Jeanne Haumesser
M1 GAED
GEOSUDS

RENDU FINAL EN ANALYSE DE DONNÉES
Parcours débutant

Séance 02 :

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

La géographie a besoin de l'outil des statistiques car elle produit des données massives et a besoin de les analyser. Toutefois pendant longtemps les relations entre ces disciplines étaient marquées par la méfiance et la sous-estimation. Or l'analyse spatiale est un courant qui assume pleinement les méthodes statistiques, et les statistiques sont indispensables pour faire de la géographie une "science des échelles".

2. Le hasard existe-il en géographie ?

Dans la philosophie, deux grandes positions philosophiques pensent le hasard : le déterminisme strict de Laplace prétend que le hasard n'existe pas, que tout a une cause, tandis que le hasard peut aussi être vu comme une "cause cachée", qui existe provisoirement. En statistiques, on trouve le hasard bénin, qui n'empêche pas la causalité et répond à la loi de probabilité normale, et le hasard sauvage, qui concerne les événements extrêmes et obéit à la loi de V. Pareto. En géographie humaine, on peut dégager des tendances globales mais il est impossible de prévoir chaque action par chaque acteur.

3. Quels sont les types d'information géographique ?

Il existe deux grands types d'information géographique : le type attributaire, qui concerne les caractéristiques d'un territoire comme sa population, ses revenus ou son climat, et le type géométrique ou morphologique, qui lui concerne la forme, la taille et la géométrie des unités spatiales. Dans un SIG, la base attributaire correspond aux variables statistiques, et les données géométriques représentent les objets spatiaux, comme les points, les lignes et les polygones.

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

La géographie a besoin d'analyse de données pour produire et collecter des données avec des nomenclatures précises (définitions, niveaux hiérarchiques), et des métadonnées (sources, dates, modes de collecte, sondage/exhaustif, fiabilité). Elle en a également besoin pour maîtriser la statistique descriptive, la statistique mathématique ou inférentielle et les différentes méthodes d'analyse de données (factorielles, classifications et régressions).

L'analyse de données apporte à la géographie la capacité à gérer la massification et l'hétérogénéité des données géographiques.

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

La statistique descriptive décrit une ou plusieurs variables, en prenant en compte les paramètres de position, de dispersion, de forme, avec l'usage de tableaux et de graphiques, pour résumer, visualiser er préparer des comparaisons et des prédictions. La statistique explicative met en relation une variable à expliquer et des variables explicatives, pour tester des hypothèses.

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

Les types de visualisation de données en géographie comprennent des graphiques univariés comme des histogrammes, des diagrammes en bâtons, des diagrammes circulaires ou sectoriels, des boîtes à moustache ou des courbes cumulées, mais aussi des graphiques multivariés et la cartographie thématique. Ces types de visualisation doivent être choisis selon le type de variable et selon si on cherche à décrire, comparer, montrer une distribution, une relation, une structure spatiale.

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

Il existe des méthodes descriptives, explicatives et de prévision. Les méthodes descriptives permettent de visualiser les données, les méthodes explicatives relient une variable à expliquer à des variables explicatives, et les méthodes de prévision permettent de modéliser les séries chronologiques.

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 ?

La population statistique est un ensemble d'unités au sens mathématique du terme. L'individu statistique est un élément de cette populations statistique, c'est une unité spatiale qu'on peut cartographier et localiser. Les caractères statistiques sont des propriétés qui sont propres à chaque individu, comme l'âge et le revenu. Les modalités statistiques sont des valeurs possibles prises par un caractère. Enfin, les deux types de caractères sont les caractères qualitatifs et quantitatifs. Les caractères quantitatifs peuvent être discrets et continus. Il existe une hierarchie entre eux.

9. Comment mesurer une amplitude et une densité ?

On mesure l'amplitude en calculant la longueur du segment qui définit une classe, en soustrayant la valeur minimale *a* de la classe à sa valeur maximale *b*.
Pour mesurer la densité, on rapporte l'effectif de la classe *ni* à l'amplitude de cette classe. Le *d* signifie densité :

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

10. À quoi servent les formules de Sturges et de Yule ?

