Madeleine Duvaux
M1 GAED parcours ACA
Parcours débutant

Rapport d'activité

Séance 2

A. Questions

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

Le positionnement est ambivalent, se voulant à la fois une discipline scientifique avec production de données mais ne s'en donnant pas les moyens en écartant les disciplines telles que les mathématiques des parcours d'enseignement de la géographie. Cependant, les statistiques sont aujourd'hui nécessaires et reconnues comme telles pour la discipline, mais ne tendent pas particulièrement à se développer au sein de l'enseignement.

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

Oui, en particulier le hasard de la contingence mais cela ne fait pas consensus parmi les géographes.

3. Quels sont les types d'information géographique ?

Les éléments de géographie humaine et les éléments de géographie physique.

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

Une connaissance mathématique et méthodologique de la production de données.
Les attributs (information thématique) et la géométrie (information spatiale)

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

La statistique descriptive se distingue de l'explicative dans son principe même, l'une étudiant les données tandis que l'autre les décrit. La statistique descriptive semble moins faire appel à des notions de mathématiques que la statistique explicative bien qu'elle puisse conduire à un approfondissement de l'interprétation des données grâce aux statistiques mathématiques. La statistique explicative prépare les données à l'étude mathématique et inférentielle en dégageant des lois de probabilité, des relations entre plusieurs informations ou encore en produisant des visualisations graphiques ou tableaux de synthèse. Cette dernière semble plus complète que la statistique descriptive qui va produire une image simplifiée de la réalité.

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

- Graphiques uni-variés
- Cartes thématiques
- Analyse factorielle

- Diagrammes multidimensionnels
- Tableaux synthétiques

Pour choisir la visualisation adaptée cela dépend de la nature de la variable, de l'objectif de l'analyse, de l'échelle spatiale et de la complexité des données

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

- Les méthodes descriptives
- Les méthodes explicatives
- Les méthodes de prévision

8. Comment définiriez-vous
a. population statistique = données pouvant être regroupées dans une même catégorie
b. individu statistique = isolation de données issues de la population statistique
c. caractères statistiques = caractéristiques de l'individu pris parmi la population statistique sur laquelle l'analyse statistique porte
d. modalités statistique = Catégorie d'un attribut indispensable pour organiser, résumer et analyser les variables qualitatives, lorsque l'on connaît ses modalités individu par individu, et devient statistiques lorsqu'elle fait l'objet d'une étude statistique

Quels sont les types de caractères ? Existe-t-il une hiérarchie entre eux ?

- variable qualitative (nominales, ordinales)
- variable quantitative (discrètes, continues)

9. Comment mesurer une amplitude et une densité ?

Pour mesurer une amplitude : longueur $b - a$ avec a la valeur minimale de la classe et b la valeur maximale

Pour mesurer une densité : le rapport entre l'effectif n_i et l'amplitude de la classe décrivant une modalité i. On appelle d la densité : $d = ni / (b - a)$

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

Il s'agit de règles pour déterminer les nombre de classes. La formule de Sturges donne un valeur approximative du nombre de classes, la formule de Yule étant une alternative. Ces équations permettent de déterminer le nombre optimal de classes avant de tracer un histogramme.

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

Un effectif n_i associé à une valeur x_i de la variable aléatoire X correspond au nombre d'apparitions de cette variable dans la population.

Pour calculer une fréquence : $f_i = n_i / n$

Pour calculer une fréquence cumulée : $\sum_{i=1}^{k} f_i = 1$

Une distribution statistique décrit la façon dont les valeurs d'une variable sont réparties dans une population ou un échantillon.

