Chapitre 1

Principes généraux de la statistique

La géographie est une discipline qui se cherche toujours. Il est fréquent qu'elle méprise les définitions mathématiques élémentaires de la statistique sous prétexte que cela n'entre traditionnellement pas dans son champ disciplinaire. Pourtant, elle produit des données massives que seul l'outil statistique permet d'étudier. Ainsi, les relations entre les deux disciplines sont très souvent tendues et complexes. Cette situation paradoxale conduit bien souvent les géographes à sous-estimer l'apport des analyses statistiques, mais, s'ils utilisent mal cet outil, comment trouver des résultats satisfaisants ? Néanmoins, il eut tout de même quelques géographes qui furent de bons statisticiens [Marchand, 1972] [Béguin, 1979] [Chadule, 1997] [Dumolard et al., 2003] [Dumolard, 2011]. Ce cours s'inspire de leurs œuvres, mais également de manuels beaucoup plus mathématiques [Spiegel, 1984] [Wonnacott et Wonnacott, 1995] [Jacquard, 2000] [Morgenthaler, 2007] [Tenenhaus, 2007] [Escofier et Pagès, 2016].

Pour commencer, dans ce premier chapitre, il faut revenir sur une confusion fréquente entre le singulier et le pluriel du terme « statistique ». **La** statistique désigne la science qui est une branche des mathématiques. La statistique est un **ensemble de méthodes** permettant de prendre de « bonnes décisions » en présence de l'incertain. **Les** statistiques renvoient à un ensemble de données concernant une catégorie de faits et utilisables par l'intermédiaire des techniques de la statistique. En géographie, on regroupe souvent les statistiques sous un vocable beaucoup plus vaste qui est celui de l'**information géographique** qui permet de classer les statistiques en fonction des objets géographiques étudiés. Par ailleurs, il ne faut pas confondre **la** statistique qui est la science, avec **une** statistique qui est l'étude particulière d'un estimateur dans le cadre de la statistique inférentielle. Il est important de fixer le vocabulaire pour un débutant, car la rencontre entre la rigueur mathématique et la souplesse du langage des sciences humaines et sociales parmi lesquelles se trouve la géographie, produit très souvent des expressions ambiguës et peu claires ; c'est un véritable *melting-pot*, pour le coup improbable, un comble

pour les statistiques...

1.1 Philosophie du hasard et géographie

Toute science recherche à réduire les éléments hasardeux au maximum. Cette position est défendue par le mouvement philosophique du **déterminisme** dont l'un des précurseurs fut P.-S. de Laplace. Le hasard n'existe pas : il existe une cause à tout. Dit autrement, dans ce cadre, hasard et déterminisme s'opposent strictement. C'est autour de ce courant que s'est construite toute la physique moderne. Par exemple, dans le cas de la gravitation d'I. Newton, établie autour de la force entre deux corps, de leurs masses, de la distance qui les sépare et de la constante gravitationnelle, d'autres paramètres interviennent, mais le hasard qu'ils engendrent est considéré comme négligeable, comme non perturbateur, c'est un **bruit**, ce qui fait que lorsque les conditions de fonctionnement de la loi, appelée ainsi parce que statistiquement hyperstable, donc causale sont réunies, il n'y a aucune raison qu'elle ne marche pas, et pourtant...

Pierre-Simon de Laplace (1749-1827)

Isaac Newton (1643-1727)

Récemment, la physique autour de la théorie du chaos a identifié des cas pour lesquels le modèle gravitionnel de I. Newton a besoin d'être corrigé en établissant de nouvelles variables. Ce qui permet d'introduire la seconde position concernant le hasard, vu comme une cause cachée à la raison humaine, cette posture admet que le hasard existe, mais, un jour, il sera expliqué et explicable. Elle s'inscrit dans un mouvement philosophique beaucoup plus vaste : celui autour du progrès de la connaissance.

Dans les modélisations mathématiques, il existe deux types de hasard, le **hasard bénin**, qui n'affecte pas l'établissement d'une causalité, et le **hasard sauvage**. Le premier possède une distribution de probabilité dite normale ; le seconde correspond à une distribution de probabilité beaucoup moins fréquente. Dans le cadre de la géographie, dès le début du XX^{e} siècle, deux grandes lois de probabilité interviennent : la loi normale et la loi de V. Pareto [Korčák, 1940] [Fréchet, 1941]. Cela étant, d'autres distributions peuvent exister. En général, tout statisticien préfère la loi normale, car elle dispose de propriétés mathématiques qui restent relativement abordables et sont simples à mettre en œuvre. En général, sans être un grand statisticien, les cours de mathématiques du collège au lycée ayant été orientés autour de la loi normale sans forcément la mentionner, font que de nombreux géographes ont conservé les vieux réflexes acquis dans ces classes, et produisent de telles lois sans le savoir. Le meilleur exemple est l'usage de la moyenne arithmétique, qui repose sur l'existence d'une loi normale [1]. Par extension, lorsque l'on entend que la moyenne (arithmétique) n'a aucun sens, c'est parce que le modèle

Vilfredo Pareto (1848-1923)

1. **Attention !** Il s'agit d'une condition nécessaire, mais pas suffisante.

statistique sous-jacent n'est pas celui de la loi normale. C'est l'occasion de souligner l'existence d'autres moyennes, qui correspondent à d'autres distributions qu'il faut connaître.