Elles permettent de déterminer le nombre optimal de classes lors de la discrétisation d'une variable quantitative, et d'éviter la perte d'information en permettant un découpage ni trop fin ni trop grossier.

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

L'effectif, ou fréquence absolue, correspond au nombre d'apparition de la variable.
La fréquence relative se calcule en faisant le rapport entre l'effectif et l'effectif total. Alors que la fréquence cumulée s'obtient en faisant la somme des fréquences associées aux valeurs inférieures ou égales à k. On peut aussi la calculer en divisant l'effectif cumulé par l'effectif total.

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

La distribution statistique représente la répartition des valeurs d'un caractère dans la population.

Séance 02 :

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

Le caractère quantitatif est plus général car les paramètres statistiques concernent principalement les variables quantitatives, et seulement ponctuellement les variables qualitatives.

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

Les caractères quantitatifs discrets sont des variables quantitatives dont les modalités sont des valeurs isolées, souvent des nombres entiers, tandis que les caractères quantitatifs continus correspondent à des variables quantitatives pouvant prendre toutes les valeurs d'un intervalle donné. La distinction entre les deux est nécessaire car les formules de calcul des moyennes notamment diffèrent entre ces deux types.

3. Paramètres de position

— Pourquoi existe-t-il plusieurs types de moyenne ?

Il existe plusieurs types de moyenne car on n'utilise pas les mêmes en fonction de la nature de la variable, et du type de relation que l'on souhaite mesurer. Par ailleurs la moyenne arithmétique varie en fonction des valeurs extrêmes ou aberrantes, ce qui nécessite d'utiliser d'autres indicateurs

— Pourquoi calculer une médiane ?

La médiane est la valeur qui partage la série de données ordonnée en deux parties égales, sans être influencée par les valeurs aberrantes, et permet de résumer les distributions fortement dissymétriques.

— Quand est-il possible de calculer un mode ?

Le mode est la modalité correspondant à l'effectif maximal, et correspond à la valeur la plus fréquente ou celle qui a la plus forte densité de probabilité. Il n'existe pas toujours, et n'est pas forcément unique s'il existe.

4. Paramètres de concentration

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

La médiale est la valeur qui permet de partager en deux parties égales la masse de la variable. Ainsi elle permet de mesurer l'importance du caractère total possédé par les individus.

L'indice de C. Gini utilise la courbe de M. O. Lorenz pour décrire et mesurer les effets de la concentration d'une population statistique.

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 à la moyenne permet de calculer la moyenne des écarts simples, qui est nulle ou très proche de zéro car les écarts négatifs et positifs se compensent, d'où l'intérêt de calculer une variance, qui elle utilise le carré des écarts pour que les valeurs ne se compensent pas. Ainsi la variance offre des propriétés mathématiques supérieures à la valeur absolue, ce qui en fait l'indicateur de dispersion par excellence. L'écart type est la racine carrée de la variance et permet de faire revenir la dispersion à la même unité que la moyenne ou l'espérance.

— Pourquoi calculer l'étendue ?

L'étendue est la différence entre la plus grande valeur et la plus petite. Elle est facile à calculer, mais limitée car elle ne dépend que des valeurs extrêmes, et donc peu utile si le nombre de données dépasse 10.

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

Les quantiles partagent la série statistique ordonnée en plusieurs parties égales, et permettent ainsi de mesurer la dispersion de la série. Les plus utilisés sont les quartiles, qui partagent la série en quatres parties égales.

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

Construire une boîte de dispersion permet de représenter schématiquement les principales caractéristiques d'une distribution, notamment avec des quartiles. Une boîte de dispersion s'interprète en comparant visuellement la position, la dispersion et la symétrie de plusieurs séries statistiques.

6. Paramètres de forme

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

Les moments absolus décrivent la distribution par rapport à l'origine en utilisant la valeur absolue de l'écart, tandis que les moments centrés décrivent la distribution par rapport à son centre, c'est-à-dire la moyenne. Ils permettent de caractériser une distribution, notamment pour calculer les coefficients de dissymétrie et d'aplatissement.