B. Résultat Code

Le corps électoral se caractérise par une mobilisation importante, bien que marquée par un nombre significatif de votes non exprimés :

- Inscrits et votants : Sur un total de 48 747 876 inscrits, 35 923 707 citoyens se sont rendus aux urnes, soit un taux de participation de 73,69 %.
- Abstention : Le taux d'abstention s'élève à 26,31 %, représentant plus de 12,8 millions d'électeurs.
- Votes blancs et nuls : Les votes blancs représentent 1,11 % des inscrits (543 609 voix) et les votes nuls 0,47 % (230 483 voix).
- Individu et Population : Ici, l'individu statistique est l'électeur, et la population est l'ensemble des inscrits.
- Caractère qualitatif nominal : Le "choix du candidat" est une variable qualitative nominale. La visualisation sous forme de tableau de synthèse permet d'identifier immédiatement le mode de la distribution (la valeur dominante), qui est ici Emmanuel Macron.
- Disparité de concentration : On observe une forte concentration des voix sur peu d'individus (les candidats de tête), ce qui pourrait être mesuré par un indice de concentration comme celui de Gini pour évaluer l'inégalité de la répartition des suffrages entre les candidats.

Séance 3

A. Questions

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

Le caractère quantitatif est le plus général puisque les paramètres statistiques concernent principalement les variables quantitatives et seulement ponctuellement les variables qualitatives.
En effet, le caractère quantitatif permet de calculer tous les paramètres (moyenne, variance, moments…), il se prête aussi à l'ensemble des opérations statistiques vues (dispersion, forme, position). Les caractères qualitatifs ne permettent quant à eux que des descripteurs limités.

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

Les caractères quantitatifs discrets correspondent aux valeurs dénombrables, séparées (comme le nombre d'enfants), leur moyenne se calcule par une somme

Les caractères quantitatif continus correspondent eux aux valeurs sur un intervalle (longueur, revenu)

On les distingue d'abord car les les formules ne sont pas les mêmes (somme vs intégrale), aussi parce que les représentations changent (histogrammes continus, classes) et enfin parce que certaines mesures comme les médianes ou les quantiles se calculent différemment pour les deux types.

3. Pourquoi existe-t-il plusieurs types de moyenne?

Le tableau du cours montre plusieurs moyennes (arithmétique, géométrique, quadratique, harmonique, mobile…) Il en existe plusieurs types afin qu'elles répondent aux situations différentes : la nature de la variable (continue ou discrète), les propriétés voulues et les contextes d'usage (vitesse —> harmonique, produits —> géométrique)

Pourquoi calculer une médiane ?

On calcule une médiane car contrairement à la moyenne elle n'est pas influencée par les valeurs extrêmes, elle convient aussi à des séries très dissymétriques, engin elle résumer la position centrale même quand la moyenne est trompeuse. On la calcule donc pour obtenir une mesure robuste et insensible aux valeurs aberrantes.

Quand est-il possible de calculer un mode?

On calcule un mode uniquement lorsque la distribution présente une valeur dominante identifiable. En effet, le mode existe lorsqu'une modalité a l'effectif maximal ou la plus grande densité. Il peut manquer ou être « non unique » (cas des séries pluri-modales) et il dépend du regroupement en classes pour les variables continues

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

La médiale partage la masse totale en deux parties égales (50 % – 50 %) , elle est toujours plus grande que la médiane et elle permet d'évaluer l'inégalité de distribution d'un caractère.

L'indice de Gini mesure la concentration d'un caractère dans la population, il montre si une petite proportion d'individus concentre une grande part de la masse totale
Il s'agit d'une mesure statistique pour mesurer l'inégalité (revenus, tailles, surfaces…).

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

La variance utilise les carrés ce qui donne des propriétés mathématiques utiles que n'a pas la valeur absolue. L'écart type correspond simplement la racine de la variance : il revient à l'unité de l'origine et est donc plus interprétable. Ainsi la variance comprend plus de rigueur mathématique tandis que l'écart type correspond à une interprétation pratique.

Pourquoi calculer l'étendue?

