Cours d'Analyse des Données

Cours 01 :Introduction génerale

Présenté par Monsieur

Hamel Elhadj

2021 /2022

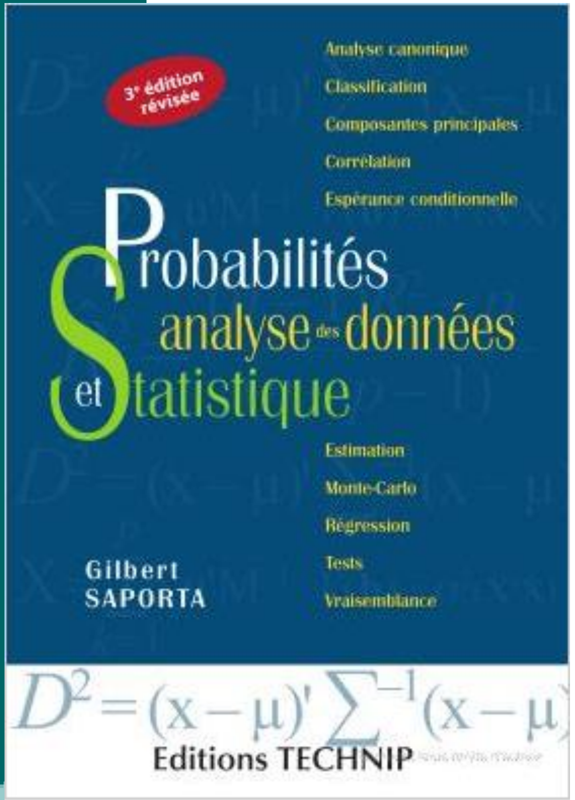
département de mathématiques
université de chlef

Email: hamel_2@yahoo.fr

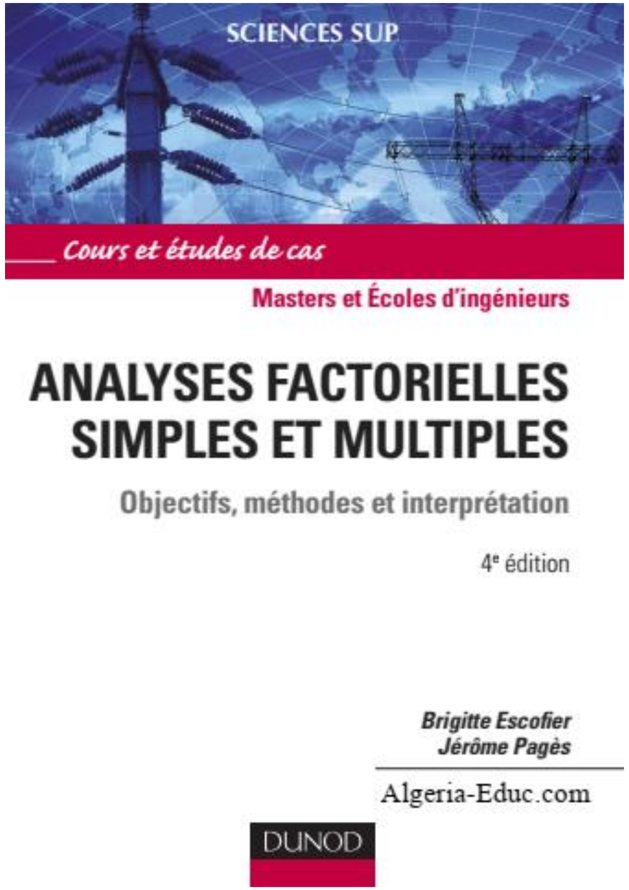
1. Introduction

L'analyse des données est aujourd'hui une technique tout à fait courante. Elle s'impose à tous ceux qui ont à manipuler des masses de données résultant d'enquêtes d'opinion, de tests expérimentaux, de mesures ou de toute autre source et qui se trouvent en présence de tableaux de chiffres afin d'en tirer des conclusions précises.

Good books


3e édition révisée
Analyse canonique
Classification
Composantes principales
Corrélation
Espérance conditionnelle
Probabilités analyse des données et Statistique
Estimation
Monte-Carlo
Régression
Tests
Vraisemblance
Gilbert SAPORTA
$D^2 = (x-\mu)' \sum^{-1} (x-\mu)$
Editions TECHNIP


L'ANALYSE DES DONNÉES


SCIENCES SUP
Cours et études de cas
Masters et Écoles d'ingénieurs
ANALYSES FACTORIELLES SIMPLES ET MULTIPLES
Objectifs, méthodes et interprétation
4e édition
Brigitte Escofier
Jérôme Pagès
Algeria-Educ.com
DUNOD


Ludovic Lebart
Alain Morineau
Marie Piron
Statistique exploratoire multidimensionnelle
DUNOD

1. Introduction

- L'analyse des données est une technique relativement ancienne 1930 (PEARSON, SPEARMAN, HOTELLING). Elle a connu cependant des développements récents entre1960-1970 du fait de l'expansion de l'informatique.

- L'analyse des données est une technique d'analyse statistique d'ensemble de données. Elle cherche à décrire des tableaux et à en exhiber des relations pertinentes. Elle se distingue de l'analyse exploratoire des données ou analyse multivariés.

- L'objectif de la démarche statistique est de faire apparaître ces liaisons. Les deux types de relations fondamentales sont les relations d'équivalence et les relations d'ordre. Ainsi, une population peut-elle être décomposée en classes hiérarchisées.

But

<u>But :</u> Synthétiser, structurer l'information contenue dans des données multidimensionnelles (n individus, p variables).

L'analyse des données multidimensionnelles permet de traiter un ensemble de variables observées sur un ensemble d'individus,
La notion de données multidimensionnelles se référant surtout au nombre important de variables et non au nombre d'individus concernés.

ELÉMENTS FONDAMENTAUX

- **les données sont vues de manière abstraites comme un nuage de points dans un espace vectoriel. On utilise (notions d'** Algèbre linéaire:

  - **<u>Des matrices</u>** qui permettent de manipuler un ensemble de variables comme un objet mathématique unique ;
  - **<u>Des valeurs et vecteurs propres</u>** qui permettent de décrire la structure d'une matrice.
  - **<u>Des métriques :</u>** permettent de définir la distance entre deux points de l'espace vectoriel ; on utilise aussi des produits scalaires.

