RAPPORT D'ACTIVITES

PARCOURS - DEBUTANT

ETUDIANTE :

TUEDOM TUEDOM ALIDA

ENSEIGNANT :

Maxime FORRIEZ

SOMMAIRE

INTRODUCTION GENERALE

Dans un contexte où les territoires sont de plus en plus étudié à travers des données chiffrées, la statistique occupe une place centrale dans la formation du géographe. Loin d'être un simple outil mathématique, elle constitue un véritable cadre de pensée permettant d'analyser, d'interpréter et de comprendre la complexité des phénomènes spatiaux. Le cours de statistiques en géographie s'inscrit ainsi dans une démarche progressive, visant de transmettre à l'étudiant des bases théoriques et méthodologiques nécessaires pour transformer des observations empiriques en connaissances scientifiques précises. Ce travail de synthèse s'inscrit précisément dans cette démarche. Il sera question de travailler sur cinq séances fondamentales du cours, depuis les principes généraux de la statistique, des fondements épistémologiques du hasard et de la donnée, jusqu'à l'analyse des structures hiérarchiques, en passant par les méthodes de description statistique, la modélisation par des lois théoriques et les principes de l'inférence statistique. L'objectif est de mettre en évidence le rôle fondamental de la statistique comme langage d'analyse du monde géographique.

SEANCE 2 : LES PRINCIPES GÉNÉRAUX DE LA STATISTIQUE

1. Quel est le positionnement de la géographie par rapport aux statistiques

La relation entre ces deux concepts a été marquée par une forme de paradoxe structurel. D'une part, la géographie avait pour but de décrire la spécificité unique des lieux, des pays et des territoires. Les géographes littéraires considéraient que les définitions mathématiques élémentaires ne relevaient pas de leur champ de compétence, créant ainsi une barrière culturelle entre l'analyse spatiale et le formalisme quantitatif.

En effet, la géographie est une science qui produit, manipule et nécessite des volumes considérables de données que ce soit de relevés météorologiques, de recensements démographiques ou d'inventaires de ressources, la matière première du géographe est tout d'abord statistique. Cette contradiction a longtemps placé la géographie dans une position inconfortable, fréquemment perçue comme une discipline avant tout descriptive ou relevant de la synthèse littéraire plutôt qu'au rang de science explicative capable de dégager des lois générales.

Aujourd'hui, sa position a radicalement évolué. L'avènement de l'ère des données massives encore appelé le Big Data et la généralisation des Systèmes d'Information Géographique ont rendu la maîtrise de la statistique incontournable. Il ne s'agit plus d'une option, mais d'une compétence importante pour la survie et la pertinence de la discipline. Aucun géographe contemporain ne peut sérieusement prétendre analyser les dynamiques territoriales complexes sans recourir à la puissance de l'outil statistique. La statistique est devenue un langage commun donc le but est de valider des hypothèses, de modéliser des phénomènes spatiaux et de transmettre à la géographie une légitimité scientifique comparable à celle des sciences économiques ou sociologiques.

2. Le hasard existe-t-il en géographie

La réponse à cette question conditionne toute la démarche scientifique. Le cours nous permet d'explorer deux grandes places philosophiques qui traversent l'histoire des sciences.

La première place est celle du déterminisme absolu. Dans ce cas, le hasard n'est qu'une illusion, un mot que nous posons sur notre ignorance. Dans cette perspective déterministe, tout événement géographique a une cause précise. Si nous ne parvenons pas à l'expliquer, ce n'est pas parce que le monde est aléatoire, mais parce que nous n'avons pas encore identifié la cause cachée. Le hasard n'est donc qu'une lacune de notre savoir qui consiste à être comblée par le progrès de la connaissance.

La seconde place, celle adoptée par la statistique moderne, est beaucoup plus productif pour la géographie humaine. Elle ne cherche pas à nier ou affirmer l'existence divine du hasard, mais elle propose un cadre méthodologique pour le gérer. Elle montre une distinction fondamentale entre le niveau local et le niveau global. C'est ici qu'intervient le raisonnement multiscalaire propre au géographe. À l'échelle de l'individu, le comportement est imprévisible, soumis au libre arbitre ou à des contingences infinies.

Toutefois, dès que l'on change d'échelle pour observer la population dans son ensemble ce hasard individuel s'efface au profit d'une régularité collective. Les comportements individuels imprévisibles s'agrègent pour former des tendances lourdes, stables et prévisibles. C'est ce que l'on appelle la loi des grands nombres. Ainsi, la statistique permet d'affirmer qu'il existe une certitude globale qui émerge de l'incertitude locale. Pour le géographe, cela signifie que l'on peut modéliser les flux de transport, les dynamiques résidentielles ou les épidémies sans avoir besoin de connaître les motivations intimes de chaque personne. Le hasard devient une composante du modèle, permettant de dégager l'action la plus probable au sein d'un territoire.

3. Quels sont les types d'information géographique

Le premier type d'information géographique concerne les séries caractérisant l'ensemble délimité, que l'on appelle généralement :

- **La base attributaire ou les données attributaires** : Il s'agit ici du contenu sémantique des lieux, de la réponse à la question Quoi ?. Ces données décrivent tout ce qui qualifie une unité territoriale donnée. En géographie humaine, ces attributs peuvent être la population totale, le taux de chômage, le niveau d'éducation, ou encore l'appartenance politique. En géographie physique, il s'agira de températures moyennes, de volumes de précipitations, de types de sols ou de débits fluviaux. Ces données sont le cœur de l'analyse statistique, les variables sur lesquelles porteront les calculs de moyenne ou de dispersion.
- **Le second type d'information géographique est les données géométriques** : Ici, nous répondons aux questions suivantes : Où et Quelle forme ?. Ces données concernent l'étude statistique de la géométrie même des unités spatiales ; la surface d'une commune, la longueur d'un réseau routier, la forme d'une parcelle, les coordonnées géographiques d'un point. Dans un système d'information géographique, cette distinction est technique mais elle est aussi conceptuelle. L'analyse spatiale naît précisément de la mise en relation de ces deux types d'informations.

4. Quels sont les besoins de la géographie au niveau de l’analyse de données

L'analyse de données en géographie répond à des besoins méthodologiques stricts qui se manifestent à deux moments clés du processus scientifique : lors de la production ou collecte de la donnée et lors de son analyse proprement dite.

- Le premier besoin est celui de la nomenclature. Avant même de commencer à compter ou mesurer, le géographe doit définir des catégories claires, exclusives et si possible hiérarchiques. Il s’agit ici de se poser des questions comme : Qu’est-ce qu'une ville ? Qu'est-ce qu'un actif ? Sans définitions préalables stables, les chiffres ne veulent rien dire. Une nomenclature hiérarchique permet en outre de naviguer entre les échelles, en agrégeant des données fines pour obtenir des synthèses globales.
- Le deuxième besoin crucial est celui des méta-données. Pour être utilisable, elle doit être accompagnée de sa carte d'identité a l’instar des interrogations suivantes : Qui l'a produite ? Quand ? Comment ? Avec quelle marge d'erreur ? elles permettent l'examen critique des sources, garantissant que le géographe ne compare pas l'incomparable.