On calcule l'étendue car elle est simple à obtenir (maximum – minimum) et parce qu'elle donne une première idée de la dispersion. Toutefois sa fiabilité reste faible surtout pour les grands effectifs puisqu'elle ne repose que sur les extrême

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

Les quantiles divisent une série en parties égales, ils permettent d'étudier la répartition interne des valeurs et de construire des indicateurs robustes.

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

La boîte de dispersion permet de visualiser rapidement à la fois la médiane, les quartiles, l'étendue,les valeurs extrêmes et aussi de comparer facilement plusieurs distributions

On peut l'interpréter ainsi :
Le rectangle : 50 % des valeurs
Ligne interne : médiane
Moustaches : valeurs minimum et maximum

Elle résume donc à la fois la position, la dispersion et l'asymétrie

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

Les moments centrés correspondent aux moments calculés par rapport à la moyenne, ils servent à mesurer la variance, l'asymétrie et l'aplatissement
Les moments absolus utilisent la valeur absolue et moins influencé par les valeurs très grandes ou très petites

Pourquoi les utiliser ?
Pour caractériser la forme de la distribution : symétrie, aplatissement, dissymétrie.

Pourquoi vérifier la symétrie d'une distribution et comment faire ?
On vérifie la symétrie d'une distribution puisque si une distribution est symétrique elle simplifie l'analyse, la moyenne, la médiane et le mode coïncident dans ce cas et aussi parce que les choix les choix statistiques (tests, modèles) dépendent de la symétrie.

On peut utiliser le coefficient d'asymétrie bêta 1
Beta 1 > 0 —> queue à droite
Beta 1 < 0 —> queue à gauche
Beta 1 = 0 —> symétrie .
Ou sinon la comparaison des paramètres :
mode ≈ médiane ≈ moyenne —> distribution symétrique.

Séance 4

A. Questions

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?

Le choix d'une loi statistique, donc d'une distribution, dépend en premier lieu de la nature du phénomène étudié, ce qui permet de choisir entre « loi discrète et loi continue ». Viennent ensuite la forme de la distribution empirique (visuellement ou statistiquement testable), les caractéristiques de la série : espérance, médiane, variance, asymétrie etc. et le nombre de paramètres de la loi, certaines lois s'adaptant davantage selon leur flexibilité .
On choisit donc une loi/distribution discrète lorsque :
- les valeurs possibles sont dénombrables, souvent limitées à des entiers
- il s'agit de comptages: nombre d'événements, de succès/échecs, d'individus
(cf lois discrètes : Binomiale, Bernoulli, Poisson, Hypergéométrique…) .

Parallèlement, on choisit une loi/distribution continue lorsque :
- la variable peut prendre toutes les valeurs d'un intervalle, non dénombrables
- il s'agit de mesures continues : temps, distance, altitude, température… (cf les lois continues : normale, log-normale, exponentielle, uniforme continue…) .

Le critère majeur est donc la nature du phénomène et la structure des valeurs observée, le tout appuyé par la forme empirique de la distribution et les paramètres statistiques.

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

Certaines lois sont particulièrement importantes la géographie, en premier lieu la loi de Poisson qui est décrite comme « indispensable pour les événements rares » et apparaît lorsque l'on compte des occurrences dans une surface ou un intervalle. Son utilité en géographie est de modéliser des événements ponctuels dans l'espace ou le temps

comme par exemple le nombre d'accidents, de séismes, d'occurrences d'un phénomène localisé.

En outre, la loi normale (Gauss) est elle, décrite comme « la plus fréquente » et constitue souvent la distribution limite de nombreux phénomènes Elle permet de nombreuses variables naturelles et sociales approximées par une normale par exemple la distribution des hauteurs, les températures, les revenus ou les rendements.

La loi log-normale est mentionnée comme essentielle pour des variables multiplicatives et asymétriques (loi de Galton-Gibrat), elle est utile pour tout ce qui est taille des villes, surface des parcelles, revenus, intensité de certains flux.

