RAPPORT D'ACTIVITÉ ANALYSE DE DONNÉES

Menard Myriam, M1 GAED - *GeoSuds*

Parcours débutant

I. Questions de cours et résultat de code

Séance 2 :

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

La statistique étant une branche des mathématiques et les statistiques renvoyant à un ensemble de fait analysable à travers des techniques de la statistique, il est fréquent que les géographes rejettent les statistiques comme n'entrant pas dans leur domaine d'étude. De ce fait, il existe dans la discipline géographique une sous estimation des apports de l'analyse statistiques (malgré des exceptions).

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

Il existe un débat en géographie autour de plusieurs conceptions du hasard. Un grand nombre de géographes et non statisticiens pensent que le hasard, en tant que phénomène aléatoire opposé au déterminisme, est à l'origine de tout, ce qui rend difficile de faire de la discipline géographique une science. L'école de l'analyse spatiale s'oppose à cette thèse et sa position oscille entre nécessité et contingence. La nécessité implique qu'un phénomène doive se produire d'une manière unique, tandis que la contingence offre la possibilité qu'un événement puisse se produire

ou non. Cela permet de réintroduire le paramètre scientifique dans la discipline géographique. La posture statistiques est la suivante : bien qu'il soit impossible de prévoir le détail des réalisations individuelles au sein d'un phénomène aléatoire, il est toujours possible de dégager une certitude global. En géographie humaine, cela signifie qu'il est impossible d'anticiper l'action de chaque acteur sur un territoire donné, mais que l'on peut en revanche établir une tendance, soit l'action la plus probable. Cette position permet de concilier la tendance générale observée à une échelle donnée avec la diversité des conditions locales, renvoyant à l'ancien raisonnement multiscalaire de la géographie, ce qui lui permet de se définir comme la science des échelles.

3. **Quels sont les types d'information géographique ?**

L'information géographique se structure autour de deux principales catégories : Les données attributaires (ou attributs) et les données géométriques. Les données attributaires (ou attributs) décrivent les caractéristiques de l'ensemble territorial clairement délimité. Elles peuvent relever de la géographie humaine (caractéristiques démographiques, économiques...) ou de la géographie physique (données climatiques...). Dans le contexte des Systèmes d'Information Géographique (S.I.G.), elles constituent la base attributaire. Les données géométriques renvoient à la morphologie et à la géométrie intrinsèque des ensembles spatiaux délimités. Elles forment les données géométriques d'un S.I.G.

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

Face à la massification de l'information géographique, le géographe se doit de mener une analyse de données rigoureuse pour comprendre la structure interne des phénomènes étudiés. Il s'agit d'abord de produire une base de données correctement documentées par des nomenclatures et des métadonnées, permettant de garantir la fiabilité, la comparabilité et la validité des observations ; puis de mobiliser les statistiques comme outil de réduction de l'incertitude, afin de détecter les tendances, les variations et les structures spatiales présentes dans les données massives qu'elle manipule (élaboration de représentations graphiques, de résumés numériques...). L'outil statistique dote donc la géographie de capacités essentielles pour interpréter les faits géographiques et permet de progresser vers des applications opérationnelles.

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

La statistique descriptive consiste à étudier d'un jeu de données afin d'en extraire des propriétés remarquables par rapport à une distribution théorique. Il s'agit de synthétiser les distributions et d'organiser les données (via des caractéristiques numériques ou graphiques) pour préparer les étapes suivantes de l'analyse, soit la comparaison et la prédiction. Toutes les variables sont traitées de manière égale. C'est une phase préparatoire à la statistique mathématique. La statistique explicative s'inscrit dans l'analyse de données, et vise à modéliser la relation entre une variable à expliquer et une ou plusieurs variables explicatives.

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

Il existe plusieurs types de visualisation de données en géographie : graphique en secteur, diagramme en bâtons (pour illustrer des données brutes), histogramme (représentation d'une distribution statistique), boite à moustache, polygone de fréquence… Le choix de la représentation s'effectue selon le type de variable statistique analysée. Par exemple, pour des données qualitatives nominales, on va utiliser un graphique en secteur ; tandis que pour des données qualitatives ordinales, on choisira un histogramme.

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

Les méthodes d'analyse de données sont : les méthodes descriptives, les méthodes explicatives et les méthodes de prévision.

Les méthodes descriptives s'appliquent lorsque toutes les variables jouent un rôle équivalent (absence de variable à expliquer). L'objectif est de synthétiser le tableau de données, d'appréhender les dimensions principales du phénomène, et de réaliser la visualisation et la classification des données. Une des méthodes descriptive est l'Analyse Factorielle en Composantes Principales (A.C.P.) pour les variables quantitatives, ou l'Analyse Factorielle des Correspondances (A.F.C.) pour les variables qualitatives.

Les méthodes explicatives visent à établir un lien fonctionnel entre une variable à expliquer et un ensemble de variables explicatives. La forme du modèle dépend de la nature de la réponse. Si elle est quantitative, le modèle est : $Y = \mathrm{f} \left({X}*{1}, \ldots{}, {X}*{k} \right) + \textrm{aléa}

$. Si elle est qualitative, le modèle est : $\Pr \left(Y = j / {X}{1}, \ldots{}, {X}{k} \right) = \mathrm{f} {j} \left({X}{1}, \ldots{}, {X}_{k} \right)$.

Les méthodes de prévision sont utilisées lors de l'étude des séries chronologiques. La prévision repose sur la construction d'un modèle qui établit une relation entre la valeur présente et les valeurs passées de la série. La formule est : ${X}*{t} = *mathrm{f}* *left(*{X}*{t - 1}, {X}_{t - 2}, \ldots{} \right) + \textrm{aléa}$.

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 est définie un ensemble au sens mathématique du terme, regroupant tous les éléments sur lesquels porte l'étude.
b. L'individu statistique (ou unité statistique) représente un élément isolé faisant partie de la population. Dans le champ de la géographie, il est couramment désigné sous le terme d'unité spatiale, car il est localisable et cartographiable.
c. Les caractères statistiques sont les particularités ou attributs de l'individu qui font l'objet de l'analyse. Lorsque les modalités sont connues pour chaque individu, le caractère est qualifié de variable statistique.
d. Les modalités statistiques correspondent aux valeurs spécifiques que peut prendre un caractère statistique. Ces modalités doivent former une partition du caractère, donc être incompatibles (disjointes) et exhaustives (couvrir tous les cas).

Il existe quatre types de caractères statistiques. Les données qualitatives nominales décrivent des états non chiffrés (comme la couleur des yeux). Elles permettent uniquement le calcul de fréquences. Les données qualitatives ordinales décrivent des relations d'ordre (comme petite, moyenne, grande). Elles permettent le calcul de la fréquence et de la médiane. Les données quantitatives discrètes (ou absolues) décrivent des listes finies et isolées de valeurs. Elles correspondent à un décompte. Les données quantitatives continues (ou relatives) décrivent des valeurs prises dans un intervalle réel. Elles impliquent l'existence d'une unité de mesure. Il n'existe pas de hiérarchie entre les variables.

9. **Comment mesurer une amplitude et une densité ?**

On obtient l'amplitude en soustrayant la valeur de sa borne minimale à celle de sa borne maximale. La densité s'obtient par la formule : d=n*i*/b-a. C'est un ratio qui rapporte l'effectif d'une classe à son amplitude. Ces deux mesures sont utilisées pour la discrétisation de caractères quantitatifs (classer les valeurs).

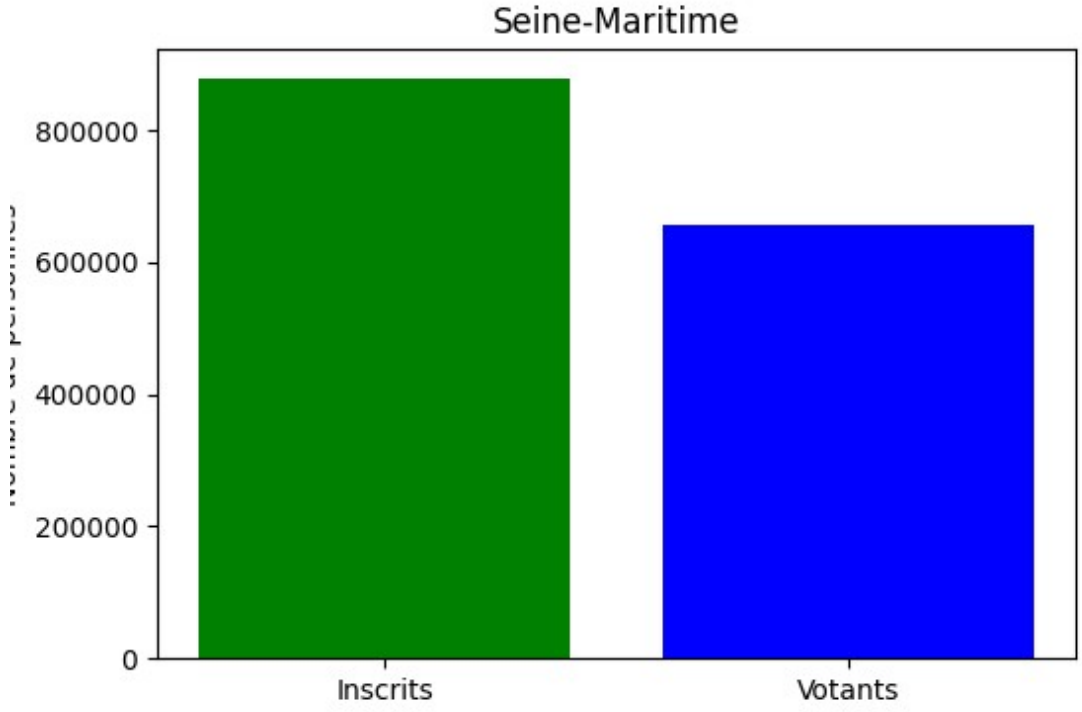
10. **À quoi servent les formules de Sturges et de Yule ?**

