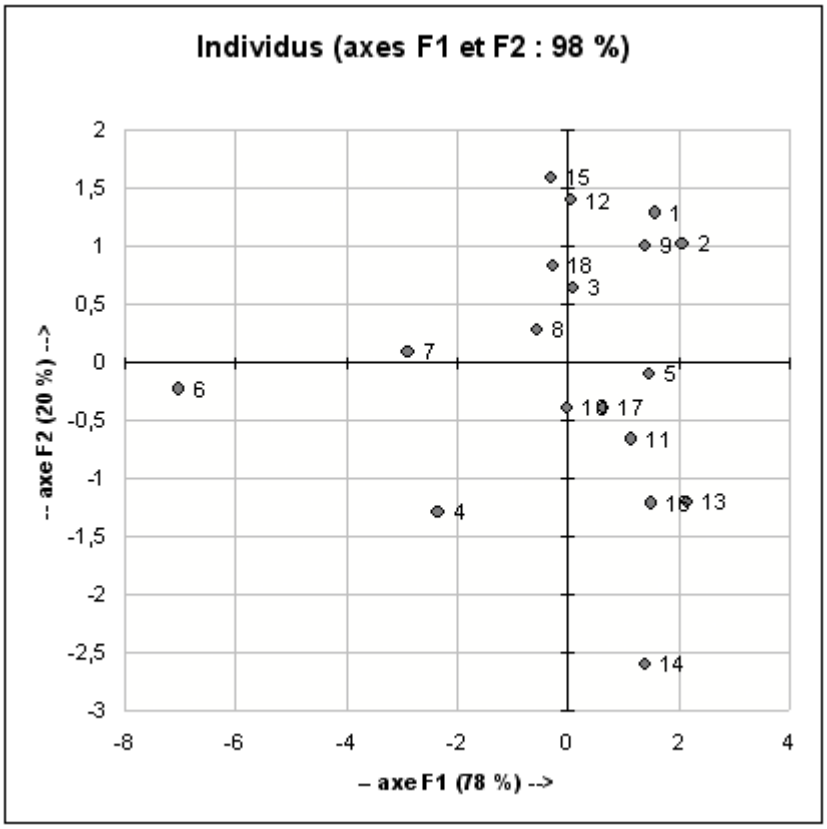

FIG. 2.25 – Représentation des individus

Ainsi sur le graphique (fig 2.25), on voit que l'élève 6 est très fort dans les matière littéraires et très mauvais pour le sciences. L'élève 14 est bon en Physique, Science Nat et Français, . . .
Afin de simplifier les analyses on représente parfois les individus et les variables sur un même graphqiue (fig 2.26), ici ce sera les matières et les élèves.
Attention : La figure 2.26 n'est très rigoureuse, car on représente sur un même graphique des objets de natures différentes : les individus qui sont des points et les variables qui sont des vecteurs (directions de l'espace des individus). Il faut donc

bien se garder d'nterpréter les proximités individus-variables.

