Séance 1: Généralités sur l'analyse Multidimensionnelle

Indices statistiques

Sébastien Gadat

Laboratoire de Statistique et Probabilités
UMR 5583 CNRS-UPS

www.lsp.ups-tlse.fr/gadat

Première partie I

Introduction à l'analyse des données

Définitions élémentaires
Variables quantitatives et Statistiques univariées
Indice de dispersion, statistiques univariées
Corrélations, statistiques bi-variées
Covariances, statistiques bi-variées
Quelques exemples
Interprétation géométrique de quelques indices
Interprétation géométrique de quelques indices
Un exemple
Stabilité des indices statistiques
Représentation graphique - Boxplot
Représentation graphique - Histogrammes

Définitions élémentaires

- Unités statistiques ou individus, numérotés de 1 à n
- Poids sur les individus
- Poids normalisés
- Objectifs :
  - Statistique inférentielle :
  - Statistique descriptive :
- Variables
  - Quantitatives
  - Qualitatives
  - Explicatives
  - À expliquer

Définitions élémentaires
Variables quantitatives et Statistiques univariées
Indice de dispersion, statistiques univariées
Corrélations, statistiques bi-variées
Covariances, statistiques bi-variées
Quelques exemples
Interprétation géométrique de quelques indices
Interprétation géométrique de quelques indices
Un exemple
Stabilité des indices statistiques
Représentation graphique - Boxplot
Représentation graphique - Histogrammes

Variables quantitatives et Statistiques univariées

- Espace des Variables $x = (x^1, \dots, x^p) \in \mathbb{R}^p$
- x_i^j est la valeur de la variable x^j pour l'individu i
- Moyenne empirique de x^j :
- Nuage d'individus X qui synthétise les données

$$X = \begin{pmatrix} x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \vdots & \ddots & \dots & \dots & \vdots \\ x_i^1 & \dots & x_i^j & \ddots & x_i^p \\ \vdots & \dots & \ddots & \ddots & \vdots \\ x_n^1 & \dots & x_i^j & \dots & x_n^p \end{pmatrix}$$

Variables quantitatives et Statistiques univariées

- Nuage centré $\bar{X}$:
- Une variable x^j est dit centré si et seulement si sa moyenne sur les individus est nulle.
- En pratique $n >> p$

Indice de dispersion, statistiques univariées

- Variance de x^j

$$\clubsuit Var(x^j) = \quad = \quad =$$

- Propriété

$$\clubsuit Var(\bar{x^j}) = Var(x^j)$$

- La variance est invariante par translation
- $\clubsuit Var(ax^j) =$
- Écart type $\clubsuit \sigma(x^j) =$
- Variable centrée réduite construite à partir de x^j

$$\clubsuit y^j =$$

Définitions élémentaires
Variables quantitatives et Statistiques univariées
Indice de dispersion, statistiques univariées
Corrélations, statistiques bi-variées
Covariances, statistiques bi-variées
Quelques exemples
Interprétation géométrique de quelques indices
Interprétation géométrique de quelques indices
Un exemple
Stabilité des indices statistiques
Représentation graphique - Boxplot
Représentation graphique - Histogrammes

Corrélations, statistiques bi-variées

- La corrélation est un indice de liaison entre 2 variables x^j et x^k défini par

$$\clubsuit \rho(x^j, x^k) =$$

- C'est un coefficient compris entre -1 et 1 :

$$-1 \leq \rho(x^j, x^k) \leq 1$$

- ρ est indépendant de l'unité de mesure, invariant par translation, et homothétie

Corrélations, statistiques bi-variées

- Dans le cas de variables réduites :

$$\clubsuit \rho(y^j, y^k) =$$

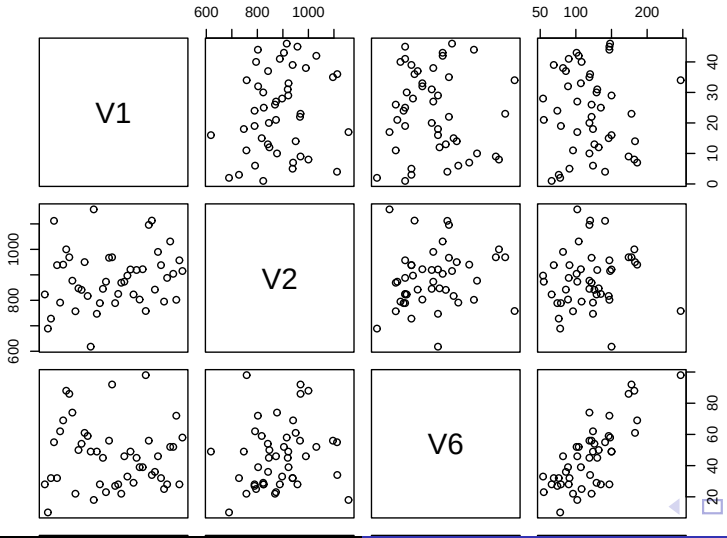
- Propriété :

$$|\rho(x^j, x^k)| = 1 \Longleftrightarrow x^j = ax^k + b \qquad \text{avec} \qquad sgn(a) = sgn(\rho)$$