- **Théorie des probabilités : nécessaire en statistique infférentielle (estimation, tests, modélisation et prévision,...).**

rappels de géométrie : produit scalaire

- produit scalaire :Le produit scalaire de deux vecteurs est le produit de la longueur de l'un par la projection de l'autre sur lui. (u.v.cos(u,v))

- Propriétés
  - Si les vecteurs sont orthogonaux le produit scalaire est nul.
  - Si les vecteurs sont colinéaires le produit scalaire est ±(u.v)
  - Si les vecteurs unitaires sont orthogonaux le produit scalaire est égal à la somme des produits des composantes correspondantes.

rappels de géométrie : projection

La projection d'un vecteur sur un axe est obtenue par le produit scalaire du vecteur par le vecteur unitaire de l'axe. Cela permet le changement d’axe de coordonnées.

rappels de géométrie: **Distance**

Dans l’espace des variables, un produit scalaire particulier, et donc une distance, s’impose.

Le choix d’une distance est toujours arbitraire dans l’espace des individus, car il est possible d’associer à chaque variable un coefficient de pondération.

rappels de géométrie : **Métrique:**

Le choix de la métrique a une influence fondamentale sur le résultat de l'analyse.

Définition : Soit M une matrice définie positive de dimension p. La fonction suivante $d_M : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+$ définit une métrique

$$d_M^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_M^2 = (\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \boldsymbol{M} (\boldsymbol{x}_i - \boldsymbol{x}_j)$$

Cette distance est appelée distance de Mahalanobis lorsque $\boldsymbol{M} = \boldsymbol{\Sigma}^{-1}$, où $\boldsymbol{\Sigma}$ est la matrice de variance-covariance des données .

Produit scalaire : La métrique définie ci-dessus dérive du produit scalaire $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle_M = \boldsymbol{x}_i^\top \boldsymbol{M} \boldsymbol{x}_j$

On dit que $\boldsymbol{x}_i$ et $\boldsymbol{x}_j$ sont M-orthogonaux si $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle_M = 0.$

Métrique

Utiliser la métrique $\boldsymbol{M} = \boldsymbol{T}^\top \boldsymbol{T}$ sur le tableau de données $\boldsymbol{X}$ est équivalent à travailler avec la métrique euclidienne sur le tableau transformé $\boldsymbol{X}\boldsymbol{T}^\top$.

Tableau transformé : Lorsqu'on travaille sur le tableau transformé comme ci-dessus, il est alors possible d'utiliser la norme euclidienne. En effet,

$$\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle_M = \boldsymbol{x}_i^\top (\boldsymbol{T}^\top \boldsymbol{T}) \boldsymbol{x}_j = (\boldsymbol{T}\boldsymbol{x}_i)^\top (\boldsymbol{T}\boldsymbol{x}_j) = \langle \boldsymbol{T}\boldsymbol{x}_i, \boldsymbol{T}\boldsymbol{x}_j \rangle_I$$

Réciproque : Pour toute matrice définie positive $\boldsymbol{M}$, il existe une matrice définie positive $\boldsymbol{T}$ telle que $\boldsymbol{M} = \boldsymbol{T}^\top \boldsymbol{T}$. On notera improprement $\boldsymbol{T} = \boldsymbol{M}^{\frac{1}{2}}$.

Appliquer préalablement la transformation $\boldsymbol{X}\boldsymbol{T}^\top \rightarrow \boldsymbol{X}$ permet de simplifier les traitements.

Métriques particulières

Métrique euclidienne : Elle est obtenue pour $M = I$.

L'une des difficultés rencontrées avec la métrique euclidienne est qu'elle privilégie les variables les plus dispersées et dépend donc de leur unité de mesure.

Métrique réduite : Elle consiste à prendre $M = D_{1/\sigma^2}$, où D_{1/σ^2} est la matrice diagonale de termes diagonaux les inverses $\frac{1}{\sigma_i^2}$ des variances des variables.

$$D_{1/\sigma^2} = \begin{pmatrix} \frac{1}{\sigma_1^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma_p^2} \end{pmatrix}$$

Cette métrique permet de s'affranchir de l'unité de mesure des variables, et de donner la même importance à chaque variable dans le calcul de la distance.

Tableau de données centrées réduites

Utiliser la métrique $\boldsymbol{M} = \boldsymbol{D}_{1/\sigma^2} = \boldsymbol{D}_{1/\sigma}^\top \boldsymbol{D}_{1/\sigma}$ sur le tableau de données $\boldsymbol{X}$ revient à travailler avec la métrique euclidienne sur le tableau transformé $\boldsymbol{X}\boldsymbol{D}_{1/\sigma}^\top$.

En effet :

$$\begin{aligned} d_M^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_M^2 &= (\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \boldsymbol{D}_{1/\sigma}^\top \boldsymbol{D}_{1/\sigma} (\boldsymbol{x}_i - \boldsymbol{x}_j) \\ &= (\boldsymbol{D}_{1/\sigma}\boldsymbol{x}_i - \boldsymbol{D}_{1/\sigma}\boldsymbol{x}_j)^\top (\boldsymbol{D}_{1/\sigma}\boldsymbol{x}_i - \boldsymbol{D}_{1/\sigma}\boldsymbol{x}_j) \end{aligned}$$

Il est équivalent de travailler avec la métrique $\boldsymbol{D}_{1/\sigma^2}$ sur le tableau $\boldsymbol{X}$, ou avec la métrique euclidienne $\boldsymbol{I}$ sur le tableau centré réduit $\boldsymbol{Z}$ composé des données :

$$z_i^j = \frac{x_i^j - \bar{x}^j}{\sigma_j}$$

Le tableau de données centré réduit $\boldsymbol{Z}$ se calcule matriciellement ainsi :

$$\boldsymbol{Z} = \boldsymbol{Y}\boldsymbol{D}_{1/\sigma} = (\boldsymbol{X} - \boldsymbol{1}\,\boldsymbol{m}^\top)\,\boldsymbol{D}_{1/\sigma}.$$

rappels sur les matrices

- Trace

La trace d'une matrice est la somme des termes de la diagonale principale.

- Valeur propre

λ est valeur propre de A <=> $\mathrm{Det}(A - \lambda I) = 0$

- Vecteur propre

V est vecteur propre de f si $f(V) = \lambda V$

- matrice diagonale

Une matrice diagonale est une matrice dont tous les termes appartiennent à la diagonale principale.