Deux autres notions sont discutées dans le cadre de la philosophie du hasard : la **nécessité** et la **contingence**. Lorsqu'une loi physique est établie, la notion de nécessité exprime l'idée qu'il est impossible que cela se passe autrement. À l'opposé, la notion de contingence, largement débattue à la fin du XIX^{e} siècle et au début du XX^{e} siècle dans le cadre de la géographie vidalienne, offre la possibilité qu'un événement se produise ou qu'il ne se produise pas. Cela s'est vérifié aussi bien en géographie physique qu'en géographie humaine. La combinaison des faits physiques dans un glacier, par exemple, ne permet pas de prévoir une avalanche. Cela étant, il est possible de déterminer un certain nombre de paramètres factuels, souvent appelés les facteurs géographiques dans la littérature, qui peuvent faire l'objet d'une étude statistique.

Avant de poursuivre, il est important de souligner que la majorité des non statisticiens affirment haut et fort que le hasard reste à l'origine de tout chose. Beaucoup de géographes défendent cette position. Un tel hasard aboutit au fait qu'il est impossible de faire de la géographie une science à proprement dit. Parmi les opposants à cette thèse se trouvent les spatialistes de l'école de l'analyse spatiale [2] qui oscillent entre nécessité et contingence. En adoptant une démarche nomothétique, qui n'est que le prolongement de la contingence vidalienne, elle tente de construire, à défaut de lois, des modèles spatiaux ou socio-spatiaux. Même si les résultats sont mitigés, cela fait de la géographie une science au sein de laquelle on peut débattre, s'opposer, se critiquer, alors que la position vidalienne a conduit la géographie à devenir une méthode plus qu'une science. C'est pourtant plutôt vers cette dernière que la géographie française s'oriente. Pour une raison historique, les géographes tiennent à leur position de « carrefour » dans laquelle les autres disciplines venant aussi bien des sciences que des sciences humaines et sociales viennent de temps à autre piocher une méthode d'analyse. L'absence de cours, dès la première de Licence de géographie, de mathématiques a conduit à cette situation paradoxale.

Au temps de la naissance de l'analyse spatiale, les données statistiques étaient sous un format peu pratique ; les traitements informatiques n'étaient qu'à leur tout début. Aujourd'hui, un développement de la science « géographie » pourrait être sans précédent, mais la disparition progressive des analystes spatiaux au sein des universités françaises fait qu'il n'en est rien. Cette politique de recrutement conduit de plus en plus à marginaliser la géographie française par rapport au reste du monde. Il n'empêche qu'aucun géographe ne peut échapper aujourd'hui aux

2. Il est important de rappeler que l'analyse spatiale concerne aussi bien la géographie physique que la géographie humaine.

approches statistiques.

Pour conclure ce point, il faut retenir que le hasard n'est qu'une vision philosophique. C'est à vous de vous positionner, mais, dans le cadre de la statistique, la posture est claire. Dans tout phénomène aléatoire, il n'est pas possible de prévoir le détail des réalisations, mais il est cependant possible de **dégager une certitude globale**. Dans le cas de la géographie humaine, il est impossible de prévoir ce que chacun des acteurs va réaliser sur un territoire donné, mais il est possible de dégager une tendance, c'est-à-dire l'action la plus probable, la plus vraisemblable. Cette position signifie que les lois du hasard qui existent n'interdisent nullement un quelconque écart à cette vraisemblance. On retrouve ici une vieille démarche géographique : celle du raisonnement multiscalaire. À l'échelle du territoire, tel ou tel élément se dégage globalement, mais, à une échelle plus locale sur ce même territoire, les conditions qui produisent la tendance générale demeurent multiples. Cela permet à la géographie de quitter son statut de méthode pour celui d'une science, la science des échelles.

Les statistiques sont par conséquent essentielles pour l'avenir de la géographie. Connaître les lois du hasard est de fait une nécessité pour comprendre l'information (massive) géographique.

1.2 Information géographique

L'information géographique se décompose en deux séries statistiques possibles. 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.

Cela paraît trivial, mais, pour faire une analyse de données, il faut des données. Le point essentiel de cette partie concerne leur production (ou leur collecte) préalable.

1.2.1 La production, ou la collecte, de données

La plupart du temps, le géographe ne produit pas ses données d'analyse. En géographie humaine, il laisse cela aux organismes publics. En géographie physique, il opère quelques mesures sur le terrain, mais il bénéficie des données offertes par les géologues, les topographes, *etc*. Cela étant, le géographe devant étudier les informations géographiques contenues dans ces jeux de données, il ne peut rester passif face à l'abondance et la massification des données touchant son objet. Il modifie notamment la nomenclature et les méta-données.

La nomenclature

Avant de réaliser une étude statistique, il faut produire des données. Pour ce, il faut constituer, ou utiliser, une **nomenclature** qui correspond à un ensemble de définitions préalables au recueil de l'**information**.