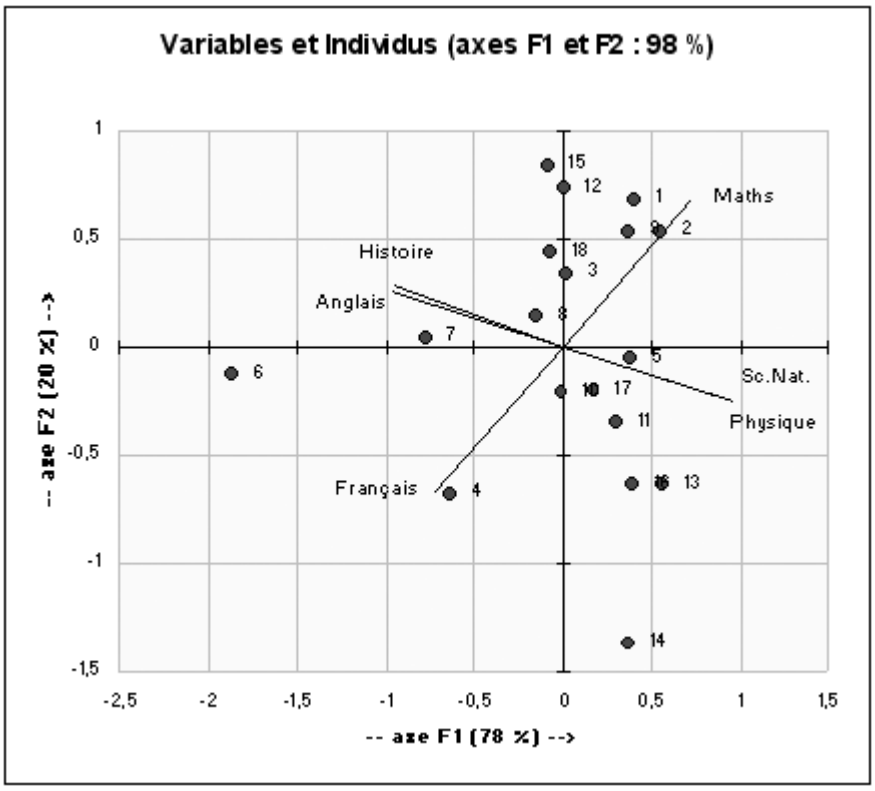

FIG. 2.26 – Représentation des individus et des variables

Le but du graphique (fig 2.25) est de faire ressortir les relations entre les variables sur un graphique de dimension deux. Les fortes corrélations entre Sc Natt (+), Physique(+) Anglais(-), Histoire (-) et Math(+) Français (-) sont immédiatement visibles.
La position des élèves donne une idée de leurs notes dans toutes les matières. Ainsi on peut faire un commentaire global sur la classe, par exemple il y a peu d'élèves qui sont exclusivement bon en lettre (4, 6, 7). Il y a peu plus d'élève qui sont exclusivement bon en science (5, 2, 9, 17, 11, 13). Pour les autres élèves leurs compétences sont plus variées.
En résumé, le premier axe sépare les littéraires des scientifiques et le second sépare les mixtes bons en Math, Histoire, Anglais des bons en Français, Physique, Sc.Nat.

2.4 Analyse multivariée de variables qualitatives ou Analyse Factorielle des Correspondances

L'analyse des correspondances est une adaptation à l'analyse en composantes principales lorsque les données à étudier sont de nature qualitative. Cette analyse prend en compte les différentes modalités de chaque variable et optimise leur représentation graphique.
Avant d'effectuer une analyse en composantes principales, nous avons vu qu'il faut s'interroger sur la nécessité de normaliser ou pas les données, ce qui revient géométriquement à définir une distance entre les variables. L'AFC définit quand à elle une distance spécifique appelée distance du χ^2 (deux variables sont d'autant plus éloignées que le χ^2 utilisé pour tester leur indépendance est élevé). L'AFC ne nécessite aucune condition particulière si ce n'est la nature qualitative des données.

2.4.1 Quelques précisions

Une analyse factorielle des correspondances est construite selon les principes d'une analyse en composantes principales. Quelques différences doivent être soulignées :

• Chaque modalité des différentes variables sera représentée par un point. Ainsi pour une variable binaire du type " possédez-vous un tracteur ? " réponse (oui = 1, non = 0), un point symbolisera les individus ayant un tracteur et un autre les individus n'en possédant pas. Une analyse en composante principale n'afficherait qu'un seul point.

• La qualité globale d'une représentation provenant d'une analyse en composantes principales peut être évaluée grâce au pourcentage de variance correspondant aux valeurs propres des composantes retenues. Pour une AFC les valeurs propres sous estiment cette variance.

• Il est possible dans une AFC d'ajouter sur le graphique des points dits "illustratifs" pour les opposer aux points dits "actifs" en ce sens qu'ils participent aux calculs de l'analyse. Les points "illustratifs" ne sont qu'une aide à l'interprétation et leur utilisation reste au choix de l'utilisateur (tous les logiciels ne permettent pas leur utilisation, l'ensemble des points étant généralement considérés comme "actifs"). Les points "illustratifs" sont introduits a posteriori, si deux composantes sont retenues à partir des points "actifs", il est possible de projeter sur le plan qu'elles engendrent de nouveaux points " illustratifs ", ils sont inclus dans le

schéma mais ne participent pas au calcul des composantes.

2.4.2 Application

On va considérer le jeu de données Notes, et prendre en compte les variables C.Math, C.Physique, C.Français, C.Sc.Nat, C.Anglais et C.Histoire. Elles représentent le classement des matières par préférence pour chaque élève (fig 2.27).