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

La symétrie permet de savoir si la distribution est équilibrée, car pour une distribution symétrique, le mode, la moyenne et la médiane sont égaux. Pour cela, il faut utiliser la mesure de la dissymétrie de Person et Fischer, qui permet de déterminer une distribution

symétrique, une distribution positive (étalée sur la droite), ou une dissymétrie négative (étalée sur la gauche.).

Séance 04 :

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 ?

Les critères que je mettrais en avant sont la nature du phénomène étudié, la forme de la distribution empirique, les principales caractéristiques de l'ensemble de données, et le nombre de paramètres des lois

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

Les lois les plus utilisées en géographie sont la loi normale ou loi de gauss, la loi de poisson et la loi de pareto, mais également

Séance 05 :

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 un sous-ensemble, appelé échantillon, au sein d'une population mère (l'ensemble total des individus étudiés). On privilégie cette approche car l'analyse d'une population entière est souvent matériellement impossible, trop coûteuse ou trop lente. Pour que les résultats soient généralisables, l'échantillon doit être représentatif et adapté à l'objet de recherche. On distingue deux grandes familles de méthodes : les méthodes aléatoires (ou probabilistes), où chaque individu est tiré au sort avec une probabilité connue (garantissant l'absence de biais), et les méthodes non aléatoires (ou empiriques). Ces dernières, comme la méthode des quotas ou l'échantillonnage systématique, visent à construire un « modèle réduit » de la population en respectant ses caractéristiques clés (âge, genre, etc.). Le choix dépend de la disponibilité d'une base de sondage et de la précision recherchée.

2. Comment définir un estimateur et une estimation ?

Dans le cadre de l'inférence statistique, l'estimateur est l'outil mathématique : c'est une fonction des observations de l'échantillon (une variable aléatoire) dont la valeur est censée être proche du paramètre réel de la population. L'estimation, quant à elle, est le résultat numérique obtenu ou la démarche globale visant à évaluer les caractéristiques inconnues de la population mère à partir des données collectées sur l'échantillon.

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

L'intervalle de fluctuation repose sur la connaissance préalable de la proportion théorique réelle de la population. Il définit, généralement au seuil de 95%, un intervalle dans lequel la proportion observée est censée se situer lorsque la valeur théorique p est connue ou

supposée. A l'inverse, l'intervalle de confiance correspond à la marge d'erreur associée à une estimation issue d'un échantillon lorsque la valeur de p est inconnue. Il permet ainsi d'encadrer l'estimation obtenue à partir de l'échantillon et d'évaluer l'incertitude liée à l'estimation des caractéristiques de la population mère.

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

Dans la théorie de l'estimation, le biais est l'écart entre l'espérance d'un estimateur et la valeur réelle du paramètre au sein de la population. Quand un estimateur est biaisé, il engendre une erreur systématique et ses valeurs tendent à se concentrer autour de son espérance mathématique plutôt qu'autour de la véritable valeur du paramètre à estimer.

5. Comment appelle-t-on une statistique travaillant sur la population totale ? Faites le lien avec la notion de données massives (*big data*) ?

On appelle une statistique portant sur l'ensemble de la population une enquête exhaustive. Elle diffère du sondage, qui repose sur l'analyse d'un échantillon représentatif de la population mère. De la même manière, les données massives se distinguent par leur volume et leur diversité particulièrement élevés, ce qui requiert des méthodes d'analyse et des infrastructures technologiques spécifiques, comme les centres de données. Dès lors, le traitement d'une enquête exhaustive s'apparente à celui des données massives, car il implique l'analyse d'un volume très important ainsi qu'une grande variété de données.

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

Le choix d'un estimateur repose sur différents critères, comme le biais, l'efficacité et la convergence. Le biais mesure l'écart entre l'espérance de l'estimateur et la valeur réelle du paramètre étudié ; un estimateur est dit sans biais lorsque ces deux valeurs coïncident. L'efficacité d'un estimateur, qui est évaluée par sa variance, est meilleure si celle-ci est faible, car la dispersion des estimations est moindre. Enfin, la convergence d'un estimateur correspond à la tendance de sa distribution à se rapprocher de la valeur réelle du paramètre quand la taille de l'échantillon augmente vers l'infini : cela constitue un indicateur essentiel de sa fiabilité et de sa représentativité.