Toutes les nomenclatures ne sont pas hiérarchiques, c'est-à-dire pouvant s'additionner ou se soustraire entre elles afin d'obtenir une information plus ou moins agrégée. Toutefois, lorsque la nomenclature est hiérarchique, le géographe peut adapter le niveau de détail à son échelle d'étude, ou, du moins, s'en approcher.

Les méta-données

Les données diffusées par les organismes qui en produisent, pour être utilisables et interprétables correctement, doivent être documentées par des **méta-données** c'est-à-dire les données décrivant les données. L'objet des méta-données est de rendre possible un **examen critique des sources** permettant d'éviter les interprétations abusives. Les méta-données recensent tout ce qui peut informer sur la fiabilité 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 ; 4. en cas de sondage, il faut les décrire les conditions précises de celui-ci ; *etc*.

1.2.2 L'analyse de données

L'analyse de données repose sur les probabilités et les statistiques. À la différence de l'étape de la production (ou de la collecte), il s'agit d'étudier la structure interne des données analysées. C'est le moment mathématique. Une fois réalisé, et ce n'est pas la chose la plus aisée à faire, il faudra confronter les résultats obtenus avec la méthodologie de production des données et avec ce que l'on connaît du phénomène étudié.

1.3 Probabilités, statistiques et analyse de données

Les trois éléments présents dans cette partie sont graduellement agencés. Pour faire des statistiques, il faut établir des probabilités. Pour faire de l'analyse des données, il faut connaître des statistiques.

1.3.1 Probabilités

Ce paragraphe effectue une présentation ultra-rapide de la notion mathématique des probabilités. Ce qu'il faut retenir, c'est que les statistiques reposent sur les règles établies par les probabilités, mais il n'est pas nécessaire de les maîtriser pour opérer des analyses statistiques, bien qu'un chapitre leur sera consacré.

Par rapport aux données réelles, les probabilités représentent la théorie. Dit autrement, elles interviennent peu dans une analyse statistique. Elles en sont son miroir. Les probabilités reposent et approfondissent les notions mathématiques d'**ensemble**, de **dénombrement**, d'**arrangement**, de **permutation** et de **combinaison**. À partir de ces notions, le domaine possède son propre vocabulaire. Par exemple, un ensemble fini est un univers. Bien entendu, d'autres outils restent nécessaires pour comprendre les équations autour des probabilités, tels que les **fonctions numériques**, la **dérivation**, la **différentiation**, les **primitives**, l'**intégration**, *etc*.

L'objectif ultime des probabilités est de déterminer les lois du hasard qui régissent tel ou tel cas théorique. Ainsi, se construisent les **lois de probabilité**. À ce niveau, il faut juste retenir qu'il existe environ une trentaine de lois principales. Cela peut paraître beaucoup, mais, au regard de tous les jeux de données existants, ce n'est pas grand-chose.

1.3.2 Statistiques

Les **statistiques** ont pour objectif d'établir la **bonne loi de probabilité** convenant à une série précise. Dit autrement, il s'agit de la mise en pratique des **distributions de probabilité** théoriques. En effet, le calcul de probabilités apporte les outils nécessaires aux techniques de la statistique mathématique, c'est-à-dire les modèles qui vont être utilisés pour décrire des phénomènes réels où le hasard intervient.

Les statistiques se divisent en deux branches :

1. la **statistique descriptive** qui consiste à étudier des données. L'objectif est de dégager des propriétés remarquables par rapport à une distribution théorique connue. Cela permet ainsi d'obtenir une image simplifiée de la réalité en mettant de l'ordre dans les données (caractéristiques numériques ou graphiques). L'intuition y joue un rôle fondamental ;

2. la **statistique mathématique** dont l'objectif est de prédire à partir des statistiques descriptives, et surtout de la distribution de probabilités théorique que l'on a établie, des *scenarii* possibles.

Statistique descriptive

La statistique descriptive décrit les données dans la population ou l'échantillon étudié. Il s'agit d'en rendre compte avec le minimum de mots, de paramètres et de graphiques, afin de :

1. résumer les distributions étudiées ;
2. préparer les comparaisons et les prédictions. Dit autrement, la statistique descriptive fait partie d'une étape préalable indispensable à la statistique mathématique et à la statistique inférentielle.

La description des données établit :

— le type de variables associées ;

— le domaine de définition, les valeurs manquantes, *etc.* ;

— les valeurs extrêmes et les valeurs aberrantes.

La description des variables quantitatives aboutit à l'identification de lois de probabilité associées à ces valeurs :

— par des lois récurrentes dans certaines situations bien précises (loi binomiale, loi de Poisson, loi normale, *etc.*) ;

— par le nombre d'observations ou de mesures.

  — Le **théorème central limite** (T.C.L.) s'applique si le nombre est supérieur à 30. On peut utiliser alors les tests paramétriques (moyenne, écart type, *etc.*)

  — Si le nombre est inférieur à 30, on peut utiliser des tests de normalité (Sharipo-Wilk, Kolmogorov-Smirnov, Lilliefors), ou, plus généralement, des tests non paramétriques (χ^2, Mann-Whitney, *etc.*).

La description des variables aboutit à un certain nombre de calculs incontournables en fonction du type de variables.

La description propose une visualisation graphique :

— un histogramme pour les variables continues ;

— une représentation sectorielle pour les variables qualitatives.

