Yanis Chappet
M1 Géopolitique - Geoint

Compte rendu du cours d'analyse de données

Maxime Forriez

Parcours débutant

<u>Séance 2 :</u> Les principes généraux de la statistique

<u>Questions de cours :</u>

Par rapport aux statistiques, la géographie se positionne comme une science avec une base commune aux statistiques en étant également productrice de données mais qui ne fait cependant pas la même analyse. Les statistiques ont alors une place centrale en géographie, car toute information géographique nécessite l'utilisation de méthodes statistiques pour être par la suite analysée géographiquement.

Le hasard en géographie est une grande question qui fait peut faire débat en société. Toutefois, en géographie, le hasard est un des paramètres importants à prendre en compte dans nos analyses en tant que géographes car on peut admettre des schémas généraux mais la géographie ne se veut tout de même pas comme une science prédictive exacte telle que la physique décrit le mouvement des planètes. La géographie donne seulement des probabilités d'occurrence d'un phénomène.

En géographie, il existe deux types d'informations géographiques. Il y a tout d'abord, l'information attributaire qui représente ainsi les données descriptives d'un territoire comme par exemple l'économie. Enfin, il y a l'information géométrique qui représente alors les structures spatiales des objets géographiques.

Au niveau de l'analyse des données, la géographie a quatre besoins fondamentaux. La géographie doit pouvoir s'appuyer sur une production et une collecte de données fiables, des métadonnées et une nomenclature cohérentes, avoir des données sur des phénomènes spatiaux divers afin de pouvoir les comparer puis les expliquer et enfin il est nécessaire de disposer d'outils informatiques capables de traiter cette information géographique.

Il existe de multiples différences entre les statistiques descriptives et explicatives. Les statistiques descriptives permettent de résumer, d'organiser et de visualiser des données brutes afin de dégager des propriétés tandis que les statistiques explicatives, permettent de donner une explication, c'est-à-dire nous faire percevoir le fonctionnement d'un phénomène spatial, d'une variable via d'autres variables.

Les types de visualisation de données en géographie correspondent à quatre types de variables. Il y a tout d'abord, les variables qualitatives nominales, puis les variables qualitatives ordinales, les variables quantitatives continues et enfin les variables quantitatives discrètes. Le choix du type de visualisation se fait en fonction de nos données (quantitative ou qualitative) et en fonction de notre travail de recherche et ce qui doit être démontré ou expliqué grâce à nos données.

En statistiques, il existe trois principales méthodes d'analyse de données. Il y a la méthode descriptive, la méthode explicative et enfin, la méthode est prédictive.

La population statistique peut se définir comme l'ensemble des objets étudiés. L'individu statistique peut se définir comme un élément de la population qui est étudiée. Le caractère est la propriété mesurée pour chaque individu. Enfin, la modalité est la valeur prise par un caractère, un caractère qui peut être de 4 types : qualitatif nominal, qualitatif ordinal (possibilité de hiérarchisation), quantitatif discret, quantitatif continu.

Calculs pour mesurer une amplitude et une densité :

Classe [a, b]

Mesure de l'amplitude : A = b − a

Mesure de la densité : d = ni / (b − a)
ni = effectif de la classe

Les formules de Sturges et Yule servent à trouver le nombre de classes (k) lors de la discrétisation d'une variable continue et d'éviter un mauvais découpage. Voici la formule de Sturges : k ≈ 1 + 3,322 × log10(n). Et la formule de Yule : k ≈ 2,5 × √[4]{n}.

Pour définir un effectif (ni), nous devons avoir le nombre d'individus qui ont la modalité i. Pour la fréquence (fi) : fi = ni / n. Pour la fréquence cumulée jusqu'à k : Fk = Σ (i=1 à k) fi = (1/n) × Σ (i=1 à k) ni. Enfin, pour la distribution statistique, elle représente la loi empirique des données, ce qui correspond à la répartition des effectifs en fonction des modalités d'un caractère.

<u>Analyse des résultats</u>

Lors de cette séance, on utilise un fichier CSV qui regroupe les résultats du premier tour de l'élection présidentielle 2022 au niveau départemental.

Après avoir chargé le fichier dans un DataFrame Pandas, on obtient le tableau suivant sur les dimensions du DataFrame. Le nombre de lignes correspond aux départements , collectivités particulières et DOM (107). Les colonnes (56) contiennent des variables d'informations électorales, démographiques et administratives. Les variables sont à la fois quantitatives avec le nombre d'exprimés mais également qualitatives avec le libellé du département.

L'analyse des types de variables permet de connaître leur nature. Il y a 38 colonnes de type *object*, donc qualitatives et 18 de types numérique qui sont des *float64* ou *int64* et donc quantitatives.