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

Il existe plusieurs méthodes pour estimer un paramètre. La plus simple, la méthode des moindres carrés, consiste à minimiser la somme des carrés des écarts entre les valeurs observées et les valeurs ajustées, c'est-à-dire les résidus. En revanche, la méthode du maximum de vraisemblance est plus générale et souvent préférable. Elle repose sur l'évaluation de la probabilité des différentes valeurs possibles du paramètre, en se basant sur les observations disponibles. L'objectif de cette méthode est de choisir la valeur du paramètre qui maximise cette vraisemblance, permettant ainsi d'inférer les paramètres de probabilité à partir de l'échantillon étudié.

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

Les tests statistiques servent à valider ou rejeter des hypothèses sur une population à partir d'un échantillon. On les construit en opposant une hypothèse nulle (H0) à une hypothèse alternative (H1). On distingue les tests paramétriques, qui supposent une distribution connue (ex: loi normale), des tests non paramétriques (ou robustes), valables quelle que soit la loi

des données. Selon l'objectif, on utilise des tests d'ajustement (adéquation à un modèle théorique), de comparaison (entre plusieurs groupes), d'indépendance (lien entre variables) ou de signification. La décision repose sur la comparaison d'un score calculé à un seuil critique.

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

Les critiques de la statistique inférentielle montrent que certaines méthodes peuvent conduire à des conclusions hâtives concernant la représentativité de l'échantillon par rapport à la population mère ou le rejet de l'hypothèse nulle. Mais malgré ces limites, leur utilité demeure certaine : la statistique inférentielle permet d'analyser les phénomènes de grande échelle à partir de données d'échantillon, tout en intégrant des outils destinés à encadrer et à maîtriser les marges d'erreur propres à chaque test.

Séance 06 :

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

La statistique ordinale s'appuie sur des variables dont les modalités peuvent être classées selon un ordre naturel (ex: petit, moyen, grand). Elle s'oppose à la statistique nominale, où les catégories ne sont pas hiérarchisées (ex: couleurs, noms de pays). En géographie, l'utilisation de rangs permet de matérialiser une hiérarchie spatiale : on peut ainsi classer des villes par importance démographique ou des régions par niveau de développement, rendant compte des structures de domination ou d'organisation d'un territoire.

2. Quel ordre est à privilégier dans les classifications ?

L'ordre croissant est généralement privilégié dans les classifications statistiques. Cet ordre dit « naturel » facilite la lecture de la série, permet d'identifier plus rapidement les valeurs aberrantes (atypiques) et aide à mettre en évidence les lois de distribution, notamment pour les phénomènes où les valeurs les plus élevées sont rares mais structurantes.

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

La corrélation des rangs évalue le degré de ressemblance entre deux classements spécifiques pour voir s'ils sont liés. La concordance, quant à elle, est une généralisation : elle examine la cohérence globale entre plus de deux classements. Elle permet de vérifier si les individus occupent des positions similaires à travers une multitude de critères ou d'évaluations différentes.

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

Le test de Spearman est une mesure globale de la corrélation basée sur la distance entre les rangs (il s'apparente à un coefficient de Pearson appliqué aux rangs). Il est très utilisé en géographie pour comparer des variables comme la population et la richesse d'un semis de villes. Le test de Kendall (τ) adopte une approche différente en examinant toutes les paires possibles d'individus pour compter celles qui sont concordantes (dans le même ordre) ou discordantes. Kendall est souvent jugé plus robuste pour les petits échantillons ou les données comportant beaucoup d'ex-æquo.

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

Ces coefficients mesurent l'association entre variables ordinales. Le Γ de Goodman-Kruskal calcule le surplus de paires concordantes sur les discordantes pour évaluer la force du lien entre deux classements (de -1 à +1). Le Q de Yule est une variante spécifique appliquée aux tableaux de contingence 2×2 (quatre cases) ; il permet de quantifier l'intensité de l'association entre deux caractères qualitatifs ordonnés, variant également entre une opposition totale et une association parfaite.