La description propose des tableaux de synthèse :

— rassemblant les paramètres caractéristiques ;

— présentant les intervalles de confiance.

En ultime étape, la statistique descriptive établit des relations entre plusieurs informations par :

— des réductions de dimensions (A.C.P., A.F.C., A.C.M., A.F.D., *etc*.) ;

— une définition de variables dépendantes (V.D.) (ou à expliquer) et de variables indépendances (V.I.) (ou explicatives).

Sciences et statistiques

À quoi servent les statistiques en sciences expérimentales [3] ? Les résultats d'une observation, d'une mesure sont rarement égaux à la valeur théorique espérée [4] par le statisticien.

Quelles sont les causes des fluctuations des valeurs mesurées ? Premièrement, les fluctuations peuvent être soit connues, soit inconnues. Secondement, elles sont soit contrôlées (hasard bénin), soit incontrôlées (hasard sauvage). D'où les deux questions fondamentales de la statistique descriptive.

1. Quel résultat doit-on prendre en compte ?
2. Quel degré de confiance peut-on accorder à la décision prise sur ce résultat ?

Pour opérer un **traitement statistique**, il faut disposer d'au moins une **variable aléatoire**. À ce stade introductif, il faut insister sur le fait que le sens large de la notion de variable aléatoire accepte de nombreuses définitions. Cela précisé, dans la plupart des cas, il s'agit de trouver la meilleure mesure possible.

En aucun cas, le traitement statistique ne peut se substituer au raisonnement explicatif. La statistique permet de :

1. créer une information systématique (observation de même nature sur un ensemble homogène d'objets), ce qui autorise les comparaisons ;
2. traiter l'information créée (résumés numériques, représentations graphiques, étude de la relation entre phénomènes) ;
3. connaître la fiabilité de l'information ;
4. progresser vers des applications opérationnelles.

La statistique ne permet pas de :

1. remplacer le raisonnement de type explicatif ;
2. remplacer la culture et les connaissances.

3. **N.B.** En procédant une analyse statistique, la science humaine et sociale qui utilise ces méthodes, adopte une posture expérimentale. Il est à noter que la géographie a été très loin dans ce domaine en proposant notamment des modèles de simulation spatiale.

4. Une **valeur théorique** est dite **espérée**. Une **valeur observée** est dite **mesurée**.

Dit autrement, la statistique permet de compléter des connaissances, mais, en aucun cas de les créér.

La statistique ne fournit que des résultats synthétiques, des visions résumées : elle peut faciliter l'interprétation, et non s'y substituer ; elle valorise la culture disciplinaire. Elle permet par conséquent d'étudier des phénomènes complexes à partir de faits constatés. On cherche alors à extraire du jeu de données statistiques des connaissances, ou des éléments significatifs, à détecter une structure dans l'aléatoire, ou encore à confronter une théorie à la réalité.

1.3.3 Analyse de données

Les méthodes statistiques d'analyse des données se divisent en trois grandes classes : (1) les méthodes descriptives ; (2) les méthodes explicatives ; (3) les méthodes de prévision.

Les méthodes descriptives

Les méthodes descriptives s'appliquent en particulier aux tableaux individus contentant k variables dans lesquels toutes les variables jouent le même rôle. Il n'y a pas de variable à expliquer Y. Il s'agit ainsi de résumer le tableau des variables $[X_1, \ldots, X_k]$, et de comprendre les grandes dimensions du phénomène étudié. L'objectif des méthodes descriptives est de visualiser et de classer les données. En France, on appelle « l'analyse des données » les méthodes descriptives multidimensionnelles (Tab. 1.1).

Méthodes de visualisation			
Nature des variables $X_1, \ldots, X_k$			
Quantitatives	Qualitatives		Mélange
	$k = 2$	$k > 2$	
Analyse factorielle en composantes principales	Analyse factorielle des correspondances	Analyse factorielle des correspondances multiples	Options dans de nombreux logiciels de statistique
Méthodes de classification			
Classification des individus	$\longrightarrow$ Classification ascendante hiérarchique (C.A.H.) des individus $\longrightarrow$ Nuées dynamiques		
Classifications des variables	C.A.H. des variables		

TABLE 1.1 – Les méthodes descriptives pour un tableau individus $\times$ variables [Tenenhaus, 2007, p. 4]

L'analyse statistique devient vite un casse-tête lorsque plusieurs variables sont en présence. Bien heureusement, il existe de nombreuses méthodes et outils (comme l'ordinateur) pour arriver à tirer des conclusions sur les relations, ou l'absence de relations, les unissant, ou pas.

L'**analyse factorielle en composantes principales**[5] (A.C.P.) permet de visualiser des données quantitatives.

L'**analyse factorielle des correspondances**[6] (A.F.C.) permet de visualiser un tableau de contingence en termes de proximités entre les lignes, les colonnes, et entre les lignes et les colonnes de deux variables qualitatives. Elle est utilisée pour les variables qualitatives. Plus largement, l'**analyse factorielle des correspondances multiples**[7] (A.C.M.) permet d'en visualiser plus de deux variables qualitatives.