Cependant, au niveau de l'analyse, les besoins changent de nature. L'analyse ne consiste pas seulement à presser un bouton sur un logiciel. Elle nécessite une compréhension approfondie de la structure interne des données. Le géographe doit étudier le moment mathématique des variables, c'est-à-dire leur distribution, leur dispersion, leurs anomalies. Le géographe doit en permanence confronter les résultats sortis des algorithmes statistiques avec sa connaissance fraiche du terrain et du phénomène étudié. Il doit faire dialoguer la mathématique avec l'expertise thématique. Un coefficient de corrélation élevé ne suffit pas à prouver une causalité ; seul le raisonnement géographique peut valider le sens de cette relation. Enfin, l'analyse requiert une maîtrise des lois du hasard pour distinguer ce qui relève d'une tendance structurelle significative de ce qui ne relève que du bruit aléatoire ou des fluctuations d'échantillonnage.

5. Quelles sont les différences entre la statistique descriptive et la statistique explicative

La différence entre la statistique descriptive et la statistique explicative aussi appelée mathématique est :

- **La statistique descriptive** : Elle cherche à résumer, synthétiser et visualiser l'information contenue dans un tableau de données, sans chercher à aller au-delà de ce qui est observé. Son but est de dégager des propriétés remarquables comme une moyenne, une fréquence,

une dispersion pour offrir une image simplifiée et intelligible de la réalité complexe. Les outils incontournables de cette approche sont les paramètres de position moyenne tels que la médiane, la dispersion écart-type, ainsi que les grandes méthodes d'analyse factorielle comme l'analyse en composantes principales ou l'analyse factorielle des correspondances. La statistique descriptive est une étape préalable indispensable.

- **La statistique explicative**, quant à elle, ne se contente pas de constater, elle cherche à comprendre, à modéliser et à prédire. Elle introduit la notion de causalité ou du moins de dépendance en distinguant une variable à expliquer (Y) de variables explicatives (X). Son but est d'ajuster un modèle mathématique à la réalité observée pour vérifier si des hypothèses théoriques tiennent la route face aux faits. Elle permet également, de généraliser des résultats obtenus sur un échantillon à l'ensemble d'une population. Les méthodes phares de cette approche sont les régressions linéaires, multiples, logistiques, les analyses de variance et les tests d'hypothèses.

6. Quelles sont les types de visualisation de données en géographie et comment les choisir

Le choix d'un type de graphique ne dépend pas des préférences de l'analyste, mais est entièrement orienté par la nature de la variable étudiée.

- **Pour les variables qualitatives nominales** comme les types d'occupation du sol ou les catégories socioprofessionnelles, l'objectif est de visualiser des fréquences, c'est-à-dire la part de chaque catégorie dans le total. La représentation graphique la plus adaptée est la représentation sectorielle qui renvoie au diagramme circulaire ou camembert. Elle permet de voir immédiatement les proportions relatives des différentes modalités.
- **Pour les variables qualitatives ordinales** des catégories ordonnées, comme une échelle de satisfaction faible, moyen, et fort, on cherche à visualiser à la fois la fréquence et la hiérarchie. L'histogramme disjoint ou diagramme en barres séparées est a adapté, car il respecte l'ordre des catégories sur l'axe des abscisses tout en montrant les effectifs en ordonnée.
- **Pour les variables quantitatives discrètes** des nombres entiers issus d'un comptage, comme le nombre d'enfants par foyer, les valeurs sont isolées. Il n'y a pas de continuité entre deux et trois enfants. La représentation graphique idéale est le diagramme en bâtons. La finesse des bâtons symbolise le caractère ponctuel et isolé de chaque valeur.

- Enfin, **pour les variables quantitatives continues** des mesures réelles comme l'âge, le revenu ou la température, la logique change. Ces variables peuvent prendre une infinité de valeurs sur un intervalle. La représentation a adopté est l'histogramme. Contrairement au diagramme en bâtons, l'histogramme est constitué de rectangles adjacents dont la surface et non seulement la hauteur est proportionnelle à l'effectif. Cette notion de surface est capitale car elle traduit l'idée de densité de probabilité sur un intervalle continu.

7. Quelles sont les méthodes d'analyse de données possibles

Le cours a dressé une liste structurée des familles de méthodes statistiques à la disposition du géographe, classées selon leur finalité.

- La première famille regroupe les **méthodes descriptives** ou analyse des données au sens strict. Elles traitent des tableaux c'est-à-dire individus et variables où toutes les variables ont le même statut. Il n'y a pas de distinction a priori entre cause et effet. Le but est de réduire la complexité, de résumer l'information et de visualiser les structures cachées. On y trouve l'analyse en composantes principales pour les variables quantitatives, qui condense l'information en nouveaux axes synthétiques, l'analyse factorielle des correspondances pour les tableaux de contingence qualitatifs, et les méthodes de classification comme la classification ascendante hiérarchique qui regroupent les individus en types homogènes.
- La deuxième famille rassemble les **méthodes explicatives**. Ici, la démarche est asymétrique : on cherche à expliquer une variable cible. Si la variable à expliquer est quantitative comme par exemple le prix de l'immobilier, on utilisera la régression linéaire simple ou multiple. Si la variable à expliquer est qualitative comme le risque de glissement de terrain, on se tournera vers la régression logistique ou l'analyse discriminante. L'analyse de la variance ou encore ANOVA permet quant à elle de tester l'effet d'une variable qualitative sur une variable quantitative.
- La troisième famille, souvent utilisée en géographie économique ou démographique, concerne les **méthodes de prévision**. Elles s'appliquent spécifiquement aux données temporelles. L'objectif est de construire un modèle qui relie le présent au passé pour extrapoler le futur. C'est le domaine de l'analyse des tendances, des cycles et de la saisonnalité.

8. Comment définiriez-vous : population, individu, caractères, modalités

La population statistique désigne l'ensemble global, au sens mathématique, sur lequel porte l'étude. En géographie, cette population peut être un ensemble d'habitants, mais aussi un ensemble de villes, de parcelles agricoles ou de jours de pluie.

L'**individu statistique** aussi appelé unité statistique ou unité spatiale en géographie est l'élément élémentaire de cette population. C'est sur lui que l'on effectue la mesure. Pour mieux l'illustrer, si la population est la France, l'individu peut être la commune ou le département.

Les caractères statistiques sont les propriétés ou les attributs sont mesurés sur chaque individu. Ce sont les variables de l'étude de l'âge, la superficie et le revenu.

Les modalités sont les différentes valeurs ou états que peut prendre un caractère. Pour qu'une analyse soit valide, les modalités d'un caractère doivent respecter deux principes : elles doivent être incompatibles c'est-à-dire un individu ne peut pas avoir deux modalités en même temps, par exemple être à la fois marié et célibataire. Elles doivent également être exhaustives c'est-à-dire que chaque individu doit pouvoir être classé dans une modalité.

