UNIVERSITE MOHAMMED PREMIER
ECOLE NATIONALE DES SCIENCES APPLIQUEES
OUJDA-MAROC

Introduction à l'analyse des données

GC3 & GI3

2016 - 2017

M. DEROUICH

Table des matières

Chapitre 1

Introduction à l'analyse des données

1.1 Introduction :

La statistique est une méthode scientifique qui consiste à réunir des données chiffrées sur des ensembles nombreux, puis à analyser, à commenter et à critiquer ces données.

Classiquement les méthodes statistiques sont employées soit pour explorer les données (nommée statistique exploratoire) soit pour prédire un comportement (nommée statistique prédictive ou décisionnelle). La statistique exploratoire s'appuie sur des techniques descriptives et graphiques. Elle est généralement décrite par la statistique descriptive qui regroupe des méthodes exploratoires simples, uni- ou bidimensionnelle (moyenne, moments, quantiles, variance, corrélation, ...) et la statistique exploratoire multidimensionnelle.

La statistique classique étudie les variables les unes après les autres, et elle construit autant de graphes (histogrammes) que de variables, mais les techniques dites d'analyse des données permettent de donner une vision globale de l'ensemble des variables.

L'analyse des données peut se définir comme l'ensemble des méthodes permettant une étude approfondie d'informations et de données de nature qualitative ou quantitative.
Dans l'analyse des données, on distingue :

- **l'analyse univariée**, qui porte sur l'étude d'une seule variable ;
- **l'analyse bivariée**, qui a pour objectif d'examiner les relations entre deux variables en même temps ;
- **l'analyse multivariée**, qui vise l'étude de plusieurs variables en même temps.

L'analyse des données recouvre principalement deux ensembles de techniques :

1. **analyses factorielles**, qui relèvent de la géométrie euclidienne et conduisent à l'extraction de valeurs et de vecteurs propres. Les méthodes les plus employées de cette technique sont :
   i) la méthode de l'**analyse en composantes principales (ACP)**
   ii) la méthode de l'**analyse factorielle des correspondances (AFC)**.
2. **classification automatique** sont caractérisées par le choix d'un indice de proximité et d'un algorithme d'agrégation ou de désagrégation qui permettent d'obtenir une partition ou arbre de classification".

1.2 Définitions

Définition 1.1 *On appelle* **variable** *toute application X définie sur P, avec P un ensemble fini appelé* **population** *ou* **univers** *; tout élément ω de P s'appelle un* **individu**.

Remarque 1.1 *X est aussi appelée* **caractère statistique**

Le caractère désigne une grandeur ou un attribut, observable sur un individu et susceptible de varier prenant ainsi différents états appelés modalités.

Définition 1.2 *On appelle* **modalité** *toute valeur : $x_i \in X(P)$ telle que : $X(P) = \{x_1, x_2, x_3, \cdots x_p\}$ avec p nombre de modalités différentes de X*

1.3 Les types de variables

Il existe deux types de variables :

1. **quantitatif :** c'est un caractère auquel on peut associer un nombre c'est-à-dire, pour simplifier, que l'on peut "mesurer".
   On distingue alors deux types de caractère quantitatif :
   – un caractère discret : c'est un caractère quantitatif qui ne prend qu'un nombre fini de valeurs. Par exemple le nombre d'enfants d'un couple.
   – un caractère continu : c'est un caractère quantitatif qui, théoriquement, peut prendre toutes les valeurs d'un intervalle de l'ensemble des nombres réels. Ses valeurs sont alors regroupées en classes $[x_{i-1}, x_i[$. Par exemple le poids ou la taille d'un individu.
2. **qualitatif :** comme la profession, la couleur des yeux, la nationalité, les groupes sanguins.

A chaque modalité du caractère X, peut correspondre un ou plusieurs individus dans l'échantillon de taille N.

Définition 1.3 *:*
On appelle **modalité** *toute valeur : $x_i \in X(P)$ telle que : $X(P) = \{x_1, x_2, x_3, \cdots x_p\}$ avec p nombre de modalités différentes de X*

A chaque modalité du caractère X, peut correspondre un ou plusieurs individus dans l'échantillon de taille N.

Définition 1.4 *:*

1/ *On appelle effectif de la modalité x_i, le nombre n_i des individus ω tel que $X(\omega) = x_i$ (Cas discret)*

2/ *On appelle effectif de la classe $[x_{i-1}, x_i[$, le nombre n_i des individus ω tel que $X(\omega) \in [x_{i-1}, x_i[$ (Cas continu)*

Remarque 1.2 *: On a $\sum_{i=1}^{p} n_i = N$. l'effectif total.*

Définition 1.5 (Série statistiqu)
Une série statistique est l'ensemble des couples $(x_i; n_i)$ ou $([a_i; a_{i+1}[; ni)$.

Définition 1.6 *:*
On appelle fréquence de la modalité x_i ou de la classe $[x_{i-1}, x_i[$, le nombre f_i tel que $f_i = \frac{n_i}{N}$

Définition 1.7 *:*

1. *On appelle effectif cumulé en x_i, le nombre $\sum_{j=1}^{i} n_j$.*
2. *On appelle fréquences cumulées en x_i, le nombre F_i tel que $F_i = \sum_{j=1}^{i} f_j$.*

Remarque 1.3 *:*
On peut noter que $\sum_{j=1}^{p} n_j = N$, taille de l'échantillon $\sum_{j=1}^{p} f_j = 1$ en effet
$\sum_{j=1}^{p} f_j = \sum_{j=1}^{p} \frac{n_i}{N} = \frac{1}{N} \sum_{j=1}^{p} n_i = \frac{1}{N} \times N = 1$

En général une série statistique à caractère discret se présente sous la forme :

et pour un caractère continue, on a la représentation suivante :

Valeurs	x_1	x_2	$\cdots$ $\cdots$	x_p
Effectifs	n_1	n_2	$\cdots$ $\cdots$	n_p
Fréquences	f_1	f_2	$\cdots$ $\cdots$	f_p

TABLE 1.1 – caractère discret

classes	$[x_0; x_1[$	$[x_1; x_2[$	$\cdots$ $\cdots$	$[x_{p-1}; x_p]$
effectifs	n_1	n_2	$\cdots$ $\cdots$	n_p
fréquences	$f_1 = \frac{n_1}{N}$	$f_2 = \frac{n_2}{N}$	$\cdots$ $\cdots$	$f_p = \frac{n_p}{N}$

TABLE 1.2 – caractère continu

Exemple 1.1 (caractère discret)
Soit un échantillon de 64 familles. Le caractère étant le nombre d'enfants par famille (tableau 2.1)

Nombre d'enfants x_i	0	1	2	3	4	5
Nombres de familles : effectif n_i	16	18	14	11	3	2
Fréquence $f_i = n_i/n$	0.25	0.281	0.218	0.172	0.047	0.031
Effectifs cumulés croissants	16	34	48	59	62	64
Effectifs cumulés décroissants	64	48	30	16	5	2

Exemple 1.2 (caractère continu)
Soit un echantillon de 80 personnes d'une collectivité portant sur la taille. On adopte un intervalle de classe de 0,05 m. (tableau 3.1)

classe	*Effectif*	*Effectif cumulé* ↑	*Effectif cumulé* ↓
[1.55 ;1.60[	*3*	*3*	*80*
[1.60 ;1.65[	*12*	*15*	*77*
[1.65 ;1.70[	*18*	*33*	*65*
[1.70 ;1.75[	*25*	*58*	*47*
[1.75 ;1.80[	*15*	*73*	*22*
[1.80 ;1.85[	*5*	*78*	*7*
[1.85 ;190[	*2*	*80*	*2*

1.4 Les représentations graphiques :

Les représentations graphiques ont l'avantage d'offrir une meilleure vue d'ensemble de la série statistique que les tableaux. Elles permettent par simple lecture, de voir les caractéristiques essentielles de la série, et aussi de comparer des séries différentes.

1.4.1 caractère discontinu

a/ Diagramme des effectifs

Lorsque le caractère est discontinu, on utilise le diagramme en bâtons : les valeurs caractère sont portes en abscisses, les effectifs correspondes sont représentées par des segment vertical dont la hauteur est proportionnelle à l'effectif. le schéma ainsi obtenu est appelé **Diagramme en bâtons des effectifs**

b/ Diagramme des fréquences

Dans ce cas, les effictifs figurants sur l'axe des ordonnnées sont remplacés par les fréquences. le schéma ainsi obtenu est appelé **Diagramme en bâtons des fréquances**

Remarque 1.4 *Si l on joint les sommets des bâtons, on obtient le polygone des effectifs ou des fréquences.*

c/ Diagramme des effectifs cumulées, Diagramme des fréquences cumulées

Le graphe de effectifs cumulées (fréquences cumulées) ne met pas en évidence les différences et ne fait pas ressortir la fréquence maximum. Pour chaque valeur du caractère la somme des fréquences cumulées croissante décroissante est évidemment égale a l effectif total. Ces graphes étant constitues par un ensemble discontinu de points(les sommets des bâtons), l interpolation entre points successifs n a pas de sens. On peut aussi bien adopter une fréquence entre les valeurs discrètes du caractère.

Exemple 1.3 *Soit la série statistique de l'exemple 1.1*

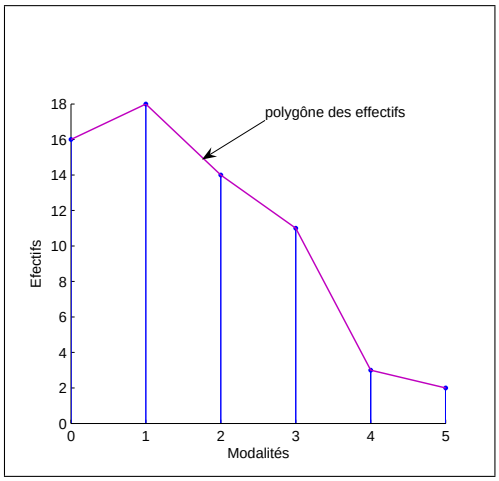


FIGURE 1.1 – Diagramme en bâtons des effectifs

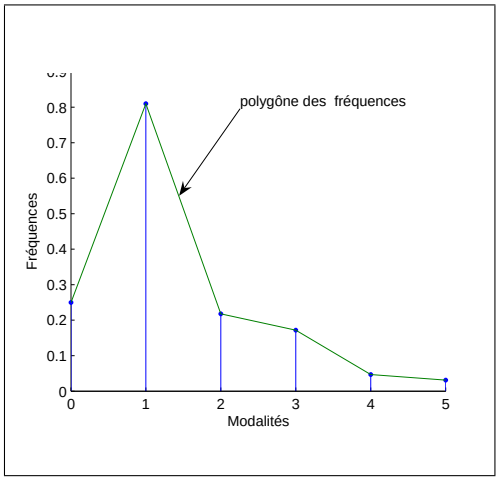


FIGURE 1.2 – **Diagramme en bâtons des fréquances**

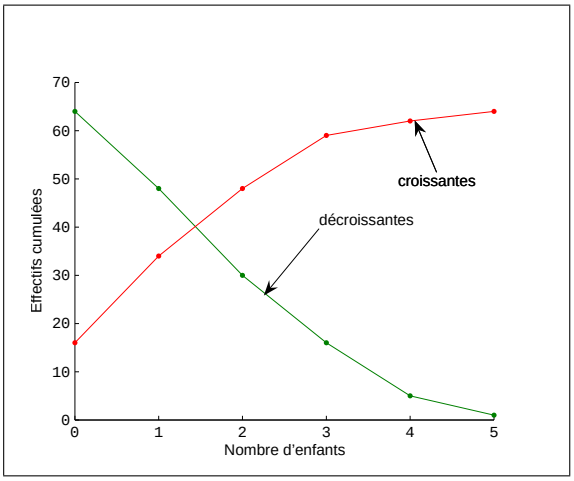


FIGURE 1.3 – **Diagramme des effectifs cumulées**

1.4.2 Caractère continu.

a/ Histogramme et polygone des effectifs

Dans le cas d'un caractère continu, on utilise l'histogramme, qui constitue une généralisation du diagramme en bâtons à la notion de classe.
Chaque classe est représentée par un rectangle dont la base est égale à l' intervalle de la classe et dont la hauteur est égale à l' effectif correspondant. L' histogramme est constitue en fait par le contour polygonal enveloppant l'ensemble de ces rectangles. Le polygone des fréquences absolues (ou des fréquences relatives) s' obtient en joignant les points dont les abscisses sont les milieux des différentes classes et dont les ordonnées sont les effectifs (ou les fréquences relatives) correspondants.

Exemple 1.4 *Soit la série statistique de l'exemple 1.2*

b/ Polygone des effectifs cumuls

En observant que les effectifs cumules croissants correspondent aux frontières supérieures des différentes classe, et les effectifs cumules décroissants aux frontières inférieures des classe.