Dimensions du DataFrame

Indicateur	Valeur
Nombre de lignes	107
Nombre de colonnes	56

On obtient alors le tableau suivant, avec le total des inscrits, qui correspond à l'ensemble des électeurs inscrits au 1er tour sur l'ensemble du territoire français qui est de 48.747.876 individus.

Total des inscrits

Indicateur	Valeur
Total des inscrits (France entière)	48 747 876

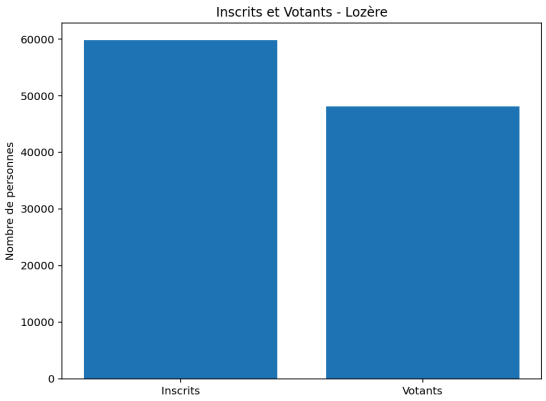
Avec une boucle conditionnelle, on calcule les effectifs sur les colonnes numériques. On obtient alors ce tableau représentant les sommes des variables quantitatives.

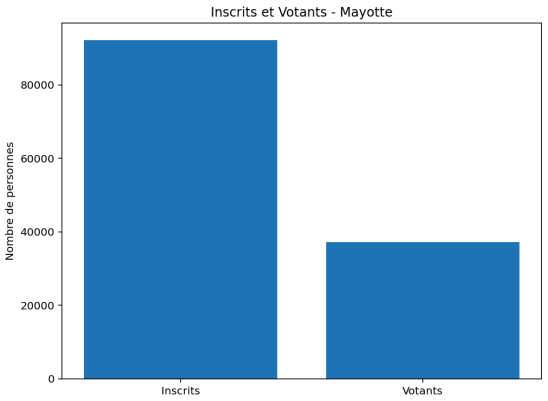
Variable	Somme
Inscrits	48 747 876
Abstentions	12 824 169
Votants	35 923 707
Blancs	543 609
Nuls	247 151
Exprimés	35 132 947
Voix (candidat 1)	197 094
Voix.1	802 422
Voix.2	9 783 058
Voix.3	1 101 387
Voix.4	8 133 828
Voix.5	2 485 226
Voix.6	7 712 520
Voix.7	616 478
Voix.8	1 627 853
Voix.9	1 679 001
Voix.10	268 904
Voix.11	725 176

On peut constater que les valeurs *Inscrits*, *Votants* et *Exprimés* ne sont pas des valeurs aberrantes. Il y a en effet une participation nationale importante avec 36 millions de votants. On constate de plus qu'il y a une abstention assez forte avec 12 millions d'inscrits.

Ensuite, le code permet de créer pour chaque département un diagramme en barres avec comme ordonnée le nombre de personnes et en abscisse le nombre d'inscrits et de votants.

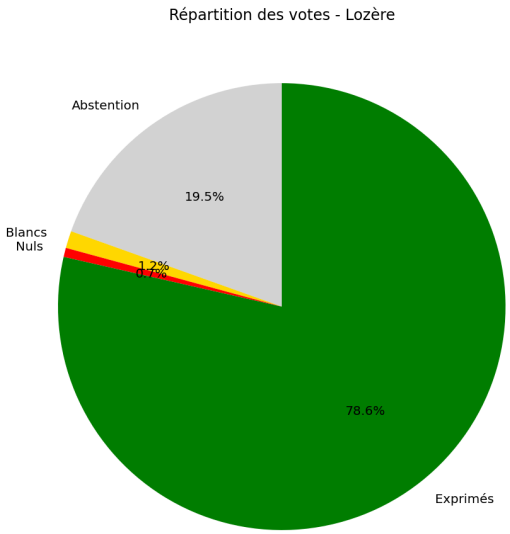
Ces diagrammes permettent de visualiser directement l'écart entre le nombre d'inscrits et celui de votants. Dans le département de la Lozère par exemple, on constate un écart assez faible, même si le nombre de votants reste inférieur au nombre d'inscrits, la proportion de votants correspond à environ ¾ des inscrits, ce qui peut être considéré comme relativement important en comparaison avec d'autres départements où la proportion de votants correspond à moins de la moitié des inscrits comme par exemple dans le département de Mayotte.