Il existe une hiérarchie implicite entre les types de caractères, fondée sur la richesse des opérations mathématiques permises. Au sommet se trouvent les variables quantitatives continues ou discrètes, car elles supportent toutes les opérations arithmétiques à savoir : addition, moyenne, variance. Viennent ensuite les variables qualitatives ordinales, qui permettent au moins un classement logique. Enfin, à la base, se trouvent les variables qualitatives nominales, qui ne sont que des étiquettes et ne permettent que des calculs de fréquence ou de mode.

9. Comment mesurer une amplitude et une densité

Lorsque l'on travaille avec des variables quantitatives continues, il est souvent nécessaire de regrouper les valeurs en classes pour rendre l'information lisible. C'est la discrétisation. Deux notions techniques sont alors primordiales : l'amplitude et la densité.

- L'**amplitude** est une mesure simple de la largeur d'une classe. Elle se définit comme la différence entre la borne supérieure et la borne inférieure de l'intervalle. C'est une étendue locale qui nous dit quelle portion de l'échelle de valeurs est couverte par une classe donnée.
- La **densité** quant à elle est une notion plus subtile mais essentielle pour la construction correcte des histogrammes. Si les classes ont des amplitudes inégales, on ne peut pas comparer directement les effectifs bruts. Pour corriger ce biais visuel, on calcule la densité, qui est le rapport entre l'effectif de la classe et son amplitude. C'est cette densité qui doit

déterminer la hauteur du rectangle dans l'histogramme. Ainsi, la surface du rectangle redonne bien l'effectif.

10. Rôle des formules de Sturges et de Yule ?

Pour éviter l'arbitraire total, les statisticiens ont mis au point des formules qui proposent un nombre de classes idéal en fonction de l'effectif total de la population.

- **La formule de Sturges** est la plus célèbre. Elle suggère que le nombre de classes doit croître de manière logarithmique avec la taille de la population.
- **La formule de Yule** propose une approche légèrement différente, souvent plus généreuse en nombre de classes pour les grands effectifs.

Ces formules ne sont pas des lois absolues, mais des guides précieux pour le géographe, lui permettant de justifier son choix de découpage et de minimiser la perte d'information inhérente à toute simplification. Une fois déterminé, on peut estimer l'amplitude moyenne idéale des classes en divisant l'étendue totale de la série par ce nombre.

11. Comment définir un effectif ? Comment calculer une fréquence et une fréquence cumulée ? Qu'est-ce qu'une distribution statistique ?

a) Définition d'un effectif

L'effectif aussi appelé **fréquence absolue**, est la donnée la plus basique ; c'est le nombre de fois qu'une valeur ou une modalité apparaît dans la population. C'est le résultat du comptage. Bien que nécessaire, l'effectif est difficile à comparer d'une étude à l'autre si la taille totale des populations diffère.

b) Comment calculer une fréquence et une fréquence cumulée ?

La fréquence ou fréquence relative s'obtient en divisant l'effectif partiel par l'effectif total de la population. La fréquence est un nombre compris entre 0 et 1 ou un pourcentage. Elle a une vertu fondamentale : elle permet de comparer la structure d'une petite population avec celle d'une grande.

La fréquence cumulée est un outil puissant qui ne s'applique qu'aux variables que l'on peut ordonner. Elle s'obtient en additionnant pas à pas les fréquences, du plus petit au plus grand. La fréquence cumulée à une valeur nous dit : quelle proportion de la population a une valeur inférieure

ou égale à la population. C'est l'outil de base pour construire les courbes de concentration et pour calculer les médianes ou les quantiles.

c) La distribution statistique

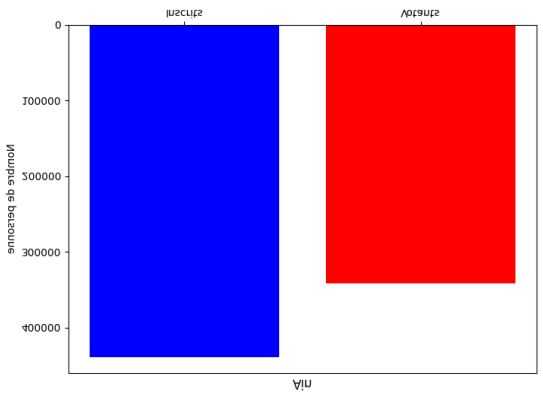
La distribution statistique est la synthèse de tout ce qui précède. La distribution n'est pas un nombre, c'est une forme. C'est la manière dont les effectifs ou les fréquences se répartissent sur l'ensemble des modalités possibles. Étudier une distribution, c'est regarder si elle est symétrique ou étalée, si elle a un seul pic ou plusieurs. C'est cette forme empirique observée que le statisticien cherchera par la suite à rapprocher d'une loi de probabilité théorique comme la loi normale pour pouvoir modéliser le phénomène. La distribution est le pont qui relie l'observation du réel à la théorie mathématique.

Au terme de cette analyse, il apparaît clairement que les principes généraux de la statistique ne sont pas de simples conventions bureaucratiques. Chaque définition, chaque distinction entre attributaire et géométrique, entre descriptif et explicatif, entre densité et amplitude porte en elle une vision du monde et une méthode scientifique. Pour le géographe débutant, l'assimilation de ces concepts est la condition sine qua non pour dépasser le stade de l'observation et entrer dans celui de l'analyse spatiale rigoureuse.

MISE EN ŒUVRE AVEC PYTHON

En effet, nous précisons que pour la manipulation python nous avons utilisé VS code vu que nous n'avons pas pu installer python via Docker.

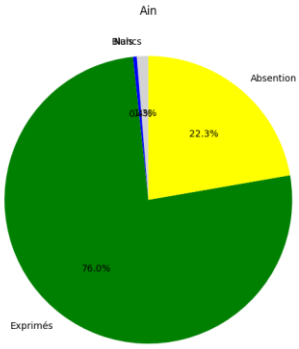
Cette séance à aboutir à l'élaboration de plusieurs résultats mais pour notre cas, nous illustrerons quelques-uns et nous les commenterons.

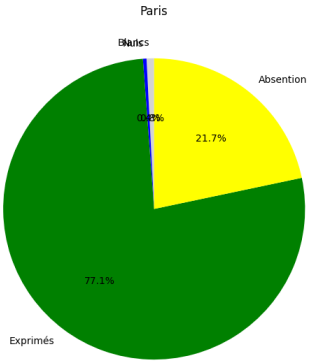


Les manipulations réalisées en Python nous ont permis de passer d'un simple fichier CSV de résultats électoraux à une analyse structurée de la participation au 1er tour de la présidentielle 2022, classé département par département. Nous avons obtenu plusieurs diagrammes. En lisant les données, en identifiant les colonnes quantitatives c'est à dire inscrits, votants, blancs, nuls, exprimés et en calculant leurs totaux, le travail met en évidence le poids

électoral de chaque territoire et la part réelle des électeurs qui se sont déplacés. Les diagrammes en barres montrent, comme pour l'Ain, un écart entre le nombre de personnes pouvant voter et celles qui ont effectivement voté, illustrant concrètement l'abstention locale.