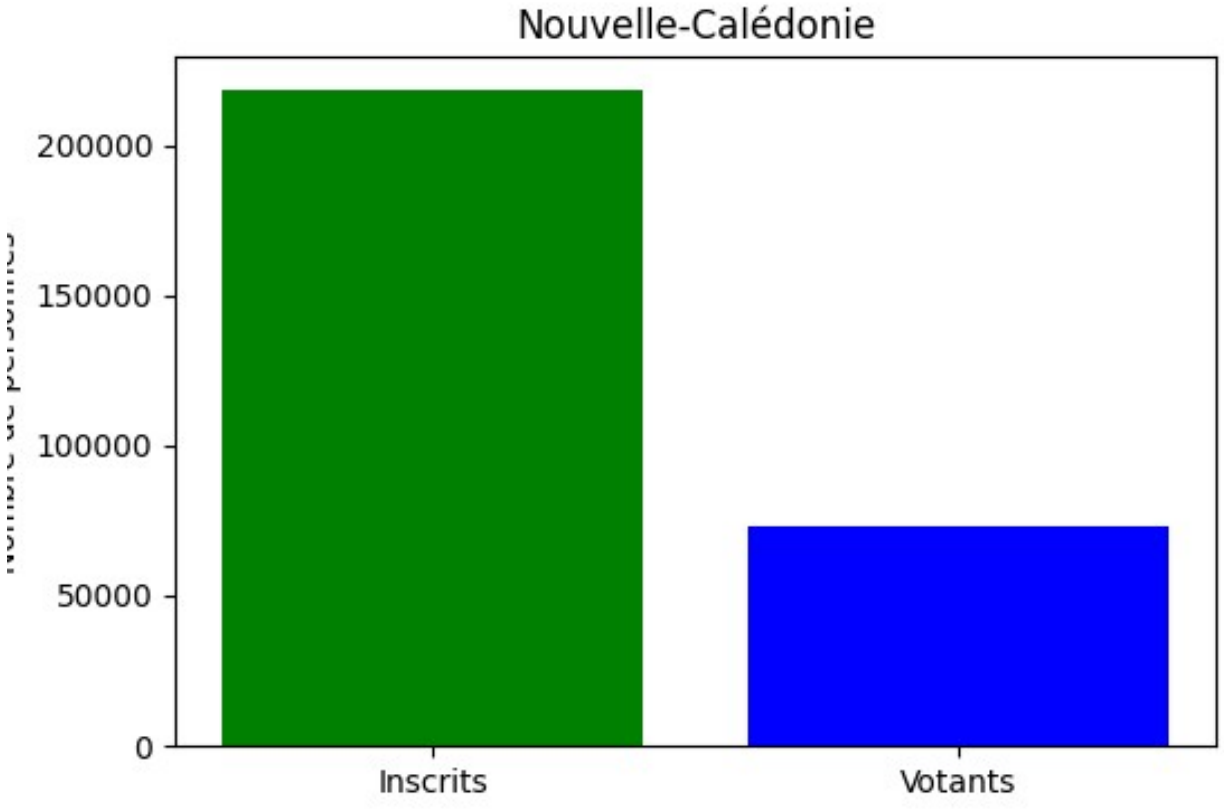
La formule de Sturges donne une valeur approximative du nombre de classe, ce qui est utile à l'analyse car un nombre de classe trop petit ou trop élevé donne une perte d'informations. La formule de Yule est une alternative à celle de Sturges.

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

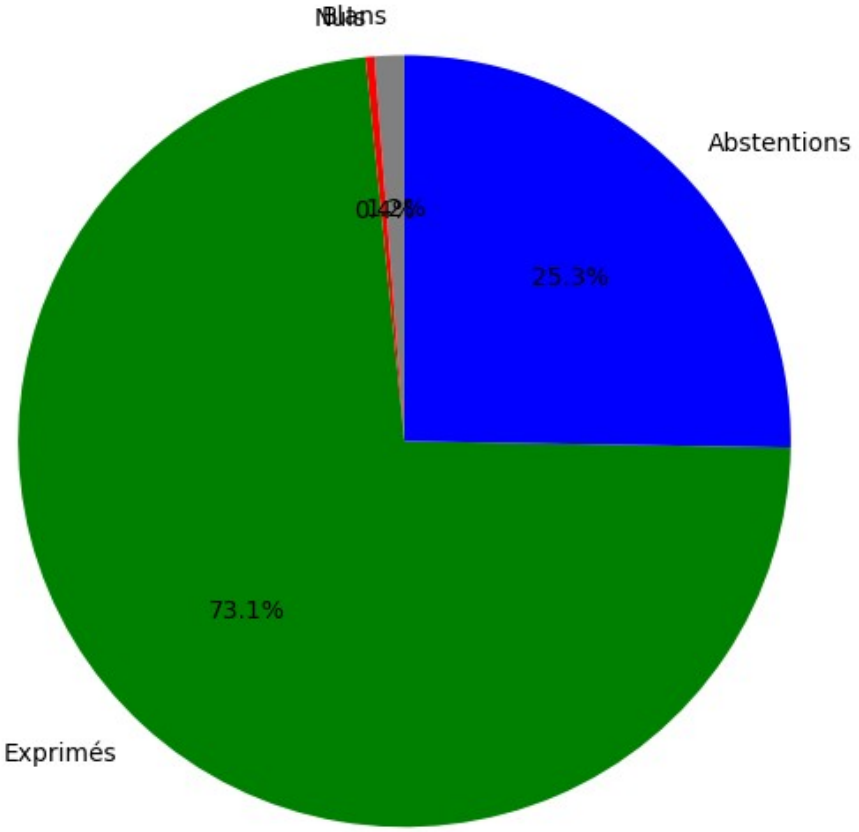
Un effectif (ou fréquence absolue) correspond au nombre de fois où une modalité ou une valeur spécifique apparaît au sein de la population statistique analysée. Une fréquence s'obtient en en divisant l'effectif de la modalité par l'effectif total de la population statistique. La valeur obtenue est comprise entre 0 et 1. La fréquence cumulée s'applique aux caractères quantitatifs ordinales, et s'obtient en additionnant les fréquences associées aux valeurs inférieures ou égales à un seuil donné. La fréquence, représentant une "probabilité réelle" observée, est l'élément qui permet d'établir une distribution statistique empirique. Cette distribution sert ensuite à déterminer et à conclure sur la loi de probabilité théorique la plus pertinente pour modéliser le phénomène étudié.

Graphiques obtenus avec le code de la séance :

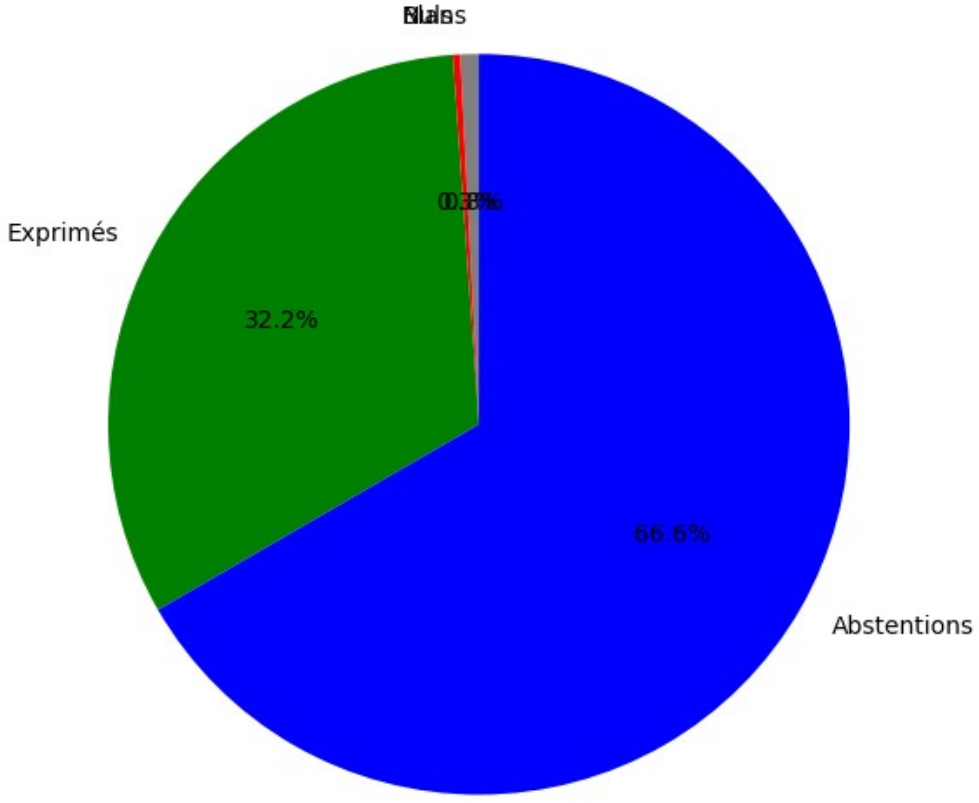




Répartition des votes - Seine-Maritime



Répartition des votes - Nouvelle-Calédonie



Le code de cette séance m'a permis d'obtenir un histogramme par département permettant de voir la proportion de votants par rapport aux inscrits, en milliers de personnes ; ainsi qu'un diagramme circulaire par département permettant de voir la répartition des votes en pourcentage. L'histogramme est assez bien sorti, mais les noms des catégories du diagramme se superposent lorsque la part est trop petite (entre votes blancs et nuls). Je n'ai pas réussi à résoudre ce problème. Les départements de France métropolitaine sont assez semblables dans la répartition des votes, mais la différence devient assez flagrante lorsque la comparaison est faite avec les départements d'Outre-Mer. J'ai mis comme exemple la Nouvelle-Calédonie car c'est le département où le pourcentage d'abstention est le plus fort. Nous pouvons dire que les départements d'Outre-Mer sont les territoires où le pouvoir politique français est le moins bien représenté, ce qui expliquerait ces résultats.

