SCI03 - Analyse de données expérimentales

Introduction à la statistique

Thierry Denœux[1]

[1]Université de Technologie de Compiègne
tél : 44 96
tdenoeux@hds.utc.fr

Automne 2014

Qu'est ce que la statistique ?

- On peut définir la Statistique comme l'activité qui consiste dans le recueil, le traitement et l'interprétation de données d'observation.
- Ces observations portent, de manière générale, sur des individus définis comme les éléments d'une certaine population.
- C'est la population, en tant qu'ensemble d'entités, qui est l'objet de l'investigation statistique, et non telle ou telle entité particulière.

Population

- Dans certains cas, la population de référence est finie : ses éléments peuvent être explicitement dénombrés.
- Souvent, une population est bien définie mais on ne peut en énumérer les éléments (ex : malades de la grippe). On parle alors de population hypothétique.
- Parfois, une population est définie par une procédure de génération de données (ex : mesure physique). L'hypothèse fondamentale est que l'expérience est répétable un nombre infini de fois.

Caractère

- Chaque individu d'une population est typiquement décrit par un ensemble de caractéristiques appelées caractères ou variables.
- Un caractère peut être soit qualitatif (par exemple : le sexe, la nationalité, l'état matrimonial d'une personne), soit quantitatif ou numérique (taille, poids, etc.).

Caractère qualitatif

- Un caractère qualitatif prend valeurs (modalités) qui ne ne sont pas de nature numérique.
- Les modalités d'un caractère qualitatif ne peuvent pas être combinées par des opérations arithmétiques, même si elles sont codées par des nombres.
- Parmi les caractères qualitatifs, on distingue
  - Les caractères ordinaux, dont les modalités sont ordonnées (par exemple : le grade pour une population de militaires).
  - Les caractères nominaux (pas de structure d'ordre).

Caractère quantitatif

- Un caractère quantitatif peut être
  - Discret (domaine fini ou dénombrable), ou
  - Continu (domaine théoriquement non dénombrable).
- La notion de caractère continu est une abstraction commode pour modéliser des grandeurs mesurées sur des échelles possédant un très grand nombre de valeurs.

Distribution

- La distribution d'un caractère X quantitatif dans une population Ω peut être décrite par la fonction $F_X : \mathbb{R} \to [0, 1]$ qui à tout réel x associe la proportion dans Ω d'individus pour lesquels on a $X \leq x$.
- Cette fonction (appelée fonction de répartition) est parfaitement définie et observable dans le cas d'une population finie.
- On suppose encore son existence dans le cas d'une population hypothétique, même si elle n'est plus observable dans ce cas.

Echantionnage

- L'étude exhaustive (par recensement) d'une population de grande taille est souvent difficile, voire impossible.
- Une méthode consiste alors à n'étudier qu'un sous-ensemble de la population totale, appelé échantillon (c'est la seule méthode envisageable dans le cas d'une population infinie).
- Le processus de sélection d'un échantillon est appelé échantillonnage.

Distribution empirique vs. théorique

- Un ensemble de n valeurs observées $x_1, \ldots, x_n$ d'une variable (scalaire ou vectorielle) sur un échantillon de taille n est appelé une distribution empirique (ou également un échantillon).
- Une distribution empirique peut être considérée comme apportant une information sur la distribution de la variable correspondante dans la population totale (distribution théorique).
- On appelle inférence statistique le processus visant à déduire des conclusions générales relatives à la population totale, à partir d'un échantillon.
- L'inférence statistique ne conduit jamais à des certitudes, mais à des conclusions possédant un certain degré de vraisemblance (utilisation des probabilités).

Statistique descriptive vs. inférentielle

- L'étude d'une distribution empirique peut être menée dans deux buts différents, bien que souvent complémentaires :
  1. Synthétiser, résumer, structurer l'information contenue dans les données, à l'aide de tableaux, de graphiques et de résumés numériques (statistique descriptive, ou exploratoire) ;
  2. Formuler et valider des hypothèses relatives à la population totale (statistique inférentielle).

Plan du cours

- Prochain chapitre : statistique descriptive, une ou deux variables ;
- Puis statistique inférentielle (estimation, tests, régression).
- Deuxième partie : analyse de données (statistique descriptive avec un grand nombre de variables).