De plus, ces manipulations nous ont également permis d'obtenir plusieurs diagrammes, mais nous prendrons deux exemples précis pour faire des commentaires. Ces diagrammes ou encore des camemberts illustre, pour chaque ville, la répartition des inscrits entre différentes catégories telles que : abstention, votes blancs, votes nuls et suffrages exprimés. Nous constatons que la plus grande part du disque est occupée par les suffrages exprimés en vert, une part importante par l'abstention en jaune, et deux parts très fines pour les bulletins blancs et nuls





Nous remarquons que les deux diagrammes montrent une situation assez proche. Dans l'Ain, l'abstention représente environ 22,3% des inscrits, alors qu'à Paris elle est d'environ 21,7%. Légèrement plus élevée dans l'Ain, l'abstention reste toutefois du même ordre de grandeur, ce qui signifie que la part d'électeurs qui ne se déplacent pas pour voter au 1er tour est importante dans les deux territoires, mais un peu plus marquée dans le département rural que dans la capitale

Au départ, la manipulation m'a paru très difficile : je ne comprenais absolument rien avec python. Je n'arrivais pas à relier le code aux résultats attendus. Pour progresser, j'ai regardé des tutoriels pour débutants sur Python a l'instar des vidéos d'introduction à Pandas et Matplotlib du type Python pour débutants, ce qui m'a aidé à comprendre pas à pas le processus de manipulation de python. J'ai aussi sollicité l'aide de mes camarades, en particulier Zara, qui m'a bien orienté dans la manipulation ; elle m'a montré comment structurer le script main.py, où placer les instructions de lecture de données et à quel moment créer les graphiques, ce qui m'a permis de mieux voir la logique générale du travail. Même si j'ai pris la main difficilement, j'ai fini par comprendre la signification des principales lignes de code, et lorsque certains passages me semblaient trop complexes, j'ai utilisé l'intelligence artificielle pour générer ou corriger des morceaux de code, avant de les relire et de les adapter à mes données.

SÉANCE 3 : LES PARAMÈTRES STATISTIQUES ÉLÉMENTAIRES

L'objectif de cette séance est de comprendre le fonctionnement et le rôle de la boîte à outils du géographe statisticien. Il s'agit ici des concepts qui permettent de réduire la complexité du réel. Ce processus de réduction s'articule autour de quatre dimensions à savoir : la hiérarchie des caractères, la position où se situe le cœur de la donnée, la concentration et la forme.

1. **Quel caractère est le plus général : le caractère quantitatif ou le caractère qualitatif? Justifier pourquoi.**

Il existe en statistique une hiérarchie implicite, qui place le caractère quantitatif au sommet de la pyramide analytique. Le caractère quantitatif est considéré comme le plus général et le plus riche. Cette primauté n'est pas un jugement de valeur sur l'importance géographique de la donnée, mais un constat sur ses propriétés opératoires.

En effet, une variable quantitative ou encore des nombres réels mesurant des quantités autorise l'intégralité des opérations mathématiques connues. On peut additionner des revenus, multiplier des surfaces, calculer des racines carrées de variances. On peut transformer une liste précise d'âges qui est une donnée quantitative en catégories c'est-à-dire jeunes, Adultes, Vieux qui représente des données qualitatif ordinal. L'inverse est impossible sans perte massive d'information ou invention de données.

À l'opposé, le caractère qualitatif nominal est restrictif. Il désigne des états ou des qualités soit une couleur, un type de sol, ou nom de ville. Sur ces données, l'arithmétique est impuissante ; l'analyse se réduit alors à des opérations logiques égal ou différent à des comptages. C'est pourquoi le quantitatif est le caractère le plus général. Il englobe les potentialités des autres types tout en offrant une profondeur d'analyse calculs de dispersion, de forme, et de modélisation probabiliste inaccessible aux simples qualités.

2. **Quel sont les caractères quantitatifs discrets et caractères quantitatifs continus? Pourquoi les distinguer.**

a) **Les caractères quantitatifs discrets**

Les caractères quantitatifs discrets sont le fruit d'un processus de dénombrement, d'un comptage. Ils évoluent dans le monde des entiers naturels. Par exemple, le nombre d'hôpitaux dans un département ou le nombre d'enfants par femme. La réalité procède ici par sauts, par bonds. Entre

la valeur deux et la valeur trois, il n'y a pas de réalité tangible. En géographie, ces variables décrivent souvent des stocks d'objets distincts.

b) Les caractères quantitatifs continus

Les caractères quantitatifs continus, à l'inverse, résultent d'une mesure physique sur une échelle fluide. Le temps, la distance, la température, la surface, le débit d'un fleuve sont des grandeurs continues. Théoriquement, entre deux valeurs, il existe toujours une infinité de valeurs intermédiaires c'est le cas de cette illustration : entre 20°C et 21°C, il y a 20,5°C, 20,55°C). Même si nos instruments de mesure ont une précision limitée qui finit par discrétiser la lecture, la grandeur sous-jacente est fluide.

Pour le discret, on utilise des bâtons et des sommes; pour le continu, on utilise des histogrammes de densité et des intégrales. Confondre les deux mène à des erreurs d'interprétation, notamment sur la notion de probabilité.

3. Les Paramètres de Position

Une fois la nature de la variable identifiée, la première tâche est de trouver le centre de la distribution. C'est le rôle des paramètres de position.

- **La Moyenne :** Si la moyenne arithmétique est la plus connue, elle n'est pas universelle. Le cours nous enseigne que la géographie physique ou les transports nécessitent d'autres outils. La moyenne géométrique est indispensable pour analyser des phénomènes multiplicatifs, comme des taux de croissance démographique cumulés sur plusieurs années. La moyenne harmonique est la seule rigoureuse pour traiter des rapports, comme des vitesses ou des densités. Faire la moyenne arithmétique de deux vitesses sur un trajet aller-retour est une erreur physique ; la moyenne harmonique rétablit la vérité du temps de parcours.

- **La Médiane :** Elle répond à une question différente telle que quelle est la valeur qui coupe la population en deux moitiés égales ? Son intérêt majeur par rapport à la moyenne réside dans sa robustesse. La moyenne est une mesure algébrique sensible aux valeurs extrêmes. La médiane, elle, ne bouge quasiment pas. Elle s'intéresse au rang, pas au poids des extrêmes.

- **Le Mode :** Le mode ou la classe modale désigne la valeur la plus fréquente du pic de la distribution. C'est le paramètre du plus grand nombre. Son immense avantage est qu'il est

calculable pour tous les types de variables, y compris qualitatives. Une distribution peut être bimodale ou plurimodale, ce qui est une information précieuse : cela révèle souvent la superposition de deux populations distinctes sur un même territoire par exemple, une structure d'âge avec un pic d'étudiants et un pic de retraités.