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

FIG. 2.27 – Préférences des élèves

Comme pour l'ACP on regarde les valeurs propres (fig 2.28 et 2.29). Mais cette fois on a beaucoup plus de variables, car pour chacune des six matières on peut avoir six préférences : a, b, c, d, e ou f. Potentiellement, on a donc 36(=6*6) variables différentes : Math.a, Math.b, . . . , Math.f, Phy.a, Phy.b,. . . Cependant, il est courant que certaines variables n'apparaissent pas comme Phy.d, ainsi le nombre de variables intervenant dans l'étude est inférieur à 36 mais il reste largement supérieur à 6.

Comme on peut le voir sur le tableau de la figure 2.28, le premier plan principal restitue tout de même 73.47% de la variance. L'information restituée par le

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

FIG. 2.28 – Valeurs propres

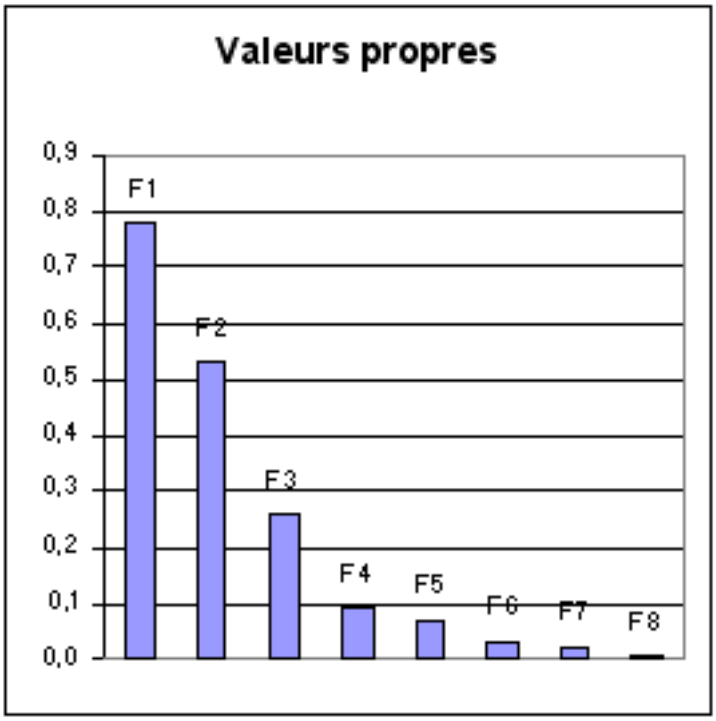

FIG. 2.29 – Histogramme des valeurs propres

premier plan principal est donc de bonne qualité.

Afin d'avoir une vision globale des classements effectués par les élèves, on projette toutes les variables sur le premier plan principal. Les coordonnées de chacune des variables permettant de réaliser une telle représentation sont données dans les tableaux de la figure 2.4.2.

Enfin la représentation des variables sur le premier plan principal est réalisée

	F1	F2	F3	F4
C.Maths - a	-0,74	-0,28	-0,07	0,06
C.Maths - b	0,52	-0,60	-0,28	-0,62
C.Maths - c	1,51	-0,39	0,33	0,32
C.Maths - d	0,96	1,14	-0,34	-0,26
C.Maths - e	0,00	2,13	-0,79	0,42
C.Maths - f	-0,85	0,50	1,65	-0,16
C.Physique - a	1,18	-0,46	0,13	0,01
C.Physique - b	-0,67	-0,61	-0,73	-0,09
C.Physique - c	-0,87	0,06	0,43	0,89
C.Physique - e	-0,50	0,89	0,71	-0,45
C.Physique - f	0,00	2,13	-0,79	0,42
C.Français - a	-0,85	0,50	1,65	-0,16
C.Français - c	0,64	1,47	-0,49	-0,04
C.Français - d	1,35	-0,42	0,24	0,14
C.Français - e	0,16	-0,67	-0,52	-0,86
C.Français - f	-0,74	-0,28	-0,07	0,06

	F1	F2	F3	F4
C.Sc. Nat. - b	1,41	-0,19	0,22	0,31
C.Sc. Nat. - c	-0,30	-0,49	-0,33	-0,33
C.Sc. Nat. - d	-0,65	0,44	0,25	0,32
C.Sc. Nat. - f	0,41	1,60	-0,08	-0,83
C.Anglais - a	0,33	2,11	-0,73	0,03
C.Anglais - b	-0,82	0,14	0,76	0,13
C.Anglais - c	-0,81	-0,54	-0,57	0,04
C.Anglais - d	-0,51	-0,59	-0,54	-0,29
C.Anglais - e	1,36	-0,29	0,19	0,14
C.Histoire - b	0,33	2,11	-0,73	0,03
C.Histoire - c	-0,79	0,50	1,51	-0,61
C.Histoire - d	-0,75	0,15	0,38	-0,36
C.Histoire - e	-0,83	-0,45	-0,35	0,35
C.Histoire - f	1,25	-0,37	0,08	-0,16

FIG. 2.30 – Coordonnées des variables

sur la figure 2.31.