Séance 3 :

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

Les variables quantitatives sont la principale préoccupation de l'étude des paramètres statistiques, qui se divisent en paramètres de position, de dispersion et de forme. C'est donc le caractère le plus général.

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

On dit d'une variable qu'elle a un caractère quantitatif discret lorsqu'elle une certaine valeur dans un intervalle donné et prend la forme d'un nombre entier. Le caractère quantitatif continue désigne une variable qui peut prendre toutes les valeurs dans l'intervalle donné. La distinction est nécessaire car les méthodes de calcul de plusieurs paramètres statistiques diffèrent selon la nature du caractère. Par exemple, le calcul de la médiane pour un caractère quantitatif discret implique l'identification d'un rang ou d'un intervalle médian, tandis que pour un caractère quantitatif continu, la médiane est définie comme la valeur m_e telle que la fréquence cumulée jusqu'à m soit égale à 1/2.

3. **Paramètres de position : Pourquoi existe-t-il plusieurs types de moyenne ? Pourquoi calculer une médiane? Quand est-il possible de calculer un mode ?**

Il existe plusieurs manières de calculer les moyennes (arithmétique, quadratique, harmonique, géométrique, mobile) en fonction de la nature de la variable étudiée. La moyenne arithmétique est sensible aux valeurs aberrantes, ce qui rend nécessaire l'existence d'autres mesures ou la suppression des valeurs extrêmes dans certains cas.

La médiane est calculée car, contrairement à la moyenne arithmétique, elle n'est pas influencée par les valeurs aberrantes. Elle est la valeur qui divise la population en deux sous-populations de probabilité équiprobable. De plus, elle résume bien les distributions fortement dissymétriques parce qu'elle est déterminée par le classement des valeurs, et non par les valeurs extrêmes elles-mêmes.

Le mode fait référence à toute modalité correspondant à l'effectif maximal ou à la densité maximale. Il s'agit de la valeur la plus fréquente (pour une variable discrète) ou de celle qui a la plus forte densité de probabilité. Le mode peut ne pas exister, et, lorsqu'il existe, il n'est pas toujours unique, ce qui donne lieu à des distributions bimodales ou plurimodales.

4. **Paramètres de concentration : Quel est l'intérêt de la médiale et de l'indice de C. Gini ?**

La médiale est la valeur centrale qui a pour intérêt de partager la masse totale de la variable en deux parties égales, chacune représentant 50 % des valeurs globales. La médiale, en comparaison avec la médiane permet de fournir une mesure de concentration. Lorsque l'écart entre la médiale et la médiane est grand par rapport à l'étendue de la distribution, la concentration est forte, et inversement, la distribution est considérée comme égalitaire. La médiale est liée à l'indice de C. Gini. La courbe de Gini a pour but de montrer comment se manifeste la concentration d'une population dans une distribution donnée. Plus la courbe s'éloigne de la diagonale, plus la concentration est importante.

5. **Paramètres de dispersion — Pourquoi calculer une variance à la place de l'écart à la moyenne ? Pourquoi la remplacer par l'écart type ? — Pourquoi calculer l'étendue ? — À quoi sert-il de créer un quantile ? Quel(s) est (sont) le(s) quantile(s) le(s) plus utilisé(s) ? — Pourquoi construire une boîte de dispersion ? Comment l'interpréter ?**

Calculer l'écart à la moyenne ne suffit pas pour mesurer la dispersion, car cette moyenne est toujours égale à zéro. La variance utilise le carré des écarts à la moyenne. La variance est considérée comme la meilleure caractéristique de dispersion car elle tient compte de toutes les données. La variance est remplacée par l'écart type car celui-ci (racine carrée de la variance) permet de ramener l'indicateur de dispersion à la même unité que la moyenne ou l'espérance. L'écart type est souvent plus pratique que la variance car il s'exprime dans la même unité que la moyenne.

L'étendue est calculée c'est un indicateur facile à obtenir. Elle représente la différence entre la plus grande et la plus petite valeur observée dans une série statistique. Cependant, l'étendue ne dépend que des valeurs extrêmes et est très peu utilisée dès que le nombre de données dépasse 10.

Les quantiles sont créés pour partager la série statistique ordonnée en parties égales. Ils servent de caractéristiques de position et sont utilisés pour définir des mesures de dispersion comme l'écart interquartile. Les quantiles les plus utilisés sont les quartiles, qui divisent la série en quatre parties de même effectif. Le deuxième quartile correspond à la médiane.

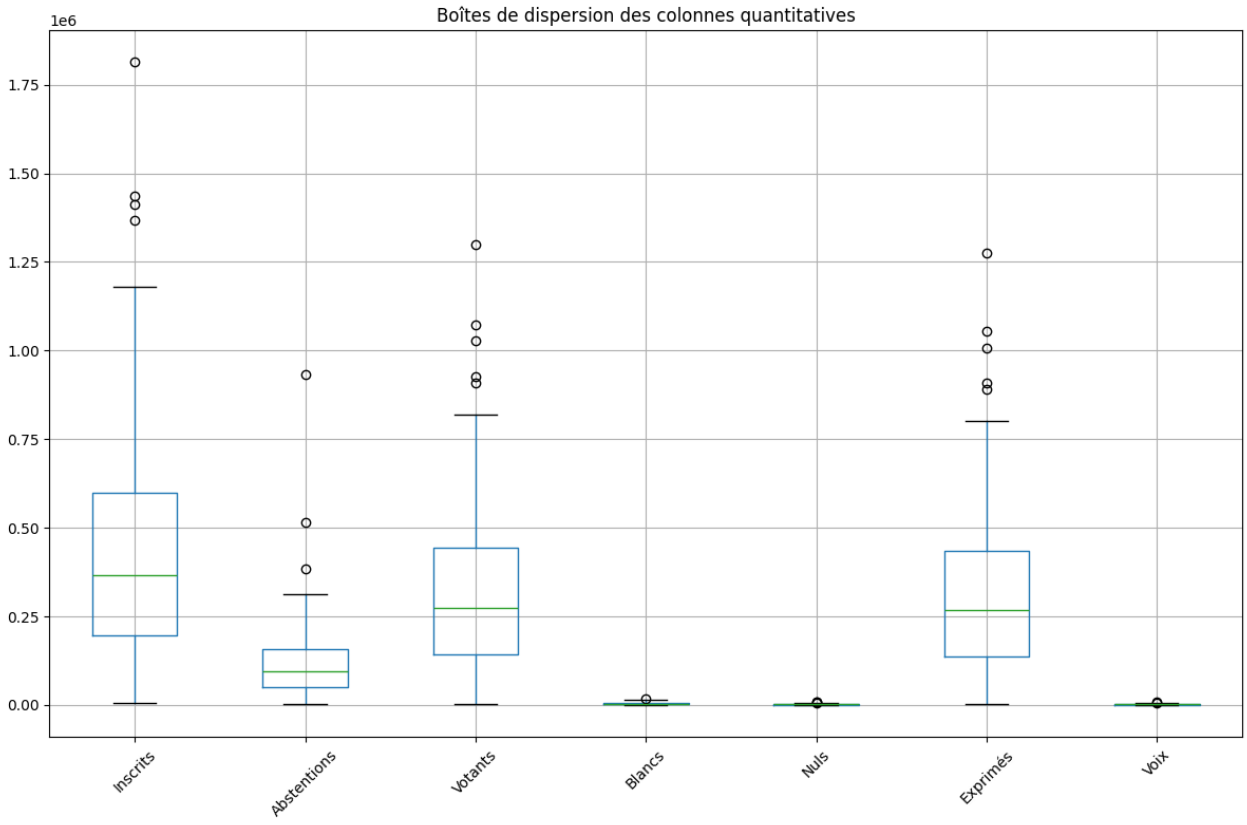
La boîte de dispersion (boîte à moustache) est construite pour représenter schématiquement les principales caractéristiques d'une distribution d'un caractère quantitatif. Elle sert principalement à comparer visuellement plusieurs séries statistiques. Pour l'interpréter, on observe le rectangle qui s'étend du premier quartile au troisième quartile. Cet intervalle interquartile contient 50 % des valeurs de la série. La médiane est marquée par un trait à l'intérieur de la boîte. Les "moustaches" représentent les segments allant des valeurs minimales et maximales jusqu'aux quartiles.