Les lois rang-taille (Zipf et Zipf–Mandelbrot) sont utilisée en géographie pour les distributions rang-taille notamment pour les tailles de villes . Elle permet de modéliser la hiérarchie urbaine et analyser la structure polarisée d'un territoire.

Enfin, La loi exponentielle est utilisée pour les processus liés au temps d'attente ou aux risques, décrite comme adaptée aux phénomènes de fiabilité et de survie .Donc utile pour les durées d'événements naturels ou techniques ou la modélisation de probabilité de défaillance, de temps entre deux occurrences.

Séance 5

A. Questions
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 prélever un sous-ensemble d'individus (échantillon) dans une population mère, de manière aléatoire ou systématique, afin d'inférer des caractéristiques de la population à partir de cet échantillon.

Pourquoi ne pas utiliser la population en entier ? L'étude de la population entière est souvent impossible ou trop coûteuse (taille trop grande, contraintes logistiques, financières, ou temporelles). L'échantillonnage permet d'obtenir des résultats fiables en étudiant un sous-ensemble représentatif de la population, appelé échantillon.

Il existe deux grandes catégories de méthodes d'échantillonnages :
Méthodes aléatoires :
• Sondage aléatoire simple (SAS) : Chaque individu a la même probabilité d'être sélectionné (équiprobabilité).
• Tirage avec ou sans remise : Avec remise, un individu peut être sélectionné plusieurs fois ; sans remise, il ne l'est qu'une fois.
• Échantillonnage systématique : Sélection d'individus selon un pas fixe (ex. : tous les 10e individus d'une liste).

• Méthode des quotas : L'échantillon respecte les proportions de sous-groupes connus dans la population (ex. : âge, sexe).

Méthodes non aléatoires :
• Échantillonnage par convenance : Sélection basée sur la facilité d'accès.
• Méthode Monte Carlo : Utilisation de simulations aléatoires pour estimer des paramètres.

Comment choisir une méthode ? Le choix dépend de :
• L'objectif de l'étude : Précision, représentativité, ou rapidité.
• La disponibilité d'une base de sondage : Liste exhaustive des individus de la population.
• Les contraintes pratiques : Coût, temps, accessibilité.
• La taille de l'échantillon : Un échantillon représentatif et aléatoire est préférable à un grand échantillon biaisé.

2. Comment définir un estimateur et une estimation ?

Estimateur Un estimateur est une variable aléatoire (fonction des données de l'échantillon) utilisée pour estimer un paramètre inconnu d'une population (ex. : moyenne, variance). Par exemple, la moyenne de l'échantillon, notée $\bar{X}$, est un estimateur de la moyenne μ de la population.

Estimation Une estimation est la valeur numérique obtenue en appliquant l'estimateur à un échantillon spécifique. Par exemple, si $\bar{X} = 5$ pour un échantillon, alors 5 est une estimation de μ.

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

L'intervalle de fluctuation est un intervalle qui encadre la fréquence observée dans un échantillon, en supposant que la proportion théorique p dans la population est connue.

Son utilisé est d'évaluer si la fréquence observée est compatible avec p. Exemple : Pour un sondage, si $p=0,3$ et $n=50$, l'intervalle de fluctuation à 95% est [0,173;0,427][0,173 ; 0,427][0,173;0,427]. Si la fréquence observée est en dehors, on remet en cause l'hypothèse sur p.

L'intervalle de confiance est une intervalle qui encadre le paramètre inconnu (ex. : p, μ) avec une certaine probabilité, calculé à partir des données de l'échantillon.
Son utilisé est d'estimer la plage de valeurs probables pour le paramètre.
Exemple : Si on observe une fréquence $f=0,4$ dans un échantillon de taille $n=100$, l'intervalle de confiance à 95% pour p pourrait être [0,30;0,50][0,30 ; 0,50] [0,30;0,50]. Il s'agit d'une différence clé.

L'intervalle de fluctuation suppose p connu et évalue la compatibilité de la fréquence observée. L'intervalle de confiance suppose p inconnu et l'estime à partir de l'échantillon.