Exemple 1.5 *A l'aide des données de l'exemple 1.2 on peut construire les polygones des effectifs cumulée suivants :*

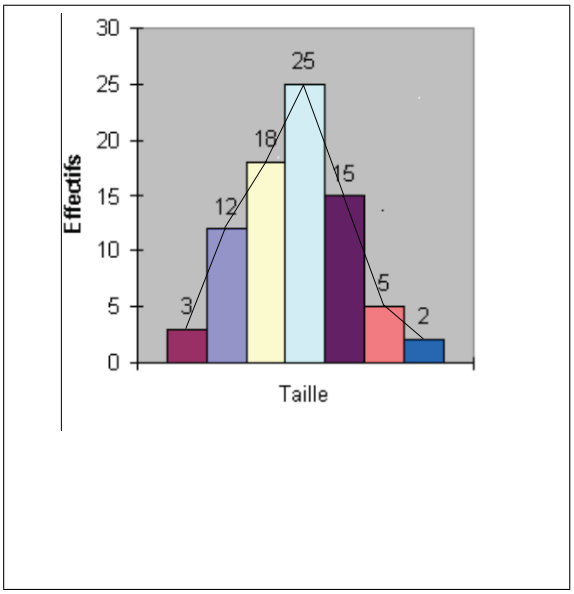


FIGURE 1.4 – **Histogramme et polygone des effectifs**

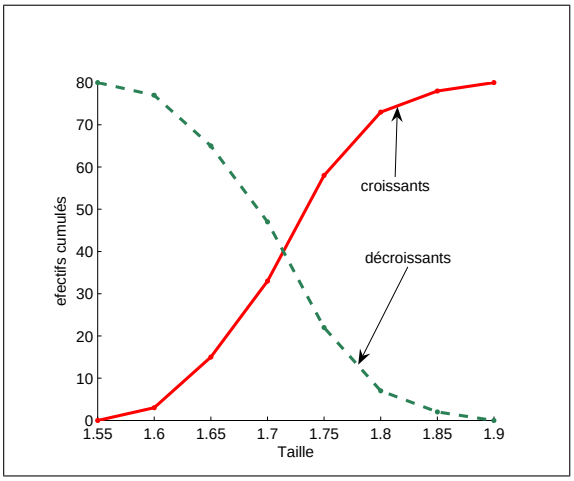


FIGURE 1.5 – **Polygones des effectifs cumulées**

1.5 Paramètres statistiques

1.5.1 Paramètres de position

a/ Dominante ou mode

Définition 1.8 *:*
Lorsque la variable est discrète, **une dominante** *ou* **mode** *est une valeur du caractère qui correspond à un effectif maximum, la série est unimodale, bimodale $\cdots$ lorsque le nombre de modes est 1, 2 $\cdots$.*
Lorsque la variable est continue, une classe modale correspondra à un effectif maximum.

Exemples.
1) Considérons la série statistique de l'exemple 1.1 : le mode est 1 enfant puisqu'il est associé à l'effectif maximum 18.
2) Considérons la série statistique de l'exemple 1.2 : la classe modale est $[1.70 - 1.75[$ puisqu'il est qui corespond à l'effectif maximum 25.

b/ Moyenne arithmétique

Définition 1.9 *:*
Lorsque la variable est discrète la moyenne arithmétique $\overline{X}$ de la série statistique est la moyenne pondérée

$$\overline{X} = \frac{\sum_{i=1}^{p} n_i x_i}{\sum_{i=1}^{p} n_i} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_p x_p}{N}$$

Lorsque la variable est continue la moyenne est :

$$\overline{X} = \frac{\sum_{i=1}^{p} n_i c_i}{\sum_{i=1}^{p} n_i}$$

où les $c_i = \dfrac{x_{i-1} + x_i}{2}$ sont les centres des classes.

Théorème 1.1 *:*
- $\overline{X + b} = \overline{X} + b$
- $\overline{aX + b} = a\overline{X} + b$

Exemples
1)Le nombre moyen d'enfants par famille de la série statistique de l'exemple 1.1 est :

$$\begin{aligned} \overline{X} &= \frac{1}{64}(16 \times 0 + 1 \times 18 + 2 \times 14 + 3 \times 11 + 4 \times 3 + 5 \times 2) \\ &= \frac{101}{64} \end{aligned}$$

2)Pour la série statistique de l'exemple 1.2, la taille moyenne d'une personne est

$$\begin{aligned}\overline{X} &= \frac{3\times 1.575+12\times 1.625+18\times 1.675+25\times 1.725}{80}\\ &\quad +\frac{15\times 1.775+5\times 1.825+2\times 1.875}{80}\\ &= \frac{137}{80}\end{aligned}$$

c/ Médiane

Définition 1.10 *:*
La médiane, Me, est la valeur du caractère pour laquelle la fréquence cumulée est égale à 0,5 ou 50%*. Elle correspond donc au centre de la série statistique classée par ordre croissant, ou à la valeur pour laquelle* 50% *des valeurs observées sont supérieures et* 50% *sont inférieures.*

Remarque 1.5 *:*

- ▶ *Dans le cas où les valeurs prises par le caractère étudié ne sont pas regroupées en classe,*
  - *si n est impair, alors* $n = 2m+1$ *et la médiane est la valeur du milieu* $Me = x_{m+1}$.
  - *si n est pair, alors* $n = 2m$ *et une médiane est une valeur quelconque entre* x_m *et* x_{m+1}*. Dans ce cas* $Me = \frac{x_m + x_{m+1}}{2}$.
- ▶ *Dans le cas où les valeurs prises par le caractère étudié sont groupées en classe, on cherche la classe contenant le* $\frac{n^e}{2}$ *individu de l'échantillon. En supposant que tous les individus de cette classe sont uniformément répartis à l'intérieur, la position exacte du* $\frac{n^e}{2}$ *individu de la façon suivante par interpolation linéaire :*

$$\frac{M_e = x_m}{\frac{N}{2} - N_m} = \frac{x_{m+1} - x_m}{N_{m+1} - N_m} \quad donc$$

$$M_e = x_m + (x_{m+1} - x_m)\left(\frac{\frac{N}{2} - N_m}{N_{m+1} - N_m}\right)$$

avec $m \in \mathbb{N}$ *telle que* $N_m \leq \frac{N}{2} < N_{m+1}$ *et* N_m *effectif cumulé de la classe* $[x_{m-1}m, x_m[$

Remarque 1.6 *Dans le cas d'une série groupée en classes, la médiane est représentée aussi par la valeur de la variable correspondant à l'intersection du polygône des effectifs cumulés avec l'horizontale représentant l'effectif moitié*

Exemples. 1) Les données suivantes représentent le capital social en 10^3 de 17 sociétés marocaines créées entre le 20 et 24-10-1995 (les valeurs du caractère étudié sont classé par ordre croissant) :

$$\overbrace{10;\ 10;\ 20;\ 20;\ 30;\ 50;\ 50;\ 50}^{};\ \overset{\text{médian}}{\underset{\downarrow}{90}}\ ;\underbrace{100;\ 100;100\,;100\,;200\,;200\,;200\,;300}.$$

Le nombre de valeurs observées est 17 (impaire). Dans ce cas, le capital médian est le neuvième (c.à.d 90) car il divise le nombre de sociétés en 2 ensembles égaux : 8 sociétés sont créées avec un capital inférieur à 90 000 DH et 8 autres sont créées avec un capital supérieur à 90 000 DH.
Donc Me=90 000

2) Les capacités de production de 10 sucreries au Maroc en 1993 (en 10^3 tonnes) sont les suivants :

$$\overbrace{30;\ 35;\ 35;\ 45};\ \ 50\ \ \overset{\text{médian}}{\underset{\downarrow}{;}}\ \ 51;\ \ \overbrace{78;\ 80;90\,;500}.$$

Le nombre de valeurs observées est 10 (paire) ; la médiane est, en effet, située ente la cinquième capacité (50) et la sixième (51). Dans ce cas-la, on peut estimer la production médiane par : $\dfrac{50+51}{2} = 50.5$

3) Médiane associée à la série statistique de l'**exemple1.1** : on a $N = 64$, les valeurs centrales sont la $32^{ème}$ valeure et la $33^{ème}$ valeure qui sont égales à 1. Donc la médiane est égale à 1.
4)Médiane associée à la série statistique de l'**exemple1.2** : on a $N = 80$ chercher la médiane revient à trouver la taille de la $40^{ème}$ personne. On procède par interpolation linéaire en utilisant le polygône des effectifs cumulés.
On a

$$\frac{Me - 1.70}{40 - 33} = \frac{1.75 - 1.70}{58 - 33}.$$

Donc

$$Me = 1.70 + \frac{7\,(1.75 - 1.70)}{25} = 1.71\,.$$

1.5.2 Caractéristiques de disperssion

Les paramètres de position sont insuffisants pour caractériser complètement une série. Par exemple, deux séries différentes ayant la même moyenne, ne se repartissent pas nécessairement de la même manière autour de cette moyenne. Elles sont plus ou moins établées, ce qui sera décrit par les caractéristiques de dispersion. Un paramètre de dispersion se rapporte a la différence

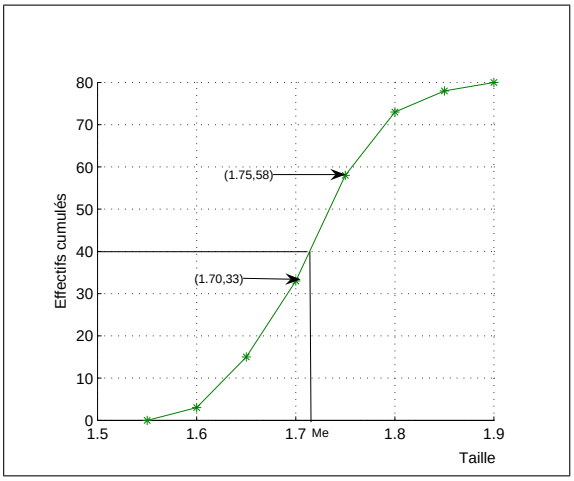


FIGURE 1.6 – **Polygones des effectifs cumulées**

de deux valeurs du caractère alors qu un paramètre de position représente une valeur du caractère.

a/ Ecart moyen arithmétique

Définition 1.11 *:*
Écart moyen arithmétique est la moyenne arithmétique des écarts par rapport à la moyenne arithmétique $\overline{X}$ *des valeurs du caractère*

$$\overline{E(X)} = \frac{1}{N}\sum_{i=1}^{p} n_i|x_i - \overline{X}|$$

Lorsque la variable est continue l'écart moyen arithmétique est :

$$\overline{E(X)} = \frac{1}{N}\sum_{i=1}^{p} n_i|c_i - \overline{X}|$$

où les $c_i = \dfrac{x_{i-1} + x_i}{2}$ *sont les centres des classes.*

b/ Variance

Définition 1.12 *:*
La variance d une série de valeurs du caractère est la moyenne arithmétique des carres des écarts de ces valeurs par rapport à leur moyenne arithmétique.

$$V(X) = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \overline{X})^2$$

Lorsque la variable est continue la variance est :

$$V(X) = \frac{1}{N}\sum_{i=1}^{p} n_i(c_i - \overline{X})^2$$

où les $c_i = \dfrac{x_{i-1} + x_i}{2}$ *sont les centres des classes.*

Théorème 1.2 *:*

- $V(X) = \dfrac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - \overline{X}^2$
- $V(X + b) = V(X)$
- $V(aX + b) = a^2 V(X)$

c/ Ecart-type

Définition 1.13 *:*
L' écart-type (ou écart quadratique moyen) est la racine carrée de la variance $\sigma = \sqrt{V}$

L'étendue d'une série :

Définition 1.14 *:*
L'étendue d'une série est la difference entre la plus grande et la plus petite valeur du caractère.

d/ D'autres définitions

Soit $(x_i, n_i)_{i\in[1,p]}$ ou $(x_i, f_i)_{i\in[1,p]}$ une série statistique discrète . Soit N son effectif total et supposons que pour tout indice i, x_i est strictement positif. On appelle

- **moyenne harmonique** de la série est le nombre h définie par

$$\frac{1}{h} = \frac{1}{N}\sum_{i=1}^{p}\frac{n_i}{x_i} = \sum_{i=1}^{p}\frac{f_i}{x_i} \quad \left(\text{moyenne de } \left(\frac{1}{x_i}\right)\right)$$

- **moyenne géométrique** de la série est le nombre g définie par : $g = \left(\prod_{i=1}^{p} x_i^{n_i}\right)^{1/N} = \prod_{i=1}^{p} x_i^{f_i}$.
- **Moment d'une série statistique :** on appelle moment d'ordre q par rapport à x_0 la moyenne arithmétique des puissances $q^{ièmes}$ des déviations des valeurs du caractère par rapport à x_0 notée :$m_q = \dfrac{1}{N}\sum_{i=1}^{p} n_i\,(x_i - x_0)^q$.

- Si $x_0 = 0$ et $q = 1$, le moment n'est rien d'autre que la moyenne.
- Si $x_0 = \overline{x}$ et $q = 2$, le moment n'est rien d'autre que la variance.

1.6 Les échelles de mesure

Il y a deux échelles de mesures pour les variables qualitatives : L'échelle nominale, L'échelle ordinale.

- L'échelle nominale permet de classer les individus dans des modalités qui sont exprimables par des noms et qui ne sont pas hiérarchisées.