Lorsque les données sont hétérogènes (certaines variables sont quantitatives, et d'autres qualitatives), il est possible d'utiliser de nombreuses procédures présentes dans les logiciels de traitement statistique. Pour mémoire, il existe l'**analyse factorielle de données mixtes** (A.F.D.M.) et l'**analyse factorielle multiple** (A.F.M.).

L'**analyse des proximités** permet de construire une carte de ces objets de telle sorte que les distances entre les objets sur cette carte soient d'autant plus petites que les objets ont un indice de proximité élevé.

Les **méthodes de classification** sont couramment utilisées en complément des procédures d'**analyse factorielle**. La plus utilisée est la classification ascendantes hiérarchiques (C.A.H.).

Les méthodes explicatives

Dans les méthodes explicatives, on cherche à relier une **variable à expliquer** Y[8] à des **variables explicatives**[9] $X_1, \ldots, X_k$[10] (Tab. 1.2). Pour cela, on dispose d'un tableau individus x variables, tableau dont les lignes représentent les valeurs des variables Y, $X_1, \ldots, X_k$ pour les différents individus étudiés. Il s'agit d'ajuster les données disponibles un modèle dont la forme dépend de la nature de la réponse Y. Si la réponse est numérique, le modèle étudié est la forme :

$$Y = \mathrm{f}\left(X_1, \ldots, X_k\right) + \text{aléa} \tag{1.1}$$

Si la réponse est qualitative, il est de la forme :

5. *Principal component factorial analysis* (P.C.A.)
6. *Factorial analysis of simple correspondences*
7. *Factorial analysis of multiple correspondences*
8. appelée aussi variable dépendante ou réponse
9. *predictor variable* ou *explanatory variable*
10. appelées aussi variables indépendantes ou prédicteurs

$$\Pr\left(Y = j/X_1, \ldots, X_k\right) = \mathrm{f}_j\left(X_1, \ldots, X_k\right) \tag{1.2}$$

Dans le cas d'une réponse numérique, la fonction f est le plus souvent linéaire. Dans le cas d'une réponse qualitative, cette fonction f doit prend en compte le fait qu'une probabilité est comprise entre 0 et 1.

Variable à expliquer	**Variables explicatives** $X_1, \ldots, X_k$		
Y	Quantitatives	Qualitatives	Mélange
Quantitative (Loi normale)	$\longrightarrow$ Régression simple $(k = 1)$ $\longrightarrow$ Régression multiple $(k > 1)$	Analyse de la variance	Modèle linéaire général
Qualitative	$\longrightarrow$ Analyse discriminante $\longrightarrow$ Régression logistique $\longrightarrow$ Segmentation		

TABLE 1.2 – Les principales méthodes explicatives [Tenenhaus, 2007, p. 2]

Les méthodes de prévision

Elles concernent l'analyse et la prévision d'une série chronologique. La prévision repose sur la construction d'un modèle reliant le présent au passé.

$$X_t = \mathrm{f}\left(X_{t-1}, X_{t-2}, \ldots\right) + \text{aléa} \tag{1.3}$$

1.4 Vocabulaire statistique

Il est temps de poser un certain nombre de définitions de notions qui seront utilisées dans tous les chapitres.

La **population statistique** [11] correspond à un ensemble au sens mathématique du terme.

Exemple spatial : le nombre d'habitants d'un territoire.

Exemple non spatial : le personnel d'une entreprise.

L'**individu statistique** [12] correspond à un élément de la population statistique. On peut aussi l'appeler **unité statistique**. Dans ce cadre, les données (attributaires) géographiques ont deux particularités :

11. *statistical population*
12. *statistical individual*

1. les « individus » statistiques sont localisables et cartographiables, appelés **unités spatiales** ;
2. les « individus » statistiques sont eux-mêmes fréquemment composés d'un ensemble de personnes, d'entreprises, de points observables, de zones plus petites, de tronçons d'un réseau, appelés **éléments de niveau inférieur**.

Exemple spatial : une ville parmi un réseau inter-urbain.

Exemple non spatial : un salarié parmi les autres.

Par ailleurs, deux types d'unités spatiales sont à considérer :

1. les unités primaires (« atomes »), c'est-à-dire les données non agrégées ;
2. les unités secondaires (« molécules »), c'est-à-dire des données agrégées issues des unités primaires.

Le ou les **caractères d'un individu** [13] sont les caractéristiques (c'est-à-dire les particularités) de l'individu pris parmi la population statistique sur laquelle l'analyse statistique porte. Si un seul caractère est étudié, on parlera d'une série numérique à une dimension. Si on possède deux caractères, on parlera d'une série numérique à deux dimensions. Par extension, si on dénombre plus de deux caractères, on parlera d'une série numérique multidimensionnelle. Pour finir, on définit plusieurs caractères statistiques en fonction de leur modalité.

Exemple spatial : un territoire contient un collection d'éléments le caractérisant.

Exemple non spatial : un salarié parmi les autres possèdent x caractéristiques (salaire, couleur des yeux, âge, nombre d'enfants).