4. Les Paramètres de Concentration

La géographie nécessite des outils spécifiques pour mesurer la concentration. Il s'agit de :

- **La médiale et la médiane :** La médiane partage l'effectif en deux c'est-à-dire 50% des individus sont en dessous. La médiale partage la masse globale de la variable en deux. Prenons l'exemple du patrimoine foncier : la médiane nous dit quelle surface possède l'individu du milieu. La médiale nous dit à partir de quel individu on a cumulé 50% de la surface totale des terres. Dans un monde parfaitement égalitaire, médiane et médiale seraient confondues. La médiale est donc presque toujours très supérieure à la médiane. L'écart entre ces deux valeurs est une mesure directe de l'inégalité structurelle.
- **L'Indice de Gini :** l'indice de Gini est géométriquement lié à la courbe de Lorenz. Si l'on trace le pourcentage cumulé de la population en abscisse et le pourcentage cumulé de la richesse en ordonnée, on obtient une courbe ventrue. L'indice de Gini calcule l'aire entre cette courbe et la diagonale de parfaite égalité. Un indice de 0 signifie l'égalité parfaite, un indice de 1 signifie l'inégalité absolue. C'est un comparateur universel pour les géographes étudiant le développement.

5. Les Paramètres de Dispersion

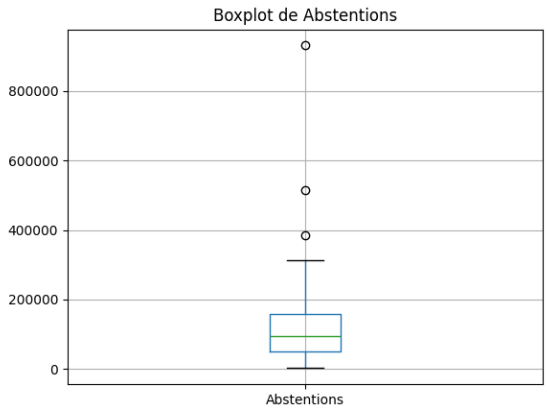
On distingue plusieurs paramètres de dispersion à savoir :

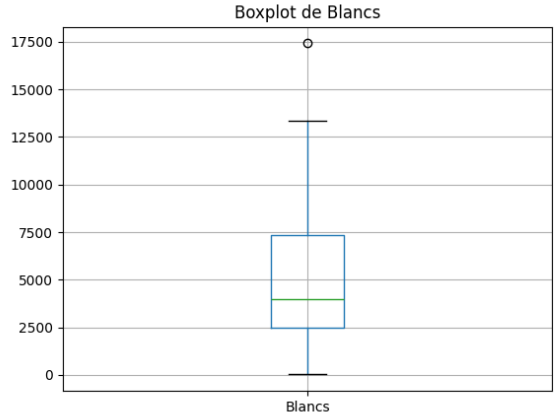
- **La Variance et écart-type :** La variance est donc la moyenne des carrés des écarts. On calcule donc sa racine carrée. L'écart-type est la mesure standard de la dispersion ; il quantifie l'hétérogénéité moyenne de la population.
- **L'Étendue :** C'est la mesure la plus intuitive c'est-à-dire le Maximum – Minimum. Elle dépend uniquement des deux valeurs les plus exceptionnelles. Une erreur de saisie sur un maximum suffit à fausser toute l'analyse. Elle est utile pour le dégrossissage mais dangereuse pour l'analyse fine.

- **Les Quantiles :** Pour analyser la dispersion interne sans être pollué par les extrêmes, on découpe la série en tranches régulières ; en cas d'illustration, nous avons : quartiles 4 tranches de 25%, déciles 10 tranches. L'écart interquartile concentre les 50% centraux de la population, offrant une mesure de dispersion.

- **La Boîte à Moustache (Box-plot) :** Inventée par Tukey, c'est la synthèse graphique ultime. Elle combine la médiane, la dispersion centrale, étendue globale et détection des valeurs aberrantes. Elle permet de comparer visuellement la structure de plusieurs distributions en un clin d'œil.

En ce qui concerne la manipulation sur python, nous avons obtenu ces résultats :





Nos résultats nous ont généré plusieurs diagrammes, nous avons pris en cas d'illustration ces deux diagrammes pour décris nos observation. Ces boîtes à moustaches montrent comment sont répartit le nombre de bulletins blancs dans les différents départements. Pour le faire, les données ont été mises dans un tableau avec Python (Pandas), puis on a choisi la colonne Blancs et on a demandé au programme de dessiner automatiquement une boxplot pour cette variable.

La boîte au centre représente la moitié des départements ; ceux qui ont un nombre de bulletins blancs. La ligne au milieu de la boîte est la médiane : cela veut dire qu'environ la moitié des départements ont moins d'environ 4 000 votes blancs, et l'autre moitié en a plus. Le point tout en haut montre un ou quelques départements où le nombre de bulletins blancs est beaucoup plus élevé que dans les autres, ce qui indique la présence de valeurs extrêmes, souvent liées aux départements très peuplés.

6. Les Paramètres de Forme

Il est crucial de vérifier la symétrie a travers le coefficient de Skewness ; la distribution penche-t-elle vers la gauche ou la droite ? la distribution est-elle pointue ou plate? Ces indicateurs ne sont pas anecdotiques. Ils servent de juge pour savoir si la distribution observée ressemble à la fameuse Loi Normale, condition sine qua non pour appliquer la plupart des tests statistiques différentielles.

La Séance 3 nous a fourni le vocabulaire grammatical de la statistique. Nous savons désormais qualifier une variable, mesurer ses inégalités internes et évaluer sa dispersion. Ces paramètres tels que la moyenne, la médiane, l'écart-type, l'indice de Gini ne sont pas des finalités en soi, mais des descripteurs qui transforment des données brutes en une image intelligible, préparant le terrain pour l'étape suivante qui est la modélisation.

SÉANCE 4 : LES DISTRIBUTIONS STATISTIQUES

Une distribution statistique est un modèle, une idéalisation de la réalité qui permet de faire des prévisions. Ce chapitre est crucial car il fait le pont entre la statistique descriptive et la statistique inférentielle.

1. **Quels critères mettriez-vous en avant pour choisir entre une distribution statistique avec des variables discrètes et une distribution statistique avec des variables continues?**

La première étape de toute modélisation est un choix binaire ; Ce choix repose sur des critères stricts qui renvoient à la nature même de la variable étudiée.