Par exemple :

a) Le genre des personnes : 1. Femme ; 2. Homme ;

b) Statut marital : célibataire ; marié ; veuf, $\cdots$

- L'échelle ordinale permet de classer les individus dans des modalités et, en plus, d'établir un ordre hiérarchique entre ces modalités. Il y a une gradation dans les modalités utilisées (Elles sont alors hiérarchisées).

Par exemple :

a) Le niveau de scolarité :
1. Primaire ; 2. Secondaire ; 3. Collégial ; 4. Universitaire.

b) Niveau d'appréciation d'un produit :
Très bonne qualité, bonne qualité, qualité moyenne, $\cdots$.

Il y a deux échelles de mesures pour les variables quantitatives : L'échelle par intervalles, L'échelle de rapport.

- L'échelle par intervalles : permet non seulement d'identifier la modalité à laquelle appartient l'unité statistique et d'établir un ordre entre les modalités observables mais aussi elle nous informe de l'écart (la distance) séparant deux modalités. Sur cette échelle le zéro est situé de manière arbitraire (une valeur de référence arbitraire, mais ne signifie pas une absence d'un caractère), comme pour la mesure des températures par exemple (échelles Celsius et Fahrenheit).
- L'échelle de rapport : possède les propriétés d'échelle d'intervalle et le zéro constitue un zéro absolu c'est-à-dire la valeur 0 indique l'absence complète du caractère considéré.

Par exemple :âge, salaire, taille, vitesse, etc...

UNIVERSITE MOHAMMED PREMIER
ECOLE NATIONALE DES SCIENCES APPLIQUEES
OUJDA-MAROC

Filière : Génie Civil 1ère année — Année universitaire : 2015-2016
Elément de module : Analyse des données — Enseignant : M. Derouich

Fiche TD $N°$: 1

Exercice 1.6.1
On considère deux séries statistiques de taille n

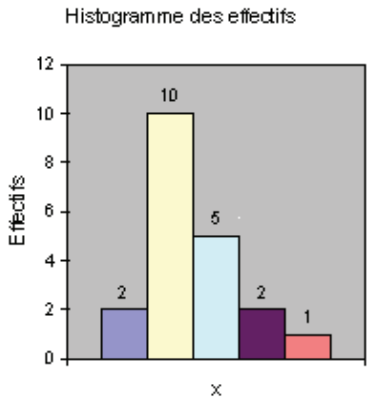
1. *Montrer que la variance d'une série* $(x_i)i = 1..n$ *est égale à* $V = \overline{x^2} - \overline{x}^2$ *où et* $\overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n} n_i x_i^2$ *et* $\overline{x}$ *est la moyenne arithmétique de la série.*
2. *Soient* (x_i) *et* (y_i) *deux séries statistiques liées par la relation suivante :* $\forall i\ y_i = \frac{x_i - a}{b}$ *avec* $b \neq 0\ a, b \in \mathbb{R}$ *Montrer les propriétés suivantes :*
   $i)\ \overline{y} = \frac{\overline{x} - a}{b} \quad ii) V(y) = \frac{V(x)}{b^2} \quad iii)\sigma(y) = \frac{\sigma(x)}{|b|}$

Exercice 1.6.2 *:*
On a relevé les nombres d'allumettes contenues respectivement dans 20 boîtes, lors d'un contrôle dans une usine de fabrication. Les résultats sont les suivants : 40, 42, 32, 38, 40, 48, 30, 38, 36, 40, 34, 40, 34, 40, 38, 40, 42, 44, 36, 42.

1. *Ranger ces résultats en classes d'intervalles de 4 allumettes, borne supérieure exclue.*
2. *Tracer l'histogramme de cette distribution.*
3. *Calculer la moyenne et l'écart type de cette série.*
4. *Calculer les moments d'ordre l, d'ordre 2 et d'ordre 3 par rapport à la valeur moyenne .*

Exercice 1.6.3 *:*
Les résultats d'un certain processus aléatoire sont des nombres entiers que l'on classe suivant l'histogramme ci-dessous.

1. *Calculer la valeur moyenne. Quel est le mode ? quelle est la médiane ?*
2. *Tracer le polygône des fréquences et le polygône des effectifs cumulés.*
3. *Retrouver la valeur de la médiane.*



Exercice 1.6.4 *:*
On reprend les données de l'exemple 1.1 du cours : effectuant le changement de variable $z = \dfrac{x - 1.71}{0.05}$

Classes	*[1.55 ;1.60[*	*[1.60 ;1.65[*	*[1.65 ;1.70[*	*[1.70 ;1.75[*	*[1.75 ;1.80[*	*[1.80 ;1.85[*	*[1.85*
Effectif	*3*	*12*	*18*	*25*	*15*	*5*	*2*

1. *Calculer* $\overline{z}$, $V(z)$ *et* $\sigma(z)$
2. *En déduire* $\overline{x}$, $V(x)$ *et* $\sigma(x)$

Chapitre 2

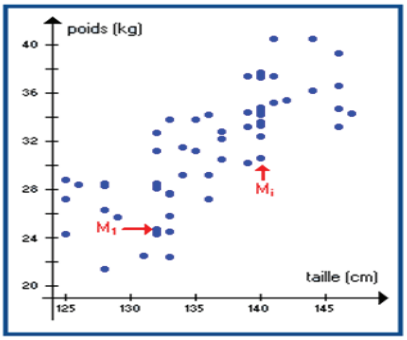
Analyse bivariée

2.1 Définitions :

On se donne une population de taille n et sur chaque élément de cette population on effectue deux observations portant sur deux caractères différents X et Y . Le problème de la corrélation consiste à chercher s'il existe une relation entre Y et Y .

Pour chaque élément de l'échantillon, on peut associer un couple de valeurs (x_i, y_i) où x_i est la valeur du caractére X et y_i est la valeur du caractére Y .

On obtient aussi un nuage de n points constituant un diagramme de dispersion.



Les résultats de ces observations peuvent être présentés sous deux formes

Données non groupées :

Individu	1	2	$\cdots$	n
Valeur X	x_1	x_2	$\cdots$	x_n
Valeur Y	y_1	y_2	$\cdots$	y_n

Données groupées :

Les valeurs prises par X et Y étant respectivement $x_l, x_2, \cdots, x_r$ et $y_l, y_2, \cdots, y_s$. n_{ij} est l'effectif des individus dont les valeurs de x et y sont respectivement x_i et y_i.

$X \diagdown Y$	y_1	y_2	$\ldots$	y_j	$\ldots$	y_s	*Totaux*
x_1	n_{11}	n_{12}	$\ldots$	n_{1j}	$\ldots$	n_{1s}	$n_{1.}$
x_2	n_{21}	n_{22}	$\ldots$	n_{2j}	$\ldots$	n_{2s}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\ldots$	n_{ij}	$\ldots$	n_{is}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	n_{r1}	n_{r2}	$\ldots$	n_{rj}	$\ldots$	n_{rs}	$n_{r.}$
Totaux	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.r}$	$\ldots$	$n_{.s}$	n

Exemple :

Soit Ω la population constituée par les quatre pays suivants : France, Allemagne, Grande bretagne et l'Italie. Notons X la production de fonte (bronze) et Y la production d'acier arrondies en millons de tonnes

Ω	*Allemagne*	*France*	*G. B.*	*Italie*
X	*27.2*	*15.9*	*17.6*	*3.5*
Y	*37.2*	*19.8*	*26.7*	*9.8*

Il est naturel de s'interroger sur la relation qui lié X et Y. On regroupe les valeurs des x_i et des y_i dans le tableau suivant :

$Y \diagdown X$	*3.5*	*15.2*	*17.6*	*27.2*	*Totaux*
9.8	*1*	*0*	*0*	*0*	*1*
19.8	*0*	*1*	*0*	*0*	*1*
26.7	*0*	*0*	*1*	*0*	*1*
37.3	*0*	*0*	*0*	*1*	*1*
Totaux	*1*	*1*	*1*	*1*	*4*

Ici, on a ce qu'on appelle données groupées.

Définitions

- Effectifs marginaux.

La somme des effectifs contenus dans la ligne de x_i est égale à l'effectif des élements dont la valeur du caractère X est x_i. Elle est notée $n_{i.}$.

$$n_{i.} = n_{i1} + \cdots + n_{is} = \sum_{j=1}^{s} n_{ij}.$$

La somme des effectifs partiels contenus dans la colonne de y_j est égale à l'effectif des élements dont la valeur du caractère Y est y_j. Elle est notée $n_{.\,j}$.

$$n_{.\,j} = n_{1j} + \cdots + n_{rj} = \sum_{i=1}^{r} n_{ij}.$$

$n_{i.}$ et $n_{.\,j}$: sont appelés effectifs partiels marginaux. On a :.

$$n = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{s} n_{.\,j} = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}.$$

- ***Fréquences marginales***

$$\begin{aligned} f_{i.} &= \frac{n_{i.}}{n} \quad \textit{fréquence marginale de } x_i. \\ f_{.\,j} &= \frac{n_{.\,j}}{n} \quad \textit{fréquence marginale de } y_j. \end{aligned}$$

On a

$$\sum_{i=1}^{p} f_{i.} = \sum_{j=1}^{q} f_{.\,j} = \sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij} = 1.$$

(f_{ij} fréquence partielle correspondant à $X = x_i$ et $Y = y_j$).
Les couples $(x_i, n_{i.})_{1\leq i\leq p}$ et $(y_j, n.\,j)_{1\leq j\leq q}$ définnissent les distributions statistiques marginales.

2.2 Tableaux de calcul

2.2.1 Données non groupées :

Comme dans le cas d'un seul caractère, on a :
Moyennes

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ et } \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Variances

$$V(X) = \sigma_X^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{X}^2 \text{ et } V(Y) = \sigma_Y^2 = \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right) - \overline{Y}^2.$$

On introduit maintenant deux nouveaux caractères qui dépendent à la fois de X et de Y.
***Covariance**. la Covariance de X et Y, notée $cov(X, Y)$, est définie par :*

$$\sigma_{XY} = cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right).$$

On montre aisément que :

$$\sigma_{XY} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{XY}.$$

Le coefficient de corrélation linéaire du couple (X,Y) noté $\rho(X,Y)$, est définis par :

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

2.2.2 Données groupées :

Plus généralement et surtout lorsque l'effectif total est grand, si $x_1,\ldots,x_p$ sont les modalités de X et $y_1,\ldots,y_q$ sont les modalités de Y, on dresse le tableau suivant :

$X \diagdown Y$	y_1	$\cdots$	y_q	$n_{i.}$	$n_{i.}x_i$	$n_{i.}x_i^2$	$\sum_{j=1}^{q} n_{ij}y_j$	$x_i \sum_{j=1}^{q} n_{ij}y_j$
x_1	n_{11}	$\cdots$	n_{1q}	$n_{1.}$	$n_{1.}x_1$	$n_{1.}x_1^2$		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$				
x_p	n_{p1}	$\cdots$	n_{pq}	$n_{p.}$	$n_{p.}x_p$	$n_{p.}x_p^2$		
$n_{..j}$	$n_{.1}$	$\cdots$	$n_{.q}$	n				
$n_{.j}y_j$	$n_{.1}y_1$	$\cdots$	$n_{.q}y_q$					
$n_{.j}y_j^2$	$n_{.1}y_1^2$	$\cdots$	$n_{.q}y_q^2$					
$\sum_{i=1}^{p} n_{ij}x_i$		$\cdots$						
$y_j \sum_{i=1}^{p} n_{ij}x_i$		$\cdots$						

Calculs

i) ***Moyennes :*** $\overline{X} = \frac{1}{n}\sum_{i=1}^{p} n_{i.}x_i \quad et \quad \overline{Y} = \frac{1}{n}\sum_{j=1}^{q} n_{.j}y_j$

ii) ***Variances :***

$V(X) = \sigma_X^2 = \left(\frac{1}{n}\sum_{i=1}^{p} n_{i.}x_i^2\right) - \overline{X}^2 \quad et \quad V(Y) = \sigma_Y^2 = \left(\frac{1}{n}\sum_{j=1}^{q} n_{.j}y_j^2\right) - \overline{Y}^2.$

iii) ***Ecart-type :*** $\sigma_X = \sqrt{V(X)} \quad et \quad \sigma_Y = \sqrt{V(Y)}$

iv) ***Covariances :***

On appelle covariance du couple (X,Y) et on le note $cov(X,Y)$ ou σ_{XY} la moyenne de $\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)$

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right).$$

On montre que :

$$\sigma_{XY} = cov(X,Y) = \left(\frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}x_i y_j\right) - \overline{XY}.$$

v) ***Coefficient de correlation linéaire :***

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

N.B *L'importance des paramètres $cov(X,Y)$ et $\rho(X,Y)$ apparaîtra quand on s'intéressera au lien (ou corrélation) éventuel entre X et Y .*

2.2.3 Propriétés :

On montre que :

- $|\rho(X,Y)| \leq 1$.
- $\rho(aX+b, a'Y+b') = \dfrac{aa'}{|aa'|}\rho(X,Y)$. *donc* $\rho(aX+b, a'Y+b') = \pm\rho(X,Y)$
- $cov(aX+b, a'Y+b') = aa'.cov(X,Y)$.