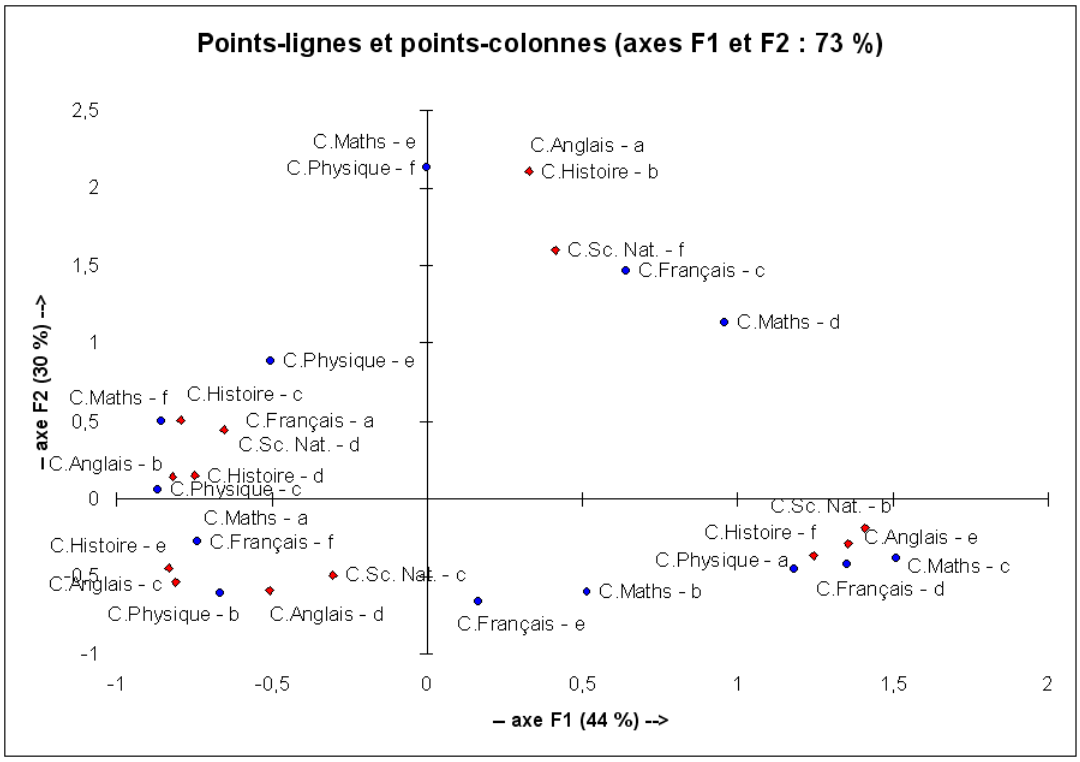

FIG. 2.31 – Graphique de l'AFC

Regardons les regroupements de variables pour les différentes modalités. Par exemple, on peut voir que les variables Ang.a et Hist.b sont très proches ce qui

sous entend que les individus qui classent l'Anglais en a classent souvent l'Histoire en b., il en est de même pour les Math en e et la Physique en f.
On peut aussi regarder le groupe de points situé dans le quart droit en bas du graphique, les variables Phy.a, Sc.Nat.b, Math.c, Français.d, Anglais.e et Hist.f sont proches. Cela donne une idée de l'ordre de préférence pour ces variables.
Afin de préciser les groupes évoqués ci-dessus nous les mettons en évidence sur la figure 2.32. Nous représentons aussi comme variables supplémentaires ou variable illustratives, Moy-Français et Moy-Math qui valent 1 si les élèves ont plus de la moyenne et 2 si les élèves ont moins de la moyenne pour la matière en question. Ces variables ne sont pas utilisées pour la construction des axes principaux, elles servent juste à faire des interprétations.