6. **Paramètres de forme : Quelle différence faites-vous entre les moments centrés et les moments absolus ? Pourquoi les utiliser ? Pourquoi vérifier la symétrie d'une distribution et comment faire ?**

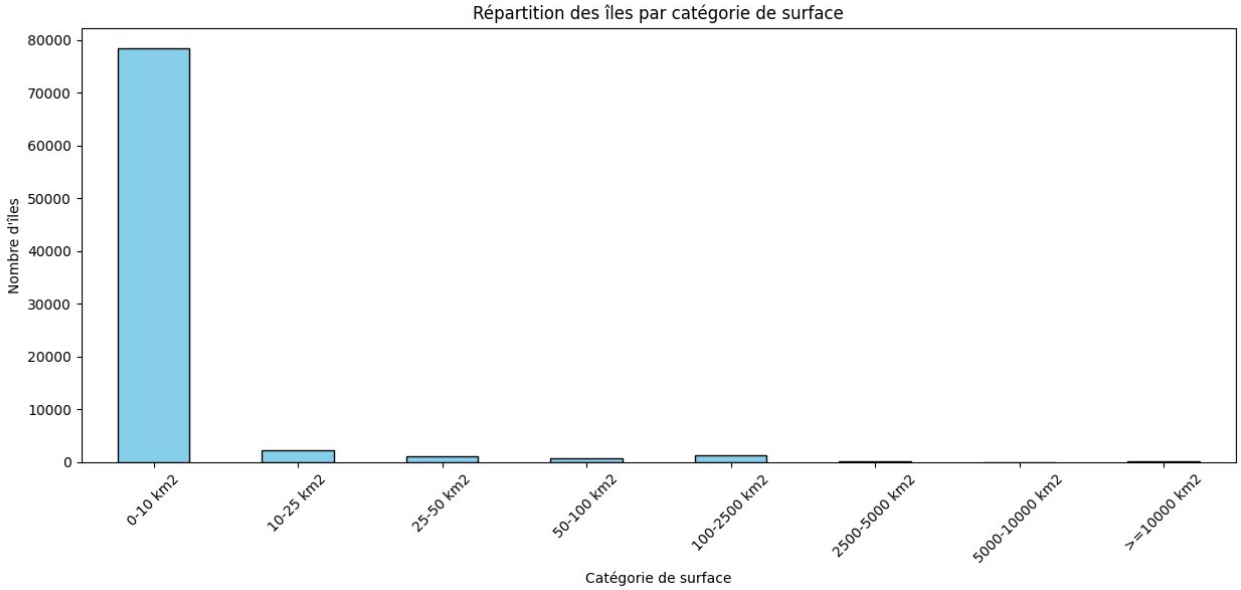
Le moment centré d'ordre tient compte du signe des écarts : les valeurs situées au-dessus de la moyenne peuvent compenser celles situées en dessous. Ces moments sont utilisés pour caractériser la variance, l'asymétrie ou l'aplatissement d'une distribution. Le moment absolu d'ordre est la moyenne des valeurs absolues des écarts à une référence (souvent la moyenne). Il ne permet pas de compensation entre les écarts positifs et négatifs, ce qui le rend particulièrement pertinent pour mesurer la dispersion réelle des observations, indépendamment de leur direction. Il mesure l'amplitude sans considération de sens. Ils sont utilisés pour caractériser une distribution et définir les paramètres de forme, notamment la symétrie et l'aplatissement de la loi de distribution.

Il est nécessaire de vérifier la symétrie d'une distribution car si elle est parfaitement symétrique, le mode, la moyenne arithmétique et la médiane sont égaux. Cette vérification caractérise la forme de la loi de distribution statistique. Pour vérifier la symétrie, on utilise la mesure de la dissymétrie. Si le résultat est inférieur à 0, alors la dissymétrie est dite positive. A l'inverse, elle est dite négative. Si le résultat est égal à 0, alors la distribution est symétrique.

<u>Résultats obtenus avec le code de la séance :</u>



Mon code a généré une seule boite à moustache, qui rassemble chaque colonne quantitative du tableau des résultats des élections. Cela peut permettre une meilleure comparaison entre les résultats de chaque colonne, mais c'est peut-être moins lisible qu'une boite à moustache par colonne. La médiane, l'étendue, les valeurs minimales et maximales, les quartiles et la moyenne sont assez visibles, la médiane étant représentée par une couleur différente. Cette boite a aussi pour intérêt de mettre en évidence les valeurs que l'on pourrait juger aberrantes. Elle permet donc une première analyse et catégorisation des valeurs quantitatives de la base de données.



L'histogramme que j'ai conçu pour répartir les îles par catégorie de surface est assez lisible et permet de catégoriser des caractères quantitatifs discrets. Nous pouvons voir qu'il y a plus de 75 000 îles comprises entre 0 et 10 km², ce qui en fait la variable statistique prépondérante de l'analyse.

Séance 4 :

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 ?**

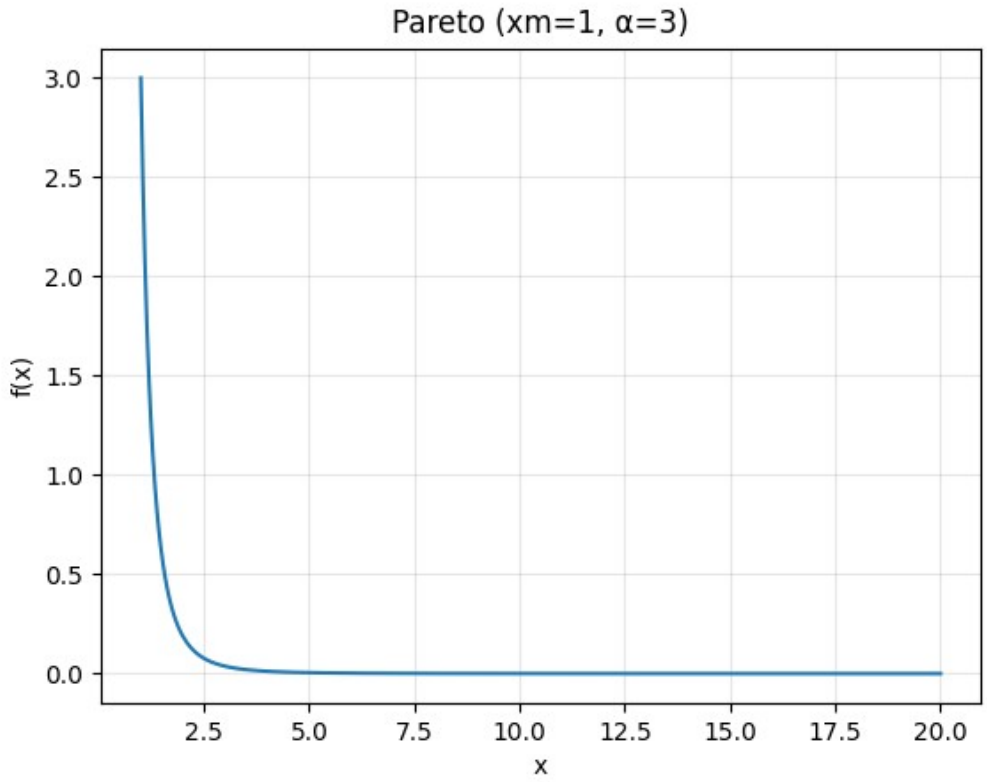
Le choix de la distribution statistique s'effectue selon la nature du phénomène étudié. Les distributions de variables discrètes sont utilisées pour modéliser des phénomènes où la variable aléatoire ne peut prendre qu'un nombre fini ou dénombrable de valeurs (souvent des nombres entiers, des comptes ou des résultats spécifiques). Par exemple, ce type de distribution sert à modéliser des sondages d'opinion ou des résultats de jeux de hasard. Les distributions de variables continues sont utilisées lorsque la variable aléatoire peut prendre toutes les valeurs au sein d'un intervalle donné.

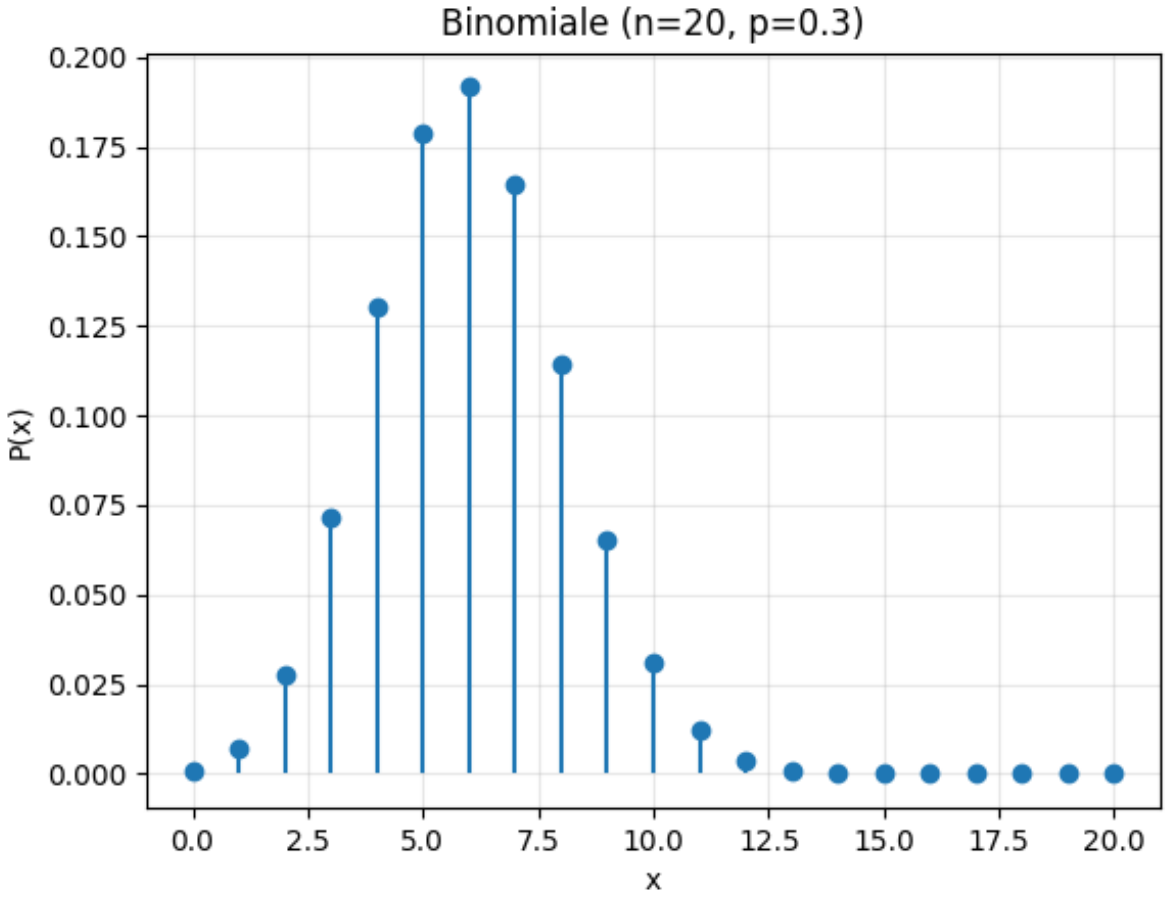
2. **Expliquez selon vous quelles sont les lois les plus utilisées en géographie ?**