Les **modalités du caractère** [14] correspondent aux valeurs prises par un caractère. Ces modalités doivent être incompatibles et exhaustives, l'objectif étant, bien évidemment, de caractériser l'appartenance, ou la non appartenance, d'un individu à une modalité. Dit autrement, à chaque individu est associée une modalité du caractère, et une seule, et tout individu de la population présente l'une des modalités du caractère. Les modalités forment une partition du caractère, car elles sont exhaustives et disjointes. On appelle un caractère « variable statistique » lorsque l'on connaît ses modalités individu par individu. Techniquement, un caractère peut être de deux natures : soit il s'agit d'une variable qualitative, soit il s'agit d'une variable quantitative. Toutefois, le caractère devient une **variable statistique** [15] (ou aléatoire), ou, pour les variables qualitatives, **valeur aléatoire** [16] lorsqu'il fait l'objet d'**une étude statistique** [17].

Il existe deux façons de définir population et caractère en géographie.

13. *statistical character*
14. *(descriptor) state*
15. *random variable*
16. *random deviate*
17. *statistical analysis*

1. La population est l'ensemble des unités spatiales, ou des attributs.
2. Le découpage territorial peut également être considéré comme le caractère ayant ces propres particularités soit relevant de la géographie humaine (population humaine, produit intérieur brut, *etc*.), soit relevant de la géographie physique (altitude, débit d'un cours d'eau, *etc*.).

Remarque fondamentale. En toute rigueur, une variable statistique ne peut jamais être continue, le degré de précision des mesures entraînant toujours des discontinuités dans les résultats. Toutefois, on considère qu'elle l'est lorsque le domaine de définition de ces modalités est l'ensemble des réels.

1.5 Caractérisation élémentaire des caractères statistiques

On dispose d'une série statistique **si** on connaît pour chaque individu la modalité (c'est-à-dire la valeur) par son **caractère** (ou le type de variable). Une série de données présente soit des caractères qualitatifs ou quantitatifs, soit des caractères discrets ou continus.

1.5.1 Typologie des variables

Savoir définir le type de la variable étudiée est fondamentale pour choisir les outils d'analyse et les lois de probabilité associée.

Il existe deux grands types de variables : les **variables qualitatives** [18] et les **variables quantitatives** [19]. Les variables qualitatives désignent une qualité. C'est une éventualité non chiffrée. Elle échappe de fait à la mesure. Leur traitement nécessite un codage préalable. L'ensemble des modalités d'un tel caractère est appelé catégorie. Numériquement, il n'est possible que d'établir des fréquences. La moyenne n'a aucun sens. Plus généralement, il n'existe aucune loi de probabilité associable à ce type de variables. Par opposition, les variables quantitatives désignent une quantité, c'est-à-dire une éventualité chiffrée. L'ensemble des modalités est appelé **valeur**. Numériquement, il existe plusieurs paramètres calculables, dont la moyenne, l'écart type, *etc*. Seules les variables quantitatives sont associables à des lois de probabilité qui aboutissent à une qualification de la distribution. En réalité, chacune de ces deux catégories se subdivisent en deux sous-catégories, ce qui fait un total de **quatre types de variables statistiques**.

Les **variables qualitatives nominales** (ou catégorielles, ou textuelles) décrivent des **états**, par exemple, la couleur des yeux. Leur but est de **qualifier** des

18. *attribute*
19. *quantitative variable*

données. Numériquement, il n'est possible que de calculer la fréquence de leurs modalités et de visualiser les résultats par un graphique en secteur. Dit autrement, la variable caractéristique est alors l'**effectif d'une modalité** qui est le nombre de fois où elle apparaît dans la population. De fait, aucun test paramétrique n'est envisageable sur ce type de variable. Néanmoins, il est possible de vérifier une liaison, une dépendance statistique, par un test du χ^2 ou d'autres tests non paramétriques.

Les **variables qualitatives ordinales**[20] décrivent des **relations**. Leur but est d'**ordonner** et de **classer** des données, par exemple, le positionnement hiérarchique. Numériquement, il est possible de calculer la fréquence et la médiane de leurs modalités et de visualiser les résultats par un histogramme disjoint. La médiane existant, des tests paramétriques sont possibles sous certaines conditions. Par exemple, le choix à une question qui propose comme réponse type : « très souvent », « souvent », « parfois », « rarement », « jamais », induit une relation d'ordre entre les réponses possibles. Parfois, on effectue des **variables de score** grâce à l'échelle de R. Likert (1, 2, 3 ,4, 5), mais le fait de remplacer les variables pour un nombre n'en fait pour autant une variable quantitative. Les tests possibles sont le test du χ^2, le test de Mann-Withney, le test de Wilcoxon.

Rensis Likert (1903-1981)

Les **variables quantitatives discrètes**[21] (ou discontinues) décrivent des **listes finies et isolées de valeurs**, comme un choix entre « oui » ou « non », le nombre d'enfants, *etc*. Leur but est de **compter** les données. Dit autrement, ces variables correspondent à un décompte dans lequel elles ne peuvent prendre que certaines valeurs bien précises. Une variable nominale est souvent sous-jacente. Il est possible de visualiser les résultats par un diagramme en bâtons. Les tests possibles sont ceux utilisant la loi normale, le test du χ^2 et les tests non paramétriques.

