QUESTIONS DE COURS PYTHON : Matthieu Maimbourg

Séance 2 :

1/ quel est le positionnement de la géographie par rapport aux statistiques ?

La relation entre la géographie et les statistiques est complexe, floue et même parfois conflictuelle. Dans l'Histoire, une partie des géographe sous-estimait le potentiel des outils statistiques, pensant qu'ils étaient trop mathématiques ou même carrément en dehors de la discipline.

Pourtant, la géographie produit et utilise des données massives, souvent impossibles à analyser sans statistiques. Les statistiques permettent de traiter, d'organiser et d'analyser cette information géographique massive, et est donc un outil indispensable pour la géographie d'aujourd'hui.

Alors certes, elles ne substituent pas au raisonnement géographique mais sont complémentaires. Elle offrent des méthodes rigoureuses qui permettent de mettre en place des automatisations, dégager des tendances, comparer des territoires et structurer l'analyse spatiale.
Ainsi, les statistiques permettent de renforcer le caractère scientifique de la géographie.

2/ Le hasard existe-t-il en géographie ?

Cette question est complexe, il faut mêler d'autres disciplines. On peut souvent penser que la question de l'existence du hasard en géographie relève de la philosophie mais à la fois aussi de la méthode scientifique.

Si on adoptait une vision plutôt déterministe, le hasard n'existe pas réellement, mais correspondrait à des causes encore inconnues.

A l'inverse on pourrait prendre un approche qui admettrait l'existence du hasard comme composante des phénomènes observés.

En géographie, il est admis que les comportements individuels ou les événements locaux ne sont pas entièrement prévisibles. Toutefois, à une échelle plus large, il est possible de dégager des régularités et des tendances statistiques. Le hasard n'empêche donc pas l'analyse scientifique (et c'est là que les statistiques entrent en jeu dans la géographie et dans d'autres disciplines) : il limite la prévision fine, mais autorise l'étude des probabilités et des structures globales des phénomènes spatiaux.

3/ Quels sont les types d'information géographique ?

L'information géographique se compose principalement de deux types complémentaires.

D'une part nous avons l'information attributaire, qui concerne les caractéristiques associées aux objets ou aux territoires étudiés, qu'elles relèvent de la géographie humaine (*population, emploi, revenus*) ou de la géographie physique (*température, précipitations, altitude*).

D'autre part, nous avons, l'information géométrique, qui renvoi à la forme, à la localisation et aux dimensions des objets géographiques, comme les surfaces, les distances ou les réseaux.

Dans les systèmes d'information géographiques (les SIG comme ArcGis ou QGIS), l'information attributaire constitue la base des analyses statistiques, tandis que l'information géométrique permet la représentation spatiale.

4/ Quels sont les besoins de la géographie au niveau de l'analyse de données ?

La géographie a fondamentalement besoin de l'analyse de données. Elle traite avec énormément de masses de données, elle en a donc besoin pour comprendre et interpréter les phénomènes complexes et multiscalaires.

L'analyse statistique permet de résumer des volumes importants de données, de comparer des territoires, de mettre en évidence des structures internes et d'identifier des relations entre plusieurs variables.

Elle offre généralement des outils pour évaluer la fiabilité de l'information et pour confronter les résultats obtenus aux conditions de production de données. De plus, lors

de l'utilisation des SIG, l'analyse de données peut permettre de créer des outils très spécifiques et très utiles à certaines situations.

Ainsi, l'analyse de données constitue une étape essentielle entre la collecte de l'information géographique et son interprétation scientifique.

5/ Quelles sont les différences entre la statistique descriptive et la statistique explicative ?

La statistique descriptive a pour objectif principal de décrire et de résumer les données observées. Elle permet de donner une image simplifiée de la réalité à l'aide d'indicateurs numériques de tableaux et de représentations graphiques.

En revanche, la statistiques explicative cherche à aller au-delà de la description en établissant des relations entre les variables. Elle vise à expliquer une variable dite dépendante à partir d'une ou plusieurs variables explicatives, à l'aide de modèles statistiques. La statistique descriptive constitue donc une étape indispensable à la statistique explicative

6/ Quelles sont les types de visualisation de données en géographie ? Comment choisir celles-ci ?

La visualisation des données en géographie dépend étroitement de la nature des variables étudiées et des objectifs de l'analyse. Les variables qualitatives sont généralement représentées à l'aide de diagrammes en secteurs ou de tableaux de fréquences, tandis que les variables quantitatives discrètes sont visualisées par des diagrammes en bâtons.