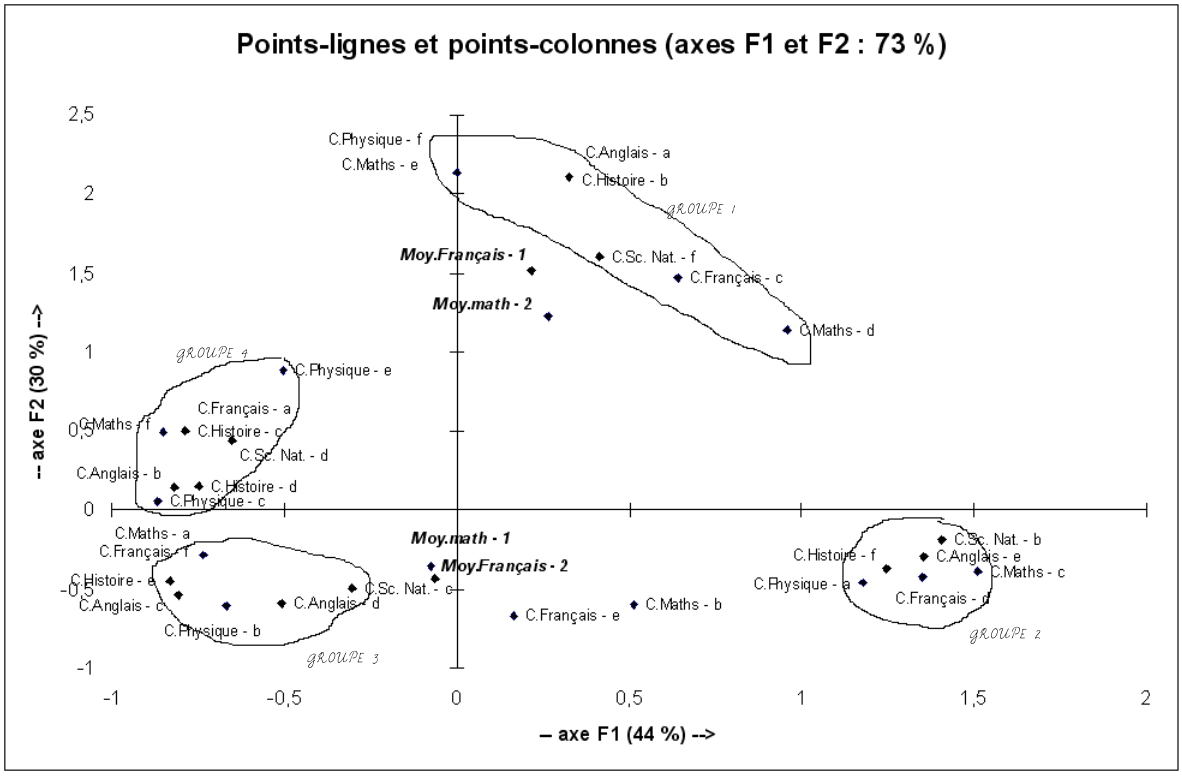

FIG. 2.32 – Graphique de l'AFC

Il ressort de l'ajout des variables supplémentaires Moyennes que les élèves qui ont la moyenne en Français et pas la moyenne en Math préfèrent les matières littéraires. Et inversément, les élèves qui ont la moyenne en Math et pas la moyenne en Français préfèrent les matières scientifiques.
Les quatres groupes mis en évidence sur la figure 2.32, peuvent se définir comme

suit :

Groupe 1	Littéraire avec préférence Anglais-Histoire
Groupe 2	Scientifque avec préférence Physique-Sc.Nat
Groupe 3	Scientifque avec préférence Math
Groupe 4	Littéraire avec préférence Français

Conclusion

Les techniques statistiques que nous avons abordé peuvent être utilisées pour décrire différents types de données et notamment les données récoltées par enquêtes. Ces méthodes sont assez élémentaires, il existe bien sûr des techniques plus avancées, cependant elles permettent de faire en général une analyse satisfaisante pour une grande variété de données.
Il est essentiel de garder à l'esprit qu'un étude statistique sera d'autant plus facile à réaliser que les données seront de bonne qualité. Et que les conclusions sont d'autant plus pertinantes que les données récoltées seront bien adaptées aux questions d'intérêt. Dans ce sens, lorsqu'on réalise une enquête, il est primordial d'accorder la plus grande attention à la rédaction du questionnaire.