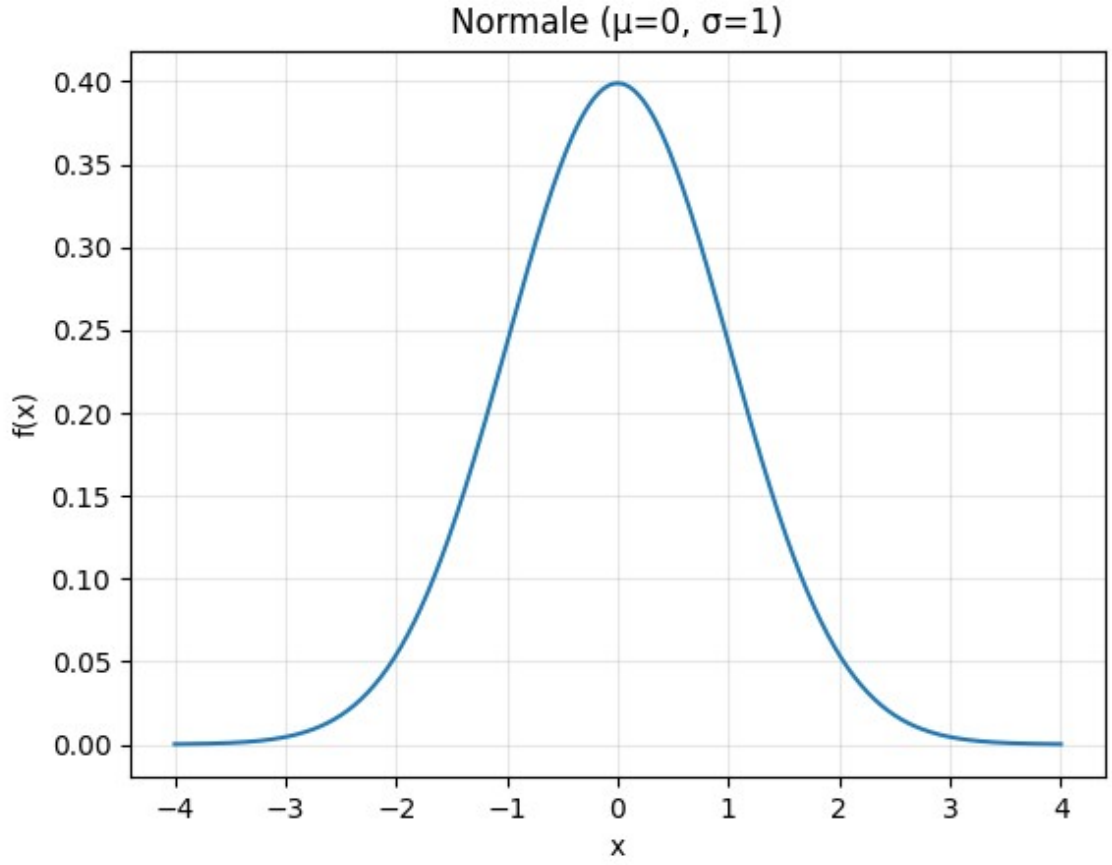
Les lois statistiques les plus utilisées en géographie concernent généralement la modélisation de phénomènes où la distribution des valeurs est fortement asymétrique ou suit des relations de rang. Concernant la loi normale, bien qu'elle soit la distribution la plus fréquente, il est reconnu dès le début du XXe siècle que les variables aléatoires concernant la fréquence du dénombrement de certains objets géographiques (comme les lacs ou les montagnes) ne suivaient pas une loi normale. Il est donc nécessaire de connaître d'autres distributions, au-delà du "bruit blanc" souvent attribué à la loi normale.

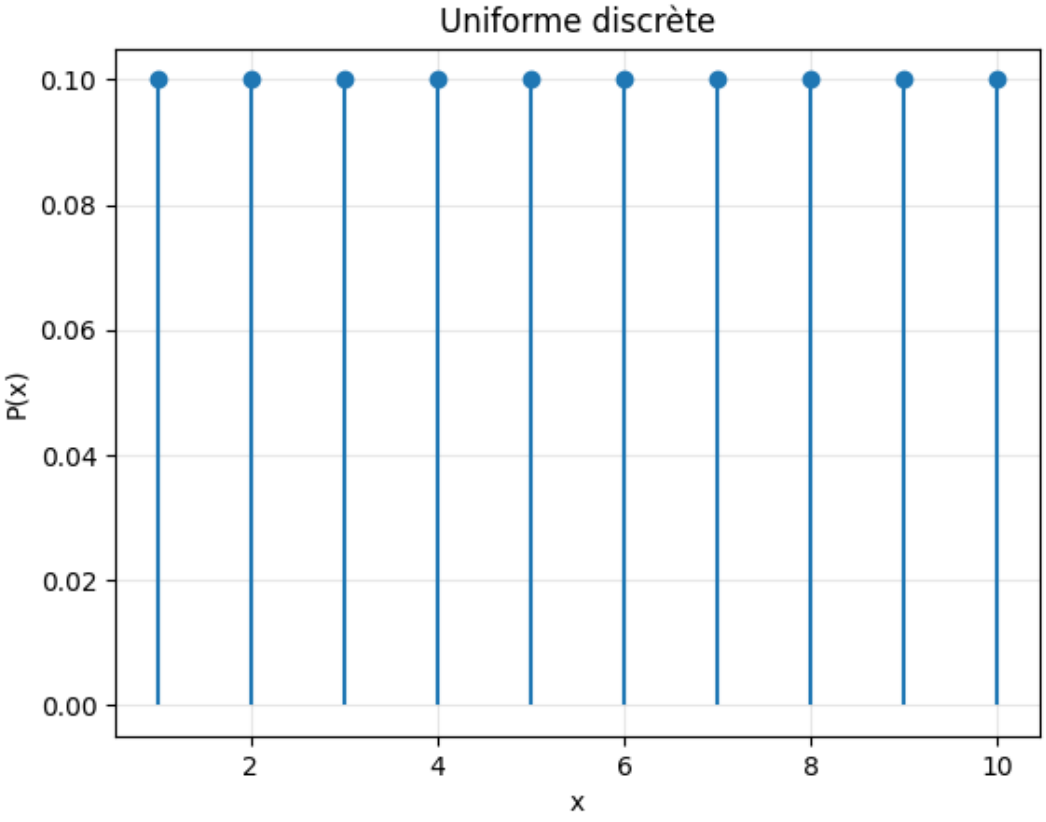
Tout d'abord, j'ai l'impression que la loi du $\chi 2$ est une distribution essentielle en statistiques. Elle est surtout pertinente pour les tests d'hypothèse et la construction d'autres lois importantes en inférence statistique. Sous cet angle, il semble que ce soit une des lois les plus utilisées. Ensuite, la loi de Zipf est utilisée dans les lois rang-taille. Il s'agit par exemple de comparer, au sein d'un territoire donné, le nombre d'habitants d'une ville avec son rang (sa classification). Cette loi permet de voir que l'occurrence d'un phénomène est inversement proportionnelle à son rang. La loi de Zipf-Mandelbrot est une généralisation de la loi de Zipf pour mieux modéliser ces phénomènes. Enfin, la loi de Benford décrit la probabilité qu'un nombre commence par un certain chiffre (sauf 0). Elle est appliquée à des données géographiques, telles que la longueur des fleuves du globe ou la superficie des pays.

<u>Résultats obtenus avec le code :</u> un graphique illustrant chaque loi, dont je met quelques exemples.


Pareto (xm=1, α=3)
f(x)
x
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Binomiale (n=20, p=0.3)
P(x)
x
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Normale (μ=0, σ=1)
f(x)
x
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−3
−2
−1
0
1
2
3
4


Uniforme discrète
P(x)
x
0.10
0.08
0.06
0.04
0.02
0.00
2
4
6
8
10

Le code que j'ai conçu à générer un graphique par loi, ce qui m'a permis de mieux comprendre le cours et de voir la manière dont s'applique chaque loi. J'ai donc un exemple de la distribution théorique de chaque loi observée, comme celle de Pareto, ou la loi normale dont j'ai mis un exemple ici. Cela permet une comparaison entre la distribution théorique et la distribution observée du phénomène étudié.

Séance 5 :

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

Un échantillon est un sous-ensemble d'une population dite mère, qui se veut représentatif (pertinent et aléatoire), sélectionné selon une variable aléatoire. On utilise un échantillon dans le cas où la populations mère est trop vaste. L'échantillon est aussi plus accessible car moins coûteux. Par exemple, concernant, l'intention de vote, on utilise un échantillon car il n'est pas possible de connaître l'opinion de plusieurs millions de personnes. Il y a deux types d'échantillons : indépendants, soit constitués par des individus différents ; ou appariés, soit constitués de mêmes individus tiré au sort est associés deux par deux.

Il existe de types de méthodes échantillonnage : aléatoire et non aléatoire. La méthode aléatoire revient donc à utiliser un tirage au sort, ce qui implique que le géographe dispose d'une base de sondage. Le tirage au sort peut être avec remise ou sans remise, selon que l'on raye ou non le numéro tiré au sort. C'est la différence entre échantillonnage non exhaustif ou exhaustif.

La méthode non aléatoire revient à essayer de construire un « modèle réduit » de la population mère. Il y a plusieurs manières de faire. L'échantillonnage systématique suppose l'existence d'une base de sondage, mais les résultats sont moins précis que les sondages aléatoires simples. La méthode par quotas permet d'obtenir des échantillons non biaisés, en respectant les proportion d'origine de la population mère. Le résultat peut donc être plus précis qu'un sondage aléatoire simple. La méthode d'échantillonnage « Monte Carlo » permet de transformer des valeurs en moyenne obtenues par un échantillon. Il s'agit de mettre en place une estimation fiable de moyenne pour une variable aléatoire. Concernant le choix de la méthode, on privilégie celle qui

garantit la meilleur chance d'obtenir l'échantillon le plus représentatif possible en fonction de notre étude.