Les variables quantitatives continues sont plutôt représentées par des histogrammes, des courbes cumulées ou des boîtes à moustaches. Pour les analyses multivariées, des graphiques factoriels ou des classifications sont utilisées.
Le choix des visualisations doit donc être guidé par le type de données, l'échelle d'analyse et la question géographique posée.

7/ Quelles sont les méthodes d'analyse de données possibles ?

Les méthodes d'analyse de données en géographie se répartissent en 3 grandes catégories.

- **Les méthodes descriptives** comme l'analyse en composantes principales ou l'analyse factorielle des correspondances, visent à résumer et à visualiser des données multidimensionnelles.
- **Les méthodes explicatives** cherchent à modéliser les relations entre une variable à expliquer et plusieurs variables explicatives, à travers des régressions ou des analyses de variance.
- **Les méthodes de prévision** s'appliquent principalement aux séries chronologiques et permettent d'anticiper l'évolution future d'un phénomène à partir de son comportement passé.

8/ Comment définiriez-vous : (a) population statistique ? (b) individu statistique ? (c) caractères statistiques ? (d) modalités statistiques ? Quels sont les types de caractères ? Existe-t-il une hiérarchie entre eux ?

A : La population statistique désigne l'ensemble des unités sur lesquelles porte l'étude, par exemple l'ensemble des communes d'un territoire.

B : L'individu statistique correspond à un élément de cette population, comme une commune prise isolément.

C : Les caractères statistiques sont les caractéristiques observées sur les individus, telles que la population, le revenu ou l'altitude.

D : Les modalités statistiques représentent les valeurs ou les catégories prises par un caractère.

Les caractères peuvent être qualitatifs ou quantitatifs, ces derniers étant discrets ou continus. Il existe une hiérarchie entre ces types de caractères, les variables quantitatives continues étant les plus riches en information, tandis que les variables qualitatives nominales sont les plus simples.

Séance 3 :

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

Le caractère qualitatif est le plus général. Toute information statistique commence par une qualification des individus étudiés, c'est-à-dire par l'attribution d'une catégorie ou d'un état. Un caractère quantitatif peut toujours être transformé en caractère qualitatif par un regroupement en classes ou en catégories (ex : classes d'âge), alors que l'inverse n'est pas possible sans perte d'information. Le caractère qualitatif constitue donc la forme la plus large et la plus fondamentale de description statistique, tandis que le caractère quantitatif correspond à un niveau plus précis et plus riche d'information.

2/ Que sont les caractères quantitatifs discrets et caractères quantitatifs continus ? Pourquoi les distinguer ?

Caractères quantitatifs discrets = variables numériques qui ne peuvent prendre qu'un nombre fini ou dénombrable de valeurs distinctes (ex : nb d'enfants).

Caractères quantitatifs continus = variables numériques qui peuvent prendre une infinité de valeurs dans un intervalle donné, comme l'âge, la température ou le revenu.

Cette distinction est essentielle car elle conditionne le choix des outils statistiques, des représentations graphiques et des lois de probabilité. Les variables continues se prêtent notamment à l'utilisation d'intégrales, d'histogrammes et de lois continues, contrairement aux variables discrètes.

3/ Paramètres de position

Pourquoi existe-t-il plusieurs types de moyenne ?

Car toutes les variables quantitatives n'ont pas la même nature ni les mêmes contraintes. La moyenne arithmétique est la plus courante, mais elle peut être inadaptée dans certaines situations (ex : vitesses, rapports ou taux). Dans ces cas, des moyennes spécifiques comme la moyenne harmonique ou la moyenne géométrique sont plus pertinentes.

Pourquoi calculer une médiane ?

La médiane permet de décrire la position centrale d'une distribution sans être influencée par les valeurs extrêmes. Contrairement à la moyenne, elle est robuste face aux valeurs aberrantes et résume efficacement les distributions dissymétriques. Elle partage la population en deux groupes de même effectif, ce qui en fait un indicateur particulièrement utile pour les revenus, les salaires ou les distributions très inégales.

Quand est-il possible de calculer un mode ?

Lorsqu'une modalité ou une valeur apparait plus fréquemment que les autres. Il est applicable aussi bien aux variables qualitatives qu'aux variables quantitatives. Toutefois, il n'existe pas toujours, et une distribution peut être unimodale, bimodale ou plurimodale. Le mode est surtout pertinent pour identifier une valeur dominante dans une distribution.

4/ Paramètres de concentration

Quel est l'intérêt de la médiane et de l'indice de C. Gini ?