Quant aux différents traitements statistiques que nous avons présenté, il est possible de rassembler les différents contextes sous la forme du tableau suivant :

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

Deuxième partie

Travaux Pratiques : Application à de “vrais” jeux de données

(logiciel : StatBox)

Chapitre 3

TP Analyse univariée

3.1 Cas d'une variable quantitative

Exercice 1

Simon Newcomb mesura le temps requis par la lumière pour se déplacer de son laboratoire de Potomac River jusqu'à Washington (fichier VitesseLumiere.xls). La distance totale est d'environ 7400 mètres. Ces données ont été utilisées pour estimer la vitesse de la lumière.

Donnez-une estimation de la vitesse. Il pourra être utile de vérifier la pertinance de toutes les données.

Exercice 2

En février 1968, Herman Bumpus recueillit des animaux échoués à Rhode Island. Il mesura sur les 49 volatiles les 5 longueurs suivantes (fichier bumpus.xls) :

LOT	longueur totale de l'oiseau
AIL	envergure des ailes
TET	longueur de la tête et du bec
HUM	longueur de l'humérus
BRE	longueur du bréchet

Faire les traitements statistiques univariés (moyenne, médiane, quartiles, histogrammes...) pour chaque variable.

Exercice 3

"Les iris de Fisher" (fichier iris.xls) sont des données proposées en 1933 par le statisticien Ronald Aylmer Fisher comme données de référence pour l'analyse discriminante et la classification. Les données correspondent à 3 espèces de fleurs (Iris setosa, Iris virginica, Iris versicolor).

Les variables mesurées sont la longueur et la largeur des sépales, la longueur et la largeur des pétales. Toutes ces variables sont exprimées en millimètres.

Faire les traitements statistiques univariés pour chaque variable.

3.2 Cas d'une variable qualitative

Exercice 4

Données "Produit Bio" (fichier pbio.xls)

Ces données sont relatives à une enquête réalisée dans des supermarchés angevins et parisiens entre 1996 et 1998 dans le but de connaître l'avis de consommateurs quant aux produits biologiques et aux produits diététiques.

Présentation du questionnaire :

1 - Matricule anonyme de la personne interrogée

2 - Connaissez-vous les produits biologiques ?

0 non réponse
1 oui
2 non

3 - Y at-il une différence entre produit biologique et produit diététique ?

0 non réponse
1 oui
2 non

4 - Avez- vous déjà consommé des produits biobiologiques ?

1 non jamais
2 oui une seule fois
3 oui rarement
4 oui de temps en temps
5 oui plusieurs fois par mois
6 oui plusieurs fois par semaine
7 ne se prononce pas

5 - Parmi les marques suivantes lesquelles connaissez-vous ?

0 non réponse
1 bio vivre
2 bjorg
3 carrefour bio
4 la vie
5 vrai
6 prosain
7 favrichon

6 - Avez-vous déjà consommé des produits " La Vie " ?

0 non réponse
1 oui une fois
2 oui occasionnellement
3 oui régulièrement
4 non jamais

7 - Sexe de la personne

1 homme
2 femme

8 - Classe d’age

1 moins de 25 ans
2 entre 25 et 35 ans
3 entre 35 et 45 ans
4 entre 45 et 55 ans
5 entre 55 et 65 ans
6 plus de 65 ans

9 - Etat-civil

1 marie
2 celibataire
3 divorcé
4 en concubinage
5 veuf
6 autre

10 - Nombre d'enfants

1 0 enfant
2 1 enfants
3 2 enfants
4 3 enfant
5 plus de 3 enfants

11 - Situation professionnelle

1 agriculteur
2 artisan
3 cadre supérieur
4 cadre moyen
5 employé
6 ouvrier
7 retraité
8 autre
9 non-réponse

12 - Classe de revenus mensuels

0 non réponse
1 moins de 5 kF
2 entre 5 et 10 kF
3 entre 10 et 15 kF
4 entre 15 et 20 kF
5 plus de 20 kF
6 ne se prononce pas

Faire les traitements univariés de base pour les différentes variables qualitatives (fréquences des modalités, graphes ...)

Chapitre 4

TP Analyse bi-variée

4.1 Cas de deux variables quantitatives

Exercice 5

Un laboratoire a analysé la contenance en sodium et calories de saucisses de hot dog (fichier HotDog.xls) suivant leur nature : boeuf, volaille, viande (mélange porc, boeuf, volaille).

Etudier les compositions suivant le type de saucisse.