Les **variables quantitatives continues**[22] décrivent des **valeurs prises dans un intervalle**, par exemple, l'âge, le salaire, *etc*. Leur but est de **mesurer** quelque chose par les données, ce qui implique que ces variables ont une unité. Il existe une infinité de valeurs prises, en général, dans un intervalle réel. On en distingue deux types :

- les **variables d'intervalle** (ou scalées) dans lesquelles le zéro ne correspond pas à une absence de la grandeur mesurée, comme la température, *etc.* ;
- les **variables de rapport** (ou métrique) dans lesquelles le zéro est naturel, c'est-à-dire qu'il correspond à l'absence du phénomène étudié, comme le temps, le volume de vente, *etc*.

Les variables quantitatives continues disposent de trois types de paramètres : 1. les paramètres de position (minimum, maximum, moyenne, médiane, mode, *etc.*) ;

20. *categorial data*
21. *discrete variate*
22. *continuous variable*

2. les paramètres de dispersion (variance, étendue, écart type, quantile, *etc*.) ; 3. les paramètres de forme (coefficients d'asymétrie, coefficient d'aplatissement, *etc*.). En général, un tableau de synthèse doit être effectué pour résumer les paramètres caractéristiques et les intervalles de confiance. Les variables peuvent être visualisées par des histogrammes, par des boîtes à moustache, par un lissage normal, *etc*. Avec des variables quantitatives continues, il est possible d'effectuer des tests paramétriques, des tests de normalité et, surtout, de comparer un échantillon par rapport à sa population d'origine. Les tests sont nombreux. Pour tester une moyenne, on utilise le test de Student, afin d'établir un intervalle de confiance et un domaine de variation pour les paramètres. Pour tester une variance, il existe le test de Fisher ou le test A.N.O.Va. On peut reprendre des tests non paramétriques tels que ceux de Mann-Withney ou de Wilcoxon. Il est difficile d'en dresser une liste exhaustive.

Remarque importante. On peut considérer une série discrète comme étant continue s'il existe un nombre de modalités importantes.

N.B. Il existe des **variables semi-quantitatives** sur lesquelles des **moyennes pondérées** peuvent être effectuées, comme la densité de population.

En résumé, identifier les variables est important pour :

1. choisir les traitements statistiques appropriés ;
2. savoir interpréter leurs résultats.

1.5.2 Discrétisation des caractères quantitatifs

La **discrétisation** est un processus ayant pour objectif de constituer des **classes**. Deux écueils sont à éviter : le découpage trop fin et le découpage trop grossier. On considère en général des intervalles du type $]a, b]$ que l'on appelle **classe statistique** [23]. Mathématiquement, cela consiste à transformer un vecteur de nombres réels en un vecteur de nombres entiers nommés **indices de classe**. De fait, cette transformation se dit en langage courant « réaliser un découpage en classes ». En statistique, discrétiser c'est à la fois réaliser cette transformation mathématique, nommer et, surtout, justifier les classes.

La série aura alors deux variables caractéristiques.

1. L'**amplitude** [24] est la longueur $b - a$ avec a la valeur minimale de la classe et b la valeur maximale. Elle concerne toujours une classe. Il faut insister sur ce point.

23. *group*, *size class*
24. *amplitude*, *range*

2. La **densité**[25] est le rapport entre l'effectif n_i et l'amplitude de la classe décrivant une modalité i. On appelle d la densité :

$$d = \frac{n_i}{b-a} \tag{1.4}$$

Les classes peuvent être d'égale amplitude, ou non. On choisit d'étudier soit le **nombre de classes**, soit l'**amplitude des classes**.

1. Le nombre de classes ne doit être ni trop petit, ni trop grand. Que ce soit l'un ou l'autre extrême, on arrive fatalement à une perte d'informations non négligeable. Évidemment, le nombre de classe dépend du nombre d'observations et de l'étalement des données. La **formule de Sturges** donne une valeur approximative du nombre k de classes :

Herbert Sturges (1892-1958)

$$k \approx 1 + 3,2222 \times \log_{10} n \tag{1.5}$$

Il est possible également d'utiliser la **formule de Yule** :

George Udney Yule (1871-1951)

$$k \approx 2,5\sqrt[4]{n} \tag{1.6}$$

2. L'amplitude des classes A est égale au rapport entre l'étendue[26] de la série des observations et du nombre de classe :

$$A = \frac{x_{\max} - x_{\min}}{k} \tag{1.7}$$

Si l'on commence par définir l'amplitude des classes, on ne doit ni choisir cette amplitude trop faible, car le nombre de classes devient évidemment trop élevé, ni choisir cette amplitude trop grande, car le nombre de classes est alors trop petit par rapport à celui que donne la formule de Sturges. Les valeurs d'une classe sont généralement assimilées à la valeur centrale (ou centre de la classe) qui est égale à :

$$A = \frac{b-a}{2} \tag{1.8}$$

avec a la valeur minimale et b la valeur maximale de l'amplitude de classe.

Dit autrement, le regroupement en classes fait perdre aux individus leur caractère propre, ainsi que les détails fins des distributions.

Graphiquement, on peut représenter un caractère continu, réparti en classe :

1. par un histogramme ;

25. *density*

26. L'étendue est la différence entre la valeur maximale et la valeur minimale de la série étudiée. L'amplitude de classe est une étendue locale.