Valeurs et vecteurs propres

- **Définition**
  un vecteur v de taille p est un vecteur propre d'une matrice A de taille p x p s'il existe λ Є C telle que

$$Av = \lambda v$$

  est une valeur propre de A associée à v.

- **Domaine**
  En général, les vecteurs propres et valeurs propres sont complexes; dans tous les cas qui nous intéressent, ils seront réels.

- **Interprétation des vecteurs propres**
  ce sont les directions dans lesquelles la matrice agit.

- **Interprétation des valeurs propres**
  c'est le facteur multiplicatif associe a une direction donnée.

Exemple: valeurs et vecteurs propres

La matrice

$$\begin{pmatrix} 5 & 1 & -1 \\ 2 & 4 & -2 \\ 1 & -1 & 3 \end{pmatrix}$$

a pour vecteurs propres

$$v_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \qquad v_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \qquad v_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

On vérifie facilement que les valeurs propres associées sont

$$\lambda_1 = 2 \qquad \lambda_2 = 4 \qquad \lambda_3 = 6$$

Cas particuliers: Valeurs et vecteurs propres

- **Matrice nulle**
  sa seule valeur propre est 0, et tout vecteur est vecteur propre.

- **Matrice identité**
  tout vecteur est vecteur propre de I avec valeur propre 1, puisque Iv = v.

- **Matrice diagonale**
  si D_λ est une matrice diagonale avec les coefficients $\lambda_1, \lambda_2, \ldots \lambda_p$, alors le i-eme vecteur coordonnée est vecteur propre de D_λ associe à la valeur propre λ_i.
  L'action d'une matrice diagonale est de multiplier chacune des coordonnées d'un vecteur par la valeur propre correspondante.

- **Matrice diagonalisable**
  c'est une matrice dont les vecteurs propres forment une base de l'espace vectoriel : tout vecteur peut être représenté de manière unique comme combinaison linéaire des vecteurs propres. Une matrice de taille p x p qui a p valeurs propres réelles distinctes est diagonalisable dans R.

Quelques matrices diagonalisables

- **Matrice symétrique**
  une matrice symétrique réelle (A' = A) possède une base de vecteurs propres orthogonaux et ses valeurs propres sont réelles

$$\left\langle v_i , v_j \right\rangle = 0 \quad si \quad i \neq j \quad et \quad \lambda_i \in \Re$$

- **Matrice M-symetrique**
  une matrice M-symetrique réelle (A'M = MA) possède une base de vecteurs propres M-orthogonaux et ses valeurs propres sont positives ou nulles

$$\left\langle v_i , v_j \right\rangle_M = 0 \quad si \quad i \neq j \quad et \quad \lambda_i \in \Re$$

- **Matrice définie positive**
  c'est une matrice symétrique dont les valeurs propres sont strictement positives et donc

$$\left\langle v_i , v_j \right\rangle = 0 \quad si \quad i \neq j \quad et \quad \lambda_i > 0$$

Analyse de la matrice notée:VM

- **Valeurs propres**
  la matrice VM est M-symetrique: elle est donc diagonalisable et ses valeurs propres λ_1, λ_2, λ_p sont réelles.

- **Vecteurs propres**
  il existe donc p vecteurs $a_{1, \ldots,} a_p$ tels que

$$VMa_i = \lambda a_i \qquad avec \quad \langle a_i , a_j \rangle_M = \begin{cases} 1 & \text{si} \quad i = j \\ 0 & \text{sinon} \end{cases}$$

  Les a_i sont les axes principaux d'inertie de VM. Ils sont M-orthonormaux.

- **Signe des valeurs propres**
  les valeurs propres de VM sont positives et on peut les classer par ordre décroissant

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$$

- **Idée du lien avec l'inertie**
  on sait que .

$$Tr(VM) = \lambda_1 + \lambda_2 + \ldots + \lambda_p$$

  Si on ne garde que les données relatives a $a_{1, \ldots,} a_p$ on gardera l'inertie $\lambda_1 + \lambda_2 + \ldots + \lambda_p$, et c'est le mieux qu'on puisse faire.

Inertie & centre de gravité

L'inertie

centre de gravité

Inerties

Inertie par rapport à un point

Définition : L'inertie du nuage de points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ en un point quelconque $\boldsymbol{a}$ est donnée par

$$I_a = \sum_{i=1}^{n} w_i \|\boldsymbol{x}_i - \boldsymbol{a}\|_M^2$$

Propriété : L'inertie du nuage de points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ en son point moyen $\boldsymbol{m}$, ou *centre de gravité*, est