Exercice 6

Reprendre la matrice de données Bumpus et faire l'analyse des variables quantitatives 2 par deux (corrélation, graphe X/Y)

4.2 Cas de deux variables qualitatives

Exercice 7

Reprendre les données de l'enquête sur les produits biologiques, et faire les tableaux de contingences.

Chapitre 5

TP Analyse multivariée

5.1 L'ACP

Exercice 8

Faire une ACP avec les données de la matrice Iris.

Exercice 9

La Direction Générale des Impots publie au Journal Officiel une Statistique Mensuelle des Vins. Le J.O. du 4 novembre 1987 publie différents tableaux afférents au mois de novembre 1986.

Le tableau qui croise des catégories de vins avec des pays exportateurs est dans le fichier VinExport.xls . L'unité commune est l'hectolitre. Les sigles se veulent explicites ; ainsi BOJO signifie Beaujolais, ANJO est mis pour Anjou et Saumur, etc. Toutefois les sigles suivants ont une signification particulière :

MOS1	:	Mousseux AOC
MOS2	:	Autres mousseux
AOCx	:	Autres AOC
XXXX	:	Autres Vdqs
RHOF	:	Cotes du Rhone forts en degré
AOCF	:	AOC forts en degré
XXXF	:	Autres forts en degré
XXFF	:	Divers très forts en degré

Exercice 10

L'objet de l'étude est de positionner un crâne fossile dont on a déterminé les mensurations pour vérifier si ce crâne fossile est celui d'un chien ou d'un loup. On mesure six caractéristiques sur des populations de chiens et de loups (fichier ChienLoup.xls).

LCB	:	longueur condylo-basale
LSM	:	longueur de la mâchoire supérieure
LBM	:	largeur bi-maxillaire
LP	:	longueur de la carnassière supérieure
LM	:	longueur de la première molaire supérieure
LAM	:	largeur de la première molaire supérieure

5.2 L'AFC

Exercice 11

Faire une AFC avec les variables de la matrice de données concernant l'enquête sur les produits biologiques.

Exercice 12

Les données communiquées par M. Tenenhaus (fichier chiens.xls) décrivent les caractéristiques de 27 races de chiens au moyen de sept variables catégorisées : taille, poids, vélocité, intelligence, affection, agressivité et fonction. Les quatre premières variables ont trois modalités chacune (petite, 1 ; moyenne, 2 ; grande, 3), les deux suivantes, deux modalités (faible, 1 ; forte, 2), et la dernière, trois modalités (compagnie, 1 ; chasse, 2 ; utilité, 3).

Faire une AFC de ces données en considérant la variable fonction comme une variable typologique (variable supplémentaire dans StatBox).

Bibliographie

[1] J-J Droesbeke, L. Lebart, *Enquêtes, modèles et applications*, Dunod, Paris (2001).

[2] A-M Dussaix, M. Deroo, *Pratique et analyse des Enquêtes par Sondage*, PUF, Paris (1980).

[3] A-M Dussais, J-M Grosbras, *Les sondages : Principes et méthodes*, PUF, Paris (1993).

[4] Ecole Française de la Papèterie et des industrie Graphique, cours Base de Données étudiants deuxième année (2003). Site : http ://cerig.efpg.inpg.fr/tutoriel/bases-de-donnees/sommaire.htm

[5] B. Falissard, *Comprendre et utiliser les statistiques dans les sciences de la vie*, Masson, Paris (1998).

[6] J-P Gouet, *L'élaboration d'un protocole d'enquête*, Institut Technique des céréales et des fourrages, Paris (1978).

[7] D. Grangé, L. Lebart, *Traitements Statistiques des Enquêtes*, , Dunod, Paris (1993).

[8] H. Immediato, "Initiation à la théorie des sondages", Cours DEUG MASS, Université Lyon 1 (2002). Site : http ://nte-serveur.univ-lyon1.fr/nte/immediato/index.htm

[9] D. Lafaye de Micheaux, "Introduction aux enquêtes par questionnaires", Cours Licence MASS, Université Montpellier 2 (2000).

[10] G. Saporta, *Probabilités, Analyse de Données et Statistique*, Technip, Paris (1990).

[11] J. Saracco, "Théorie des sondages", Cours Licence MASS, Université Montpellier 2 (2000).

[12] G. W. Snedecor, W. G. Cochran, *Méthodes Statistiques*, Association de Coordination Technique Agricole, Paris (1971).

[13] Encyclopaedia Universalis (1996).