

STATISTIQUES DESCRIPTIVES: INTRODUCTION

1. CONTEXTE

La statistique est l'activité ou méthode scientifique qui consiste à recueillir des données d'observation, suivie d'un traitement de ces données afin de les analyser et en obtenir une interprétation. Ces observations portent sur des *individus* ou *unités statistiques*, qui sont définis comme les éléments d'une *population*. L'étude statistique porte sur la population, en tant qu'ensemble d'entités, et non sur un individu particulier. Par exemple, une population peut être l'ensemble des pays d'un continent, et les individus les pays eux-même. Historiquement, le vocabulaire employé vient des recensements démographiques.

Une Population peut être plus ou moins bien définis. Dans l'exemple de l'ensemble des pays, la population est connue de manière exhaustive. Souvent, la notion de population doit être défini de manière plus abstraite. Par exemple, si l'on veut étudier statistiquement la mise au point d'un vaccin, la population sera l'ensemble des malades actuels et à venir. On parle alors de *population hypothétique*.

Pour étudier une population, on va donc se concentrer sur un sous-ensemble restreint à quelques individus : le choix de ces individus s'appelle *l'échantillonnage*. La méthode d'échantillonnage doit assurer une certaine représentativité de la population globale. On se limitera dans ce cours à deux types d'échantillonnage :

- (1) L'échantillonnage exhaustif ou *recensement* : chaque individu de la population est connu.
- (2) L'échantillonnage aléatoire : les individus sont tirés aléatoirement parmi la population.

2. VARIABLES STATISTIQUES

Un individu est décrit par un ensemble de variables (ou caractères). Ces dernières peuvent être de deux sortes : qualitatives ou quantitatives.

2.1. Variables qualitatives. Ces caractères ne sont pas de nature numérique et aucune opération arithmétique n'est possible (même si, parfois, elles peuvent être codées par un nombre). Les valeurs prises par ces variables sont appelées *modalités*. On peut distinguer deux types de variables qualitatives :

les variables nominales: ce sont les variables sur lesquels on ne peut faire ni opération arithmétique, ni comparaison. L'échelle nominale est utilisée pour représenter les variables dont les catégories ne sont pas naturellement ordonnées. On peut coder ces catégories par des nombres. Exemple de variables nominales : l'état civil, le sexe...

les variables ordinales: L'échelle ordinale est utilisée pour représenter des variables dont les catégories sont naturellement ordonnées ou en relation les unes par rapport aux autres. Les modalités sont non-quantitatives et indiquent uniquement une position dans une série ordonnée (on ne peut

pas mesurer la différence qui existe entre deux positions). Par exemple, un classement de préférence ou par jugement comme «j'aime un peu», «beaucoup», «pas du tout».

2.2. Variables quantitatives. Les modalités sont de nature numériques et l'on peut effectuer des opérations arithmétiques et des classements. Elles peuvent être de deux sortes :

discrète: les modalités prennent leurs valeurs dans un ensemble fini ou dénombrable. La mesure est donc toujours exacte. Par exemple, le nombre d'enfant par famille est une variable discrète.

continue: ici les variables prennent des valeurs qui peuvent être arbitrairement proche les unes des autres, et une valeur peut être aussi précise que l'on veut. En réalité, comme les mesures sont faites en précision finie, l'échelle continue est une abstraction commode pour modéliser les échelles possédant un grand nombre de valeurs (qui sont théoriquement aussi proche qu'on veut les unes des autres). Par exemple, la taille est une variable continue.

3. MÉTHODOLOGIE

Une étude statistique débute par la collecte des données : les observations «brutes» sont obtenues après enquête, mesures etc. C'est à ce niveau qu'intervient les méthodes d'*échantillonnages* (non abordées dans ce cours).

Une fois les données collectées, et avant d'apporter des réponses précises aux questions posées au préalable de l'étude, il convient d'analyser ces données. Cette analyse a pour but de synthétiser, résumer et structurer l'information contenue dans les données à l'aide de tableaux, graphiques et résumés numériques. C'est l'objet de la *statistique descriptive*, ou exploratoire, qui est l'objet de ce cours. Cette analyse se fait sur *l'échantillon* qu'on a à disposition.

Cette description des données n'est en général pas suffisante. Une bonne étude statistique consiste en formuler et valider des hypothèses relative à *la population* totale. Le but est d'étendre les résultats observés précédemment à toute la population, en étudiant le risque d'erreur possible. C'est le but de *l'inférence statistique* (qui ne sera pas abordée dans ce cours).

4. LIEN STATISTIQUE/PROBABILITÉ

| Probabilité | Statistique |
|-----------------------------------|---|
| Espace fondamental | Population |
| Epreuve | Tirage (d'un individu), expérimentation |
| Evènement élémentaire | Individu, observation |
| Variable aléatoire | Caractère |
| Epreuves répétées | Echantillonnage |
| Nbre de répétitions d'une épreuve | Taille de l'échantillon, effectif total |
| Probabilité | Fréquence observée |
| Loi de probabilité | Distribution observée ou loi empirique |
| Espérance mathématique | Moyenne observée |
| Variance | Variance observée |