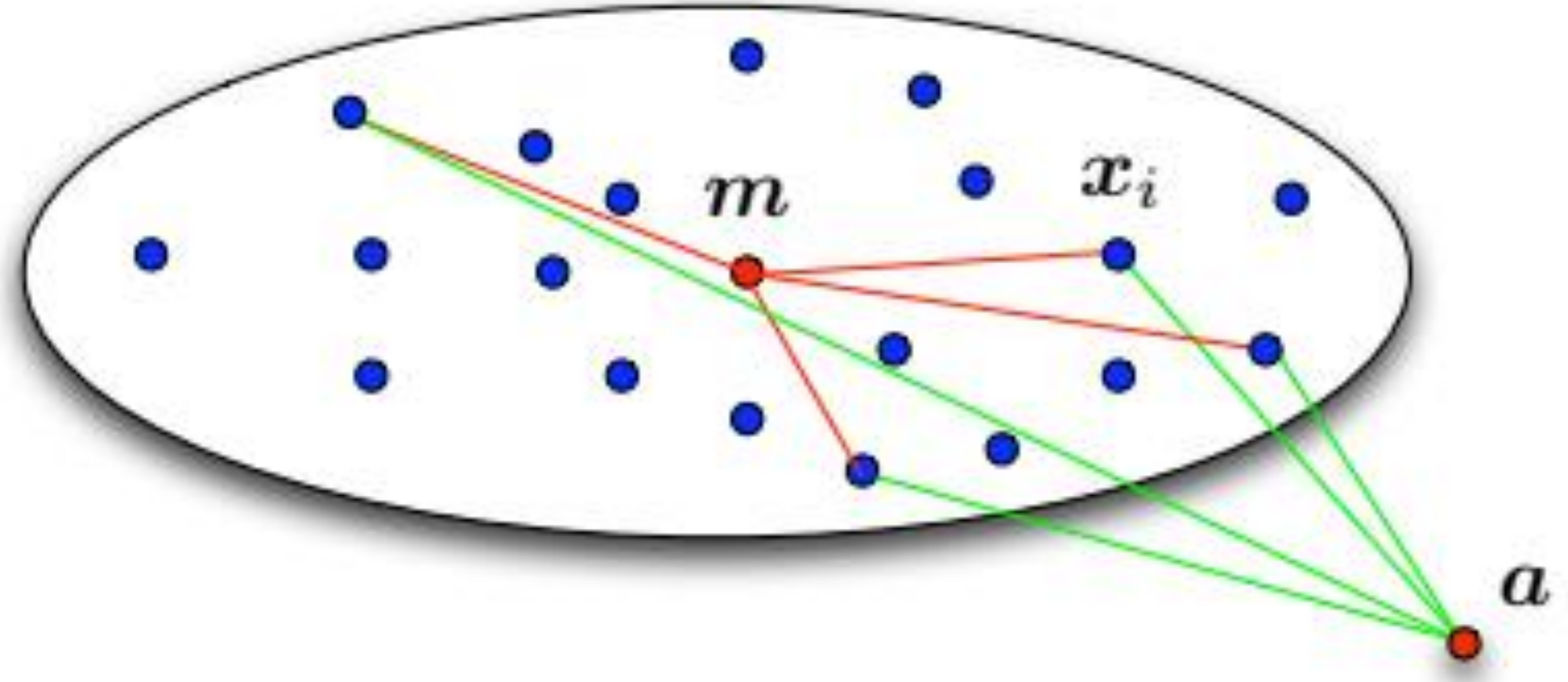
$$\boxed{\begin{aligned} I_m = \sum_{i=1}^{n} p_i \|\boldsymbol{x}_i - \boldsymbol{m}\|_M^2 &= \frac{1}{2} \sum_{i,j=1}^{n} w_i\, w_j \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_M^2 \\ &= \mathrm{Trace}(\boldsymbol{\Sigma} \boldsymbol{M}) \end{aligned}}$$

Cette propriété a été démontrée au chapitre précédent.

Propriété : Pour un tableau de données centrées réduites, on a

$$I_m = \mathrm{Trace}(\boldsymbol{R}) = p$$

Inertie par rapport à un point



INERTIES

Théorème de Huygens

Propriété : Soit $\boldsymbol{m}$ le centre de gravité du nuage de points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, et $\boldsymbol{a}$ un point quelconque de $\mathbb{R}^p$. L'inertie I_a du nuage au point $\boldsymbol{a}$ est donnée par

$$\boxed{I_a = I_m + d_M^2(\boldsymbol{a}, \boldsymbol{m})}$$

En conséquence, I_a est minimum pour $\boldsymbol{a} = \boldsymbol{m}$.

Démonstration :

$$\begin{aligned}
I_a &= \sum_{i=1}^{n} p_i \, \|(\boldsymbol{x}_i - \boldsymbol{m}) - (\boldsymbol{a} - \boldsymbol{m})\|_M^2 \\
&= \sum_{i=1}^{n} p_i \, \|\boldsymbol{x}_i - \boldsymbol{m}\|_M^2 + \sum_{i=1}^{n} p_i \, \|(\boldsymbol{a} - \boldsymbol{m})\|_M^2 - 2 \sum_{i=1}^{n} p_i \, \langle \boldsymbol{x}_i - \boldsymbol{m}, \boldsymbol{a} - \boldsymbol{m} \rangle_M \\
&= I_m + d_M^2(\boldsymbol{a}, \boldsymbol{m})
\end{aligned}$$

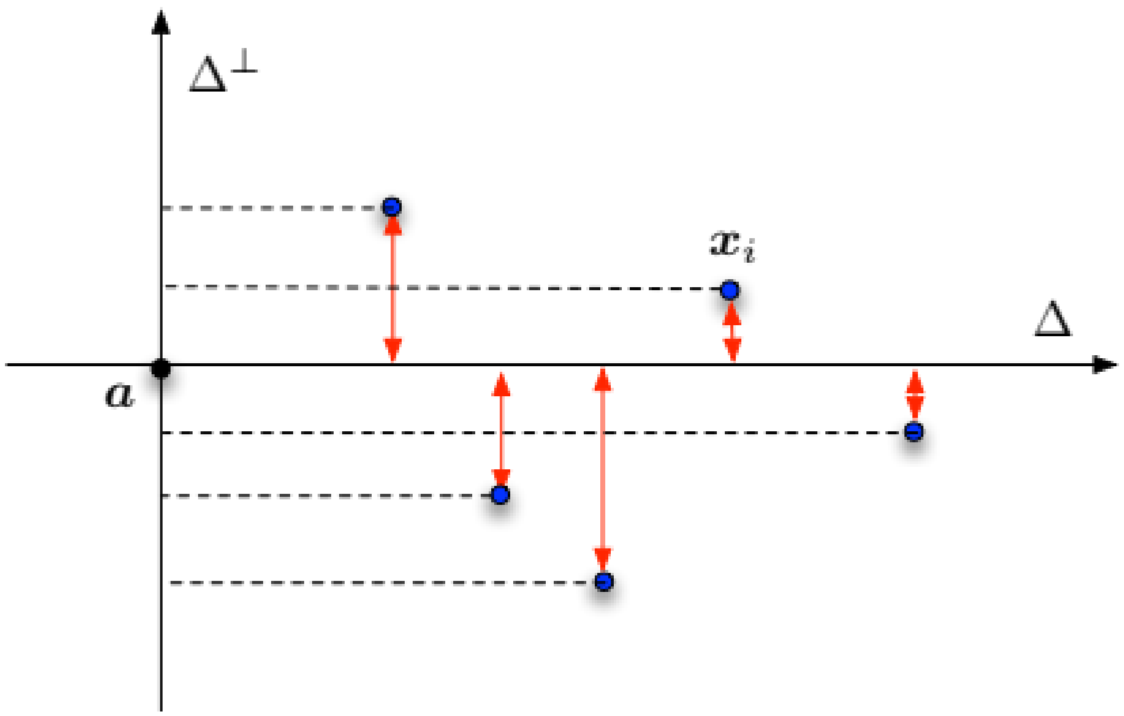
Inerties

Inertie par rapport à un axe

Définition : L'inertie du nuage de points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ par rapport à un axe Δ est définie par

$$I_\Delta = \sum_{i=1}^{n} p_i \, d^2_M(\boldsymbol{x}_i, \Delta)$$

Cette inertie quantifie la dispersion du nuage des individus autour de Δ.