2. **Comment définir un estimateur et une estimation ?**

L'estimation et l'estimateur sont des concepts utilisés pour tirer des conclusions sur les caractéristiques d'une population mère à partir de l'étude d'un échantillon. Un estimateur est une variable aléatoire qui sert de fonction des données d'un échantillon. Il est construit de manière à ce que sa valeur soit la plus proche possible de la vraie valeur du paramètre que l'on cherche à déterminer dans la population. Il peut être sans biais, soit sans différence entre son espérance mathématique et la valeur du paramètre à estimer dans la population ; convergent, s'il converge en probabilité vers le paramètre lorsque la taille l'échantillon tend vers l'infini ; efficace**,** s'il est sans biais et s'il présente la variance la plus petite parmi tous les estimateurs sans biais. La moyenne est un estimateur sans biais et convergent par exemple. L'estimation est le processus statistique par lequel on cherche à obtenir des valeurs approchées des caractéristiques (paramètres) d'une population à partir des résultats obtenus sur un échantillon aléatoire. L'estimation ne se limite pas à donner une valeur ponctuelle mais vise également à évaluer la fiabilité de cette valeur en construisant un intervalle de confiance. Un intervalle de confiance est un intervalle qui a une forte probabilité de contenir la vraie valeur du paramètre. De fait, l'estimateur est l'outil mathématique, tandis que l'estimation est le résultat chiffré obtenu en appliquant cet outil aux données spécifiques de l'échantillon.

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

La distinction entre l'intervalle de fluctuation et l'intervalle de confiance se fonde sur l'orientation de l'incertitude et sur ce qui est connu *a priori*. L'intervalle de fluctuation est un outil d'échantillonnage utilisé lorsque la vraie proportion théorique de la population est supposée connue. Son objectif est de prédire ou d'encadrer la fréquence observée d'un échantillon d'une certaine taille, afin de vérifier si le résultat de cet échantillon est compatible avec le paramètre théorique connu. Si la fréquence observée tombe en dehors de cet intervalle, l'hypothèse de la proportion connue est rejetée. L'intervalle de confiance est quant à lui utilisé lorsque le paramètre de la population est inconnu. Sa fonction est d'encadrer la vraie valeur inconnue de ce paramètre

de la population en se basant sur les données obtenues de l'échantillon,. L'intervalle de confiance est donc un intervalle aléatoire qui a une forte probabilité (selon le niveau de confiance) de contenir la valeur exacte du paramètre. Finalement, l'intervalle de fluctuation permet de juger la conformité d'un échantillon à une population de référence connue, tandis que l'intervalle de confiance permet d'estimer et d'évaluer la fiabilité d'un paramètre inconnu de la population à partir des données de l'échantillon.

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

Dans la théorie de l'estimation, un biais est défini comme la différence entre l'espérance mathématique de l'estimateur et la vraie valeur du paramètre à estimer dans la population. Ce décalage peut aussi être défini comme une erreur d'estimation. Lorsqu'un estimateur est biaisé, le biais est alors considéré comme une erreur systématique. Cela signifie que l'estimateur va varier autour de son espérance mathématique, au lieu de varier autour de la valeur exacte du paramètre. Le concept de biais est permet d'évaluer la précision d'un estimateur, qui est mesurée par l'erreur quadratique moyenne.

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

Une statistique travaillant sur la population totale est un recensement, donc une enquête exhaustive.
En statistique, le recours au recensement est souvent difficile en terme logistique et financier. De plus, dans de nombreux cas, il est théoriquement impossible d'obtenir l'information totale d'une population. C'est pourquoi on utilise des procédés d'échantillonnage. Le terme de données massives décrit la collection, la gestion est l'analyse d'une base de données d'un volume important et de sources hétérogènes. Cet ensemble de données est trop vaste pour être analysé de manière pertinente, c'est là le lien avec l'impraticabilité du recensement en statistique.

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

L'objectif de la théorie de l'estimation est de sélectionner le meilleur estimateur parmi toutes les statistiques possibles pour approcher le plus fidèlement possible le paramètre exact de la population. L'enjeu principal du choix de l'estimateur est de s'assurer que l'estimation obtenue à partir de l'échantillon soit la plus fiable possible pour pouvoir être inférée (étendue) à la population mère. Ce choix est déterminant pour la précision de l'estimation. De plus, le choix est lié à l'utilisation maximale de l'information contenue dans l'échantillon. Un estimateur se doit d'être un résumé ou une statistique dérivée de l'échantillon. Enfin, le choix de l'estimateur est aussi contraint par le contexte et la qualité des données. Il se doit d'être peu sensible aux données aberrantes qui peuvent invalider les résultats d'une analyse statistique (par exemple, la moyenne et la variance sont plus sensibles aux données aberrantes que la médiane et les quartiles).

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

La théorie de l'estimation propose plusieurs approches pour déterminer les valeurs des paramètres inconnus d'une population à partir des données d'un échantillon.

L'estimation ponctuelle consiste à donner une valeur approchée spécifique du paramètre de la population en utilisant une fonction des données de l'échantillon, appelée estimateur. Par exemple, la moyenne calculée sur l'échantillon est une estimation ponctuelle de la moyenne de la population. Dans le cas de l'estimation par intervalle de confiance, l'objectif n'est pas de fournir une seule valeur, mais de construire un intervalle de confiance qui a de « grandes chances » (associées à un niveau de confiance) de contenir la vraie valeur du paramètre.

La méthode du Maximum de Vraisemblance (M.V.) consiste à choisir l'estimateur qui maximise la fonction de vraisemblance, c'est-à-dire la valeur du paramètre qui rend l'événement observé (l'échantillon) le plus probable. Cette méthode nécessite de présupposé que l'événement qui s'est produit était le plus probable.

La méthode des moindres carrés est utilisée lorsque les quantités à estimer sont des espérances.

La méthode du *bootstrap* utilise les méthodes d'échantillonnage de Monte-Carlo. Elle consiste à remanier plusieurs fois l'échantillon de départ dans le but d'obtenir un intervalle de confiance. Elle est utile pour estimer des paramètres pour lesquels les calculs analytiques sont difficiles ou impossibles (comme la médiane ou les quartiles). Les méthodes de Monte-Carlo, en général, proposent une transformation des valeurs en moyenne obtenues par un échantillon et fournissent

souvent le seul moyen pratique pour déterminer les distributions d'échantillonnage dans des situations mathématiquement intraitables.

Le choix de la méthode d'estimation d'un paramètre dépend de plusieurs critères : l'absence de biais, l'efficacité (soit celui qui possède la variance la plus petite), la convergence (la distribution devant se concentrer autour de la valeur inconnue du paramètre lorsque la taille de l'échantillon augmente), l'exhaustivité et la robustesse (peu de sensibilité face aux données aberrantes). Sélectionner une méthode d'estimation revient à opérer un arbitrage entre ces critères, en fonction de la loi de probabilité supposée des données et de l'importance d'éviter le biais ou de garantir la plus faible variance. L'objectif est toujours de s'assurer que l'estimation ponctuelle est la plus proche possible du paramètre exact, quel que soit l'échantillon.

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

Le test statistique est une méthode de calcul, dont le résultat permet de déterminer si une série statistique d'observations est compatible avec une loi de probabilité entièrement spécifique. Il permet donc de vérifier une hypothèse, c'est-à-dire de vérifier si le résultat de l'analyse statistique est en accord avec la distribution théorique. Le test permet de choisir de manière pertinente entre deux hypothèses statistiques (un postulat au sujet d'une population mère, que l'on vérifie par l'analyse d'un échantillon).

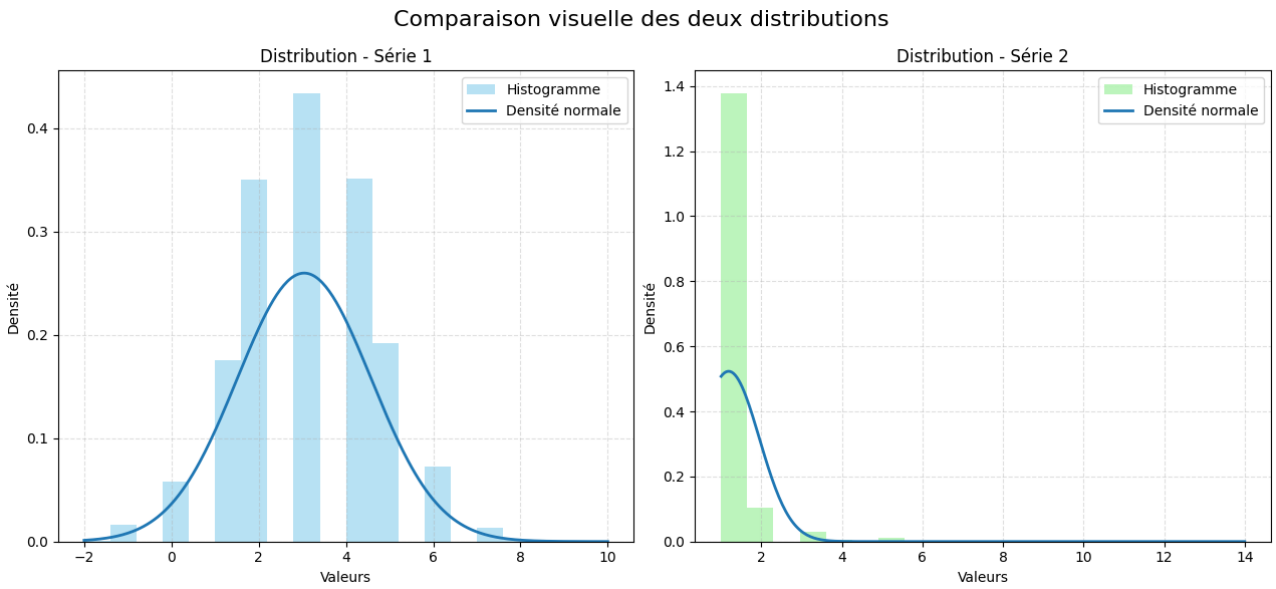
La mise en œuvre d'un test statistique se doit de prendre une décision statistique en contrôlant le risque d'erreur. Il faut d'abord formuler les hypothèses, choisir le seuil de signification (c'est-à-dire fixer le risque d'erreur de première espèce), et définir la statistique de test. Ensuite, il faut spécifier la loi de probabilité de la statistique en supposant que l'hypothèse nulle est vraie, définir la région critique et calculer la valeur numérique de la statistique du test à partir des données de l'échantillon, qui permet d'aboutir à une conclusion.