- **Critère ontologique :** Le critère principal est la nature de la variable. Si le phénomène étudié provient d'un comptage d'événements distincts, cela montre que nous sommes dans le domaine discret. Par exemple, le nombre de séismes dans une année, le nombre de clients dans une file d'attente, ou le nombre d'espèces végétales sur une parcelle. Ici, la variable ne peut prendre que des valeurs entières soit 0, 1, 2.... La probabilité est ponctuelle ; on peut calculer la probabilité exacte qu'il y ait 3 séismes. Les lois associées sont la loi de Bernoulli, la loi Binomiale ou la loi de Poisson.
- **Critère métrologique :** À l'inverse, si le phénomène procède d'une mesure sur une échelle fluide, nous sommes dans le domaine continu tels que le temps, l'espace, la température, les débits fluviaux. Ici, la probabilité qu'une valeur soit exactement égale à un chiffre précis est mathématiquement nulle. Les lois associées sont la Loi Normale, la loi Exponentielle ou la loi Log-Normale.
- **La zone grise et l'approximation :** Il existe cependant un critère pragmatique lié à la taille de l'échantillon et à la richesse des valeurs. Parfois, une variable est théoriquement discrète comme la population d'une ville, on ne peut pas avoir 1,5 habitant, mais elle prend un tel nombre de valeurs différentes de 100 à 10 millions qu'il devient plus commode de la traiter comme une variable continue. Les théorèmes de convergence permettent d'approximer une loi discrète comme binomiale par une loi continue lorsque l'effectif est grand. C'est un choix de simplification mathématique.

2. **Expliquez selon vous quelles sont les lois les plus utilisées en géographie?**

La géographie, en tant que science de l'espace et des sociétés, mobilise un spectre spécifique de lois de probabilité.

- **La Loi Normale ou Loi de Gauss-Laplace :** Grâce au Théorème Central Limite, tout phénomène résultant de l'addition d'un grand nombre de petites causes indépendantes et faibles tend vers une loi Normale. En géographie physique précisément en climatologie, hydrologie et pédologie, elle est omniprésente car les phénomènes naturels résultent de multiples facteurs aléatoires qui se compensent. Elle modélise un monde où les extrêmes sont rares et où la majorité des individus se regroupent autour de la moyenne. Elle sert de référence absolue pour la détection d'anomalies et pour la construction des intervalles de confiance.

- **La Loi de Poisson ou loi de la rareté spatiale** : elle gère les événements rares et les processus de comptage. En géographie, elle est fondamentale pour l'analyse spatiale des semis de points. Elle sert de modèle nul pour tester les répartitions spatiales. Si l'on observe la localisation de commerces, d'incendies ou de cas de maladies, on compare la réalité à une distribution de Poisson. Si la réalité correspond à un Poisson, la répartition est aléatoire. Si elle s'en écarte, cela révèle une structure spatiale. C'est la loi de base des SIG pour les analyses de densité.

- **Les Lois de Pareto et de Zipf ou encore les Lois Puissance :** C'est ici que la géographie humaine se distingue. Contrairement au monde physique souvent centré sur la moyenne. La loi de Zipf, décrit des distributions extrêmement hiérarchisées. Si l'on classe les villes d'un pays par taille, on observe souvent que la deuxième ville est deux fois plus petite que la première, la troisième trois fois plus petite. Il y a très peu d'événements énormes les métropoles géantes et une multitude d'événements petits. Contrairement à la loi Normale qui écrase les écarts, ces lois modélisent des systèmes où les inégalités sont structurelles et invariance d'échelle. Elles sont indispensables pour comprendre les réseaux urbains et les dynamiques économiques.

- **La Loi Log-Normale :** Proche de la loi Normale, elle s'applique aux variables qui ne peuvent pas être négatives et qui résultent de processus multiplicatifs et non additifs. En géographie, elle décrit souvent la distribution des revenus, la taille des exploitations agricoles ou la taille des villes dans certains systèmes. Elle est asymétrique, et s’étale toujours vers la droite.

- **Les Lois de Valeurs Extrêmes Gumbel et Fréchet :** Elle s'applique particulièrement dans le domaine de la géographie du risque**.** La géographie des risques ne peut se contenter de la moyenne. La loi Normale sous-estime gravement la probabilité de ces catastrophes. Les lois de valeurs extrêmes comme Gumbel sont spécifiquement conçues pour modéliser les queues de distribution, c'est-à-dire les événements très rares mais aux conséquences maximales. Elles permettent de dimensionner les digues ou les normes parasismiques.

Parvenu au terme de cette séance qui nous a fait basculer de l'observation à la théorisation, nous retenons que choisir une loi de distribution, ce n'est pas seulement faire un ajustement de courbe sur un ordinateur ; c'est faire une hypothèse forte sur la nature du processus qui a engendré les données. Dire cela suit une loi Normale, c'est dire ce phénomène est le fruit du hasard additif et stable. Dire cela suit une loi de Zipf, c'est dire ce système est hiérarchique et inégalitaire. Les distributions statistiques sont les voies théoriques à travers lesquelles le géographe déchiffre l'ordre caché du territoire.

SÉANCE 5 : LES STATISTIQUES INFÉRENTIELLES

L'inférence est l'art de l'induction. C'est la méthode scientifique qui permet de généraliser des résultats obtenus sur une petite partie qui est l'échantillon à l'ensemble de la population. En géographie, où les populations sont souvent nombreuses, il est impossible de tout mesurer. Nous devons donc apprendre à travailler avec l'information partielle, à gérer l'incertitude et à quantifier le risque d'erreur. Ce chapitre détaille les mécanismes de l'échantillonnage, de l'estimation et de la décision par les tests.

1. Comment définir l'échantillonnage. Pourquoi ne pas utiliser la population en entier? Quelles sont les méthodes d'échantillonnage? Comment les choisir?

La première question fondamentale est celle de la légitimité de l'échantillon. Pourquoi ne pas toujours travailler sur la population entière, ce que l'on appelle un recensement ? La réponse est pragmatique. Un recensement exhaustif est une opération financièrement exorbitante et temporellement longue. De plus, pour certaines populations comme l'ensemble des poissons d'un océan ou l'ensemble des pièces produites par une usine à l'infini, l'exhaustivité est techniquement impossible. L'échantillonnage est donc un compromis nécessaire : on accepte de perdre un peu de précision pour gagner en faisabilité. Mais pour que ce pari fonctionne, l'échantillon doit être une image fidèle, une maquette de la réalité. C'est la notion de représentativité. Pour l'atteindre, il existe deux grandes familles de méthodes.

La première méthode est : l'**échantillonnage probabiliste ou aléatoire** : Elle repose sur le hasard pur. Si chaque individu de la population a une chance connue et non nulle d'être tiré au sort, alors les lois mathématiques garantissent que, sur un grand nombre, les biais s'annulent. C'est la seule méthode qui permet de calculer scientifiquement une marge d'erreur.

La seconde méthode est empirique : Ici, on force l'échantillon à ressembler à la population sur des critères connus comme le sexe, l'âge, la profession. On construit artificiellement un mini-modèle de la société. C'est efficace, mais mathématiquement moins rigoureux car on ne maîtrise pas les biais cachés.

2. Comment définir un estimateur et une estimation?

L'**estimateur** est un objet mathématique abstrait. C'est une fonction, une formule, une variable aléatoire. En tant que variable aléatoire, l'estimateur a des propriétés théoriques : il a une espérance, une variance, une loi de probabilité. Il existe avant même que l'on ait récolté la moindre donnée.