La médiane permet de mesurer la concentration d'une variable en tenant compte non seulement des effectifs, mais aussi de la masse totale des valeurs. Elle indique le point à partir duquel 50% de la valeur globale est atteinte.

L'indice de Gini, lui, synthétise le degré d'inégalité d'une distribution en mesurant l'écart par rapport à une situation d'égalité parfaite. Ces deux indicateurs sont essentiels pour analyser les phénomènes d'inégalité, notamment dans l'étude des revenus ou des répartitions spatiales.

5/ Paramètres de dispersion

Pourquoi calculer une variance à la place de l'écart à la moyenne ? Pourquoi la remplacer par l'écart type ?

L'écart simple à la moyenne n'est pas exploitable car les écarts positifs et négatifs se compensent. La variance résout ce problème en utilisant le carré des écarts, ce qui permet de mesurer la dispersion réelle des données. Toutefois, la variance est exprimée dans une unité au carré, ce qui complique son interprétation. L'écart type, qui est la racine carrée de la variance, permet de revenir à l'unité initiale de la variable et facilite la lecture des résultats.

Pourquoi calculer l'étendue ?

L'étendue donne une première indication simple de la dispersion en mesurant l'écart entre la valeur maximale et la valeur minimale. Elle permet d'appréhender rapidement l'amplitude globale d'une distribution, même si elle reste un indicateur fragile car elle ne repose que sur deux valeurs extrêmes.

À quoi sert-il de créer un quantile ? Quel(s) est (sont) le(s) quantile(s) le(s) plus utilisé(s) ?

Les quantiles servent à découper une distribution ordonnée en parts égales d'effectif. Ils permettent de mieux comprendre la structure interne des données. Les quantiles les plus utilisés sont les quartiles, notamment le premier quartile, la médiane (2eme quartile) et le 3eme quartile, ainsi que les déciles et les centiles dans certaines analyses plus fines.

Pourquoi construire une boîte de dispersion ? Comment l'interpréter ?

La boîte de dispersion permet de représenter graphiquement la médiane, les quartiles, l'étendue et les éventuelles valeurs extrêmes. Elle facilite la comparaison entre plusieurs distributions et donne une vision synthétique de la dispersion et de la symétrie des données. Une boite étirée indique une forte dispersion, tandis qu'une médiane décentrée signale une dissymétrie.

6/ Paramètres de forme

Quelles différence faites-vous entre les moments centrés et les moments absolus ? Pourquoi les utiliser ?

Les moments absolus mesurent les caractéristiques globales d'une distribution à partir des valeurs brutes, tandis que les moments centrés mesurent ces caractéristiques par rapport à la moyenne ? Les moments centrés sont particulièrement utiles car ils permettent d'analyser la dispersion, l'asymétrie et l'aplatissement d'une distribution. Ils offrent une description la plus fine de la forme des données.

Pourquoi vérifier la symétrie d'une distribution et comment faire ?

Cela permet de savoir si les données sont équilibrées autour de leur valeur centrale ou si elles présentent une dissymétrie. Cette vérification est importante pour choisir les bons indicateurs statistiques et les bons modèles. Elle peut se faire graphiquement (histogramme, boîte à moustaches) ou numériquement grâce au coefficient d'asymétrie. Une distribution symétrique présente une moyenne, une médiane et un mode proches.

Séance 4 :

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

Pour choisir entre une distribution statistique à variables discrète et une distributions à variables continues, je mettrais en avant surtout la nature du phénomène étudié ainsi que le niveau de précision des données disponibles.

En effet, les variables discrètes sont adaptées lorsque le phénomène observé correspond à un dénombrement d'unités distinctes, comme un nombre d'habitants. Dans ce cas les valeurs possibles sont séparées et finies, ce qui justifie l'utilisation de distributions discrètes.

A l'inverse, les variables continues sont plus pertinentes lorsque le phénomène peut prendre une infinité de valeurs dans un intervalle donné, comme un revenu, une température ou une distance.

Le choix dépend également de l'objectif de l'analyse : une variable continue permet une analyse plus fine et plus précise, mais nécessite parfois une discrétisation en classes pour faciliter la représentation graphique ou l'interprétation.

Enfin, les contraintes méthodologiques et les outils statistiques disponibles influencent ce choix, certaines lois et représentations étant spécifiques aux variables continues ou discrètes.

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

Selon moi, en géographie, les lois (statistiques) les plus utilisées sont celles qui permettent de décrire des phénomènes spatiaux complexes tout en prenant en compte leur variabilité.