Ces formules sont utilisables pour simplifier les calculs.

ii) En effectuant. les changement de variables suivants : $X' = \dfrac{X-c}{d}$ *de et* $Y' = \dfrac{Y-c'}{d'}$ *avec* $d, d' \neq 0$*, on obtient :*

$cov(X',Y') = \dfrac{1}{dd'}cov(X,Y)$.

$\rho(X',Y') = \dfrac{|dd'|}{dd'}$ *donc* $\rho(X',Y') = \pm\rho(X,Y)$.

2.2.4 Distribution conditionnelle, indépendance

La fréquence conditionnelle de x_i sachant y_j (y_j réalisé)

$$f_{i/j} = \frac{n_{ij}}{n_{.j}} = \frac{f_{ij}}{f_{.j}}$$

Où n_{ij} est l'effectif correspondant à $X = x_i$ et $n_{.j}$ l'effectif partiel marginal de y_j.

On a $f_{j/i} = \dfrac{n_{ij}}{n_{i.}} = \dfrac{f_{ij}}{f_{i.}}$.

Ainsi $f_{ij} = f_{i.} \times f_{j/i} = f_{.j} \times f_{i/j}$.

Définition 2.1 *Deux variables statistiques X et Y sont dites statistiquement* **indépendantes** *si et seulement si, pour chacune des deux variables, les distributions conditionnelles sont identiques à la distribution marginale :*

$$f_{i/j} = f_{i\cdot} \quad ou \quad f_{j/i} = f_{\cdot j} \forall (i,j)$$

Conséquence *: Les caractères X et Y sont* **indépendants** *si et seulement si*

$$\forall (i,j) \quad f_{ij} = f_{i\cdot} \times f_{\cdot j}$$

Application :

Sur le tableau suivant figure l'âge de la mère (x) et le poids de l'enfant (y) pour un échantillon de 40 naissances , présentés avec un groupement à deux dimensions en classe d'âge de 5 ans et en classe de poids de 500g

	2500	3000	3500	4000	4500	$n_{i.}$
20	1	5	4	2	-	12
25	2	3	5	1	-	11
30	1	2	2	1	-	6
35	-	3	3	1	1	8
40	-	2	-	1	-	3
$n_{.j}$	4	15	14	6	1	40

$n_{l3} = 4$ *signifie qu'il ya 4 enfants dont l'âge de la mère est 20 ans et dont le poids est 3500g. Il y a 6 mères dont l'âge est 30 ans. Il Y a 14 enfants dont le poids est 3500g.*

2.3 Ajustement linéaire et corrélation

2.3.1 Introduction

On considère une population de taille (effectif total) n sur laquelle on définit une statistique double X et Y. (On éffectue sur Ω deux observations portant sur 2 caractères différents).
Le problème qui se pose est celui qui consiste à rechercher s'il existe une relation entre X et Y.
A chaque élement ω_i de l'échantillon, on associe un couple de valeurs (x_i, y_i) qu'on représente graphiquement par un point $M_i(x_i, y_i)$ du plan. Et on obtient ainsi un nuage de n points qui constitue ce qu'on appelle un diagralmme de disperssion.

Ajuster *un ensemble de points consiste à déterminer une courne* (C) *simple aussi proche que possible des points* M_i*. L'ajustement linéaire est le cas où* (C) *est une droite.*

2.3.2 Définition

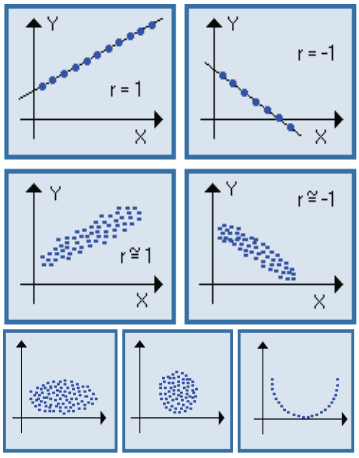
On dit qu'il y a correlation entre deux caractères observés sur une même population lorsque les variations des deux caractères se produisent dans le même sens ou lorsque les variations sont de sens contraires.

Nuage de points *:* ***Diagramme de disperssion***

L'existence d'une correlation peut-être décelée (détectée) graphiquement. La forme du nuage de points formé par les points $M_i\,(x_i, y_i)$ *nous permettent de constater si les caractères* X *et* Y *sont en correlation ou non.*

Définition. *On dit qu'une correlation (lorqu'elle existe) qui lié 2 caractères* X *et* Y *est positive ou directe si* Y *croît en même temps que* X *Si* Y *décroît lorque* X *croît, la correlation est dite inverse ou négative.*

Exemples de différents nuages



a- Coefficient de correlation

Rappel. *Le coefficient de correlation du couple* (X, Y) *la quantité*

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$$

(n *représente le nombre de couple d'observations* (x_i, y_i)*)*

ρ est compris entre -1 et 1.

** Si les caractères X et Y sont indépendants alors $\rho = 0$. Cependant la réciproque n'est pas necessairement vraie. Si $\rho = 0$, on dit qu'il y a correlation nulle entre X et Y ; la liaison entre X et Y peut être de forme outre que linéaire.*

** Si $0 < \rho < 1$, la correlation est positive (X et Y varient dans le même sens) La valeur $\rho = +1$ indique une relation linéaire parfaite $Y = aX + b$ avec $a > 0$. C'est un cas extrème très peu rencontré en pratique.*

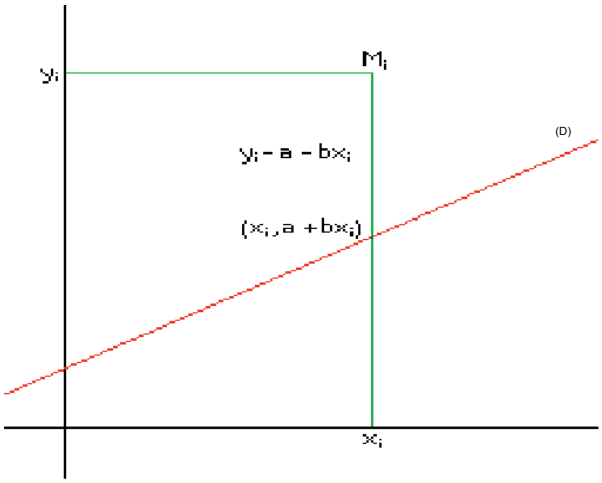
** Si $-1 < \rho < 0$, la correlation est négative (X et Y varient dans le sens contraire) La valeur $\rho = -1$ indique une relation linéaire parfaite $Y = aX + b$ avec $a < 0$. C'est un cas extrème très peu rencontré en pratique.*

Formule pratique de ρ pour le calcul

$$\rho = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\left[(\sum_{i=1}^{n} x_i^2) - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right]\left[(\sum_{i=1}^{n} y_i^2) - \frac{(\sum_{i=1}^{n} y_i)^2}{n}\right]}}.$$

b- Méthode des moindres carrés

Soit M_{ij} un point de cordonnées (x_i, y_i) On appelle distance de M_{ij} parallélernent à l'axe (oy) à la droite $(\triangle)$ d'équation $y = ax + b$, le réel positif $d_{ij} = |y_i - ax_i - b|$ (Attention ! il ne s'agit pas des distances des points M_i à la droite $(\triangle)$.)



Le problème admet une solution unique solution du système linéaire issue de l'annulation des dérivées partielles premières de la fonction $\delta(a,b)$. *Il s'agit de résoudre*

$$\begin{cases} \dfrac{\partial \delta}{\partial a} = -2\sum_{i=1}^{n} x_i \left(y_i - ax_i - b\right) = 0 \\ \dfrac{\partial \delta}{\partial b} = -2\sum_{i=1}^{n} \left(y_i - ax_i - b\right) = 0 \end{cases}$$

Ce système en a *et* b *admet pour solution unique*

$$a = \frac{cov(X,Y)}{V(X)} \text{ et } b = \overline{y} - a\overline{x}.$$

Théorème 2.1 *Soit* $M_i\left(x_i, y_i\right)_{1\leq i\leq n}$ *un ensemble fini de points fixes du plan euclidien où* x_i *sont les modalités d'un caractère* X *et* y_i *celles d'un autre caractère* Y *définis sur une même population* Ω.
Soient $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$, $V(X) = \dfrac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$ *et* $cov(X,Y) = \dfrac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$.

La droite d'équation $y - \overline{y} = a\left(x - \overline{x}\right)$ *où* $a = \dfrac{cov(X,Y)}{V(X)}$ *est la droite de régression de* y *en* x *et est notée* $D_{y/x}$.

Remarque 2.1 *i) On définit de même la droite de régression de* x *en* y *notée* $D_{x/y}$ *et d'équation*

$$x - \overline{x} = a'\left(y - \overline{y}\right) \text{ avec } a' = \frac{cov(X,Y)}{V(Y)}.$$

ii) Les droites $D_{y/x}$ et $D_{x/y}$ passent par le point $G(\overline{x}, \overline{y})$.

La méthode des moindres carrés consiste à chercher les valeures de a et b qui mininisent $\delta(a,b) = \sum_{i,j} (y_j - ax_i - b)^2 = \sum_{i,j} d_{ij}^2$.
On montre que $\delta(a,b)$ est minimun pour

$$a = \frac{cov(X,Y)}{V(X)} \text{ et } b = \overline{y} - a\overline{x}. \tag{2.1}$$

La droite $(\triangle)$ d'équation $y = ax + b$ où a et b vérifient (2.1) s'appelle droite de regression de y en x. Cette droite passe par le point $G(\overline{X}, \overline{Y})$ (car :$\overline{Y} = a\overline{X} + b$).

De même la droite de regression de x en y à pour équation : $x = a'y + b'$ où $a' = \dfrac{cov(x,y)}{V(y)}$ et $b' = x - a'y$
$(\triangle')$ passe aussi par $G(\overline{X}, \overline{Y})$.
Les valeurs de a' et b' sont celles qui minimisent : $\delta(a,b) = \sum_{i,j} (y_j - ax_i - b)^2 = \sum_{i,j} d_{ij}^2$.
Ces droites de régression sont aussi appelées droites des moindres carrés.

c- Ajustement et corrélation

Les droites de regression $(\triangle)$ et $(\triangle')$ ayant pour équations : $y = ax + b$, $x = a'y + b'$ ont les propriétés suivantes :

- *elles passent toutes les deux par le point $G(\overline{X}, \overline{Y})$ appelé point moyen de la statistique.*
- *les pentes des deux droites sont de même signe celui de la covariance et sont respectivement a et $\dfrac{1}{a'}$*
- *$aa' = \dfrac{cov(x,y)^2}{v(x)v(y)} = (\rho(x,y))^2$*
- *$(\triangle)$ et $(\triangle')$ sont confondues si elles ont la même pente (car elles passent toutes les deux par $G(\overline{X}, \overline{Y})$).*
- *dans ce cas $a = \dfrac{1}{a'}$ c.à.d. $aa' = 1$ donc $\rho(x,y) = l$. les points M_{ij} sont alors alignés.*
- *La corrélation linéaire est d'autant bonne (ou forte) que le coefficient de corrélation ρ est proche en valeur absolue de 1 ($\rho \simeq 1 \Longleftrightarrow a \simeq \dfrac{1}{a'}$).*
- *si ρ est proche de zéro, on dit qu'il y a corrélation linéaire très mauvaise entre X et Y. il faudrait alors approcher le nuage des points M_{ij} par une courbe.*

2.3.3 Exemple (Ajustement exponentiel) :

Exemple 2.1 *La statistique suivante indique l'évolution de la consommation d'énergie éléctrique dans un pays exprimée en TWh*

Année	*Consommation*
1949	30
1953	41
1957	56
1961	73
1965	97
1969	123
1973	165
1977	207

La relation qui lie la consommation au temps (année) est de type exponentiel. Déterminons la droite de régression de $Y = \log y$ *en* x *On effectue le changement de variable* $X = \dfrac{x - 1961}{4}$*, on a*

X_k	$Y_k = \log y_k$
-3	*1.417*
-2	*1.613*
-1	*1.748*
0	*1.863*
1	*1.987*
2	*2.019*
3	*2.217*
4	*2.316*

On en déduit $\overline{X} = \dfrac{1}{2}$, $\overline{Y} = \dfrac{15.18}{8} \simeq 1.8975$

$$\sigma_X^2 = 5.25 \quad \sigma_Y^2 \simeq 0.0794$$

$$\text{et } \sigma_{XY} \simeq 0.64.$$

La droite $Y = AX + B$ *est définie par*

$$\begin{aligned} A &= \frac{\sigma_{XY}}{\sigma_X^2} \simeq \frac{0.64}{5.25} \simeq 0.12 \\ B &= \overline{Y} - A\overline{X} \simeq 1.836. \end{aligned}$$

Remarquons que l'on a :

$$\rho_{XY} \simeq \frac{0.64}{\sqrt{5.25 \times 0.0794}} \simeq 0.99$$

ce qui justifie la recherche d'un ajustement exponentiel.