L'estimation, quant à elle, est la concrétisation de cet outil. C'est la valeur numérique précise obtenue après avoir appliqué la formule aux données d'un échantillon spécifique. Si je calcule la moyenne des âges de 100 personnes et que je trouve 42 ans, ce chiffre est une estimation. L'estimateur est la balance ou encore l'instrument de mesure ; l'estimation est le poids affiché. Tout l'enjeu de la statistique est de construire de des estimateurs fiables pour obtenir des poids justes.

3. Comment distinguer l'intervalle de fluctuation et l'intervalle de confiance?

L'intervalle de fluctuation relève d'une logique déductive c'est-à-dire du général au particulier. C'est une prédiction théorique. Par exemple, si je sais qu'une pièce est équilibrée, je peux prédire que sur 100 lancers, j'aurai 95% de chances d'obtenir entre 40 et 60 piles.

L'**intervalle de confiance** relève d'une logique inductive du particulier au général. C'est la situation réelle de l'enquêteur. Je ne connais pas la vérité, je n'ai que mon échantillon. Je me demande alors : Au vu de mon résultat, où se trouve vraisemblablement la vraie valeur inconnue de la population ?. Je construis une fourchette autour de mon estimation. Je ne peux jamais être sûr à 100%, donc j'associe à cet intervalle un niveau de confiance généralement 95%. Cela signifie que si je répétais mon enquête 100 fois, la vraie valeur serait capturée par mon intervalle 95 fois. Elle ne donne jamais de valeur exacte, mais une plage de vraisemblance.

4. Qu'est-ce qu'un biais dans la théorie de l'estimation?

Comment savoir si notre balance est bonne ? On juge un estimateur sur deux critères principaux.

- Le premier est **l'absence de Biais** : Un biais est une erreur systématique. Un estimateur est biaisé si, en moyenne, il vise à côté de la cible. Par exemple, si je veux estimer la taille moyenne des Français mais que je ne mesure que des joueurs de basket, mon estimateur est biaisé. Mathématiquement, un estimateur est sans biais si son espérance mathématique est égale à la vraie valeur du paramètre recherché.
- Le second critère est **la Convergence** : C'est la capacité de l'estimateur à devenir de plus en plus précis à mesure que l'on accumule des données. Si j'interroge 10 personnes, mon estimation sera floue. Si j'en interroge 10 millions, mon estimation doit être quasi-parfaite. Si un estimateur ne converge pas vers la vérité avec l'augmentation de la taille de l'échantillon, il est inutile. L'estimateur idéal est donc sans biais et convergent, souvent qualifié de variance minimale.

5. **Comment appelle-t-on une statistique travaillant sur la population totale? Faites le lien avec la notion de données massives1?**

Avec l'avènement du Big Data, on entend parfois que l'échantillonnage serait obsolète car on disposerait désormais N qui est egale la population totale. C'est une illusion dangereuse. D'abord, travailler sur la population totale relève de la statistique descriptive, pas inférentielle, puisqu'il n'y a plus d'incertitude d'échantillonnage. Mais surtout, les données massives sont souvent des données d'échantillons géants non contrôlés. Les données Twitter ou GPS ne sont pas la population totale, ce sont des échantillons biaisés. Le Big Data ne supprime pas le besoin de théorie de l'estimation ; au contraire, il rend la traque des biais encore plus cruciale, car avec des millions de données, un petit biais initial devient une erreur grave à l'arrivée.

6. **Quelles sont les méthodes d'estimation d'un paramètre? Comment en sélectionner une?**

La méthode la plus utilisée est celle du **Maximum de Vraisemblance**. Elle consiste à se poser la question suivante : Quelle est la valeur du paramètre inconnu qui rendrait mon échantillon observé le plus probable ?. C'est une méthode puissante qui fournit souvent des estimateurs efficaces.

Une autre méthode courante, notamment en régression, est la méthode des **Moindres Carrés**, qui cherche à minimiser la somme des erreurs au carré entre le modèle et la réalité.

7. **Quels sont les tests statistiques existants? À quoi servent-ils? Comment créer un test?**

Pour valider une découverte scientifique, on utilise des tests d'hypothèse. Le mécanisme est toujours le même et s'apparente à un procès judiciaire. C'est ce que l'on veut souvent réfuter. En face, on pose l'hypothèse alternative ; on calcule ensuite la probabilité d'observer nos données si elle était vraie. C'est la fameuse p-value. Si cette probabilité est inférieure à un seuil conventionnel de 5%, on considère que ce n'est pas raisonnable d'attribuer cela au hasard. On rejette et on déclare que le résultat significatif. Il existe une zoologie des tests :

- **Les tests de conformité** qui consiste à comparer une moyenne à une norme,
- **Tests d'homogénéité** qui permet de comparer deux populations entre elles,
- **Tests d'indépendance** qui renvoie au Chi-2.

- **Les tests paramétriques** comme le test de Student, il est puissants mais exigeants ils requièrent la normalité des données.
- **Les tests non-paramétriques** comme Mann-Whitney, plus robustes et applicables à des données qualitatives ou non-normales.

8. Que pensez-vous des critiques de la statistique inférentielle?

Elle repose sur des conventions arbitraires. Elle peut confondre la significativité statistique et la significativité pratique. Avec un échantillon assez grand, tout devient statistiquement significatif, même des écarts ridicules. De plus, l'inférence ne prouve jamais la causalité ; elle ne fait que rejeter l'hypothèse du hasard. Le géographe doit donc toujours interpréter les résultats statistiques à la lumière de sa connaissance du terrain.

En définitive, Nous avons compris que la vérité statistique n'est pas absolue, mais probabiliste. L'échantillonnage, l'estimation et les tests sont des outils rigoureux pour gérer le risque d'erreur inhérent à toute généralisation. Ils permettent de transformer une observation locale en une loi générale, à condition de respecter scrupuleusement les conditions de validité (représentativité, absence de biais, normalité).

SÉANCE 6 : LA STATISTIQUE D'ORDRE DES VARIABLES QUALITATIVES

En géographie, comme dans beaucoup de sciences sociales, l'information pertinente ne réside pas toujours dans la valeur absolue, mais dans la position relative. Savoir que New York a 8,4 millions d'habitants est une information ; savoir qu'elle est la première ville des Etats Unis est une information d'une autre nature, structurelle. C'est le domaine de la statistique ordinale, la statistique des rangs et des classements. Cette séance explique comment matérialiser l'ordre, comment comparer des hiérarchies et comment traiter des données qualitatives qui résistent à l'arithmétique classique.

1. Qu'est-ce qu'une statistique ordinale? À quelle autre statistique catégorielle s'oppose-t-elle? Quel type de variables utilise-t-elle? En quoi cela peut matérialiser une hiérarchie spatiale?

La statistique cardinale s'intéresse à l'intensité de la mesure. Elle suppose que l'écart entre 10 et 11 est le même qu'entre 100 et 101. C'est une échelle de rapport.