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

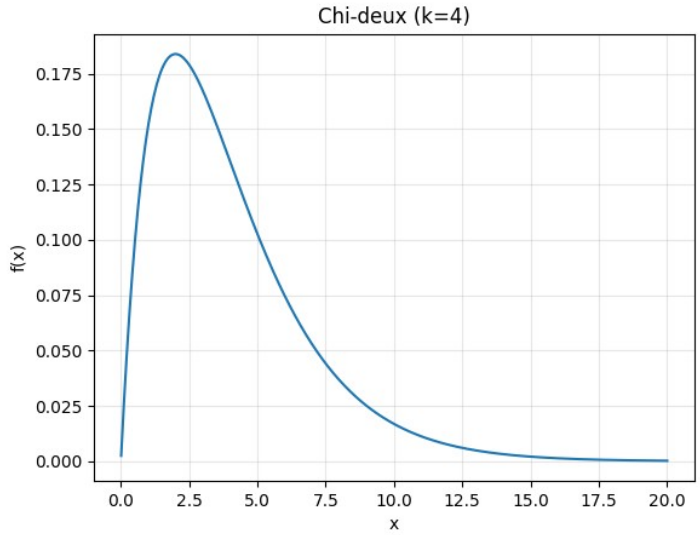
Il me semble que les critiques adressées à la statistique inférentielle sont en grande partie fondées, car cette approche repose sur un ensemble de conditions rarement parfaitement réunies dans la pratique. En effet, l'inférence dépend d'échantillons soumis aux fluctuations d'échantillonnage et

dont la représentativité n'est jamais totalement garantie, ce qui limite la solidité des conclusions que l'on peut en tirer. Elle se révèle particulièrement vulnérable aux valeurs aberrantes, susceptibles d'invalider des estimateurs centraux. De plus, les tests d'hypothèse comportent des risques incompressibles d'erreurs de première et de seconde espèce et s'appuient sur des seuils de signification arbitraires, ce qui nourrit une critique légitime de leur rigidité. En termes d'analyse scientifique, il semble aussi que la manière dont les résultats sont présentés dans les articles scientifiques rend souvent difficile une méta-analyse. Pour autant, malgré ces limites, je pense que la statistique inférentielle reste un outil indispensable car elle permet d'accéder à des informations autrement inaccessibles, notamment lorsque l'observation exhaustive d'une population est impossible. Les critiques ne doivent donc pas conduire à son rejet, mais plutôt à une utilisation éclairée et rigoureuse, conforme aux précautions méthodologiques.

Comparaison des lois obtenues avec le code :



Avec mon code, j'ai donc obtenu une graphique permettant de comparer les deux distributions côte à côte. Pour déterminer laquelle des deux distributions correspond à la loi normale, j'ai pu m'aider des distributions théoriques de la séance précédente. Je pense donc que la série 1 correspond à la loi normale car sa courbe suit celle de la distribution théorique de la loi normale. Quant à la deuxième série, je pense qu'elle correspond à la distribution du chi-2.



(graphique de la distribution théorique du chi-2 obtenu à la séance 4)

Séance 6 :

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 concerne les variables qualitatives qui peuvent être ordonnées, c'est-à-dire classées dans un ordre croissant ou décroissant, comme petite, moyenne, grande par exemple. On peut l'opposer à la statistique nominale, qui concerne des variable qualitatives sans ordre naturel, et exploite l'appartenance à une catégorie (par exemple, la couleur des yeux) et non pas un rang, comme la statistique ordinale.

La statistique ordinale est donc un outil pour matérialiser une hiérarchie spatiale. Elle permet d'effectuer régulièrement (rythme annuel, mensuel ou hebdomadaire) d'un certain nombre de classements utilisant des objets géographiques. L'objectif de ces classements est de matérialiser la hiérarchie en montrant quelle entité géographique a descendu, stagné ou monté dans le classement. Cela établit un lien spontané entre l'ordination et la variable quantitative dans le cadre d'études géographiques.

2. **Quel ordre est à privilégier dans les classifications ?**

L'ordre à privilégié dans les classifications pour la statistique d'ordre est l'ordre croissant dit ordre naturel, ce qui permet d'identifier facilement les valeurs aberrantes et d'estimer facilement un coefficient dans le cas des tests de corrélation des rangs, comme le test de Kendall par exemple.

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

La corrélation des rangs et la concordance de classements renvoient toutes deux à l'analyse de données ordinales, mais se distinguent par leur objectif. La corrélation des rangs, mesurée notamment par les coefficients de Spearman et de Kendall, cherche à évaluer le degré de dépendance entre deux séries de rangs. Elle indique dans quelle mesure deux classements ordonnent les mêmes objets de façon similaire, en s'appuyant sur la covariance des rangs ou sur le décompte des couples concordants et discordants. La concordance de classements vise quant à elle à mesurer le niveau d'accord global entre plusieurs classements portant sur les mêmes objets. Ainsi, la corrélation des rangs concerne principalement la comparaison deux à deux de classements afin de déterminer leur similarité, tandis que la concordance de classements s'attache à mesurer

l'accord collectif entre plusieurs classements et à tester s'ils expriment une structure hiérarchique commune.

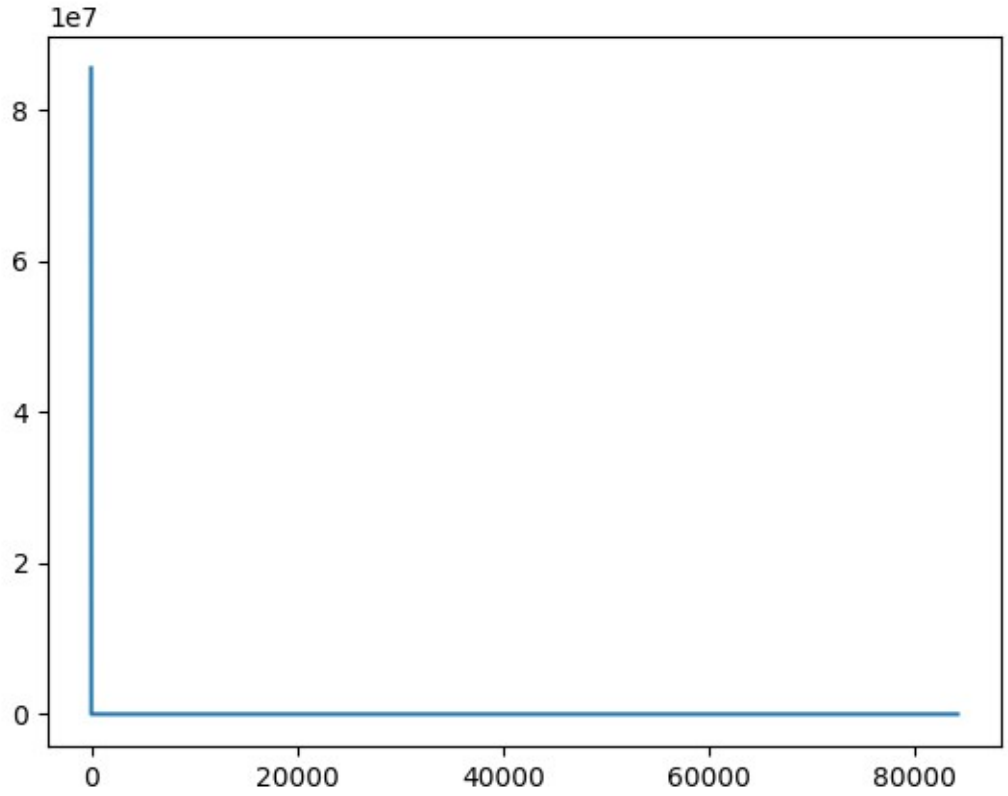
4. **Quelle est la différence entre les tests de Spearman et de Kendall ?**

Les tests de Spearman et de Kendall permettent une comparaison entre classements afin d'évaluer la dépendance entre deux variables ordinales. Le test de Spearman transforme les valeurs en rangs et cherche ensuite à établir une corrélation, fondée sur la covariance entre les rangs : il mesure donc principalement la linéarité monotone entre deux classements. Le test de Kendall repose sur une comparaison de chaque paire d'objets afin de déterminer si les deux classements respectent le même ordre (paires concordantes) ou l'inversent (paires discordantes). Le résultat de ce teste est un coefficient mesurant la proportion de concordance entre les deux ordres.