2.4 4- Conclusion

L'étude des séries statistiques à deux variables permet de mettre en rapport deux caractères afin de pouvoir déterminer une valeur manquante ou de prévoir une tendance. Néanmoins, deux caractères peuvent avoir un très fort coefficient de corrélation sans pour autant être réellement liés.

Fiche TD N° : 2

Exercice 2.4.1 :

1. *Montrer que :* $|\rho(X,Y)| \leq 1$.
2. *On considère deux séries statistiques* (x_i) *et* (y_i) *de taille n*
   Soient α_i *et* β_i *deux séries statistiques liées aux séries statistiques* (x_i) *et* (y_i) *par les relations suivantes :*
   $\forall i\ \alpha_i = \dfrac{x_i - c}{d}$ *avec* $d \neq 0\ c, d \in \mathbb{R}$
   $\forall i\ \beta_i = \dfrac{y_i - c'}{d'}$ *avec* $d' \neq 0\ c', d' \in \mathbb{R}$
   Montrer les propriétés suivantes :
   $i)\ cov(\alpha,\beta) = \dfrac{1}{dd'} cov(x,y)$ *et* $ii)\ \rho(\alpha,\beta) = \dfrac{|dd'|}{dd'}\rho(x,y)$

Exercice 2.4.2 :
Le tableau suivant représente des âges de patients X et les pressions systoliques Y de 9 malades.

L'âge X	*56*	*42*	*72*	*36*	*63*	*47*	*55*	*49*	*38*
Tension artérielle Y	*147*	*125*	*160*	*118*	*149*	*128*	*150*	*145*	*115*

1. *Calculer la moyenne et l'écart-type de chacun des deux caractère X et Y.*
2. *Calculer la covariance et le coefficient de corrélation du couple (X,Y). Que peut-on conclure ?*
3. *Trouver la droite de régression de X en Y.*
4. *Lorsque l'âge est 75 ans , quelle Tension artérielle Y peut-on prévoir ?*

Exercice 2.4.3 :
sur un échantillon de 100 étudiants, on relevé la taille X en centimètre, ainsi que le poids Y en kilogrammes comme l'indique le tableau suivant

$X \diagdown Y$	$[50,60[$	$[60,70[$	$[70,80[$	$[80,90[$	$[90,100[$
$[150,160[$	*10*	*3*	*1*	*0*	*0*
$[160,170[$	*2*	*12*	*6*	*7*	*2*
$[170,180[$	*1*	*7*	*11*	*17*	*4*
$[180,190[$	*0*	*2*	*2*	*4*	*9*

1. *Calculer la moyenne et l'écart-type de chacun des deux caractères X et Y*
2. *Calculer la covariance et le coefficient de corrélation du couple (X,Y). Que peut-on conclure ?*
3. *Trouver la droite de régression de Y en X.*

Exercice 2.4.4 *:*
sur un échantillon de 100 foyers, on a relevé le revenu mensuel moyen X en DH, ainsi le nombre de pièces habitées Y

$X \backslash Y$	*1*	*2*	*3*	*4*
$[1000, 2000[$	*6*	*3*	*1*	*0*
$[2000, 3000[$	*4*	*11*	*3*	*1*
$[3000, 4000[$	*1*	*10*	*16*	*3*
$[4000, 5000[$	*0*	*5*	*13*	*7*
$[5000, 6000[$	*0*	*1*	*5*	*10*

1. *Calculer la moyenne et l'écart-type de chacun des deux caractères X et Y*
2. *Calculer la covariance et le coefficient de corrélation du couple (X,Y). Que peut-on conclure ?*
3. *Trouver la droite de régression de Y en X.*

Chapitre 3

Analyse en composante principale (ACP)

3.1 Préambule

L'ACP propose, à partir d'un tableau de données relatives à p variables quantitatives portant sur n unités (individus), des représentations géométriques de ces unités et de ces variables.

Pour **les unités**, *on cherche si l'on peut distinguer des groupes en regardant quelles sont les unités qui se ressemblent, celles qui se distinguent des autres...*

Pour **les variables**, *on cherche quelles sont celles qui sont trés corrélées entre elles, celles qui, au contraire ne sont pas corrélées aux autres...*

3.2 Tableau de données

On dispose d'un tableau de données relatives à p variables quantitatives $x_1, \ldots, x_p$ *portant sur n unités (individus).*

Le tableau des données X *a la forme suivante :*

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ \vdots & \vdots & & \vdots \\ x_{i1} & x_{i2} & \ldots & x_{ip} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}.$$

x_{ij} *est la valeur de la variable* x_j *pour l'unité* i.

On peut représenter chaque unité par le vecteur de ses mesures sur les p

variables :

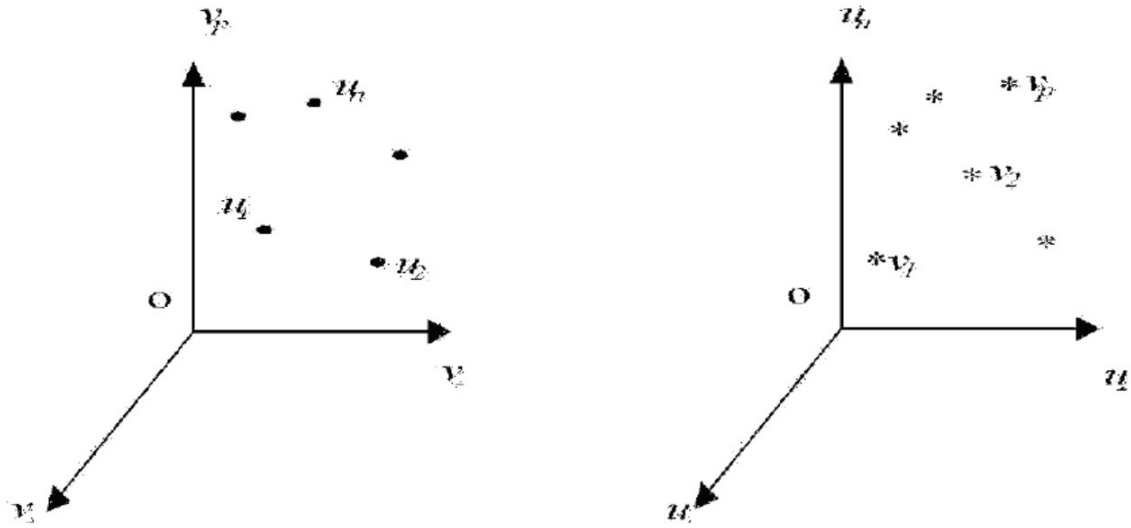
$$u_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \in \mathbb{R}^p.$$

L'ensemble des points qui représentent les unités est appelé **"nuage des individus"**, *qui est l'ensemble* $\{u_1, \ldots, u_n\}$.

Aussi, on peut représenter chaque variable par un vecteur de $\mathbb{R}^n$ *dont les composantes sont les valeurs de la variable pour les n unités :*

$$x_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix} \in \mathbb{R}^n.$$

L'ensemble des points qui représentent les variables est appelé **"nuage des variables"**, *qui est l'ensemble* $\{x_1, \ldots, x_p\}$.



3.3 Choix d'une distance

Pour faire une représentation géométrique, on doit choisir une distance entre deux points de l'espace.

La distance utilisée par l'ACP dans l'espace $\mathbb{R}^p$ *où sont représentées les unités, est la distance euclidienne classique.*

La distance entre deux unités u_i et $u_{i'}$ est

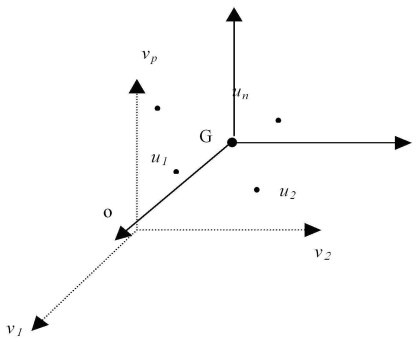
$$d^2(u_i, u_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Avec cette distance, toutes les variables jouent le même rôle et les axes définis par les variables constituent une base orthogonale.

3.4 Choix de l'origine

Le point O correspondant au vecteur de coordonnées toutes nulles n'est pas une origine satisfaisante, car si les coordonnées des points du nuage des individus sont grandes, le nuage est éloigné de cette origine.
On choisit comme origine le point

$$G = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ij} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ip} \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_j \\ \vdots \\ \bar{x}_p \end{pmatrix}$$



On travaille avec le tableau des données centrées :

$$X^c = \begin{pmatrix} x_{11} - \bar{x}_1 & \dots & x_{1p} - \bar{x}_p \\ \vdots & & \vdots \\ x_{i1} - \bar{x}_1 & \dots & x_{ip} - \bar{x}_p \\ \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & \dots & x_{np} - \bar{x}_p \end{pmatrix}.$$

Le vecteur des coordonnées centrées de l'unité u_i et le vecteur des coordonnées centrées de la variable x_j sont respectivement :

$$u_i^c = \begin{pmatrix} x_{i1} - \bar{x}_1 \\ x_{i2} - \bar{x}_2 \\ \vdots \\ x_{ip} - \bar{x}_p \end{pmatrix} \in \mathbb{R}^p, \quad x_j^c = \begin{pmatrix} x_{1j} - \bar{x}_j \\ x_{2j} - \bar{x}_j \\ \vdots \\ x_{nj} - \bar{x}_j \end{pmatrix} \in \mathbb{R}^n.$$

3.5 Matrice de variance

Les méthodes factorielles et leurs représentations géométriques utilisent la relation entre géométrie euclidienne et statistique empirique. Les statistiques élémentaires empiriques calculées sur n unités ont chacune leur correspondant géométrique dans un repère donné.

Pour un ensemble quelconque de variables $x_1, x_2, \cdots, x_m$ *:*

- *Variance et carré de la norme :* $Var(x_j) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2 = \frac{1}{n}\|\overrightarrow{Ox_j}\|^2$
- *Covariance et produit scalaire :*

$$cov(x_k, x_l) = \frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \overline{x_k})(x_{il} - \overline{x_l}) = \frac{1}{n} < \overrightarrow{Ox_k}, \overrightarrow{Ox_l} >$$

- *Coefficient de corrélation linéaire et cosinus d'angle :*

$$\rho(x_k, x_l) = \frac{cov(x_k, x_l)}{\sigma_{x_k}\sigma_{x_l}} = \frac{< \overrightarrow{Ox_k}, \overrightarrow{Ox_l} >}{\|\overrightarrow{Ox_k}\|\|\overrightarrow{Ox_l}\|} = \cos(\overrightarrow{Ox_k}, \overrightarrow{Ox_l})$$

.

Définition 3.1 *: On appelle matrice de variance la matrice symétrique* Σ *contenant les variances* $Var(x_j)$ *sur la diagonale et les covariances* $cov(x_k, x_l)$ *en dehors de la diagonale (ligne k colonne l pour* $cov(x_k, x_l)$*).*

$$\Sigma = \begin{pmatrix} Var(x_1) & cov(x_1, x_2 & \dots & cov(x_1, x_n) \\ cov(x_2, x_1) & Var(x_2) & \dots & cov(x_2, x_n) \\ \vdots & \vdots & & \vdots \\ cov(x_n, x_1) & cov(x_n, x_2) & \dots & Var(x_n) \end{pmatrix}.$$

Cette matrice s'écrit : $\Sigma = \frac{1}{n}(X^c)^t X^c$

De même, on définit le coefficient de corrélation linéaire entre les variables x_k *et* x_l *par* $\rho(x_k, x_l) = \frac{cov(x_k, x_l)}{\sigma_{x_k}\sigma_{x_l}}$.
Ce coefficient exprime le niveau de corrélation (linéaire) entre les variables x_k *et* x_l *: plus il est proche de 1, plus les variables sont corrélées positivement, plus il est proche de -1, plus elles sont corrélées négativement. Un coefficient de corrélation nul indique l'absence de corrélation linéaire.*

En divisant chaque colonne j du tableau centré X^c *par l'écart-type* σ_{x_j} *de la variable* x_j *, on construit le tableau Z des données centrées réduites :* $z_{ij} = \frac{x_{ij} - \overline{x_j}}{\sigma_{x_j}}$ *. La matrice Z s'exprime en fonction de* X^c *par* $Z = D_{1/\sigma}X^c$

où $D_{1/\sigma}$ est la matrice diagonale contenant $\frac{1}{\sigma_{x_1}}, \cdots, \frac{1}{\sigma_{x_p}}$ sur sa diagonale. Le terme réduit signifie que les variances des variables z_j sont égales à 1.

La matrice $R = D_{1/\sigma}\Sigma D_{1/\sigma}$ est dite de corrélation. Regroupant les coefficients de corrélation linéaire entre les p variables prises deux à deux, elle résume la structure des dépendances linéaires entre les p variables. Elle est symétrique et sa diagonale est composée de 1.

3.6 Moments d'inertie

3.6.1 Inertie totale du nuage des individus

Définition 3.2 *: Le moment d'inertie du nuage des individus par rapport au centre de gravité G est :*

$$I_G = \frac{1}{n}\sum_{i=1}^{n} d^2(G, u_i) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \bar{x}_j)^2$$

C'est une mesure de la dispersion du nuage des individus par rapport à son centre de gravité.