La **statistique ordinale**, elle, s'intéresse au rang. Elle transforme les données brutes en une liste ordonnée. Ici, l'écart entre le premier et le deuxième n'a pas besoin d'être égal à l'écart entre le deuxième et le troisième. Dans une course, le premier peut arriver 1 heure avant le second, et le second 1 seconde avant le troisième ; le classement reste 1, 2, 3.

Cette approche est cruciale en géographie pour l'analyse des hiérarchies. Les systèmes territoriaux sont souvent structurés par des relations de domination ou de primauté. La statistique ordinale permet aussi de traiter des données subjectives issues d'enquêtes à travers les échelles de satisfaction telles que : « très satisfait », « satisfait », « pas satisfait » en les transformant en objets mathématiques rigoureux sans leur imposer une métrique artificielle.

2. Quel ordre est à privilégier dans les classifications?

La géographie utilise très souvent la convention inverse pour ses classements hiérarchiques, notamment dans le cadre de la loi de Zipf. Dans ce contexte, premier rang est attribué à l'entité la plus grande, le deuxième rang à la suivante. L'ordre est donc décroissant par rapport à la valeur de la variable qui est la population, mais croissant par rapport au rang. Cette inversion est importante à garder en tête lors de l'interprétation des corrélations : plus le rang augmente, plus la taille diminue. Le géographe doit toujours expliciter le sens de son classement.

3. Quelle est la différence entre une corrélation des rangs et une concordance de classements?

La **Corrélation des Rangs** transpose la logique métrique aux rangs. Elle calcule la différence entre le rang d'un individu dans la première liste et son rang dans la seconde liste. Elle cherche à voir si, globalement, les rangs sont proches. C'est une vision géométrique de la proximité.

La **Concordance des Classements** adopte une logique purement combinatoire et probabiliste. Elle ne regarde pas la distance entre les rangs, mais l'ordre relatif. Elle prend toutes les paires d'individus possibles et pose la question : « Est-ce que A est classé avant B dans les deux listes ? » Si oui, la paire est concordante. Si l'ordre s'inverse, la paire est discordante. C'est une vision structurelle de la cohérence.

4. Quelle est la différence entre les tests de Spearman et de Kendal?

Le Rho de Spearman est historiquement le premier. Il se calcule comme un coefficient de corrélation classique sur les rangs. Il est simple à calculer et donne beaucoup de poids aux gros écarts de classement. Il est très utilisé, mais il traîne avec lui l'héritage de la statistique paramétrique.

Le Tau de Kendall est conceptuellement supérieur pour des données ordinales pures. Il se calcule en faisant le bilan des paires concordantes moins les paires discordantes, divisé par le nombre total de paires. Son interprétation est très intuitive en termes de probabilités ; c'est la probabilité que deux individus pris au hasard soient classés dans le même ordre, moins la probabilité qu'ils soient inversés. Le Tau de Kendall est souvent plus faible en valeur absolue que le Rho de Spearman, mais il est plus robuste et gère mieux les ex-aequo.

5. À quoi servent les coefficients de Goodman-Kursdal et de Yule?

Le Gamma de Goodman-Kruskal généralisent la logique de concordance de Kendall. Ils ignorent les ex-aequo pour se concentrer sur la capacité prédictive de l'ordre strict.

Le Q de Yule est un cas particulier. Spécifiquement conçu pour les tableaux de deux variables binaires. Il mesure l'intensité de la liaison entre deux attributs dichotomiques par exemple : Nord/Sud et Riche/Pauvre). Ces coefficients permettent de quantifier des relations qualitatives sans avoir besoin de passer par des régressions linéaires complexes et souvent inadaptées à la nature qualitative des données.

En définitif nous retenons que la statistique ordinale offre au géographe des outils puissants pour analyser les structures, les hiérarchies et les relations d'ordre qui organisent l'espace. Savoir qu'une ville est "première" ou "seconde", savoir que deux phénomènes varient dans le "même sens" (concordance), c'est souvent comprendre l'essentiel de la dynamique territoriale. En maîtrisant ces outils de rangs, l'analyste complète sa palette et devient capable de traiter toute la diversité de l'information géographique, du qualitatif subjectif au quantitatif le plus dur.

CONCLUSION GÉNÉRALE DU RAPPORT

Ce travail de synthèse approfondie, mené à travers les cinq séances fondamentales du cours, met en lumière la cohérence intellectuelle de l'enseignement statistique en géographie. Nous sommes partis des définitions épistémologiques du hasard et de la donnée (Séance 2), nous avons appris à décrire et résumer l'information (Séance 3), à modéliser les structures par des lois théoriques (Séance 4), à généraliser nos observations par l'inférence (Séance 5), pour finir par l'analyse des structures hiérarchiques (Séance 6).

Ce parcours démontre que la statistique n'est pas une fin en soi, mais un langage. C'est le langage qui permet au géographe de dialoguer avec la complexité du monde, de transformer des observations éparses en connaissances validées, et de dépasser l'intuition pour atteindre la démonstration. L'étudiant en géographie sort de cet enseignement non pas transformé en mathématicien, mais armé d'un esprit critique aiguisé face aux chiffres, conscient des limites de ses outils (biais, intervalles, risques d'erreur) et capable de choisir la méthode adaptée pour faire parler les territoires.

DIFFICULTES RENCONTREES

Les différentes séances du cours se sont révélées difficile à appréhender dès le départ, que soit sur le plan conceptuel que pratique. Les séances de manipulation ont elles aussi été complexes, nécessitant une adaptation progressive à une nouvelle manière de raisonner et de travailler avec les données. J'ai toutefois pu m'appuyer sur l'entraide de plusieurs camarades de promotion, dont les échanges ont joué un rôle essentiel dans la compréhension des exercices et des méthodes employées. De plus, En ce qui concerne le déroulement des cours en présentiel, l'idée d'avoir mis les questions par séance de cours sont intéressante, mais elle ne correspond pas toujours à la façon dont mes difficultés apparaissent. Pour la plupart du temps, je comprenais vraiment où j'étais bloquer après la séance, en retravaillant seule, je n'arrivais pas toujours à formuler le problème ou à identifier d'où vient l'erreur. Du coup, j'ai parfois eu l'impression de ne pas profiter pleinement de ces temps d'échange, non pas par manque d'intérêt, mais parce que les questions me viennent trop tard. Par ailleurs, je n'ai pas encore le réflexe d'utiliser Discord. Malgré cela, je retiens quelque chose d'important : une première initiation à Python, une vision plus concrète de ce que signifie analyser des données, et l'envie d'intégrer cet outil dans mon parcours pour la suite.

Par ailleurs, j'ai également eu recours à des outils d'assistance tels que ChatGPT afin de faire certaines manipulations, de mieux comprendre la logique du code et de surmonter des blocages ponctuels. Venant d'un système de formation différent, l'acquisition des compétences en Python et en analyse de données n'a pas été immédiate et a demandé un effort d'adaptation important. Néanmoins, cette démarche progressive, fondée sur l'apprentissage, l'entraide et la recherche autonome de solutions, m'a permis de consolider peu à peu mes acquis.