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

Un biais est la différence entre l’espérance de l’estimateur et la vraie valeur du paramètre :

$$\text{Biais}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

Sans biais :

$$\mathbb{E}(\hat{\theta}) = \theta$$

Biaisé :

$$\mathbb{E}(\hat{\theta}) \neq \theta$$

Conséquences : un biais systématique fausse les estimations. Par exemple, un estimateur qui sous-estime toujours la moyenne est peu fiable.

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

Statistique exhaustive (recensement) : Une enquête exhaustive (ou recensement) consiste à étudier tous les individus d’une population. Elle fournit des résultats exacts, sans inférence.
Exemple : Un recensement national pour compter la population totale.

Il y a un lien avec les données massives (Big Data). Les données massives permettent parfois d’analyser des populations entières (ex. : traces numériques, capteurs IoT), éliminant le besoin d’échantillonnage. Les avantages sont la précision, absence de biais d’échantillonnage. En revanche, les défis sont le coût de collecte, stockage, traitement, et respect de la vie privée.

Comparaison des deux méthodes :
Échantillonnage : Moins coûteux, mais résultats approximatifs (sous réserve de représentativité).
Données massives : Potentiellement exhaustives, mais complexes à gérer.

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

Le choix d'un estimateur est central en statistique inférentielle, car on cherche à reconstituer les paramètres inconnus d'une population mère à partir d'un échantillon, nécessairement imparfait.

Les principaux enjeux sont :

Limiter l'erreur d'estimation

Un échantillon ne fournit qu'une information partielle : toute estimation est affectée par des fluctuations d'échantillonnage. L'enjeu est donc de minimiser l'écart entre l'estimation et la vraie valeur du paramètre.

Les biais et variance

Un estimateur peut être biaisé, si son espérance diffère du paramètre réel. Il peut aussi avoir une variance élevée, avec des estimations très dispersées.

Le critère fondamental reste l'erreur quadratique moyenne (ERQM), qui combine biais et variance : [ERQM(\hat\theta)=V(\hat\theta)+(\text{biais})^2]. L'enjeu est donc de choisir u estimateur sans biais ou faiblement biaisé et de variance minimale.

Les convergence et consistance

Un bon estimateur doit converger vers le vrai paramètre lorsque la taille de l'échantillon augmente, et voir son biais et sa variance tendre vers zéro. Cela garantit la fiabilité à long terme de l'inférence statistique.

L'enjeu global est donc d'obtenir une estimation fiable, précise, stable et interprétable, malgré l'incertitude inhérente à l'échantillonnage.

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

L'estimation ponctuelle

Elle consiste à associer au paramètre inconnu θ une statistique calculée sur l'échantillon :

- La moyenne empirique pour μ ;
- La variance empirique (corrigée) pour σ^2 ;
- La fréquence pour une proportion p.

Elle fournit une valeur unique mais sans indication directe sur l'incertitude.

L’estimation par intervalle (intervalle de confiance)

Elle encadre le paramètre θ avec un niveau de confiance donné, par exemple :

[\hat\mu \pm t_{\alpha/2}\frac{\hat\sigma}{\sqrt{n}}]

Cette méthode est essentielle pour quantifier l’incertitude liée à l’échantillonnage.

Les méthodes fondées sur la vraisemblance

La vraisemblance est un principe fondamental de l’inférence. Elle permet un choix d’estimateurs maximisant l’information contenue dans l’échantillon. Elle fait aussi le lien avec les statistiques exhaustives et l’information de Fisher.

Les estimateurs robustes

Il s’agit donc de la médiane, des quartiles, des moyennes tronquées, ainsi que des M-estimateurs (Huber, bicarré).

Pour sélectionner une méthode d’estimation, il faut prendre en compte :

- La nature du paramètre (moyenne, variance, proportion) ;
- Les propriétés statistiques de l’estimateur : sans biais, variance minimale, convergent ;
- La taille de l’échantillon ;
- La sensibilité aux valeurs aberrantes ;
- La quantité d’informations conservée (la statistique exhaustive est préférable).