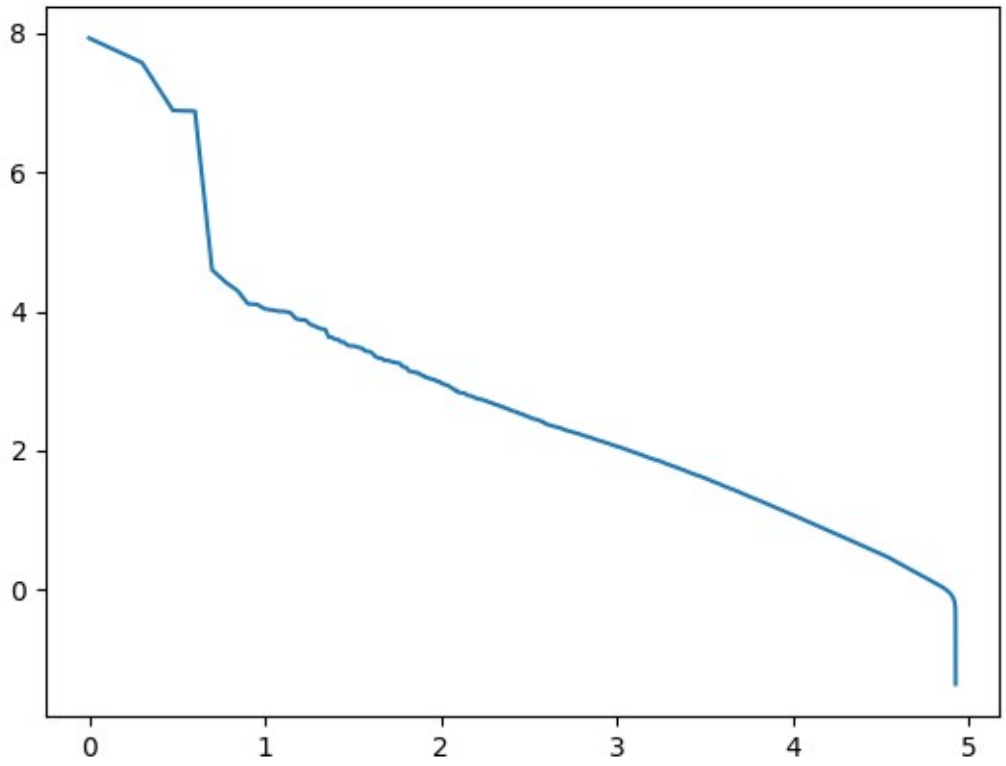
5. **À quoi servent les coefficients de Goodman-Kursdal et de Yule ?**

Le coefficient de Goodman-Kruskal permet de mesurer l'association entre deux variable en quantifiant le « surplus » de paires concordantes par rapport aux paires discordantes. Il est calculé comme une proportion et s'applique aux variables qualitatives ordonnées. Le coefficient de Yule est une sous catégorie du premier destiné à être appliqué aux tableaux de contingence afin d'évaluer la fréquence de deux évènements binaires. Ces deux coefficients ont donc pour fonction principale de mesurer l'intensité et la direction d'une association entre variables qualitatives, en fournissant des indicateurs non paramétriques adaptés à des données qui ne peuvent être traitées par des corrélations classiques.

Résultats obtenus avec le code :



Cet premier graphique représente la loi rang–taille des surfaces des îles en échelle linéaire. La courbe est dominée par les valeurs les plus élevées ce qui provoque un écrasement de l'ensemble des autres valeurs. Ces valeurs élevées sont celles des continents, que l'on a ajouté dans le code. Cette forte dissymétrie rend la lecture de la distribution impossible et empêche toute interprétation de la hiérarchie spatiale des surfaces. On ne peut pas comparer les rangs.



Le second graphique représente la même loi rang–taille après la conversion des axes en logarithme. Cela permet de réduire les écarts relatifs entre les valeurs et de rendre la courbe lisible. La décroissance régulière met en évidence une organisation hiérarchique des surfaces, caractérisée par quelques unités très grandes (les continents ajoutés) et une majorité d'unités de petite taille (les îles), ce qui correspond à une structure spatiale fortement hiérarchisée.

II. Réflexion sur les sciences des données et humanités numériques

Les sciences des données sont selon moi un champ d'étude aujourd'hui essentiel dans les sciences sociales en général. Je vais poursuivre cette réflexion sous un angle géographique, en tant qu'étudiante dans un master GAED. Pour analyser, étudier ou démontrer un phénomène, il est nécessaire de constituer une base de faits recueillis sur le terrain d'étude ou via des sources numériques, qu'il s'agisse de données textuelles, spatiales ou statistiques. Ces données doivent ensuite être traitées, structurées et interprétées pour produire des connaissances. Ainsi, pour analyser efficacement ces jeux de données, un géographe se doit d'avoir des connaissances, même partielles, en science des données, notamment en traitement statistique, en modélisation et en programmation. Ces compétences permettent de répondre à la fois à des questions de terrain et à des problématiques plus larges liées aux sociétés contemporaines. Cette articulation est d'ailleurs également valorisée dans les formations académiques qui combinent sciences humaines et datascience, comme certains masters en Humanités numériques qui intègrent explicitement l'analyse de données et les méthodes computationnelles dans les sciences humaines et sociales.

La géographie, par essence, est une discipline ouverte et interdisciplinaire. Elle se situe au carrefour des sciences humaines, des sciences de la nature et des sciences formelles, et doit pouvoir intégrer des outils mathématiques et numériques pour enrichir ses approches. Refuser d'intégrer ces outils reviendrait à limiter notre capacité à comprendre des phénomènes géographiques complexes, tels que les nouvelles formes d'interactions sociales rendues visibles par des données massives, ou encore le numérique en tant que nouvel espace d'appropriation. Ces nouveaux enjeux peuvent participer à l'élaboration de nouveaux champs de réflexions, notamment en géographie sociale. Dans ce sens, les sciences des données ne doivent pas être perçues comme une simple technique auxiliaire, mais comme une dimension essentielle des démarches analytiques contemporaines.

C'est ici que l'articulation avec les humanités numériques s'avère particulièrement pertinente. Les humanités numériques constituent un champ de recherche et d'enseignement en plein développement dans les sciences humaines et sociales, qui vise à repenser les méthodes traditionnelles à la lumière des technologies numériques et des grandes masses de données qu'elles génèrent. Ce champ inclut l'étude, la gestion et l'analyse de corpus numériques, ainsi que la réflexion sur les formats, les standards et les modalités techniques d'une gestion scientifique des

objets numériques. Il s'agit non seulement d'apprendre à traiter des données structurées ou non structurées, comme des textes, des images ou des métadonnées, mais aussi d'interroger la manière dont ces données sont produites, représentées et utilisées.

Par ailleurs, le rapprochement entre sciences des données et humanités numériques invite à une réflexion critique sur les données elles-mêmes : leur production, leurs biais, leur représentativité et leur usage éthique. Dans un contexte où les technologies numériques façonnent de plus en plus les pratiques sociales et les territoires, il est essentiel de ne pas réduire les phénomènes sociaux à des simples métriques. L'analyse des données doit s'accompagner d'une conscience épistémologique de leurs limites et des enjeux sociaux qu'elles recouvrent. Cette réflexion devrait être intégrée directement dans le champ d'étude des humanités numériques, en questionnant notamment la disponibilité, l'accessibilité et l'interprétation des données numériques.

Ainsi, pour la géographie contemporaine, les sciences des données et les humanités numériques ne devraient pas être des domaines isolés ou annexes ; elles devraient plutôt être intégrées de manière systématique dans les démarches de recherche et d'enseignement. Cette intégration favorise une meilleure compréhension des rapports entre les sociétés et leurs environnements numériques et territoriaux, et permet de renouveler les approches méthodologiques, tout en inscrivant la géographie dans un cadre interdisciplinaire ouvert aux défis conceptuels et techniques du XXI^{e} siècle.

III. Problèmes et difficultés d'apprentissage

Personnellement, je ne connaissais rien au codage avant ce semestre et j'appréhendais ce cours, n'étant pas très à l'aise avec l'informatique. Pour commencer, je ne savais pas comment installer des outils de codage sur mon ordinateur, ni lesquels. J'ai commencée toute seule et j'ai eu beaucoup de mal avec l'installation de docker, malgré avoir regardé des tutoriels sur internet, je ne comprenais pas comment je devais installer python depuis cette application. J'ai donc demandé de l'aide à Zara Huston, qui était dans mon groupe du parcours débutant. Elle m'a vraiment beaucoup aidé tout au long du semestre. Elle a réussi à installer python sans passer par docker sur mon ordinateur et m'a conseillé d'installer VS Code pour visualiser mon code et mes graphiques. C'était une application que je ne connaissais pas et que je n'aurai pas installé si elle ne m'avais pas aidé. Elle m'a aussi installé les bibliothèques pour le code, telle que Panda, et m'a expliqué ce que c'était. J'vais l'impression d'être perdue en démarrant le codage, et je dois avouer que la pédagogie inversée n'a pas été la méthode la plus pertinente pour des gens qui entraient pour la première fois dans le monde du codage avec tout à apprendre. J'ai eu recours à l'IA pour m'aider sur certains points de code qui me dépassaient. Cela m'a tout de même appris à reconnaître mes erreurs de code les plus courante et à les résoudre moi-même par la suite.

C'est aussi Zara qui m'a appris à copier le chemin d'accès du dossier dans le terminal pour que Python trouve les datas. C'est grâce à elle que j'ai pu commencer les exercices de manipulations sur Python, car au début je n'arrivais pas à ouvrir les dossiers de data avec Python. C'est encore une fois Zara qui m'a aidé. Concernant le codage, j'ai trouvé que c'était des exercices assez difficiles pour des débutants. Le premier cours (séance 2) était assez compréhensible pour moi, mais j'ai trouvé la suite assez difficile et dense, notamment avec toutes les formules mathématiques que je ne comprenais pas, étant donné que je n'avais pas de mathématiques depuis la première année de lycée. Avec la pédagogie inversée, j'avais l'impression d'être un peu livrée à moi même face au codage. Zara a tout de même organisé des séances de travail en groupe pour nous aider.

Ce cours m'a permis de découvrir Python et son utilisation. J'ai compris en quoi connaître le code pouvait être d'une très grande aide en analyse géographique, et comment il permet une meilleure interprétation des données. Je pense donc que je poursuivrai cet apprentissage, mais de manière beaucoup plus progressive afin de bien intégrer les bases du codage.