Si ce moment d'inertie est grand, cela signifie que le nuage est très dispersé ; s'il est petit, alors le nuage est très concentré autour de son centre de gravité.

Remarque 3.1 *I_G peut aussi s'écrire sous la forme suivante :*

$$I_G = \sum_{j=1}^{p}\left[\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2\right] = \sum_{j=1}^{p} var(x_j),$$

où $Var(x_j)$ est la variance de la variable x_j. Sous cette forme, l'inertie totale est égale à la trace de la matrice de covariance $\sum$ des p variables. ($I_G = trace(\Sigma)$)

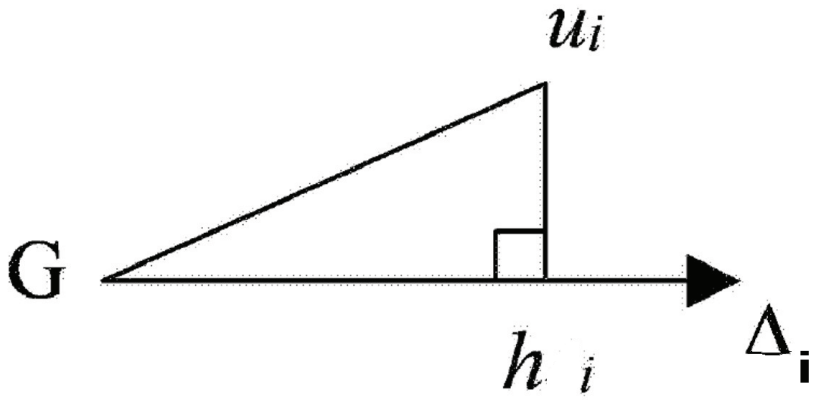
3.6.2 Inertie du nuage des individus par rapport à un axe passant par G

Définition 3.3 *: L'inertie du nuage des individus par rapport à un axe Δ passant par G est égale à :*

$$I_\Delta = \frac{1}{n}\sum_{i=1}^{n} d^2(h_i, u_i)$$

où h_i est la projection orthogonale de u_i sur l'axe Δ.

Cette inertie mesure la proximité à l'axe Δ du nuage des individus.

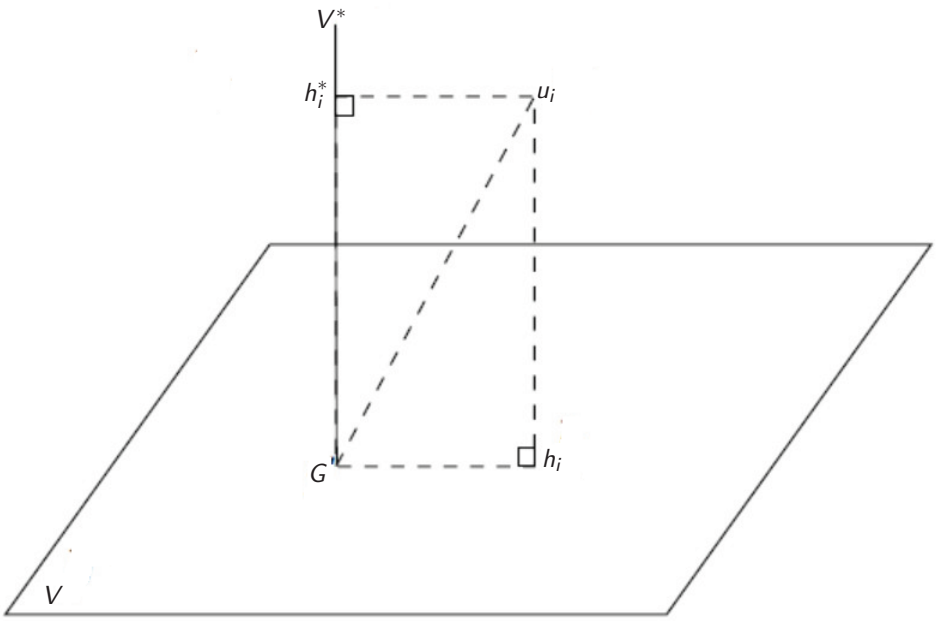


3.6.3 Inertie du nuage des individus par rapport à un sous-espace vectoriel V passant par G

Définition 3.4 *: L'inertie du nuage des individus par rapport à un sous-espace vectoriel V passant par G est égale à :*

$$I_V = \frac{1}{n}\sum_{i=1}^{n} d^2(h_i, u_i),$$

où h_i

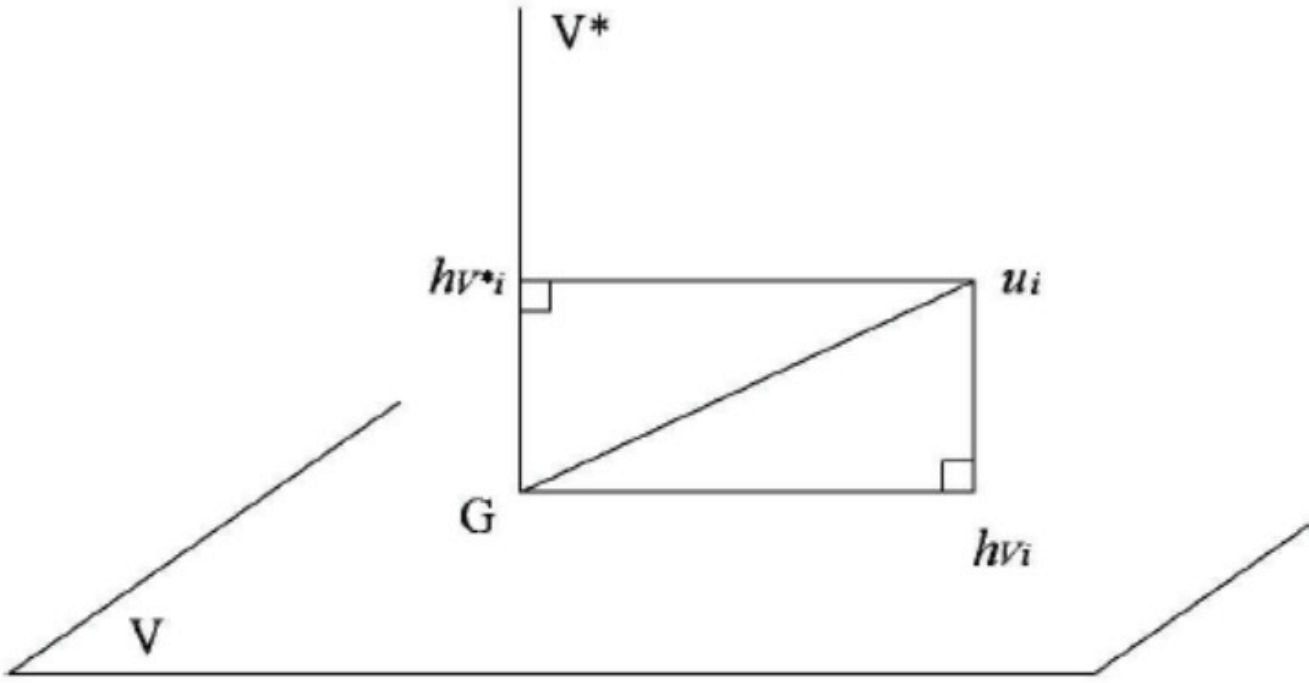


3.6.4 Décomposition de l'inertie totale

Si on note V^* le complémentaire orthogonal d'un espace vectoriel V dans $\mathbb{R}^p$ et h_i^* la projection orthogonale de u_i sur V^*, en appliquant le théorème de Pythagore, on peut écrire :

$$d^2(h_i, u_i) + d^2(h_i^*, u_i) = d^2(G, u_i) = d^2(G, h_i) + d^2(G, h_i^*),$$

où h_i est la projection orthogonale de u_i sur V, et h_i^* est la projection orthogonale de u_i sur V^*.
Par exemple dans un espace de dimension 3
On en déduit que :



$$I_G = I_V + I_{V^*},$$

c'est le théorème de Huygens.
Cas général :

Théorème 3.1 *Si on décompose l'espace $\mathbb{R}^p$ comme la somme de sous-espaces de dimension 1 et orthogonaux entre eux (i.e. $\mathbb{R}^p = \Delta_1 \oplus \Delta_2 \oplus \ldots \oplus \Delta_p$), alors*

$$I_G = I_{\Delta_1^*} + I_{\Delta_2^*} + \ldots + I_{\Delta_p^*}.$$

3.7 Dérivation matricielle (Rappel)

Définition 3.5 *:*

Soient une forme linéaire $f : \mathbb{R}^k \to \mathbb{R}$ *et* $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix} \in \mathbb{R}^k$.

On appelle dérivée de f *en* x *et on note* $\dfrac{\partial f}{\partial x}$ *ou* $\nabla f(x)$ *ou encore* $f'(x)$ *le vecteur colonne des dérivées partielles de* f *par rapport aux* x_i *:*

$$\frac{\partial f}{\partial x} = \begin{pmatrix} \dfrac{\partial f}{\partial x_1} \\ \dfrac{\partial f}{\partial x_2} \\ \vdots \\ \dfrac{\partial f}{\partial x_k} \end{pmatrix}$$

3.7.1 Dérivation de formes linéaires

$$\text{Soit } f : \begin{cases} \mathbb{R}^k & \to & \mathbb{R} \\ x = (x_1, x_2, \cdots, x_k)^t & \mapsto & \displaystyle\sum_{i=1}^{k} a_i x_i \end{cases}$$

Remarque 3.2 *:*

$$\sum_{i=1}^{k} a_i x_i =< x, a >=< a, x >= a^t x = x^t a$$

Proposition 3.1 *:*

$$\forall a, x \in \mathbb{R}^k \quad on\ a \quad \frac{\partial a^t x}{\partial x} = \frac{\partial x^t a}{\partial x} = a$$

en effet On a $\forall 1 \leq i \leq k \quad \dfrac{\partial f}{\partial x_i} = a_i$ *donc* $\dfrac{\partial f}{\partial x} = a$

3.7.2 Dérivation d'une forme quadratique

Définition 3.6 *:*

Une forme quadratique est un polynôme homogène de degré 2 avec un nombre

quelconque de variables :

$$f : \begin{cases} \mathbb{R}^k & \to & \mathbb{R} \\ x = (x_1, x_2, \cdots, x_k)^t & \mapsto & \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij} x_i x_j \end{cases}$$

En notant $A = (a_{ij}) \in \mathcal{M}_{k,k}(\mathbb{R})$, *f s'écrit :*

$$f : \begin{cases} \mathbb{R}^k & \to & \mathbb{R} \\ x = (x_1, x_2, \cdots, x_n)^t & \mapsto & x^t A x \end{cases}$$

Remarque 3.3 *:*

$$x^t A x =< x, Ax >=< Ax, x >$$

Proposition 3.2 *:*

$$\forall x \in \mathbb{R}^k, A = (a_{ij}) \in \mathcal{M}_{k,k}(\mathbb{R}) \frac{\partial x^t A x}{\partial x} = (A + A^t)x$$

Remarque 3.4 *:*

- *Si A est symétrique, i.e : si* $A^t = A$, *on a :* $\frac{\partial x^t A x}{\partial x} = 2Ax.$
- *En particulier, pour* $A = I_k$, *on a :* $\frac{\partial x^t x}{\partial x} = 2x.$

3.8 Recherche des axes principaux

3.8.1 Recherche du premier axe principal Δ_1 passant par G d'inertie minimum

On cherche un axe Δ_1 *passant par G d'inertie* I_{Δ_1} *minimum car c'est l'axe le plus proche de l'ensemble des points du nuage des individus.*

La droite Δ_1 *s'appelle le premier axe principal.*

Si on utilise la relation entre les inerties, donnée au paragraphe précédent, rechercher Δ_1 *tel que* I_{Δ_1} *est minimum, est équivalent à chercher* Δ_1 *tel que* $I_{\Delta_1^*}$ *est maximum.*

On définit l'axe Δ_1 *par son vecteur directeur unitaire* $\overrightarrow{Ga_1}$.

Problème :

Trouver $\overrightarrow{Ga_1}$ *tel que* $I_{\Delta_1^*}$ *est maximum sous la contrainte que* $||\overrightarrow{Ga_1}||^2 = 1$.

On définit l'axe Δ_1 *par son vecteur directeur unitaire* $\overrightarrow{Ga_1}$.

Problème :

Trouver $\overrightarrow{Ga_1}$ *tel que* $I_{\Delta_1^*}$ *est maximum sous la contrainte que* $||\overrightarrow{Ga_1}||^2 = 1$.

