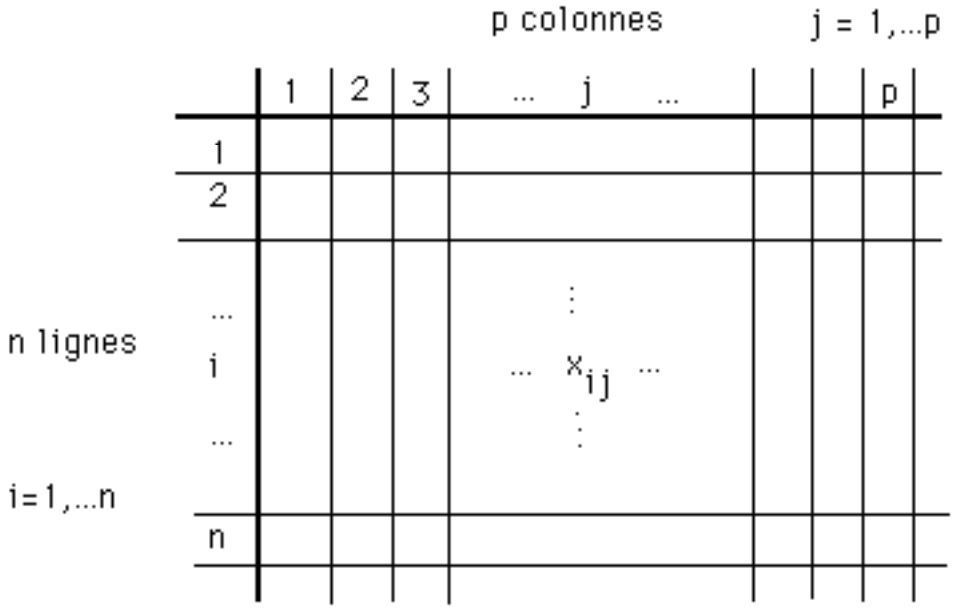
STRUCTURE DES TABLEAUX DE DONNEES [1]

De façon typique, le type d'étude envisagé produit un tableau de données à deux entrées, généralement d'assez grandes dimensions. Nous adopterons la notation générale **X** pour désigner ce tableau, et les éléments qui le composent seront indicés de la façon suivante :



Dans la suite du cours, ce tableau de données sera défini comme une matrice, de dimensions n x p. D'un point de vue statistique, il peut être considéré comme une table de contingence.

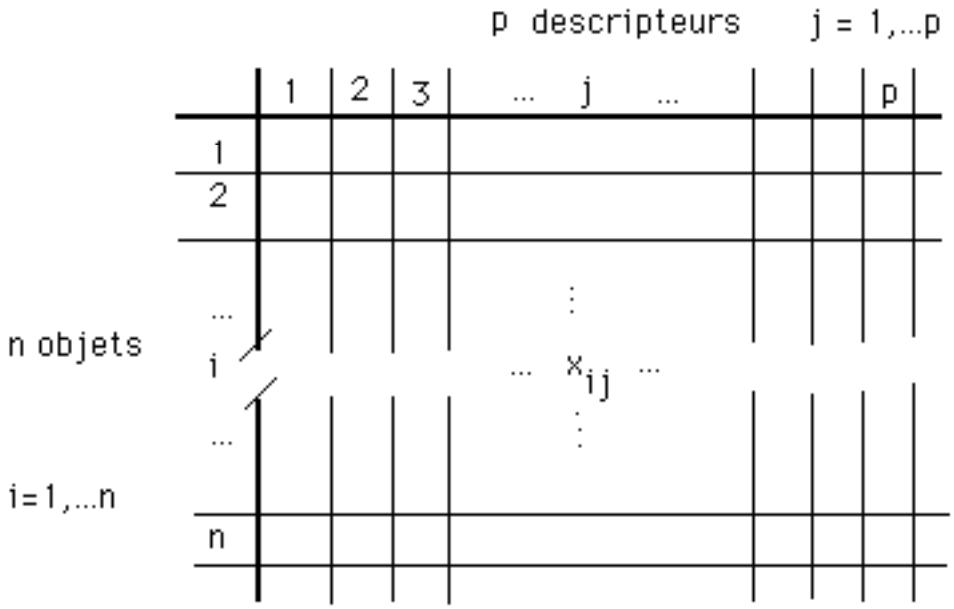
1.Tableau à une ou deux entrées fixes

Distinguons deux catégories de problèmes susceptibles de produire un tableau de données présentant les caractéristiques définies ci-dessus:

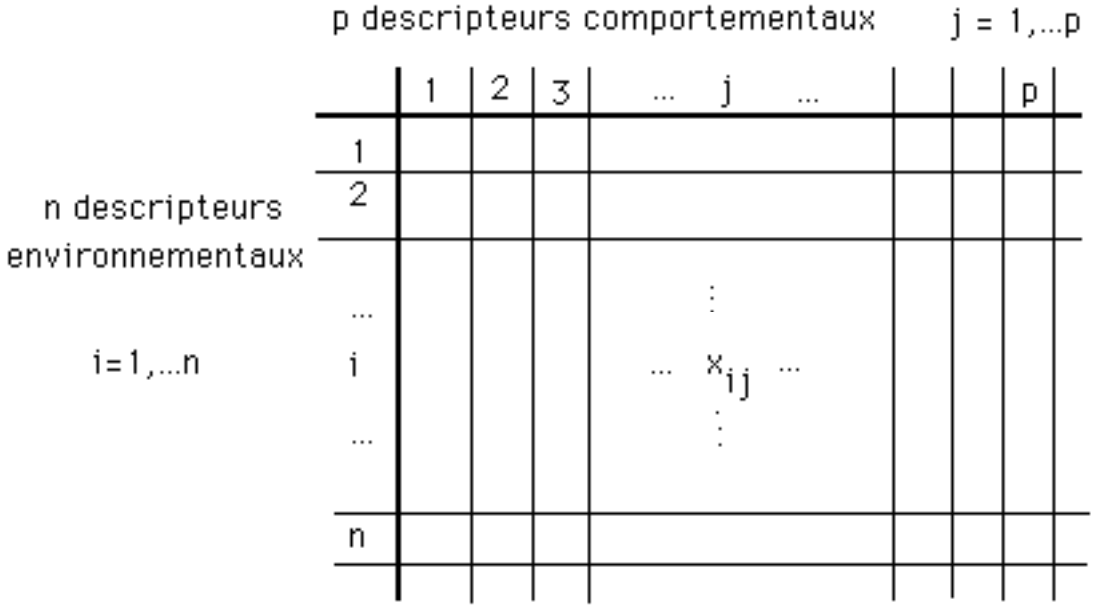
- une dimension du tableau est fixée a priori, la seconde étant fonction de l'échantillonnage,
- les deux dimensions sont fixées a priori.

1. L'exemple de l'anthropologiste illustre la première catégorie de tableau. Au départ, les mesures qu'il effectue sont définies : chaque mâchoire sera caractérisée par une série de descripteurs (je préfère ce terme de façon à ne parler de variable que pour désigner la caractéristique numérique mesurée), qui correspondront, par convention, aux colonnes du tableau : p descripteurs. Chaque observation réalisée ajoutera une ligne au tableau, que nous appellerons par convention un objet. Il y aura donc d'autant plus de lignes au tableau que l'échantillon comprendra d'objets.