Définitions élémentaires
Variables quantitatives et Statistiques univariées
Indice de dispersion, statistiques univariées
Corrélations, statistiques bi-variées
Covariances, statistiques bi-variées
Quelques exemples
Interprétation géométrique de quelques indices
Interprétation géométrique de quelques indices
Un exemple
Stabilité des indices statistiques
Représentation graphique - Boxplot
Représentation graphique - Histogrammes

Corrélations, statistiques bi-variées

- Interprétation du coefficient ρ :
  - Si $\rho(x^j, x^k)$ proche de 1, x^j et x^k sont fortement corrélées
  - Si $\rho(x^j, x^k)$ proche de -1, x^j et x^k sont fortement anti-corrélées
  - Si $\rho(x^j, x^k)$ proche de 0, x^j et x^k sont non-corrélées
- La corrélation mesure la dépendance linéaire entre x^j et x^k



Définitions élémentaires
Variables quantitatives et Statistiques univariées
Indice de dispersion, statistiques univariées
Corrélations, statistiques bi-variées
Covariances, statistiques bi-variées
Quelques exemples
Interprétation géométrique de quelques indices
Interprétation géométrique de quelques indices
Un exemple
Stabilité des indices statistiques
Représentation graphique - Boxplot
Représentation graphique - Histogrammes

Covariances, statistiques bi-variées

- On appelle covariance de x^j et x^k la quantité

$$Cov(x^j, x^k) =$$

- La covariance est une forme bilinéaire symétrique positive.
- On a bien sûr :

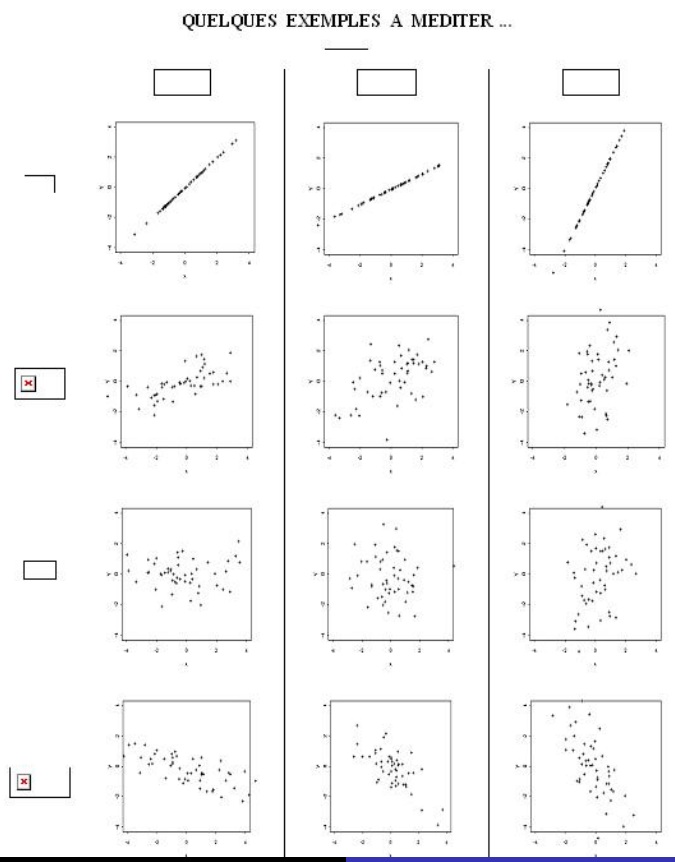
$$Cov(x^j, x^j) = Var(x^j)$$

Covariances, statistiques bi-variées

On synthétise ces informations bivariées dans la matrice de Variance / Covariance symétrique définie positive :

$$V = \begin{pmatrix} Var(x^1) & \dots & Cov(x^1;x^j) & \dots & Cov(x^1;x^p) \\ \vdots & \ddots & \dots & \dots & \vdots \\ Cov(x^i;x^1) & \dots & Cov(x^i;x^j) & \ddots & Cov(x^i;x^p) \\ \vdots & \dots & \ddots & \ddots & \vdots \\ Cov(x^p;x^1) & \dots & Cov(x^p;x^j) & \dots & Cov(x^p;x^p) \end{pmatrix}$$

Quelques exemples



Définitions élémentaires
Variables quantitatives et Statistiques univariées
Indice de dispersion, statistiques univariées
Corrélations, statistiques bi-variées
Covariances, statistiques bi-variées
Quelques exemples
Interprétation géométrique de quelques indices
Interprétation géométrique de quelques indices
Un exemple
Stabilité des indices statistiques
Représentation graphique - Boxplot
Représentation graphique - Histogrammes

Interprétation géométrique de quelques indices statistiques

On munit $\mathbb{R}^p$ du produit scalaire défini par

$$\langle x^j, x^k \rangle = \sum_{i=1}^{n} p_i x_i^j x_i^k$$

La norme associée vaut

$$\|x^j\| = \langle x^j, x^j \rangle^{1/2}$$

La moyenne empirique est donnée par

$$\bar{x^j} = \langle x^j, \mathbf{1_n} \rangle$$

x^j centrée si et seulement si x^j orthogonal à $\mathbf{1_n}$.