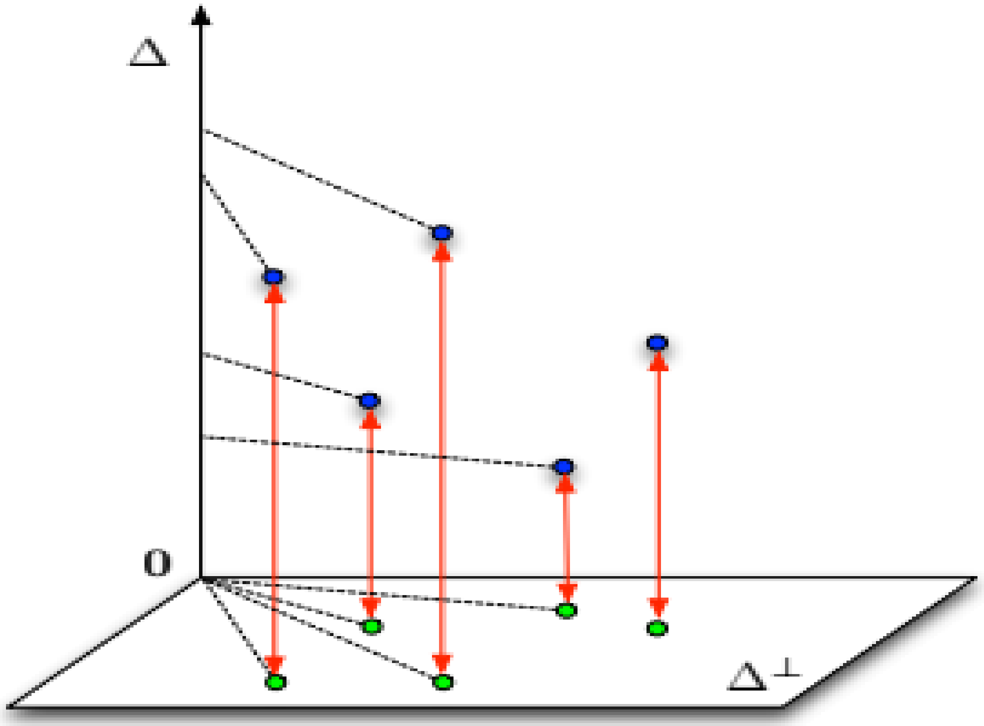


Inertie par rapport à un sous-espace affine

Définition : L'inertie du nuage de points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ par rapport à un sous-espace affine $\mathcal{F}$ est définie par

$$I_{\mathcal{F}} = \sum_{i=1}^{n} p_i \, d_M^2(\boldsymbol{x}_i, \mathcal{F})$$

Cette inertie quantifie la dispersion du nuage des individus dans $\mathcal{F}^{\perp}$.



Inertie

- **Définition**
  l'inertie en un point a du nuage de points est

$$I_a = \sum_{i=1}^{n} p_i \left\| e_i - a \right\|_M^2 = \sum_{i=1}^{n} p_i (e_i - a)' M (e_i - a)$$

- **Autres relations**
  l'inertie totale Ig est la moitie de la moyenne des carres des distances entre les individus

$$2I_g = \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \left\| e_i - e_j \right\|_M^2$$

- L'inertie totale est aussi donnée par la trace de la matrice MV (la trace d'une matrice étant la somme de ses éléments diagonaux).

$$I_g = Tr(MV)$$

rappels de statistique descriptive

- **La Statistique Descriptive** est l'ensemble des méthodes et techniques permettant de présenter, de décrire, de résumer, des données nombreuses et variées.
- **population statistique** est l'ensemble étudié dont les éléments sont des individus ou unités statistiques.
- **Recensement**
  étude de tous les individus d'une population donnée.
- **Sondage**
  étude d'une partie seulement d'une population appelée échantillon.
- **Echantillon** est un ensemble d'individus extraits d'une population initiale de manière aléatoire de façon à ce qu'il soit représentatif de cette population
- **Caractère** est l'aspect des individus que l'on étudie