[1] P. Dagnélie. L'analyse statistique à plusieurs variables

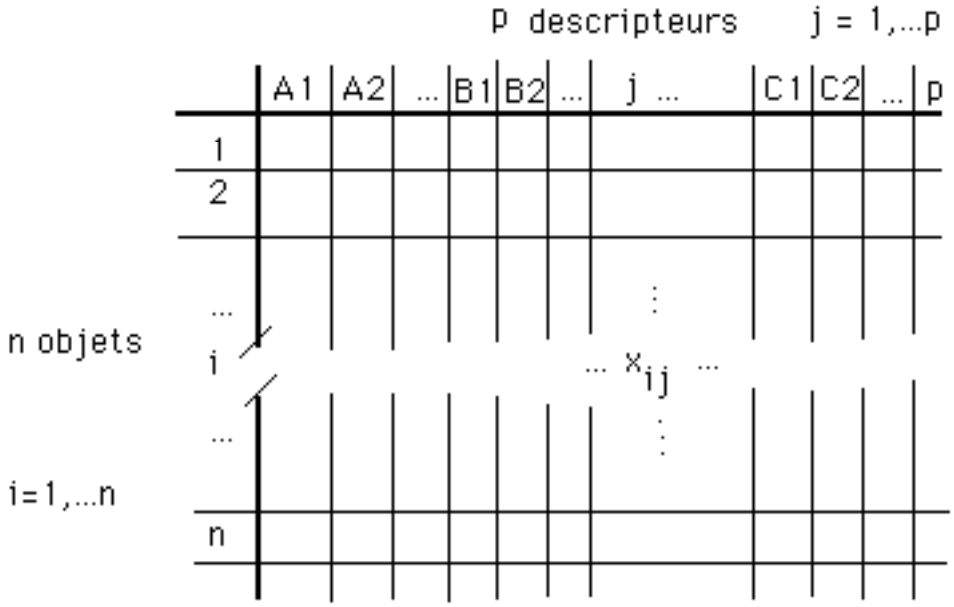


2. Envisageons maintenant une autre expérience, qui consiste à étudier le comportement du porcelet en fonction de son environnement d'élevage. L'expérimentateur définit d'une part un ensemble de comportements qu'il peut distinguer par son observation de l'animal : il mange, il dort, il se roule par terre, il est apathique, agressif, curieux, il mord, il grogne …, et d'autre part un ensemble de caractéristiques de son environnement : il fait chaud, l'éclairage est permanent, les animaux ont trop peu d'espace, la litière est souillée, il n'y a pas de litière, il y a des objets pour attirer son attention … Les descripteurs comportementaux constitueront une entrée du tableau, les descripteurs environnementaux constitueront l'autre entrée. Dans ce cas, le nombre de lignes et de colonnes du tableau sont fixées dès le départ, et les observations consistent à mesurer la fréquence des comportements observés dans les différents environnements.



2. Tableaux complets et incomplets

Certains types d'expériences génèrent naturellement des tableaux complets, d'autres des tableaux incomplets. Hormis un problème de donnée manquante accidentel, les deux types d'expériences décrites ci-dessus font partie de la première catégorie. Envisageons à présent une expérience menée en clinique, au cours de laquelle on suit un certain nombre de patients (objets) pour une série de paramètres cliniques (pression sanguine, globules blancs, urée, cholestérol, réflexes …) pour différentes périodes : une période d'observation (descripteurs notés A1, A2 …), une période préopératoire (descripteurs notés B1, B2…), une période postopératoire (descripteurs notés C1, C2 …)



Par principe, cette expérience va générer un tableau incomplet, car une série de patients seront opérés d'urgence et n'auront pas été caractérisés avant l'opération, d'autres ne seront pas opérés, d'autres éventuellement ne survivront pas à l'opération.

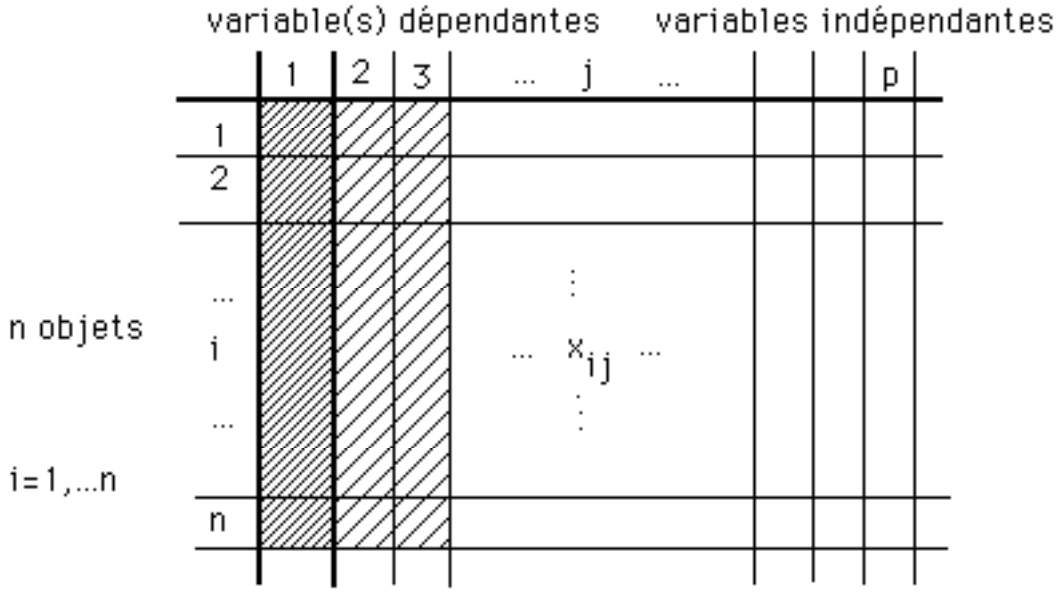
3.Tableaux structurés et tableaux non structurés