On privilégie un estimateur sans biais, convergent, efficace et robuste, adapté au contexte empirique.

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

À quoi servent les tests statistiques ?

Les tests statistiques permettent de prendre une décision sous incertitude, à partir d’un échantillon, concernant une paramètre de la population ou une hypothèse théorique (égalité, différence, effet).

Ce sont une application centrale de l'inférence statistique.

Les principaux types de test sont :

- Les tests sur la moyenne (la loi normale et la loi de Student) ;
- Les tests sur une proportion ;
- Les tests basés sur les intervalles de fluctuation ;

- Les tests fondés sur la statistique de l'échantillon et sa loi asymptotique (le théorème central limite).

Comment créer un test statistique ?

Pour créer un test statistique, on commence par formuler les hypothèses, avec H_0 l'hypothèse nulle et H_1 l'hypothèse alternative. Puis, on choisit une statistique de test en fonction de la moyenne ; de la proportion et de la variance. Il faut ensuite déterminer sa loi sous H_0 : normale ? Student ? asymptotique ?, avant de fixer un risque α (souvent 5 %). Les étapes qui suivent sont de définir une région critique, et de comparer la statistique observée au seuil critique. Enfin, il faut décider si l'on rejette ou non H_0.

Ainsi, un test est une procédure décisionnelle rigoureuse, fondée sur les propriétés probabilistes des estimateurs.

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

Plusieurs critiques peuvent être faites à l'encontre de la statistique inférentielle. Tout d'abord, es résultats sont probabilistes, jamais certains. Elle est aussi très dépendante aux hypothèses de modèle (normalité, indépendance, etc.), ainsi que sensible aux échantillons biaisés. Enfin, la statistique inférentielle est très vulnérable aux valeurs aberrantes si les outils ne sont pas adaptés.

Néanmoins, la statistique inférentielle intègre ses propres limites, grâce à l'estimation de l'erreur, aux tests avec un risque contrôlé, aux informations de Fisher et à la borne de Cramer-Rao, aux estimateurs robustes et enfin grâce aux intervalles de confiance.

La statistique inférentielle est une science de la décision raisonnée sous incertitude et non une science de certitude. Les critiques sont pertinentes lorsqu'on oublie ses hypothèses, mais injustes lorsqu'on ignore la rigueur méthodologique qu'elle impose. Elle reste indispensable dès lors qu'un recensement exhaustif est impossible.

B. Résultat Code

1. Analyse de la Loi Normale

- Structure de la distribution : La courbe est symétrique et "en cloche", centrée sur une moyenne de 100111.
- Paramètres de dispersion : L'écart-type est de 152. Les données montrent que la grande majorité des observations se concentrent autour de la moyenne (entre 85 et 115)

2. Analyse du Test de Normalité

Le test de Shapiro-Wilk, est utilisé pour vérifier statistiquement si une série de données suit effectivement une loi normale.

- Résultat du test : La statistique W est de 0,988 avec une p-value de 0,4525.
- Interprétation : Puisque la p-value est largement supérieure au seuil classique de 0,05, on ne peut pas rejeter l'hypothèse nulle (H_0). Cela signifie que les données sont considérées comme suivant une distribution normale.

Cette analyse synthétise les données de vos deux nouveaux documents en les mettant en relation directe avec les concepts de statistique et de géographie abordés dans votre rapport d'activité.

Séance 6

A. Questions

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

Une statistique d'ordre ou statistique ordinale est le cœur de la géographie humaine. De manière annuelle, mensuelle, voire hebdomadaire, un certain nombre de classements est opéré en utilisant des objets géographiques.
Leur objectif commun est de montrer quelle entité a descendu, stagné ou monté dans le classement.
La statistique ordinale s'oppose principalement à la statistique nominale.
La statistique ordinale utilise des variables qualitatives ordinales. Cela peut matérialiser une hiérarchie spatiale car la statistique ordinale permet d'établir un ordre entre les espaces : ordre entre le centre et la périphérie, un ordre du centre urbain dense jusqu'à la ruralité...