rappels de statistique descriptive

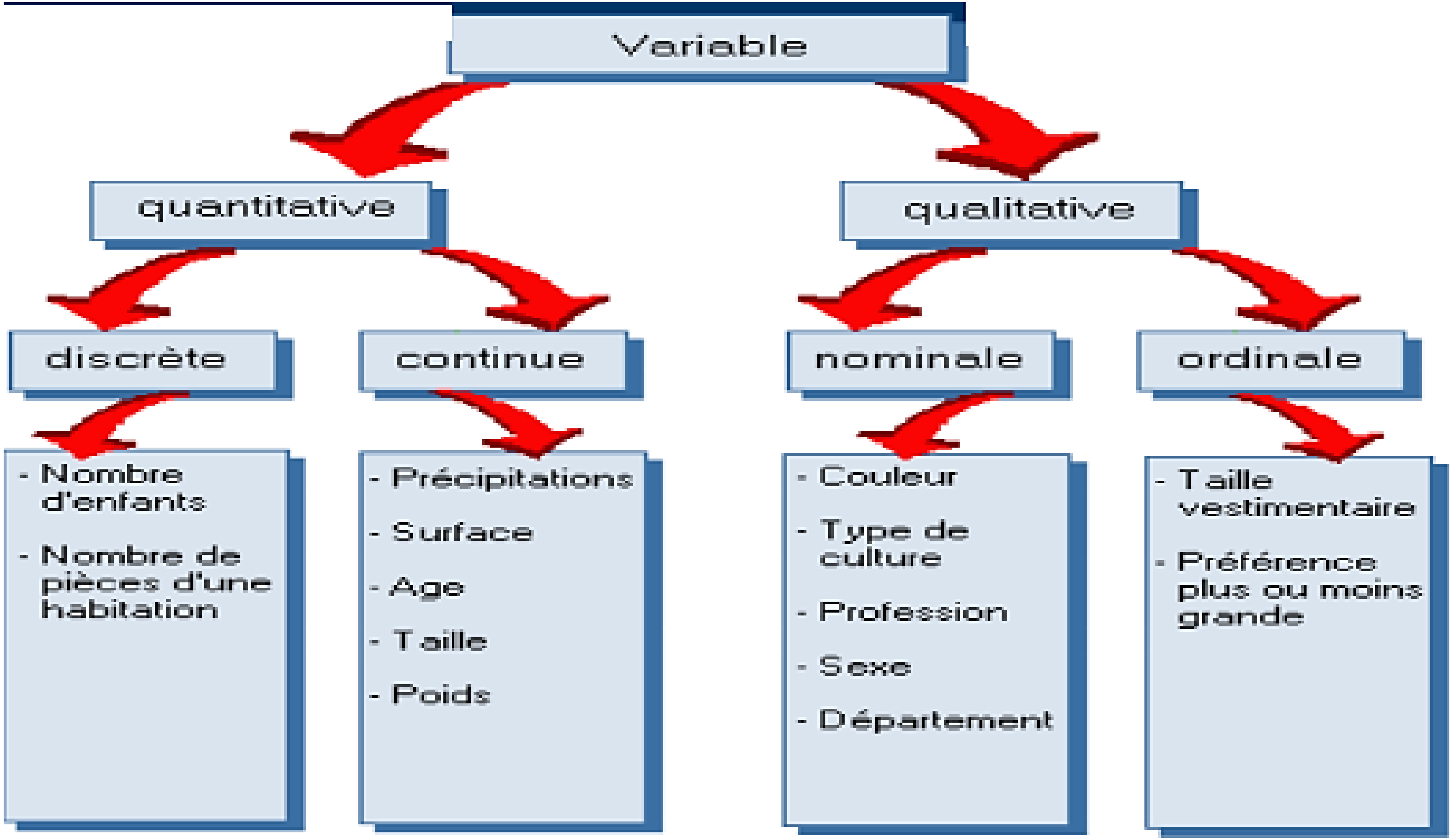
- **Nature du caractère**
  - **quantitatives**: nombres sur lesquels les opérations usuelles (somme, moyenne,...) ont un sens ; elles peuvent être discrètes (ex : nombre d'éléments dans un ensemble) ou continues (ex: prix, taille) ;

  La variable peut alors être discrète ou continue selon la nature de l'ensemble des valeurs qu'elle est susceptible de prendre (valeurs isolées ou intervalle).

  - **qualitatives**: appartenance a une catégorie donnée ; elles peuvent être nominales (ex : sexe, CSP) ou ordinales quand les catégories sont ordonnées (ex : très résistant, assez résistant, peu résistant)

On distingue des variables qualitatives ordinales ou nominales, selon que les modalités peuvent être naturellement ordonnées ou pas.

Une variable est ordinale si l'ensemble des catégories est munie d'un ordre total si non elle est nominale


Variable
quantitative
qualitative
discrète
continue
nominale
ordinale
- Nombre d'enfants
- Nombre de pièces d'une habitation
- Précipitations
- Surface
- Age
- Taille
- Poids
- Couleur
- Type de culture
- Profession
- Sexe
- Département
- Taille vestimentaire
- Préférence plus ou moins grande

paramètres de position et dispersion

Introduction

on dispose d'une série d'indicateurs qui ne donne qu'une vue partielle des données : effectif, moyenne, médiane, variance, écart type, minimum, maximum, étendue, 1er quartile, 3eme quartile, ...

Ces indicateurs mesurent principalement la tendance centrale et la dispersion. On utilisera principalement la moyenne, la variance et l'écart type.

paramètres de position :Moyenne arithmétique

- **Définition : La moyenne arithmétique** d'une série brute numérique x_1 , x_2 , ... , x_n est le quotient de la somme des observations par leur nombre
- On note

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

ou pour des données pondérées

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

- **Propriétés**
  la moyenne arithmétique est une mesure de tendance centrale qui dépend de toutes les observations et est sensible aux valeurs extrêmes. Elle est très utilisée a cause de ses bonnes propriétés mathématiques.

Statistiques de tendance centrale

- **La moyenne**, notée $\bar{x}$, est définie par : $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

- **La médiane**, notée *Med x*, est un nombre réel tel qu'au moins la moitié des données sont ≤ *Medx* et au moins la moitié des données sont ≥ *Medx*.

- **Le mode**, noté *Modex,* est la valeur la plus fréquente à l'intérieur de l'ensemble des données.

Contrairement à la moyenne, la médiane et le mode ne sont pas toujours uniques.

Statistiques de tendance centrale

- **Exemple :** on lance 20 fois un dé équilibré. On obtient les résultats :

face	1	2	3	4	5	6
effectif	3	1	4	5	5	2

- $\bar{x} = 3.7$, $\quad Med_x = 4 \quad Mode_x = 4$ (ou 5).

Exemple : l'analyse du plomb dans un échantillon d'eau potable par absorption atomique donne les résultats suivants en ppm :

19.4 19.5 19.6 19.8 20.1 20.3

Calcul de la moyenne :

$$\overline{X} = \frac{19.4 + 19.5 + 19.6 + 19.8 + 20.1 + 20.3}{6} = 19.8\,ppm$$

$$\text{mediane} = \frac{19.6 + 19.8}{2} = 19.7\,ppm$$

paramètres de dispersion: Variance et écart-type

- **Définition** : calculés généralement en complément de la moyenne, pour mesurer la plus ou moins grande dispersion autour de celle-ci

la variance de x est définie par

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad ou \qquad s_x^2 = \sum_{i=1}^{n} p_i (x_i - \bar{x})^2$$

L'écart type s_x est la racine carrée de la variance.

- **Propriétés**

La variance satisfait la formule suivante

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

La variance est « la moyenne des carres moins le carre de la moyenne ». L'ecart-type, qui a la même unité que x, est une mesure de dispersion.