Dans le cadre de ce cours, nous nous limiterons essentiellement à l'analyse de tableaux complets, dont une seule dimension est fixée par la définition de l'expérience. Parmi ce type de tableaux, nous pouvons encore définir une série de catégories, suivant la structure qui peut être dégagée au sein des lignes et/ou des colonnes :

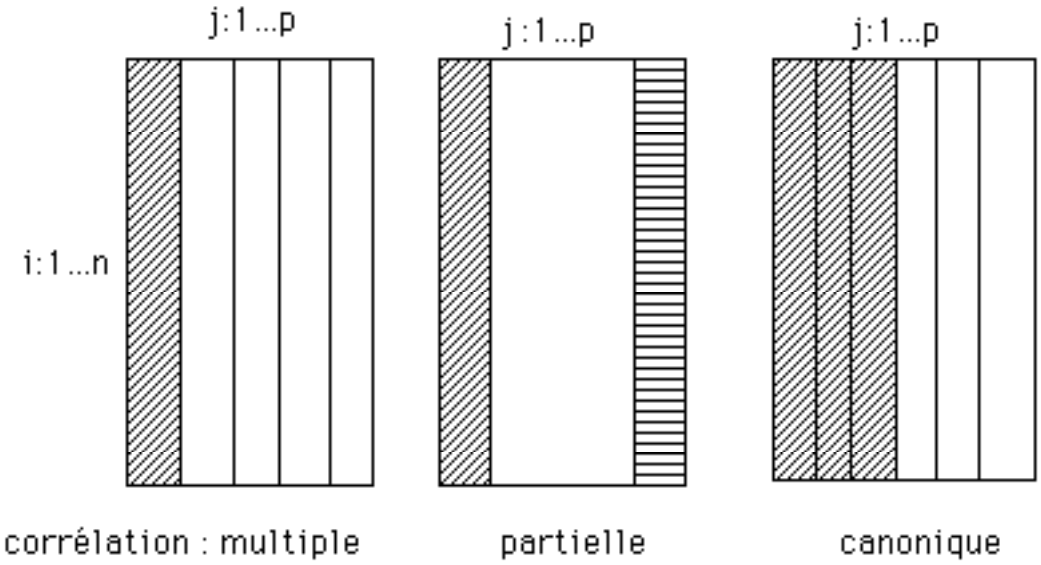
1- les descripteurs sont subdivisés en plusieurs groupes
2- les objets sont subdivisés en plusieurs groupes
3- il n'y a aucune structure dans le tableau

1. Envisageons une étude au cours de laquelle on a mesuré, pour une série d'années (objets), plusieurs variables météorologiques (descripteurs) : pluviosité, insolation, température moyenne en juillet … et le rendement observé pour une culture de maïs, en moyenne pour la région considérée, exprimé en kg/ha.

Dans cet exemple, deux type de descripteurs sont envisagés : une variable dépendante, le rendement, et un groupe de variables indépendantes, les variables météorologiques. Le but de l'expérience est ici d'établir une relation entre la variable dépendante et les variables indépendantes, dans le but de pouvoir prédire le rendement, en fonction des données climatiques. Le type d'analyse à envisager pour analyser les données est la **régression multiple**, qui est une généralisation de la régression à deux variables. Ce type d'approche peut également être généralisé à un ensemble de variables dépendantes, à mettre en relation à un autre ensemble de variables indépendantes. On parlera alors d'un système d'équations de **régression simultanée.** Le tableau associé à ce type de données aura la forme générale suivante :



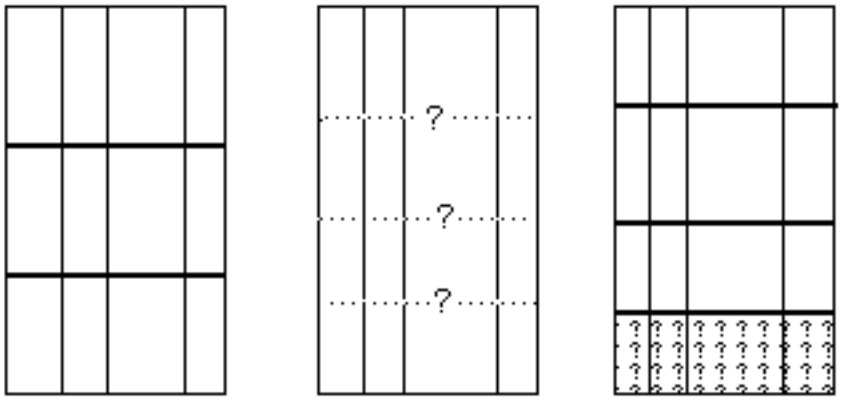
D'autres approches qui sont une généralisation de la notion de corrélation, sont associées à ce type de problème. **Le coefficient de corrélation multiple** permet d'apprécier l'intensité de la relation qui unit une variable dépendante et un groupe de variable indépendantes. **Le coefficient de corrélation partielle** permet de déterminer l'intensité de la relation qui unit deux variables, abstraction faite de la relation qui les lie chacune à une troisième. Enfin **le coefficient de corrélation canonique** permet d'apprécier l'intensité de la relation qui unit un groupe de variables dépendantes et un groupe de variable indépendantes.