La loi normale me paraît centrale, car de nombreux phénomènes géographiques résultent de la combinaison de plusieurs facteurs indépendants, ce qui conduit souvent à des

distributions proches de la normalité, notamment pour des variables physiques ou socio-économiques agrégées.

J'estime également que la loi de Poisson est fréquemment utilisée en géographie pour analyser des événements rares ou des phénomènes de comptage localisées dans l'espace, comme la répartition d'équipements ou certains événements ponctuels.

Par ailleurs, les lois log-normales et plus généralement les distributions asymétriques me semblent particulièrement adaptées à l'étude des revenus, des tailles de villes ou des hiérarchies urbaines, qui présentent souvent de fortes inégalités.

Enfin, en géographie humaine et urbaine, je pense que des régularités empiriques comme les lois de rang-taille sont essentielles pour comprendre l'organisation et la structuration des systèmes territoriaux.

Séance 5 :

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

L'échantillonnage consiste à sélectionner une partie de la population statistique afin d'en tirer des informations sur l'ensemble. L'utilisation de la population entière n'est pas toujours possible en raison de contrainte de coût, de temps, de disponibilité des données ou de taille excessive de la population étudiée.

L'échantillonnage permet ainsi de produire des résultats exploitables tout en limitant ces contraintes. Il existe plusieurs méthodes d'échantillonnage, notamment l'échantillonnage aléatoire simple, l'échantillonnage stratifié, l'échantillonnage systématique et l'échantillonnage par quotas.

Le choix de la méthode dépend de la structure de la population, de l'objectif de l'étude et du niveau de prévision attendu, l'enjeu principal étant la représentativité de l'échantillon.

2/ Comment définir un estimateur et une estimation ?

Un estimateur est une statistiques calculée à partir d'un échantillon et destinée à approcher la valeur inconnue d'un paramètre de la population. Il s'agit d'une variable aléatoire dont la valeur dépend de l'échantillon observé.

L'estimation correspond à la valeur numérique obtenue lorsque l'estimateur est appliqué aux données.

L'estimateur relève donc du cadre théorique, tandis que l'estimation est le résultat concret issu des observations.

3/ Comment distingueriez-vous l'intervalle de fluctuation et l'intervalle de confiance ?

L'intervalle de fluctuation est construit à partir d'une hypothèse supposée vraie sur la population et permet de vérifier si une valeur observée est compatible avec cette hypothèse. Il est principalement utilisé dans le cadre des tests statistiques.

L'intervalle de confiance, lui, est calculé à partir de l'échantillon et sert à encadrer une valeur inconnue du paramètre de la population avec un niveau de confiance donné. L'intervalle de fluctuation est donc lié à une hypothèse préalable, tandis que l'intervalle de confiance vise à estimer un paramètre inconnu.

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

Un biais correspond à un écart systématique entre l'espérance mathématique d'un estimateur et la valeur réelle du paramètre de la population. Un estimateur est dit biaisé lorsqu'il tend à surestimer ou sous-estimer le paramètre de manière régulière. Le biais peut être lié à la méthode d'échantillonnage, à la taille de l'échantillon ou au choix de l'estimateur. Il constitue un problème majeur car il affecte la fiabilité des estimations.

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

Une statistique calculée sur la population totale relève de la statistique descriptive exhaustive. Dans ce cas, il n'existe pas d'erreur d'échantillonnage, puisque l'ensemble des individus est observé. Cette approche fait le lien avec la notion de données massives, ou *big data*, qui permettent parfois d'analyser des volumes de données très importants, proche de l'exhaustivité. Toutefois, même dans ce contexte, des biais ou des problèmes de qualité des données peuvent subsister.

6/ Quels sont les enjeux autour du choix d'un estimateur ?

Le choix d'un estimateur est un enjeu central de la statistique inférentielle, car il conditionne la qualité des résultats obtenus. Un estimateur doit présenter de bonnes propriétés statistiques, notamment une faible variance, un biais limité et une bonne convergence vers le paramètre réel lorsque la taille de l'échantillon augmente. Il doit également être adapté à la nature des données et aux hypothèses formulées sur la population étudiée.

7/ Quelles sont les méthodes d'estimation d'un paramètre ? Comment en sélectionner une ?

Il existe plusieurs méthodes d'estimation d'un paramètre, parmi lesquelles la méthode des moments, méthode du maximum de vraisemblance et les méthodes bayésiennes. Le choix d'une méthode dépend de la loi supposée de la population, de la taille de l'échantillon, des hypothèses retenues et des objectifs de l'analyse ? une méthode est sélectionnée en fonction de sa cohérence théorique et de ses performances statistiques.