Expressions algébriques de $I_{\Delta_1^*}$ **et de** $||\overrightarrow{Ga_1}||^2$

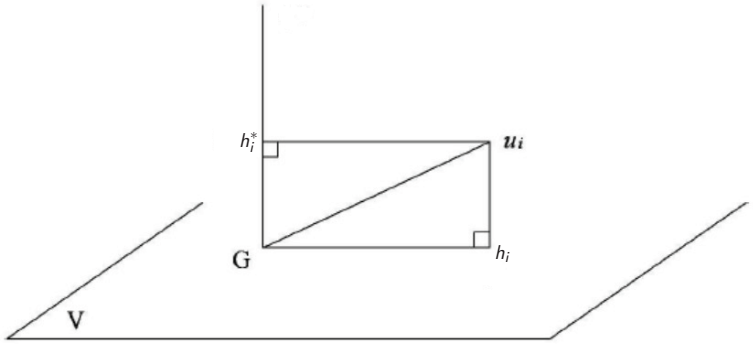
Proposition 3.3

$$I_{\Delta_1^*} = a_1^t \Sigma a_1 \quad et \quad ||\overrightarrow{Ga_1}||^2 = a_1^t a_1.$$

Avec, Σ est la matrice de variance-covariance de $x_1, \ldots, x_p$.

Preuve : *On a*

$$< \overrightarrow{Gu_i}, \overrightarrow{Ga_1} >= \|\overrightarrow{Ga_1}\|\|\overrightarrow{Gh_i}\| = 1 \times \|\overrightarrow{Gh_i}\| = d(G, h_i) = d(u_i, h_i^*)$$



$$\Longrightarrow d^2(u_i, h_i^*) = d^2(G, h_i) =< \overrightarrow{Gu_i}, \overrightarrow{Ga_1} >^2 = a_1^t u_i^c (u_i^c)^t a_1,$$

par symétrie du produit sclaire.

Ainsi

$$I_{\Delta_1^*} = \frac{1}{n} \sum_{i=1}^{n} a_1^t u_i^c (u_i^c)^t a_1 = a_1^t \left[\frac{1}{n} \sum_{i=1}^{n} u_i^c (u_i^c)^t \right] a_1$$

Par suite

$$I_{\Delta_1^*} = a_1^t \Sigma a_1$$

et $||\overrightarrow{Ga_1}||^2 = a_1^t a_1$.

Recherche du maximum

Le problème à résoudre : trouver a_1 tel que $a_1^t \Sigma a_1$ soit maximum avec la contrainte $a_1^t a_1 = 1$.

C'est un problème de recherche d'un optimum d'une fonction de plusieurs variables liées par une contrainte (les inconnues sont les composantes de a_1).

La méthode des multiplicateurs de Lagrange peut alors être utilisée.

Dans le cas de la recherche de a_1, il faut calculer les dérivées partielles de :

$$g(a_1) = g(a_{11}, a_{12}, ..., a_{1p}) = a_1^t \Sigma a_1 - \lambda_1 (a_1^t a_1 - 1).$$

En utilisant la dérivée matricielle, on obtient :

$$\frac{\partial g(a_1)}{\partial a_1} = 2\Sigma a_1 - 2\lambda_1 a_1 = 0.$$

Le système à résoudre est :

$$\begin{cases} \Sigma a_1 - \lambda_1 a_1 &= 0 \ (1) \\ a_1^t a_1 - 1 &= 0 \ (2) \end{cases}$$

De l'équation matricielle (1) *de ce système on déduit que* a_1 *est vecteur propre de la matrice* Σ *associé à la valeur propre* λ_1*.*

En multipliant à gauche par a_1^t *les deux membres de l'équation* (1)*, on obtient :*

$$a_1^t \Sigma a_1 - \lambda_1 a_1^t a_1 = 0$$

et en utilisant l'équation (2) *on trouve que :*

$$a_1^t \Sigma a_1 = \lambda_1.$$

Le premier membre de l'équation précédente est égal à l'inertie $I_{\Delta_1^*}$ *qui doit être maximum. Cela signifie que la valeur propre* λ_1 *est la plus grande valeur propre de la matrice de covariance* Σ *; cette valeur propre est égale à l'inertie portée par l'axe* Δ_1^**.*

Proposition 3.4 *:*
L'axe Δ_1 *pour lequel le nuage des individus a l'inertie minimum a comme vecteur directeur unitaire le premier vecteur propre associé à la plus grande valeur propre de la matrice de covariance* Σ*.*

3.8.2 Recherche du deuxième axe principal Δ_2 passant par G, orthogonal à Δ_1 et d'inertie minimum

On recherche ensuite un deuxième axe Δ_2 *orthogonal au premier axe principal* Δ_1 *et d'inertie minimum.*

La droite Δ_2 *s'appelle le deuxième axe principal.*

On peut, comme dans le paragraphe précédent, définir l'axe Δ_2 *passant par* G *par son vecteur directeur unitaire* a_2*. L'inertie du nuage des individus par rapport à son complémentaire orthogonal est égale à :*

$$I_{\Delta_2^*} = a_2^t \Sigma a_2$$

elle doit être maximum avec les deux contraintes suivantes :

$$a_2^t a_2 = 1 \ et \ a_2^t a_1 = 0.$$

En appliquant la méthode des multiplicateurs de Lagrange, cette fois avec deux contraintes, on trouve que a_2 est le vecteur propre de Σ correspondant à la deuxième plus grande valeur propre. On peut montrer que le plan défini par les axes Δ_1 et Δ_2 est le sous-espace de dimension 2 qui génère un inertie minimum du nuage.

3.8.3 Recherche des axes principaux suivants

On peut rechercher de nouveaux axes principaux en suivant la même procédure.

Les vecteurs des nouveaux axes principaux sont tous des vecteurs propres de Σ correspondant aux valeurs propres ordonnées. La matrice de covariance Σ étant une matrice symétrique réelle, elles possède p vecteurs propres réels, formant une base orthogonale de $\mathbb{R}^p$.

$$\begin{cases} \Delta_1 \perp \Delta_2 \perp \ldots \perp \Delta_p \\ a_1 \perp a_2 \perp \ldots \perp a_p \\ \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \\ I_{\Delta_1^*} \geq I_{\Delta_2^*} \geq \ldots \geq I_{\Delta_p^*} \end{cases}$$

On passera de la base orthogonale initiale des variables centrées à la nouvelle base orthogonale des vecteurs propres de Σ.

3.9 Contributions des axes à l'inertie totale

On utilise le théorème de Huygens pour décomposer l'inertie totale du nuage des individus

$$I_G = I_{\Delta_1^*} + I_{\Delta_2^*} + \ldots + I_{\Delta_p^*} = \lambda_1 + \ldots + \lambda_p.$$

Définition 3.7
La contribution absolue *de l'axe Δ_k à l'inertie totale du nuage des individus est égale à :*

$$ca(\Delta_k / I_G) = \lambda_k.$$

Sa contribution relative *est égale à :*

$$cr(\Delta_k / I_G) = \frac{\lambda_k}{\lambda_1 + \ldots + \lambda_p}$$

On emploie souvent l'expression "pourcentage d'inertie expliquée par Δ_k".

On peut étendre ces définitions à tous les sous-espaces engendrés par les nouveaux axes. Ainsi, le pourcentage d'inertie expliqué par le plan engendré par les deux premiers axes Δ_1 et Δ_2 est égal à :

$$cr(\Delta_1 \oplus \Delta_2 / I_G) = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \ldots + \lambda_p}$$

Ces pourcentages d'inertie sont des indicateurs qui rendent compte de la part de variabilité du nuage des individus expliquée par ces sous-espaces. Si les dernières valeurs propres ont des valeurs faibles, on pourra négliger la variabilité qu'expliquent les axes correspondants.

3.10 Composantes principales

3.10.1 Expression des composantes principales

A chaque axe pricipal est associé une variable appelée "composante principale".

Définition 3.8 *:*
La composante principale C_k *est le vecteur des cordonnées des projections des individus sur l'axe* Δ_k*. Elle s'écrit sous forme de combinaison linéaire des variables initiales (centrée).*

$$C_k = X_c \; a_k \in \mathbb{R}^n$$

Ce sont les coordonnées des n individus sur le kème axe principal.

Proposition 3.5 *La kème composante principale est*

$$C_k = \begin{pmatrix} c_k^1 \\ c_k^2 \\ \vdots \\ c_k^n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{p} x_{1j}^c a_j^k \\ \sum_{j=1}^{p} x_{2j}^c a_j^k \\ \vdots \\ \sum_{j=1}^{p} x_{nj}^c a_j^k \end{pmatrix} \in \mathbb{R}^n.$$

Preuve :*en effet*

$$C_k = \begin{pmatrix} c_k^1 \\ c_k^2 \\ \vdots \\ c_k^n \end{pmatrix} = \begin{pmatrix} x_{11}^c & x_{12}^c & \dots & x_{1p}^c \\ x_{21}^c & x_{22}^c & \dots & x_{2p}^c \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1}^c & x_{n2}^c & \dots & x_{np}^c \end{pmatrix} . \begin{pmatrix} a_1^k \\ a_2^k \\ \vdots \\ a_p^k \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{p} x_{1j}^c a_j^k \\ \sum_{j=1}^{p} x_{2j}^c a_j^k \\ \vdots \\ \sum_{j=1}^{p} x_{nj}^c a_j^k \end{pmatrix}$$

3.10.2 Propriétés des composantes principales

Proposition 3.6

- *Chaque composante principale C_k est une variable centrée.*
- *La variance d'une composante principale C_k est égale à l'inertie apportée par l'axe principal qui lui est associé.*
  Ainsi, $var(C_1) = \lambda_1$, $var(C_2) = \lambda_2$, ..., $var(C_p) = \lambda_p$.
- *Les composantes principales sont non corrélées deux à deux car les axes principaux associés sont orthogonaux.*

Preuve :

$$\overline{C_k} = \frac{1}{n}\sum_{i=1}^{n} c_k^i = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^c a_j^k = \frac{1}{n}\sum_{j=1}^{p} a_j^k \sum_{i=1}^{n} x_{ij}^c = 0$$

$$Var(C_k) = \frac{1}{n} C_k^t C_k = \frac{1}{n}(X^c.a_k)^t(X^c.a_k) = \frac{1}{n} a_k^t (X^c)^t X^c.a_k = \lambda_k$$

$$\begin{aligned} \rho(C_k, C_l) &= < C_k, C_l >= C_k^t C_l = (X^c.a_k)^t.(X^c.a_l) \\ &= a_k^t.(X^c)^t.X^c.a_l = n a_k^t.\Sigma.a_l = n a_k^t.\lambda_l a_l \\ &= n\lambda_l a_k^t.a_l = 0 \end{aligned}$$

$$\Longrightarrow C_k \perp C_l$$

3.11 Représentation des individus

3.11.1 Coordonnées des individus

Dans la nouvelle base $(a_1, \ldots, a_p)$, les coordonnées de l'individu i sont

$$\begin{pmatrix} c_1^i \\ c_2^i \\ \vdots \\ c_p^i \end{pmatrix} \in \mathbb{R}^p.$$

Ici,c_j^i est le ième terme de la composante principale C_j.

3.11.2 Qualité de la représentation des individus et l'étude de la proximité entre les individus

Si les projections des individus sont éloignés sur un axe (ou sur un plan), alors les points représentants ces individus sont éloignés dans l'espace. Cependant, deux individus dont les projections sont proches sur un axe (ou sur un plan) peuvent ne pas être proches dans l'espace.

Pour interpréter correctement la proximité des projections de deux individus sur un plan, il faut s'assurer que ces individus sont bien représentés dans l'axe (ou dans le plan). Pour que l'individu i soit bien représenté sur un axe (ou sur un plan, ou un sous-espace), il faut que l'angle entre le vecteur $\overrightarrow{Gu_i}$ et l'axe (ou le plan, ou le sous-espace) soit petit (le cosinus est proche de 1).

Représentation par rapport à un axe

Proposition 3.7 *Le carré du cosinus de l'angle α_{ik} entre $\overrightarrow{Gu_i}$ et un axe Δ_k de vecteur directeur unitaire a_k est égal à :*

$$\cos^2(\alpha_{ik}) = \frac{< \overrightarrow{Gu_i}\overrightarrow{Ga_k} >^2}{||\overrightarrow{Gu_i}||^2} = \frac{a_k^t \; u_i^c \; (u_i^c)^t \; a_k}{(u_i^c)^t \; u_i^c}.$$

Représentation par rapport à un plan

En utilisant le théorème de Pythagore, on peut montrer que le carré du cosinus de l'angle d'un vecteur avec un plan engendré par deux vecteurs orthogonaux, est égal à la somme des carrés des cosinus des angles du vecteur avec chacun des deux vecteurs qui engendrent le plan. Ainsi, le carré du cosinus de l'angle $\alpha_{ikk'}$ entre $\overrightarrow{Gu_i}$ et le plan engendré par deux axes Δ_k et $\Delta_{k'}$ est

$$\cos^2 \alpha_{ikk'} = \cos^2 \alpha_{ik} + \cos^2 \alpha_{ik'}$$

Remarque :

Cette propriété se généralise à l'angle d'un vecteur avec un sous-espace de dimension quelconque (≥ 2).

Interprétation de la Représentation

Si le carré du cosinus de l'angle entre $\overrightarrow{Gu_i}$ et l'axe (ou le plan, ou le sous- espace) est proche de 1, alors l'individu i est bien représenté par sa projection sur l'axe (ou le plan, ou le sous-espace). Et si deux individus sont bien représentés en projection sur un axe (ou un plan, ou un sous-espace) et ont des projections proches, alors ils sont proches dans l'espace.

