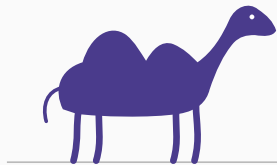


Introduction à l'analyse des données

Chapitre 1 : Voir en deux dimensions ce qui en a beaucoup



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Prérequis

Une photographie de chameau

Le même problème, en gestion

L'objet du cours

Ce qu'il faut retenir

Prérequis

Le cours qui précède

Statistique descriptive, en particulier la **covariance** et la **corrélation** — ce cours n'en est que le prolongement à plus de deux variables.

Les notions mobilisées

Moyenne, variance, écart-type, covariance, corrélation. La variable centrée réduite, qui commande tout le reste.

Les trois premiers chapitres les rappellent : ce n'est pas une formalité, c'est là que se joue la compréhension de la suite.

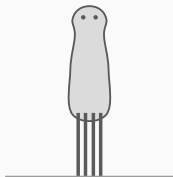
Ce cours ne suppose pas l'algèbre linéaire

Vecteur propre et valeur propre sont **rappelés au moment où ils servent**, sans démonstration.

C'est le parti pris du niveau Licence : montrer ce que font les méthodes. *Analyse des données avancée*, au Master, reprend l'ensemble et démontre pourquoi elles le font — et là, l'algèbre linéaire est exigée.

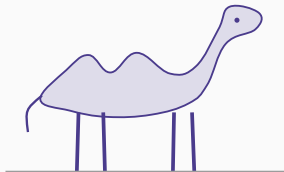
Une photographie de chameau

Le problème du photographe



vue de face

qu'est-ce que c'est ?



vue de profil

deux bosses, une encolure : aucun doute

Le même animal, le même appareil : seul l'angle a changé.

Un chameau a *trois* dimensions ; une photographie n'en a que *deux*. Passer de l'un à l'autre détruit forcément de l'information — et le photographe choisit l'angle qui en détruit le moins.

L'analyse des données fait exactement cela, sur un tableau de trente variables au lieu d'un animal.

Ce que l'image montre

Le même animal, deux clichés

De **face**, un chameau photographié en noir et blanc est une colonne étroite surmontée d'une tache. Rien ne permet de reconnaître l'animal : ni les bosses, ni l'encolure, ni la ligne du dos.

De **profil**, il n'y a plus le moindre doute. Deux bosses, un long cou, une démarche caractéristique.

Pourtant, c'est le même animal, le même appareil, la même lumière. **Seul l'angle de prise de vue a changé.**

Pourquoi l'angle décide de tout

Un chameau est un objet à **trois dimensions**. Une photographie est une surface à **deux dimensions**.

Passer de trois à deux **détruit nécessairement** de l'information : c'est le prix de l'aplatissement.

La seule question qui reste au photographe est donc : **quel angle en détruit le moins ?**

Le métier du photographe, en une phrase

Il tourne autour de son sujet et cherche la position d'où la silhouette reste la plus reconnaissable — non pas *toute* l'information, ce qui est impossible, mais **la plus grande part possible**.

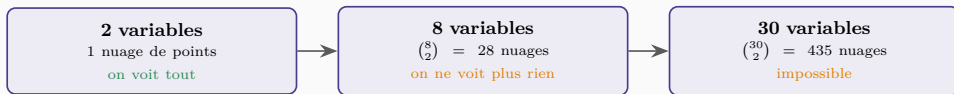
Le même problème, en gestion

La situation ordinaire d'une entreprise

Une enseigne dispose d'un fichier de **trois cents clients**, chacun décrit par une **douzaine de variables** : montant dépensé, fréquence de visite, ancienneté, notes de satisfaction sur plusieurs critères, part des achats en promotion.

La direction pose une question simple : **quels sont les profils de nos clients?**

Le tableau, lui, ne répond rien. Trois cents lignes de chiffres ne se regardent pas.