Définitions élémentaires
Variables quantitatives et Statistiques univariées
Indice de dispersion, statistiques univariées
Corrélations, statistiques bi-variées
Covariances, statistiques bi-variées
Quelques exemples
Interprétation géométrique de quelques indices
Interprétation géométrique de quelques indices
Un exemple
Stabilité des indices statistiques
Représentation graphique - Boxplot
Représentation graphique - Histogrammes

Interprétation géométrique de quelques indices statistiques

- Décomposition orthogonale de x^j :

$$x^j = \bar{x^j}\mathbf{1_n} + \underbrace{(x^j - \bar{x^j}\mathbf{1_n})}_{\tilde{x^j} \in \mathbf{1_n}^\perp}$$

- $Var(x^j) = \|\tilde{x^j}\|^2 \qquad \sigma(x^j) = \|\tilde{x^j}\|$
- $Cov(x^j; x^k) = \langle \tilde{x^j}; \tilde{x^k} \rangle$
- $\rho(x^j; x^k) = \cos(\tilde{x^j}; \tilde{x^k})$

Un exemple

On se donne 5 individus équipondérés décrits par deux variables

$$X = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 4 & 0 \\ 7 & 1 \\ 7 & -1 \end{pmatrix}$$

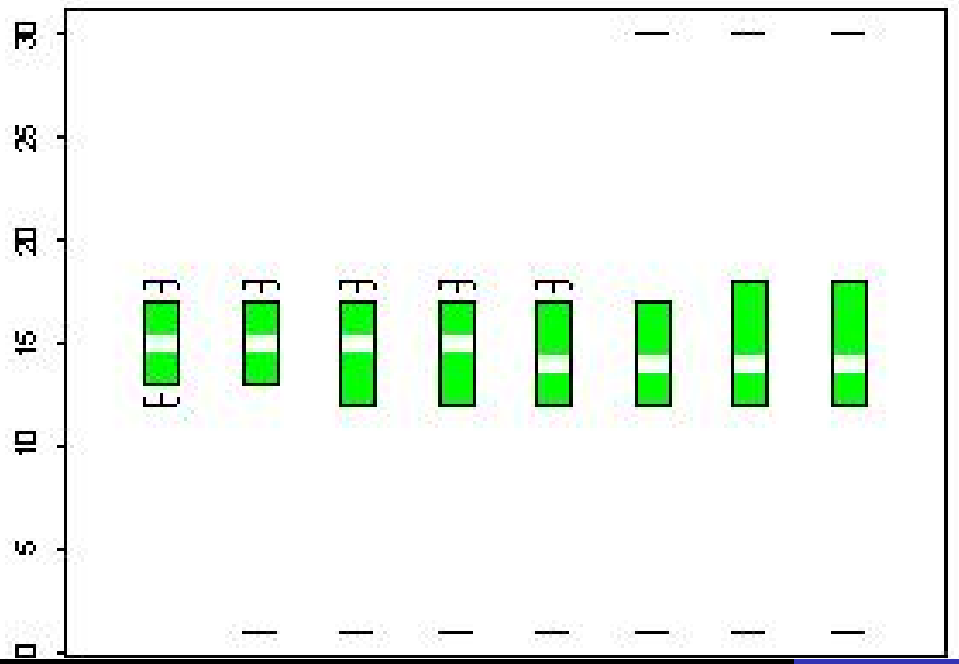
Pour chacune des variables, calculer leurs moyennes, donner le tableau centré associé, la matrice de variance/covariance ainsi que celle des corrélations.

Stabilité des indices statistiques

- Ces indices ne sont que des résumés numériques. Ils sont très sensibles aux ”outlayers”. Représenter les données dans une ”bonne” base sera une nécessité.
- Moyenne plus stable que la variance.
- Médiane plus stable que les quartiles.
- Médiane plus stable que la moyenne.
- Quartiles plus stables que la variance.
- Avant des analyses complexes, une étude unidimensionnelle (et bidimensionnelle) s’impose en général

Définitions élémentaires
Variables quantitatives et Statistiques univariées
Indice de dispersion, statistiques univariées
Corrélations, statistiques bi-variées
Covariances, statistiques bi-variées
Quelques exemples
Interprétation géométrique de quelques indices
Interprétation géométrique de quelques indices
Un exemple
Stabilité des indices statistiques
Représentation graphique - Boxplot
Représentation graphique - Histogrammes

Représentation graphique - Boxplot



Définitions élémentaires
Variables quantitatives et Statistiques univariées
Indice de dispersion, statistiques univariées
Corrélations, statistiques bi-variées
Covariances, statistiques bi-variées
Quelques exemples
Interprétation géométrique de quelques indices
Interprétation géométrique de quelques indices
Un exemple
Stabilité des indices statistiques
Représentation graphique - Boxplot
Représentation graphique - Histogrammes

Représentation graphique -Histogrammes