Distribution statistique à deux variables: Mesure de liaison entre deux variables

- Relations entre deux caractères quantitatifs

- Covariance

- Coefficient de corrélation linéaire de BRAVAIS-PEARSON

- relations entre deux caractères qualitatifs

- Le khi-deux

- relations entre caractères quantitatifs et qualitatifs

- Le rapport de corrélation théorique

- Le rapport de corrélation empirique

Distribution statistique à deux variables: Mesure de liaison entre deux variables

- Définitions: la **covariance** observée entre deux variables x et y est

$$s_{xy} = \sum_{i=1}^{n} p_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} p_i x_i y_i - \overline{xy}$$

et le **cœfficient** de r de Bravais-Pearson ou coefficient de corrélation est donnée par

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} p_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} p_i (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} p_i (y_i - \bar{y})^2}}$$

Propriétés du coefficient de corrélation

- **La covariance est positive** si X et Y ont tendance à varier dans le même sens, et négative si elles ont tendance à varier en sens contraire
- **La covariance** ne dépend pas de l'origine choisie pour X et Y, mais dépend des unités de mesure. C'est pourquoi, pour mesurer l'aspect plus ou moins "allongé" du nuage dans une direction, par un coefficient sans unité : C'est le **coefficient de corrélation** linéaire
- Ce coefficient, symétrique en X et Y, indépendant des unités choisies pour X et Y, et de l'origine, est toujours **compris entre - 1 et 1**.

- $|r_{xy}| = 1$ si et seulement si x et y sont **linéairement liées**
En particulier, $r_{xx} = 1$.

- si $r_{xy} = 0$, on dit que les variables sont **de -corrélées ou indépendants**.

Exemple

L'analyse de la régression linéaire simple

Exemple 1 d'illustration

À partir des données ci-dessous, déterminez les estimations ponctuelles des paramètres de la droite de régression selon la méthode des moindres carrés :

ε

y_i	x_i
24	10
40	20
36	30
45	40
55	50

L'analyse de la régression linéaire simple

Exemple d'illustration : réponse

y_i	x_i	$x_i y_i$	x_i^2
24	10	240	100
40	20	800	400
36	30	1080	900
45	40	1800	1600
55	50	2750	2500
$\bar{y} = \frac{\sum y_i}{5} = 40$	$\bar{x} = \frac{\sum x_i}{5} = 30$	$\sum x_i y_i = 6670$	$\sum x_i^2 = 5500$

ε

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{6670 - 5\times 30\times 40}{5500 - 5\times (30)^2} = 0.67$$

$$b_0 = \bar{y} - b_1\bar{x} = 40 - 0.67\times 30 = 19.9$$

$$\hat{Y} = 19.9 + 0.67X$$

Le coefficient de corrélation de l'échantillon

Le coefficient de corrélation peut être déterminé de la manière suivante (ou encore en prenant la racine carrée du coefficient de détermination):

$$r_{XY}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2.\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{n\sum x_i y_i-\left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\sum x_i^2-\left(\sum x_i\right)^2}\sqrt{n\sum y_i^2-\left(\sum y_i\right)^2}}$$

- **On a toujours:** $-1\leq r_{XY}\leq 1$
- **Si** $r_{XY}=\pm 1$ **alors il existe une relation linéaire exacte entre X et Y**
- **Si** $r_{XY}=0$ **alors soit que X et Y sont indépendantes, soit qu'il y a une dépendance non linéaire entre les deux variables**
- **Si** $r_{XY}\neq 0$ **ou** $r_{XY}\neq\pm 1$ **alors il existe une relation linéaire plus ou moins forte entre X et Y**
- **Le coefficient de corrélation permet de voir s'il est facile d'approcher les données par une droite.**

Le coefficient de corrélation de l'échantillon

Toujours en utilisant l'exemple numérique de la publicité et les ventes d'autos, mesurez le degré de dépendance linéaire entre X et Y.