Et le problème n'est pas seulement le nombre de graphiques.

Vingt-huit nuages pris deux à deux ne montrent que des liaisons *par paires*. Une structure qui n'apparaît qu'en combinant quatre variables reste invisible sur les vingt-huit.

L'analyse des données ne fait pas plus de graphiques : elle en fait un seul, et le bon.

Pourquoi les graphiques usuels échouent

Le décompte

Avec p variables, on peut tracer

$$\binom{p}{2} = \frac{p(p-1)}{2} \text{ nuages de points}$$

Variables	Nuages à examiner
2	1
8	28
12	66
30	435

Intuition

Même en les examinant tous, on ne verrait que des liaisons **deux à deux**.

Une structure qui n'apparaît qu'en **combinant** quatre variables — des clients à la fois fidèles, peu sensibles au prix et anciens — reste invisible sur chacun des quatre cent trente-cinq nuages.

L'analyse des données ne fait donc pas plus de graphiques : elle en fait un seul, et le bon.

L'objet du cours

Deux problèmes, une seule méthode

Le chameau	Le fichier clients
un objet à 3 dimensions	un tableau à p variables
une feuille à 2 dimensions	un graphique plan
l'angle de prise de vue	les axes factoriels
la silhouette reconnaissable	la structure des données
ce que la photo perd	l' inertie non représentée

Analyse des données

L'**analyse des données** est l'ensemble des méthodes qui **réduisent la dimension** d'un tableau, en cherchant la représentation plane qui conserve la plus grande part de l'information.

Trois choses, dans cet ordre

1. **Mesurer l'information** contenue dans un tableau — sans quoi on ne peut pas dire qu'on en conserve beaucoup ou peu. Ce sera l'**inertie**.
2. **Trouver le meilleur angle** — c'est-à-dire les axes qui conservent le maximum de cette inertie.
3. **Lire la représentation obtenue**, ce qui est un travail à part entière.

Une mise au point utile dès maintenant

L'analyse des données ne **teste** rien. Elle ne calcule aucune p -value, ne rejette aucune hypothèse, ne conclut à aucune significativité.

Elle **explore** : elle fait apparaître des structures qu'on ne soupçonnait pas, et qu'il appartiendra ensuite de confirmer par d'autres moyens.

C'est pourquoi on l'appelle aussi **statistique exploratoire**, et pourquoi elle vient *avant* l'inférence, jamais après.

Démarche confirmatoire

Statistique inférentielle, économétrie

une hypothèse **posée d'avance**



les données tranchent

on sait ce qu'on cherche

Démarche exploratoire

Analyse des données

aucune hypothèse préalable



les données suggèrent

on cherche ce qu'on ne sait pas

Les deux démarches ne s'opposent pas : elles se suivent.

L'exploration fait *naître* les hypothèses ; la confirmation les *juge*. Une segmentation obtenue par classification n'est pas un résultat prouvé : c'est une hypothèse de travail, qu'il faudra tester sur d'autres données.

Le danger est de présenter comme démontré ce qui n'a été qu'observé. Aucune *p*-value ne sortira de ce cours ; ce n'est pas un oubli.

Ce qu'il faut retenir

Cinq points

1. Un objet à trois dimensions photographié sur une feuille **perd** de l'information : la seule question est **combien**.
2. Le photographe cherche l'**angle** qui en perd le moins; l'analyse des données cherche les **axes**.
3. Avec p variables, il y a $p(p - 1)/2$ nuages — et ils ne montrent que des liaisons **deux à deux**.
4. La méthode ne fait pas plus de graphiques : **elle en fait un seul, et le bon**.
5. C'est une démarche **exploratoire**, qui décrit et suggère, sans tester.

La suite

Avant de construire quoi que ce soit, il faut situer ce cours par rapport à celui que vous connaissez déjà — la statistique descriptive. C'est l'objet du chapitre 2, et la réponse est plus simple qu'on ne l'imagine.