2. Envisageons une expérience au cours de laquelle on veut comparer les moyennes de différentes catégories de poissons (objets) vivants dans différentes conditions de température, quant à la longueur, le poids et la fécondité des individus (descripteurs). Les objets sont donc répartis en différentes catégories, qui sont des échantillons représentatifs de chacune de ces populations. La technique appropriée pour analyser les données est une technique d'inférence, soit une généralisation du test de t (deux populations) **(T^2)**, soit une généralisation de l'analyse de la variance, à plusieurs variables **(MANOVA).**

   Une autre possibilité est que les objets soient répartis en différents groupes, mais que leur répartition entre ces groupes soit inconnue. Le but de l'expérimentateur est d'obtenir une classification des objets. Envisageons que l'on cherche à établir une typologie des cours d'eaux sur base de leur peuplement en invertébrés benthiques. En considérant l'abondance des différentes espèces (descripteurs), dans les différentes stations (objets), on établira une **classification (cluster analysis)** des objets, en déterminant des groupes distincts de stations semblables.

   Si enfin on dispose d'un certain nombre d'objets, caractérisés par différents descripteurs, et répartis dans différents groupes connus, le problème peut se poser de classer des individus nouveaux dans un de ces groupes, en fonction de leur plus grande ressemblance avec les individus constituant les groupes. Il s'agit alors d'un problème de **classement**, et les techniques appropriées sont les techniques **d'analyse discriminante.**



3. Enfin, une dernière catégorie d'analyse se rapporte aux tableaux dans lesquels on ne considère aucune structure a priori ni au niveau des descripteurs, ni au niveau des objets. Cette catégorie répond essentiellement aux analyses que nous avons qualifiées de « génératrices d'hypothèses » dans notre introduction. Le but de l'expérimentateur est essentiellement de structurer ces données de façon à pouvoir en obtenir une représentation synthétique la plus compréhensible possible, afin de pouvoir les visualiser et les comprendre. Deux techniques se distinguent essentiellement, **l'analyse en composantes principales**, plus spécialement développée pour l'analyse de tableaux dont une seule entrée est fixée, et **l'analyse des correspondances**, pour l'analyse de tableaux dont les deux entrées sont fixées.

DONNEES QUANTITATIVES ET QUALITATIVES

Les mesures expérimentales peuvent générer des données qualitatives ou quantitatives, selon la nature des variables envisagées. Parfois l'expérience est conçue de telle façon qu'elle implique les deux types de mesure. Envisageons que nous entreprenions de caractériser un grand nombre de variétés de haricots par un maximum de paramètres, afin d'en établir une classification qui permettra par la suite de sélectionner la variété la mieux adaptée à être cultivée dans une région déterminée. Les descripteurs vont caractériser d'une part les paramètres de l'environnement dans laquelle la variété se développe le mieux : altitude, latitude, ensoleillement, humidité, pluviosité, nature du sol … et des caractéristiques du plant de haricot : durée de floraison, hauteur du plant, nombre de graines… Mais par ailleurs d'autres facteurs devront être pris en considération, par exemple parce que dans tel pays d'Amérique latine, les indigènes ne mangeront jamais des haricots à grains blancs … Envisageons donc également les descripteurs suivants : présence de points colorés sur la graine, couleur de la graine, orientation de la tige… Ceci sont les variables qualitatives. Pour être incorporées dans une analyse numérique, elles devront nécessairement être codées, chaque variable posant un problème spécifique à l'expérimentateur :

- présence de points colorés sur la graine :

codé par une variable binaire

(non : 0, oui : 1)

- orientation de la tige :

codé par une variable ordinale

(0 : tige droite, 0.5 : tige légèrement courbée; 1 : tige très courbée)

- couleur de la graine:

codé par un ensemble de variables disjonctives

brun	1	0	0
noir	0	1	0
pourpre:	0	0	1
brun&noir	0.5	0.5	0