Analyse des données : introduction

Vincent Audigier
vincent.audigier@lecnam.net

CNAM, Paris

STA101 2019-2020

Science statistique

- La science statistique est l'ensemble des méthodes permettant d'analyser un ensemble d'observations (ou de données)
  - méthodes relevant des mathématiques
  - méthodes relevant des outils informatiques
- Deux grandes classes de méthodes
  - statistique descriptive ou exploratoire
  - statistique inférentielle

Science statistique

- La science statistique est l'ensemble des méthodes permettant d'analyser un ensemble d'observations (ou de données)
  - méthodes relevant des mathématiques
  - méthodes relevant des outils informatiques
- Deux grandes classes de méthodes
  - statistique descriptive ou exploratoire
  - statistique inférentielle

Analyse exploratoire multidimensionnelle

Les origines des méthodes d'analyse des données

Des besoins

- la démarche scientifique s'appuient sur des faits, collectés sous forme de données (expériences, observations)
- traiter cette information potentiellement importante
- synthèse pour communiquer

Avant l'apparition de l'ordinateur

- résumés à la main
- machines à calculer mécaniques

A partir des années 50

- démocratisation de l'ordinateur
- premiers langages de programmation évolués (e.g. Fortran, 1954)
- JP Benzécri fonde l'analyse des données "à la Française"

FIG.: Jean-Paul Benzécri

L'analyse de données "à la Française"

Une famille de techniques

- pour résumer, décrire un tableau de données **multidimensionelles**
- pour visualiser
- géométriques

Parmi elles, les **méthodes factorielles**

- Analyse Factorielle des Correspondances (données textuelles)
- Analyse en Composantes Principales
- Analyse des Correspondances Multiples
- Analyse Factorielle des Données Mixtes

et les méthodes de **classification** non-supervisée

- méthodes de partitionnement
- classification hiérarchique

Elles font partie des méthodes de fouille (data-mining)

Applications

Vaste champ d'applications

- enquêtes d'opinion
- biologie
- économie
- santé
- sociologie
- etc

Logiciels

L'analyse de données est indissociable de l'outil informatique

- R
  - libre
  - très riche
  - grande communauté
  - interfaces type "clic-bouton" à FactoMineR (FactoShiny, Rcmdr)
- SPAD
  - gratuit pour les inscrits à STA101
  - clic-bouton
- SPSS, SAS, Statistica, ...

Quelques références

G. Saporta.
Probabilités, analyse des données et statistique.
Editions Technip, 2011.

F. Husson, S. Lê, and J. Pagès.
Analyse de données avec R.
Presses universitaires de Rennes, 2009.

L. Lebart, A. Morineau, and M. Piron.
Statistique exploratoire multidimensionnelle: visualisations et inférences en fouille de données.
Sciences Sup. Mathématiques. Dunod, 2006.

P-A Cornillon, A. Guyader, F. Husson, N. Jégou, J. Josse, M. Kloareg, E. Matzner-Løber, and L. Rouvière.
Statistiques avec R.
Presses universitaires de Rennes, 2012.

- Chaîne youtube de F. Husson
  `https://www.youtube.com/user/HussonFrancois/videos`
- Le site du package FactoMineR
  `http://factominer.free.fr/index_fr.html`