Chapitre 1 – Présentation de l'analyse des données

I Statistique descriptive et statistique inférentielle

1 Statistique descriptive

Observation d'individus décrits par un grand nombre de caractères.

Ensemble de méthodes pour analyser ces individus.

Méthodes descriptives : pas d'hypothèses probabilistes, pas de possibilité de vérification.

Résultat : valeurs de paramètres (moyenne, écart-type) des caractères étudiés sur la population globale.

2 Statistique inférentielle

Recherche de caractéristiques sur une population via la construction d'un échantillon sur lequel on étudie des caractères.

Formalisation : pour un caractère, les observations $(x_1, \ldots, x_n)$ sont des réalisations d'un ensemble de variables aléatoires $(\mathrm{X}_1, \ldots, \mathrm{X}_n)$.

Hypothèses probabilistes restrictives sur ces variables : indépendance, même loi, type de loi (normale, Poisson, binomiale)...

Résultat : estimation des paramètres d'une population (moyenne, écart-type...). Détermination d'intervalles de confiance pour donner la précision des estimations obtenues.

Éventuellement, construction de tests statistiques sur des grandeurs se référant à une population ou une variable aléatoire X.

3 Analyse des données

Fondements théoriques dans les années 1930 et fort développement depuis les années 1960 (JP Benzécri), surtout avec l'essor des ordinateurs (beaucoup de calculs pour les applications pratiques).

Trois questions à résoudre :
- Quels sont les grands types de données à étudier ?
- Comment sont-elles représentées ?
- Comment mesurer la dépendance entre deux caractères ?

II Types de données

1 Individus et variables

Individu : entité de base sur laquelle un observateur réalise des mesures (employé, client, étudiant, ville...).

Variable (ou caractère) : caractéristique étudiée sur la population et recueillir pour chaque individu (question d'une enquête, salaire, ancienneté, diplôme, note à un examen, âge, niveau d'étude...)

2 Types de caractères

Modalités d'un caractère : différentes valeurs prises par le caractère (distinctes et exhaustives).

Caractère quantitatif : si ses modalités sont des valeurs numériques sur lesquelles les opérations algébriques ont un sens (note, âge, taille, poids, salaire...).
Caractère quantitatif discret : si l'on peut énumérer toutes les valeurs que peut prendre le caractère.
Caractère quantitatif continu : s'il peut prendre n'importe quelle valeur dans un intervalle.

Caractère qualitatif : dans le cas contraire (profession, diplôme, département, niveau hiérarchique...).
Caractère qualitatif ordinal : s'il existe une relation d'ordre entre les modalités prises par le caractère.
Caractère qualitatif nominal : s'il n'existe pas de relation d'ordre entre les différentes modalités prises par le caractère.

III Présentation des données

1 Tableaux à double entrée

Individus/caractères : individus en ligne (n) et variables en colonne (p).

À l'intersection de la ligne i et de la colonne j : caractéristique relative à l'individu i pour le caractère j.

Caractères qualitatifs : représentation sous forme disjonctive complète.

2 Tableaux de contingence

Fréquences d'association entre les modalités de deux caractères qualitatifs.

3 Tableaux de proximité

Mesures de ressemblance ou de dissemblance entre tous les objets pris deux à deux.

IV Principales méthodes d'analyse des données

1 Méthodes factorielles

Analyse par réduction des dimensions : regroupement des individus en fonction de leurs ressemblances à l'aide d'une représentation graphique déformée (sur deux ou trois axes).

Analyse en composantes principales : caractères quantitatifs.

Projection des individus et variables dans un espace géométrique, puis transformations des données pour les visualiser dans un espace de petite dimension (un plan, souvent) en perdant le minimum d'informations.

Idée principale : transformation de variables liées entre elles (corrélées) en nouvelles variables (composantes principales) décorrélées. Ces nouvelles variables doivent capturer le maximum de variance des variables initiales.

Mesure de la qualité de la représentation par le calcul de la contribution de l'inertie de chaque composante à l'inertie totale.

Analyse factorielle des correspondances : 2 caractères qualitatifs (tableaux de contingence).

Analyse des correspondances multiples : caractères qualitatifs (extension de l'AFC).

Analyse discriminante : variable qualitative à k modalités et variables quantitatives explicatives.

Aspect descriptif : détermination des meilleurs combinaisons linéaires de variables explicatives pour séparer au mieux les k catégories, et représentation graphique.

Aspect décisionnel : attribution d'une catégorie à un nouvel individu en fonction des résultats précédents.

2 Méthodes de classification

Regroupement des individus en un nombre restreint de classes homogènes.

Méthodes hiérarchiques, ou non.

V Suggestions de bibliographie

- *Probabilités, analyse des données et statistiques*, Gilbert Saporta, Technip.
- *L'analyse des données*, Jean-Marie Bouroche et Gilbert Saporta, Que sais-je ?, PUF.
- *Analyses factorielles simples et multiples*, Brigitte Escofier et Jérome Pagès, Dunod.
- *Analyses factorielles simples*, Xavier Bry, Economica.
- *Analyses factorielles multiples*, Xavier Bry, Economica.
- *Data Mining et statistique décisionnelle*, Stéphane Tufféry, Technip.

VI Travaux pratiques avec SAS

L'énoncé de chaque séance est accessible à l'adresse `www.di.ens.fr/users/ccheval/SAS`.

Le lien pour télécharger SAS (via une machine virtuelle, et dans une version un tout petit peu moins complète que celle de l'université) sur votre machine personnelle est le suivant :

`www.sas.com/en_us/software/university-edition/download-software.html`

Toutes les explications pour l'installation sont ici, elles sont à suivre pas à pas :

`support.sas.com/software/products/university-edition/docs/en/SASUniversityEditionQuickStartVirtualBox.pdf`