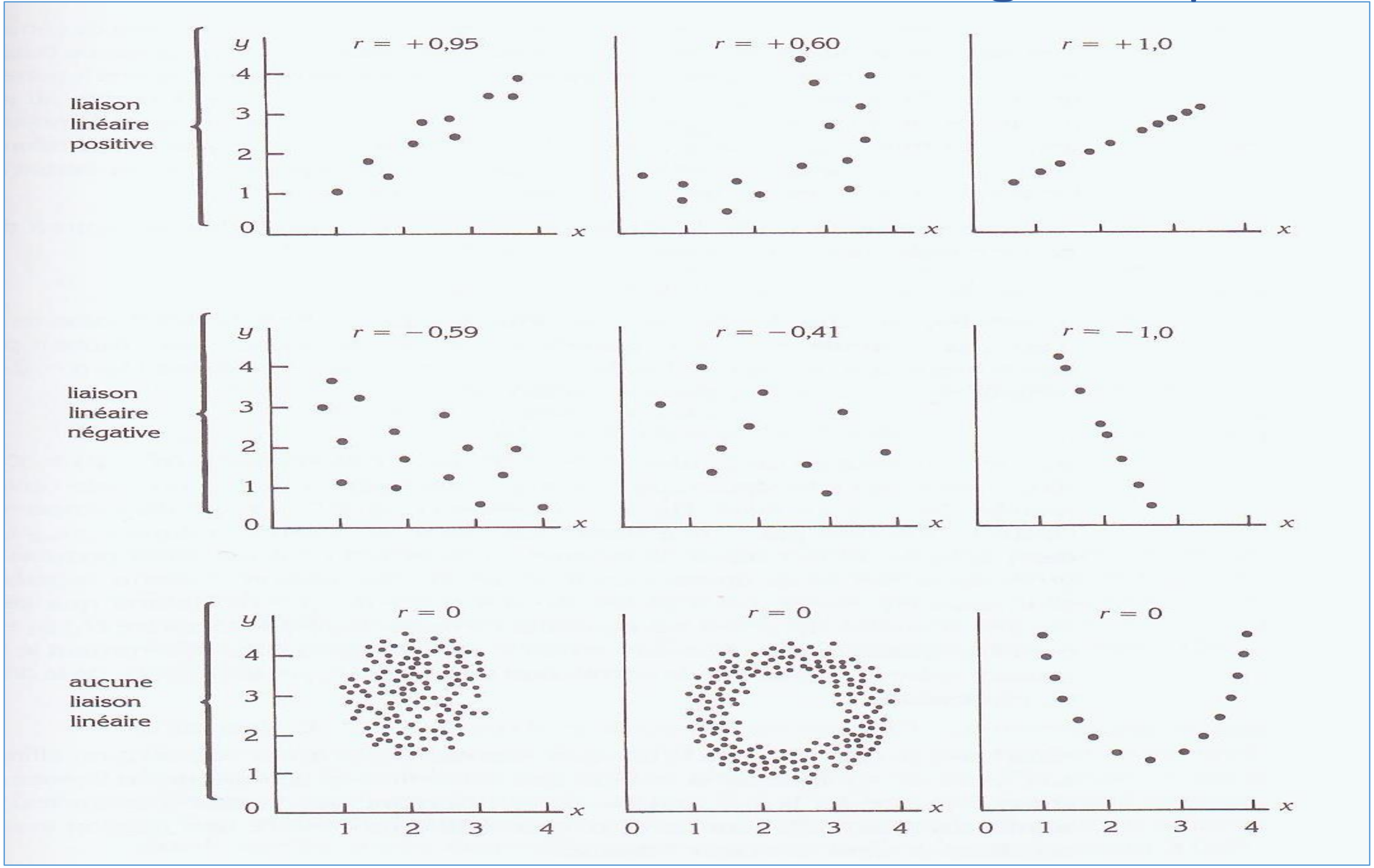
Réponse

Les dépenses en publicité et les ventes varient dans le même sens

$$r_{XY}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2.\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{\sum_{i=1}^{10}(x_i-3,3)(y_i-46,35)}{\sqrt{\sum_{i=1}^{10}(x_i-3,3)^2.\sum_{i=1}^{10}(y_i-46,35)^2}}=\frac{75,45}{\sqrt{19,10*307,53}}=0,9845$$

Il existe une relation linéaire très forte entre les dépenses en publicité et les ventes

Coefficient de corrélation et nuage de points



Corrélation et liaison significative

- **Problème**
  A partir de quelle valeur de r_{xy} peut-on considérer que les variables x et y sont liées?

- **Domaine d'application**
  on se place dans le cas ou le nombre d'individus est n > 30.

- **Méthode**
  si x et y sont deux variables gaussiennes indépendantes, alors on peut montrer que

$$\frac{(n-2)r_{xy}^2}{1-r_{xy}^2}$$

suit une loi de Fischer-Snedecor F(1; n-2). Le résultat est valable dans le cas non gaussien pour n > 30.

Le test

- on se fixe un risque d'erreur (0,01 ou 0,05 en général) et on calcule la probabilité

$$P\left(F(1,n-2) > \frac{(n-2)r_{xy}^2}{1-r_{xy}^2}\right) = \pi$$

- Si $\pi < \alpha$ on considère que l'événement est trop improbable et que donc que l'hypothèse originale d'indépendance doit être rejetée au seuil . On trouvera en général ces valeurs dans une table pré-calculée de la loi F.

les tableaux

les populations comprennent des individus distingués selon un certain nombre de variables. ces informations sont rassemblées dans des tableaux de base croisant individus et variables.

ces tableaux peuvent s'interpréter de deux façons, un nuage d'individus dans un ensemble de variables ou un nuage de variables dans un ensemble d'individus.

exemple :Tableau de données

- Pour n individus et p variables, on a le tableau
  X est une matrice rectangulaire a n lignes et p colonnes

$$X = (x^1, \ldots, x^p) = \begin{bmatrix} x_1^1 & x_1^2 & & \cdots & & x_1^p \\ x_2^1 & x_2^2 & & & & \\ & & \ddots & & & \\ \vdots & & & x_i^j & & \vdots \\ & & & & \ddots & \\ x_n^1 & & & \cdots & & x_n^p \end{bmatrix}$$

Vecteurs variable et individu

- **Variable**
  Une colonne du tableau

$$x^j = \begin{bmatrix} x_1^j \\ x_2^j \\ \\ x_n^j \end{bmatrix}$$

- **Individu**
  Une ligne du tableau

$$e_i' = (x_i^1 \quad x_i^2 \qquad x_i^p)$$

les tableaux

- Tableaux individus x variables quantitatives

- Tableaux logiques ou booléens ou binaires

- Tableaux disjonctifs complet : individu x variable à chaque modalité, placée en colonne, correspond une variable indicatrice. c'est la juxtaposition de plusieurs tableaux logiques.

x'x est une matrice diagonale dont les éléments sont les effectifs de chaque modalité.

-TABLEAUX PRÉSENCE ABSENCE

-TABLEAUX DE DONNÉES ORDINALES OU DE PRÉFÉRENCES INDIVIDUS X OBJETS À CLASSER. UNE CASE CORRESPOND À UNE NOTE VARIANT DE 1 AU NOMBRE D'OBJETS À CLASSER

-TABLEAU DE DISTANCES OU DE PROXIMITÉS : INDIVIDUS X INDIVIDUS IL PRÉSENTE LES DISTANCES ENTRE LES INDIVIDUS. CES TABLEAUX SONT SYMÉTRIQUE AUTOUR DE LA DIAGONALE PRINCIPALE.

-TABLEAUX DE CONTINGENCE : VARIABLE X VARIABLE IL CROISE LES MODALITÉS DE DEUX VARIABLES QUALITATIVES

-TABLEAUX DE BURT : IL CROISE LES MODALITÉS DE PLUS DE 2 VARIABLES QUALITATIVES. IL EST SYMÉTRIQUE.

-TABLEAUX DES RANGS

-TABLEAUX HÉTÉROGÈNES OU MIXTES INDIVIDUS X VARIABLES LES VARIABLES SONT DE DIFFÉRENTES NATURES SOIT LES VARIABLES SONT DÉJÀ DES CLASSEMENTS, SOIT POUR LES VARIABLES QUANTITATIVES ON REMPLACE LES VALEURS PAR LEUR RANG

LES ANALYSES FACTORIELLES

1- L'ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

2- L'ANALYSE FACTORIELLE DES CORRESPONDANCES (AFC)

3- L'ANALYSE DES CORRESPONDANCES MULTIPLES ACM

4- L'ANALYSE FACTORIELLE DES SIMILARITÉS (OU DE DISSIMILARITÉS) ET DES PRÉFÉRENCES

5- L'ANALYSE DISCRIMINANTE (AFD)

6- L'ANALYSE DES MESURES CONJOINTES

7- L'ANALYSE CANONIQUE

LES MÉTHODES DE CLASSIFICATION

1- L'ANALYSE NON HIÉRARCHIQUE
2- L'ANALYSE HIÉRARCHIQUE