2. par un polygone de fréquence ;
3. par une courbe cumulative décroissante ;
4. par des diagrammes à secteurs circulaires ;
5. par des diagrammes à rectangles horizontaux.

N.B. La discrétisation permet également de préparer le croisement entre deux données quantitatives et des données qualitatives. Les classes définies devient alors des catégories.

1.5.3 Effectifs et fréquences

Avec les effectifs et les fréquences, un pont s'établit entre probabilité et statistique.

Effectifs et fréquences simples

L'**effectif**[27] (ou **fréquence absolue**) n_i associé à une valeur x_i de la variable aléatoire X correspond au nombre d'apparitions de cette variable dans la population.

Pour une modalité d'un caractère (ou pour une classe statistique), la **fréquence relative**[28] est le rapport entre l'effectif n_i par l'effectif total n (correspond à la somme de tous n_i). La fréquence de la modalité i d'un caractère, notée f_i, vaut :

$$f_i = \frac{n_i}{n} \tag{1.9}$$

Effectifs et fréquences cumulées

Pour un caractère quantitatif, les valeurs (ou les classes) peuvent être rangées par ordre croissant. Dans ce cas, il est possible de calculer un effectif et une fréquence cumulés

L'**effectif cumulé** (ou fréquence cumulée absolue) jusqu'à k modalités est la somme des effectifs associés aux valeurs du caractère qui sont inférieures ou égales à k.

$$f_i = \sum_{i=1}^{k} n_i \leq k \tag{1.10}$$

27. *number*
28. *frequency*

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.

$$\sum_{i=1}^{k} f_i = \sum_{i=1}^{k} \frac{n_i}{n} = \frac{1}{n} \sum_{i=1}^{k} n_i \leq k \tag{1.11}$$

De fait, pour toute série à k modalités, la formule fondamentale suivante apparaît :

$$\sum_{i=1}^{k} f_i = 1 \tag{1.12}$$

Remarque importante. Il ne faut pas confondre lors d'une analyse de données l'effectif total n et le nombre de modalités k. Au stade exploratoire des données, il est fréquent que l'on confonde les deux nombres entiers. De fait, il est rare que $n = k$, c'est-à-dire que le nombre de modalités soit égal à l'effectif total, puisque, dans la plupart des études, il s'agit de réduire le nombre n à quelques éléments agrégés significatifs représentés par le nombre k. Ainsi, dans la pratique, on trouve $k \leq n$. On retrouve ici l'idée que les statistiques ne disent rien au sujet de chaque caractère, mais informent sur les grandes tendances observées.

La fréquence, étant une valeur comprise entre 0 et 1, sert à établir une probabilité réelle dans le sens où elle est observée. Elle permet d'établir une **distribution statistique empirique** de laquelle il sera possible de conclure sur le type de loi de probabilité utilisée. Pour construire la distribution, il faut préablement connaître les paramètres statistiques élémentaires.

Bibliographie

[Béguin, 1979] BÉGUIN, H. (1979). Méthodes d'analyse géographique quantitative. Litec, Paris.

[Chadule, 1997] CHADULE, G. (1997). Initiation aux pratiques statistiques en géographie. Masson, Paris.

[Dumolard, 2011] DUMOLARD, P. (2011). Données géographiques. Analyse statistique multivariée. Lavoisier - Hermès, Paris.

[Dumolard et al., 2003] DUMOLARD, P., DUBUS, N. et CHARLEUX, L. (2003). Les statistiques en géographie. Atout géographie. Belin, Paris.

[Escofier et Pagès, 2016] ESCOFIER, B. et PAGÈS, J. (2016). Analyses factorielles simples et multiples. Cours et études de cas. Sciences sup. Dunod, Paris.

[Fréchet, 1941] FRÉCHET, M. (1941). Sur la loi de répartition de certaines grandeurs géographiques. Journal de la société statistique de Paris, 82:114–122.

[Jacquard, 2000] JACQUARD, A. (2000). Les probabilités. Que sais-je ? n°1571. PUF, Paris. réédition de 1974.

[Korčák, 1940] KORČÁK, J. (1940). Deux types fondamentaux de distribution statistique. Bulletin de l'Institut international de statistique, 30(3):295–299. Rapports et communications présentés à la XXIVe session de l'Institut international de statistique, Prague, 1938, 2e partie.

[Marchand, 1972] MARCHAND, B. (1972). L'usage des statistiques en géographie. L'espace géographique, 1(2):79–100.

[Morgenthaler, 2007] MORGENTHALER, S. (2007). Introduction à la statistique. Enseignement des mathématiques. Presses polytechniques et universitaires romandes, Lausanne. 3e édition augmentée.

[Spiegel, 1984] SPIEGEL, M. R. (1984). Théorie et applications de la statistique. Série Schaum. McGraw-Hill, Paris. réédition de 1972.

[Tenenhaus, 2007] TENENHAUS, M. (2007). Statistique. Méthodes pour décrire, expliquer et prévoir. Dunod, Paris.

[Wonnacott et Wonnacott, 1995] WONNACOTT, T. H. et WONNACOTT, R. J. (1995). Statistique. Économie – Gestion – Sciences – Médecine. Économica, Paris.