8/ Quels sont les tests statistiques existants ? À quoi servent-ils ? Comment créer un test ?

Les tests statistiques servent à prendre une décision à partir de données observées en évaluant la validité d'une hypothèse. Ils sont utilisés pour comparer des paramètres, tester des proportions, analyser des dépendances ou vérifier des distributions. La construction d'un test repose sur la formulation d'une hypothèse nulle, le choix d'une statistique de test, la connaissance de sa loi de probabilité et la fixation d'un seul de décision. Le test permet alors d'accepter ou de rejeter l'hypothèse nulle.

9/ Que pensez-vous des critiques de la statistique inférentielle ?

Les critiques de la statistique inférentielle soulignent principalement les risques liés à une utilisation mécanique des tests et des seuils de significativité, ainsi que la dépendance des résultats aux hypothèses formulées.

Selon moi, ces critiques sont pertinentes lorsqu'elles dénoncent un manque de recul méthodologique. Toutefois, elles ne remettent pas en cause la statistique inférentielle en tant qu'outil scientifique, mais plutôt ses usages abusifs. Utilisée de manière rigoureuse et critique, elle demeure indispensable pour analyser des phénomènes à partir d'échantillons.

Séance 6 :

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

Une statistique ordinale est une statistique appliquée à des variables qualitatives ordonnées, c'est-à-dire des variables pour lesquelles les modalités peuvent être

classées selon un ordre logique ou naturel. Elle s'oppose à la statistique nominale, qui concerne des variables qualitatives sans ordre intrinsèque entre les modalités.

La statistique ordinale utilise donc des variables catégorielles ordonnées, appelées variables ordinales. En géographie, ce type de statistique permet de mettre en évidence des hiérarchies spatiales, par exemple à travers des classements de villes, de territoires ou d'unités spatiales selon leur importance démographique, économique ou fonctionnelle. L'ordination rend ainsi visibles les rapports de la domination, de centralité ou de différenciation au sein de l'espace géographique.

2/ Quel ordre est à privilégier dans les classifications ?

Dans les classification statistiques, l'ordre à privilégier est l'ordre croissant, également appelé ordre naturel. Cet ordre facilite l'interprétation des résultats et permet d'identifier plus aisément les valeurs extrêmes, qu'elles soient trop faibles ou trop élevées. En géographie, cet ordre est généralement utilisé pour analyser des distributions spatiales, et repérer des valeurs aberrantes ou des phénomènes remarquables. Il existe toutefois certaines exceptions, notamment dans le cadre de la loi rang-taille, où l'ordre décroissant peut être privilégié pour des raisons analytiques spécifiques.

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

La corrélation des rangs mesure le degré de relation entre deux variables ordinales issues de deux classements portant sur les mêmes individus ou objets. Elle cherche à déterminer si les positions relatives des individus sont similaires dans les deux classements.

A concordance de classements vise à évaluer le degré d'accord global entre plusieurs classements, éventuellement supérieur à deux.

Alors que la corrélation des rangs se concentre sur la relation entre deux séries ordonnées, la concordance permet de mesurer la cohérence d'ensemble entre plusieurs ordonnancements appliqués aux même objets.

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

Le test de Spearman repose sur un calcul d'un coefficient de corrélation appliqué aux rangs, dérivé du coefficient de corrélation linéaire classique. Il mesure la force et le sens de la relation monotone entre deux classements et est particulièrement adapté lorsque les rangs sont sans ex aequo.

Le test de Kendall, quant à lui se base sur le comptage des paries concordantes et discordantes entre deux classements. Il est souvent considéré comme plus robuste,

notamment pour de petits effectifs ou en présence d'ex aequo, et peut être généralisé à plus de deux classements.

Les deux test sont non paramétriques et permettent de tester l'indépendance entre variables ordinales.

5/ À quoi servent les coefficients de Goodman-Kruskal et de Yule ?

Le coefficient de Goodman-Kruskal sert à mesurer l'intensité de l'association entre deux variables ordinales en comparant le nombre de paires concordantes et de paires discordantes. Il indique le surplus de concordances par rapport aux discordances et varie entre -1 et +1.

Le coefficient Q de Yule est un cas particulier du coefficient de Goodman-Kruskal, appliqué exclusivement aux tableaux de contingence de dimension 2x2. Il mesure l'association entre deux variables dichotomiques et permet d'évaluer la force et le sens de leur relation. Ces coefficients sont particulièrement utiles en géographie pour analyser des relations ordinales ou binaires entre phénomènes spatiaux.