Une variable ordinale peut de fait matérialiser une hiérarchie spatiale dès lors que ses catégories représentent un niveau, un rang ou une intensité appliquée à des lieux, des territoires ou des zones.

Cartographiquement, l'ordinal produit des cartes en classes ordonnées. Cette statistique permet aussi de traduire des relations de domination ou de centralité. Enfin, elle établit des niveaux ou des rangs territoriaux.
En géographie physique, les lois d'ordre servent notamment à étudier la hauteur maximale des crus d'un cours d'eau, l'intensité du plus fort tremblement de terre dans une zone sismique donnée. Pour ce qui est de la géographie humaine, leur utilisation découle du fait de l'apparition plus ou moins spontanée de hiérarchies au sein des sociétés et des espaces étudiés.

2. Quel ordre est à privilégier dans les classifications ?

L'ordre à privilégier est l'ordre croissant ou ordre naturel. Il existe des exceptions en géographies telles que la loi dite rang-taille. L'ordination permet ainsi de rechercher les valeurs aberrantes, trop grandes ou trop petites, d'une série d'observations.

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

La corrélation des rangs mesure la force et le sens de la relation entre deux séries de rangs (comparaison de deux variables ordonnées, on cherche à savoir si les classements sont proportionnellement liés) tandis que la concordance des classements mesure à quel point plusieurs classements sont identiques (possibilité de comparer plus de deux classements, recherche de l'accord général entre plusieurs classements).

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

Le test de Spearman mesure une corrélation monotone, compare les rangs transformés des deux variables, puis calcule une corrélation sur les rangs tandis que le test de Kendall mesure la probabilité d'accord entre les paires d'observations, il compare un par un tous les couples concordants et discordants en se basant sur la circonstance du classement et non sur les écarts contrairement au teste de Spearman. Le test de Spearman est par ailleurs plus sensible aux valeurs aberrantes et aux ex aequo.

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

Le coefficient de Goodman-Kursdal se base sur la différence entre les paires concordantes et les paires discordantes. Il calcule le surplus de paires concordantes par rapport aux paires discordante en exerçant une proportion. Le coefficient de Yule quant à lui est un cas particulier du coefficient de Goodman-Kursdal en ce qu'il est appliqué dans

le cas des matrices 2x2. Il est nécessaire de construire la table de contingence qui évalue la fréquence des événements.

B. Résultat Code

Le document croise deux types d'information : la géométrie (information spatiale) et les attributs (information thématique). Le tableau permet de mettre en œuvre la statistique ordinale pour matérialiser des hiérarchies mondiales :
- croissance et rangs : quels États stagnent ou progressent dans le classement mondial du PNB ou de l'IDH
- loi de puissance : permet de mettre en avant la distribution des richesses ou des populations entre les États qui suit une loi normale typique des structures polarisées en géographie.

Conclusion

Les humanités numériques ne sont pas une simple application des mathématiques à la géographie. Il s'agit davantage d'une démarche qui utilise la statistique inférentielle et l'échantillonnage pour répondre à des contraintes de coût et de temps, tout en s'ouvrant aux perspectives des Big Data. L'enjeu reste la maîtrise de ces outils, pour garantir une analyse à la fois solide sur le plan technique et significative sur le plan social.

Concernant le retour d'expérience du travail effectué, je suis reconnaissante d'avoir pu bénéficier de l'aide de camarades car seule je n'aurais honnêtement pas réussi, bien que les explications étaient claires et l'aide de l'IA précieuse. Je suis cependant ravie d'avoir pu mettre un pied dans le monde du code, cela me paraît moins obscure et me donne l'envie d'en apprendre davantage.