

Bien qu'elle produise des données massives que seul l'outil statistique permet d'étudier, la géographie à tendance à mépriser ce terme. En effet, les géographes ont historiquement un rapport compliqué à la statistique. Ces statistiques sont pourtant essentielles pour analyser l'information géographique, ainsi que pour le développement scientifique de la discipline.

La question du hasard est assez ambivalente en géographie, dans la mesure où deux positions philosophiques s'opposent. Pour le déterminisme, le hasard n'existe pas car il existe une cause à tout, tandis qu'une autre posture admet que le hasard existe, mais sera toujours, d'un moment ou un autre, explicable. On note deux types de hasard dans les modélisations mathématiques, à savoir le hasard bénin et le hasard sauvage.

L'information géographique peut se décomposer en deux séries statistiques, à savoir par la caractérisation de l'ensemble délimité (données humaines ou physiques), ainsi que par l'étude de la morphologie de ces ensembles.

Les statistiques présentent un outil essentiel en ce qui concerne l'avenir et la compréhension de l'information géographique. En effet, l'analyse de données repose sur les probabilités et les statistiques, deux éléments essentiels d'étude de la structure interne des données analysées.

La statistique descriptive repose sur la synthétisation et la description de données, afin d'obtenir une image plus simplifiée de la réalité. La statistique explicative (ou mathématique), quant à elle, prévoit des scénarios possibles, et repose donc sur la compréhension des données. Elle se fait à partir des données et de la distribution de probabilité théorique associée.

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 ordinaires (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).

On note trois grandes classes en méthodes statistiques d'analyse des données :

- la méthode descriptive : permettant de résumer et d'organiser l'information
- la méthode explicative : cherchant à établir les relations entre les variables
- la méthode de prévision : permettant d'anticiper l'évolution d'une donnée dans le temps

À cela s'ajoutent différents outils ou techniques, comme l'ACP, l'AFC ou encore l'ACM (pour n'en citer que quelques exemples), permettant de représenter ou réduire la complexité des données.

En statistique, la population correspond à l'ensemble des éléments que l'on observe. L'individu est le plus petit élément qui compose cette population. En géographie, ces individus proviennent souvent d'un découpage spatial : ce sont alors des unités spatiales. Le ou les caractères d'un individu désignent les informations ou caractéristiques que l'on mesure. Les modalités d'un caractère sont les différentes valeurs possibles qu'il peut prendre.

Il existe deux grandes natures de caractères :

- les variables qualitatives, qui décrivent un état ou une catégorie;
- les variables quantitatives, qui expriment une quantité issue d'une mesure ou d'un dénombrement (avec leurs quatre sous-catégories vues auparavant).

Enfin, les modalités constituent une partition du caractère : elles doivent être exhaustives et mutuellement exclusives.

L'amplitude d'une classe correspond à la différence entre ces bornes :  $b - a$ , où  $a$  est la valeur minimale et  $b$  la valeur maximale.

La densité, noté  $d$ , est obtenue en divisant l'effectif de la classe  $n$  par son amplitude. Elle exprime donc la concentration des valeurs dans cette classe.

La formule de Sturges permet de donner une valeur approximative du nombre  $k$  de classes pour construire un histogramme.

La formule Yule repose sur la même idée : elle sert également à déterminer un nombre de classes adapté à la distribution étudiée.

L'effectif  $n$  correspond au nombre d'individus associés à une valeur ou une classe.

On obtient la fréquence d'une valeur en divisant son nombre d'occurrence par le nombre total d'occurrences. Quant à la fréquence cumulative, elle représente la somme des fréquences des valeurs qui sont inférieures ou égales à  $k$ .

Ces fréquences servent à établir une distribution statistique empirique, c'est-à-dire un tableau qui regroupe les classes ou valeurs potentielles ainsi que leurs occurrences et/ou fréquences d'apparition.