Remarque : *Si un individu est très proche du centre de gravité dans l'espace, c'est-à-dire si $||\overrightarrow{Gu_i}||^2$ est très petit, le point représentant cet individu sur un axe (ou un plan, ou un sous-espace) sera bien représenté.*

3.11.3 Interprétation des axes principaux en fonction des individus

Un individu contribue à la confection d'un axe lorsque sa projection sur cet axe sera éloignée du centre de gravité du nuage. Inversement, un individu dont la projection sur un axe sera proche du centre de gravité contribue faiblement à l'inertie portée par cette axe.

3.12 Représentation des variables

3.12.1 Coordonnées des variables

Soient $C_1, \ldots, C_p$ les composantes principales. Il est intéressant de voir comment les anciennes variables x_j^c sont liées à ces composantes principales C_k ; on calcule alors les corrélations entre C_k et x_j^c.

La représentation des anciennes variables se fera en prenant comme coordonnées leurs coefficients de corrélation avec les composantes principales. On obtient alors ce que l'on appelle le "cercle des corrélations".

Proposition 3.8 *le coefficient de corrélation entre C_k et x_j^c est*

$$cor(C_k, x_j^c) = \sqrt{\lambda_k}\frac{a_{kj}}{\sqrt{var(x_j)}},$$

où a_{kj} est la $j$ème coordonnée du vecteur directeur unitaire a_k de Δ_k.

Preuve : *On a $x_j^c = x^c.e_j$ avec e_j est le $j^{ème}$ vecteur de la base canonique de $\mathbb{R}^n$*

Donc

$$\begin{aligned} Cov(C_k, x_j^c) &= \frac{1}{n} < C_k, x_j^c >= \frac{1}{n} C_k^t . x_j^c \\ &= \frac{1}{n} a_k^t (X^c)^t X^c e_j = a_k^t \Sigma e_j \\ &= \lambda_k a_k^t e_j = \lambda_k a_{kj} \end{aligned}$$

$$\Longrightarrow \rho(C_k, x_j^c) = \frac{Cov(C_k, x_j^c)}{\sqrt{var(C_k)}\sqrt{var(x_j)}} = \frac{\lambda_k a_{kj}}{\sqrt{\lambda_k}\sqrt{var(x_j)}} = \sqrt{\lambda_k}\frac{a_{kj}}{\sqrt{var(x_j)}}$$

3.12.2 Qualité de la représentation des variables

Une variable sera d'autant mieux représentée sur un axe que sa corrélation avec la composante principale correspondante est en valeur absolue proche de 1.

Ainsi, une variable sera bien représentée sur un plan si elle est proche du bord du cercle des corrélations.

3.12.3 Interprétation des axes principaux en fonction des anciennes variables

Une variable x_j explique d'autant mieux un axe principal qu'elle est fortement corrélée avec la composante principale correspondant à cet axe.

3.12.4 Etude des liaisons entre les variables

Sur le graphique du cercle des corrélations, on peut aussi interpréter les positions des anciennes variables les unes par rapport aux autres en termes de corrélations.

- *Deux points très proches entre elles et très proches du cercle des corrélations, donc bien représentées dans le plan, seront très corrélées positivement entre elles.*
- *Si elles sont proches du cercle, mais dans des positions symétriques par rapport à l'origine, elles seront très corrélées négativement.*
- *Deux variables proches du cercle des corrélations et dont les vecteurs qui les joignent à l'origine forment un angle droit, ne seront pas corrélées entre elles.*

3.13 Interprétation

A partir des relations données précédemment, nous pouvons définir quelques règles pour l'interprétation :

- *Il est naturel de commencer l'examen détaillé des graphiques par les variables parce qu'elles sont moins nombreuses et plus chargée de sens que les individus*
- *Interprétation axe par axe : Interpréter un axe factoriel consiste à donner un " sens " à l'axe en recensant les variables les plus liés à chaque axe. Deux situations typiques peuvent se produire :*
  1. *Toutes les variables très liées au facteur sont situées d'un même coté de l'axe. Le facteur apparait alors comme une synthèse entre ces variables.*
  2. *les variables très liées au facteur présentent une coordonnée positive pours les unes et négative pour les autres. Il faut alors rechercher un dénominateur commun qui, à la fois, relies les variables situées de même côté et oppose les variables situées de part et d'autre de l'origine.*

Par exemple, supposons que les variables soient des notes dans différentes matières : un facteur peut traduire l'opposition entre matières scientifiques et matières littéraire.

- *Interprétation par plans : Le plans factoriel apport le pouvoir synthétique du graphique, et la prise en compte simultanée de deux dimensions qui donne une image plus fidèle des données et peut aussi suggérer d'interpréter d'autre directions que les axes factoriels. Il est utile de représenter en plus des points (variables) : " Le cercle de rayons 1, ou cercle de corrélations. " Les vecteurs joignant l'origine aux points variables afin de visualiser les angles qui mesurent la liaison entre variables.*
- *Un individu sera du côté des variables pour lesquelles il a de fortes valeurs, inversement il sera du côté opposé des variables pour lesquelles il a de faibles valeurs.*
- *Plus les valeurs d'un individu sont fortes pour une variable plus il sera éloigné de l'origine suivant l'axe factoriel décrivant le mieux cette variable.*
- *Deux individus à une même extrémité d'un axe (i.e. éloignés de l'origine) sont proches (i.e. se ressemblent). - Deux variables très corrélées positivement sont du même côté sur un axe.*
- *Il n'est pas possible d'interpréter la position d'un individu par rapport à une seule variable, et réciproquement, il n'est pas possible d'interpréter la position d'une variable par rapport à un seul individu. Les interprétations doivent se faire de manière globale.*

3.14 ACP normée

Jusqu'à présent, on a étudié **l'ACP simple**, *pour laquelle, tous les individus ont le même poids et toutes les variables sont traitées de façon symétrique (elles jouent le même rôle) et les axes principaux sont issus de la matrice de covariance des variables. Cela pose parfois certains problèmes :*

- *le premier problème : les anciennes variables sont hétérogènes, comme par exemple des poids, des tailles et des âges ;*
- *le deuxième problème : si on change d'unités sur ces variables, on peut changer complètement les résultats de l'ACP ;*
- *le dernier problème : une variable contribue d'autant plus à la confection des premiers axes que sa variance est forte.*

Pour échapper à tous ces problèmes, on travaile sur des variables centrées et réduites.

Cela revient à faire la même analyse que pour l'ACP simple, mais à choisir une autre distance euclidienne entre les individus que la distance euclidienne classique. La distance choisie est :

$$d^2(u_i, u_{i'}) = \sum_{j=1}^{p} \frac{1}{\sigma_j^2}(x_{ij} - x_{i'j})^2$$

Si on reprend tous les calculs de l'ACP simple, mais en remplaçant les variables de départ par les variables centrées réduites, on voit que ce n'est plus la matrice de covariance, mais la matrice de corrélation R qui intervient pour la recherche des nouveaux axes.

Puisque la matrice de corrélation R n'a que des 1 *sur sa diagonale principale, sa trace est égale à p. Par suite, l'inertie totale du nuage des individus dans $\mathbb{R}^p$ est égale à p.*

3.15 Récapitulatif : démarche d'une ACP

- *Tableau de données (n unités/p variables quantitatives).*
- *Calcul de la matrice de covariance (pour une ACP simple) ou la matrice de corrélation (pour une ACP normée).*
- *Recherche des valeurs propres $\lambda_1, \ldots, \lambda_p$.*
- *Recherche des vecteurs propres $a_1 \in \mathbb{R}^n, \ldots, a_p \in \mathbb{R}^n$.*
- *Calucl du pourcentage d'inertie ou de contribution de l'axe à la variation totale (pourcentage expliqué par les axes principaux).*
- *Détermination des composantes principales $C_1 \in \mathbb{R}^n, \ldots, C_p \in \mathbb{R}^n$.*
- *Etude du nuage des variables : Coordonnées des variables sur les axes principaux (corrélation entre les variables et les composantes princi-*

pales) ; qualité de la représentation des variables, étude de liaison entre les variables.

- *Etude du nuage des individus : Coordonnées des individus sur les axes principaux, qualité de la représentation des individus, étude de la proximité.*

Fiche TD N° : 4

Exercice 3.15.1 *:*

Soit V le tableau suivant : $V = \frac{1}{4}\begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$.

1. *Calculer le trace de V*
2. *Déterminer les différentes valeurs propres de V*
3. *Trouver un vecteur propre associé à la valeur propre $\lambda = 0$*
4. *Déterminer des vecteurs propres orthonormés de V associés aux valeurs propres de V différentes de 0*

Exercice 3.15.2 *:*

Soit X le tableau suivant : $X = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}$.

On considère les six vecteurs $x_1, x_2, \cdots, x_6$ de $\mathbb{R}^3$(muni de la métrique euclidien usuelle) dont les composantes sont données par les lignes de X . On pose $m_i = 1/6$ pours $i = 1; 2; \cdots, 6$

1. *Montrer que le nuage N de six vecteurs est centré à l'origine de $\mathbb{R}^3$. Calculer $V = 1/6 \sum_{i=1}^{6} x_i x_i'$*
2. *Calculer I_G, moment d'inertie de N par rapport à l'origine*
3. *Déterminer les différentes valeurs propres de V*
4. *Déduire la dimension de l'espace qui contient le nuage des individus*
5. *Trouver un vecteur propre associé à la valeur propre $\lambda = 0$*
6. *Déterminer des vecteurs propres orthonormés de V associés aux valeurs propres de V différentes de 0, et représenter N dans le sous-espace de $\mathbb{R}^3$ engendré par ces vecteurs.*

Exercice 3.15.3 *:*

On considère Le tableau Y de notes sur 20 obtenues par 9 élèves en mathématiques, physique , français, et anglais.(n=9 individus , p=4 variables) :

	mathématiques	*physique*	*français*	*anglais*
Jean	*6.0*	*6.0*	*5.0*	*5.5*
Aline	*8.0*	*8.0*	*8.0*	*8.0*
Annie	*6.0*	*7.0*	*11.0*	*9.5*
Monique	*14.5*	*14.5*	*15.5*	*15.0*
Didier	*14.0*	*14.0*	*12.0*	*12.5*
André	*11.0*	*10.0*	*5.5*	*7.0*
Pierre	*5.5*	*7.0*	*14.0*	*11.5*
Brigitte	*13.0*	*12.5*	*8.5*	*9.5*
Evelyne	*9.0*	*9.5*	*12.5*	*12.0*

1. *montrer que le centre de gravité est donné par le vecteur :* $G = \begin{pmatrix} 9.67 \\ 9.83 \\ 10.22 \\ 10.06 \end{pmatrix}$.
2. *On désire soumettre le tableau Y à un ACP. Pour cela on est conduit à rechercher les vecteur propre de la matrice* $V = 1/9X'X$ *des variances-covariances des cinq variables, qui est*

$$V = \begin{array}{|c|c|c|c|} \hline 11.389 & 9.917 & 2.657 & 4.824 \\ \hline 9.917 & 8.944 & 4.120 & 5.481 \\ \hline 2.657 & 4.120 & 12.062 & 9.293 \\ \hline 4.824 & 5.481 & 9.293 & 7.914 \\ \hline \end{array}$$

   a- *Indiquer la transformation qui permet de passer de la matrice Y à la matrice X. Calculer la première ligne de X*

   b- *Calculer* I_G*, moment d'inertie de N par rapport à l'origine*

   c- *Les deux plus grandes valeurs propres de la matrice V des variances-covariances sont* $\lambda_1 = 28.253, \lambda_2 = 12.075$ *. Quels sont les taux d'inertie expliquée par chacun des deux axes factoriels correspondant ? En limitant la représentation à l'espace des 2 premiers facteurs. Quel est le taux d'inertie totale expliquée par cette représentation ? que peut-on en conclure ?*

   d- *Les deux vecteurs propres normés de V sont donnés dans le tableau ci-dessous :*

	1	*2*
Maths	*0.515*	*-0.567*
Physique	*0.507*	*-0.372*
Français	*0.492*	*0.650*
Anglais	*0.485*	*0.323*

Calculer les coordonnées de " Jean " sur les deux axes factoriels.

3. *Calculer les coefficients de corrélation linéaire entre le premier facteur et les 5 variables*
4. *Les corrélations entre les variables et les autres facteurs sont données ci-dessous*

	1	*2*	*3*	*4*
math	*0.81*	*-0.584*	*0.01*	*-0.02*
phys	*0.90*	*-0.432*	*-0.03*	*0.02*
fran	*0.75*	*0.651*	*-0.02*	*-0.01*
ang	*0.92*	*0.399*	*0.04*	*0.02*

   Donner brièvement un interprétation possible pour les 2 facteurs.
5. *En utilisant les résultats obtenus à la première et à la troisième question, calculer le carré du cosinus de l'angle α entre $\overrightarrow{Gu_1}$ et un axe Δ_1 de vecteur directeur unitaire a_1 (l'indice ponctuel de la représentation de " Jean " sur le premier axe factoriel). Puis sur le plan des deux premiers facteurs. Conclure.*