

Séance 2

Les relations entre les deux disciplines sont très souvent tendues et complexes.” En effet la géographie a une relation assez paradoxale et complexe avec la statistique car il s'agit d'une matière du domaine des mathématiques, ce qui pousse la géographie à “mépriser” les définitions mathématiques élémentaires de la statistique sous prétexte que les statistiques ne rentrent pas dans son champ de domaine (quand la géographie se considère comme une matière beaucoup plus du domaine des sciences sociale). Néanmoins étant donné que cette discipline (la géographie) produit une quantité importante de données elle a besoin de la statistique pour le traitement de ces données.

Cela étant, il existe des géographes qui furent de bons statisticiens.

Ainsi, la position de la géographie par rapport à la statistique est ambivalente : elle dépend de la statistique pour traiter ses données, mais hésite à intégrer pleinement la rigueur mathématique que cette discipline requiert.

En géographie, notamment depuis le XX^e siècle, le hasard est intégré par les modèles statistiques. Deux formes sont distinguées : le hasard bénin (lié à la loi normale) et le hasard sauvage (associé à des distributions comme celle de Pareto).

Il existe 2 types d'informations en géographie : *D'une part, il peut s'agir pour une entrée territoriale claire et précise d'étudier tout ce qui peut caractériser l'ensemble délimité par des éléments de géographie humaine (population humaine, caractéristiques sociales, caractéristiques économiques, etc.), ou de géographie physique (température, volume des précipitations, etc.). D'autre part, il peut s'agir d'étudier la morphologie même des ensembles délimités. De fait, la géométrie des ensembles géographiques peut faire l'objet d'une étude statistique. Les premières correspondent à l'objet de cette initiation. Dans le cadre d'un système d'information géographique (S.I.G.), elles définissent la base attributaire(ou les attributs). Les secondes caractérisent les données géométriques du S.I.G. Elles feront l'objet d'une autre initiation.*

Pour faire de l'analyse de données il faut des données donc il faut les collecter et/ou les produire. La plupart du temps, que ce soit en géographie physique ou humaine, la donnée est déjà produite, que ce soit dans les organismes ou chez les géologues etc. Néanmoins une fois ces données acquise il doit modifier deux éléments :

- La nomenclature : *elle correspond à un ensemble de définitions préalables au recueil de l'information.*
- La méta données : Ce sont “des données décrivant des données”. *Elles sont souvent présentées sous forme de fiches décrivant : 1. les définitions et nomenclatures utilisées; 2. les lieux et dates de l'observation; 3. le fait que l'observation a été exhaustive ou qu'un sondage a été opéré; etc*

La statistique descriptive a pour but de décrire de manière simplifiée les données en dégagant “des propriétés remarquables par rapport à une distribution théorique connue.”

La statistique explicative permet quant à elle de trouver une explication à ce que l'on observe et déterminer les causes et les facteurs qui influence un phénomène

Ainsi la différence notable que l'on peut voir entre ces deux types de statistiques c'est que la première a un but de description et de simplification et le deuxième à un but d'explication de phénomène.

La visualisation de données, qui se fait sous la forme de graphiques, dépend du type de variable étudiée. Elle peut reposer sur un histogramme pour les valeurs en continues, ou sur une représentation sectorielle pour les variables qualitatives. On note plusieurs choix de

graphiques selon la nature de la variable. En effet, elle peut reposer sur le choix de variables qualitatives nominales, souvent représentées par un graphique en secteurs, ou ordinales (histogramme disjoint). Elle peut également reposer sur le choix de variables s'appuyant sur des dénombrements, à savoir par le choix de variables quantitatives discrètes (diagramme en bâtons), ou bien continues (histogrammes, boîtes à moustaches, lissage normal).

Les méthodes statistiques d'analyse des données se divisent en trois grandes classes :

- (1) *les méthodes descriptives* : L'objectif des méthodes descriptives est de visualiser et de classer les données. Elles s'appliquent en particulier aux tableaux individus contenant k variables dans lesquels toutes les variables jouent le même rôle.
- (2) *les méthodes explicatives* : elles cherchent à expliquer une variable (appelée variable à expliquer, ou variable dépendante Y) à partir de plusieurs variables explicatives (appelées X_1, X_2, \dots, X_k).
- (3) *les méthodes de prévision* : Ces méthodes permettent d'anticiper une valeur future d'une variable, à partir de données passées ou de relations connues. Elle cherche non pas à expliquer un phénomène (comme les méthodes explicatives), mais à prévoir ce qui va se produire.

La population statistique désigne l'ensemble des individus sur lesquels porte l'étude

L'individu statistique correspond à un élément de la population statistique. On peut aussi l'appeler unité statistique.

Les caractères statistiques sont les caractéristiques (c'est-à-dire les particularités) de l'individu pris parmi la population statistique sur laquelle l'analyse statistique porte.

Les modalités sont simplement les valeurs que peut prendre un caractère. Par exemple pour le caractère sexe → modalités : homme, femme

Une série de données présente soit des caractères qualitatifs (désignant une qualité, une éventualité non chiffrée) ou quantitatifs (désignant une quantité, une éventualité chiffrée), soit des caractères discrets (listes finies et isolées de valeurs dont le but est de compter les données) ou continus (valeurs prises dans un intervalle dont le but est de mesurer quelque chose par des données).

Oui, il existe une hiérarchie entre ces valeurs.

1. *L'amplitude* 24 est la longueur $b - a$ avec a la valeur minimale de la classe et b la valeur maximale. Il s'agit donc de la différence entre ces deux valeurs.
2. *La densité* est le rapport entre l'effectif n_i et l'amplitude de la classe décrivant une modalité i . On appelle d la densité

$$d = \frac{n_i}{b - a}$$

La règle de Sturges permet de déterminer le nombre de classes (pour construire un histogramme) lorsque le nombre total d'observations est connu.

La formule de Yule quant à elle permet de déterminer l'amplitude des classes d'un histogramme.

L'effectif correspond au nombre d'apparitions d'une variable dans la population.

La fréquence cumulée relative jusqu'à k modalités s'obtient en additionnant les fréquences associées aux valeurs inférieures ou égales à k , ou en divisant l'effectif cumulé par l'effectif total.

La distribution statistique est une manière d'organiser et de présenter des données d'une variable pour voir comment elles se répartissent.