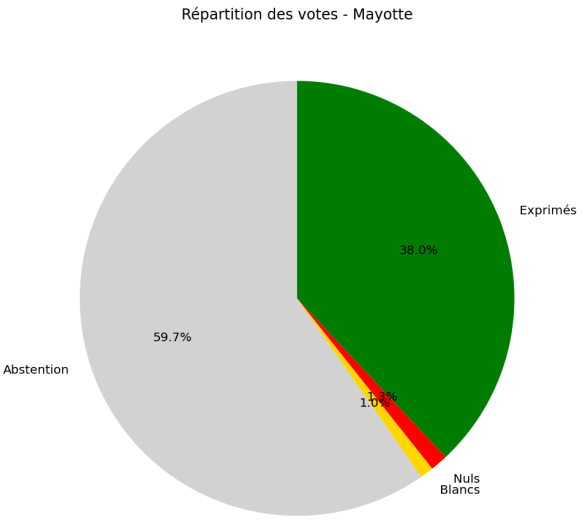




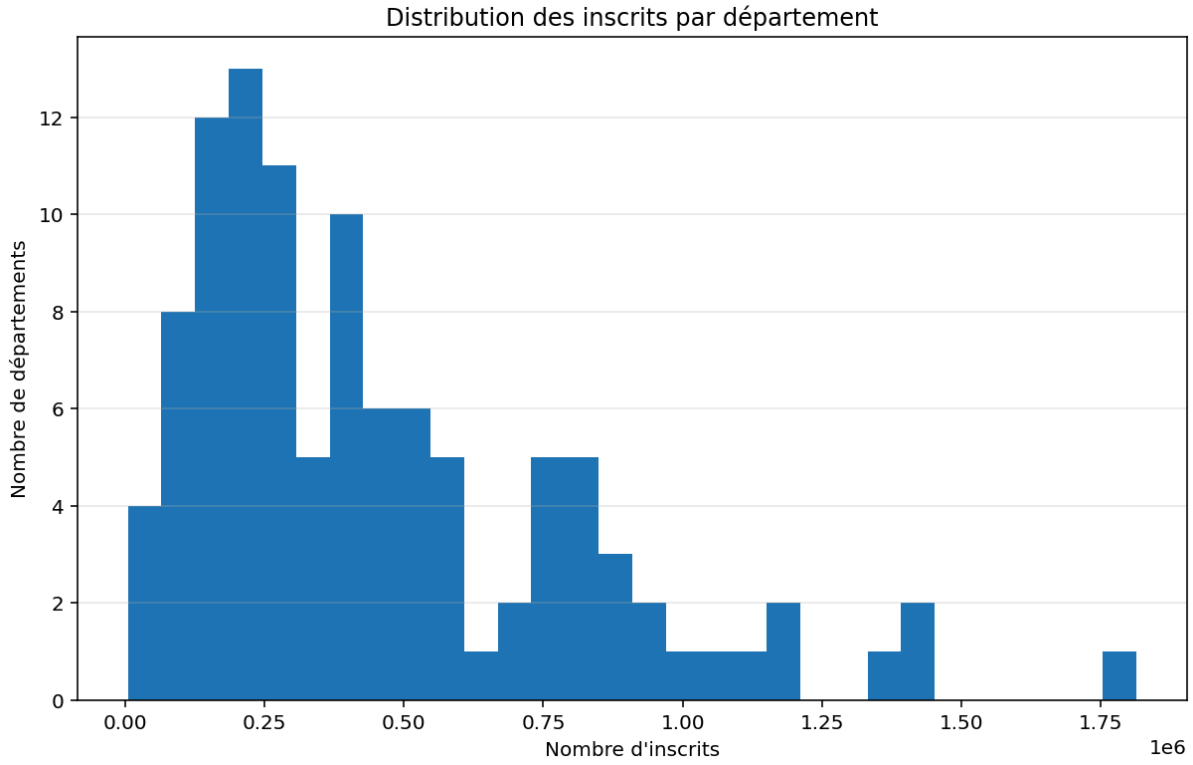
Par la suite, on crée un diagramme circulaire par département en distinguant les bulletins blancs, les bulletins nuls, les suffrages exprimés et l'abstention.

On constate que l'abstention en Lozère est de 19,5% et de 59,7% à Mayotte avec des bulletins blancs et nuls pour environ 2% chacun. Ce qui nous permet ainsi d'approcher de manière plus fine nos résultats et de bien mettre en évidence les proportions pour faciliter les comparaisons entre départements.





Enfin, on crée un histogramme de la distribution des inscrits. Nous pouvons ainsi constater une distribution asymétrique du nombre d'inscrits par départements. En effet, la majorité des départements ont entre 100.000 et 400.000 inscrits. Il y a alors une distribution asymétrique vers la droite avec beaucoup de départements de taille moyenne et quelques départements très peuplés.



<u>Séance 3 :</u> Les paramètres statistiques élémentaires

<u>Questions de cours</u>

Le caractère le plus général est le caractère quantitatif car il permet de mesurer des valeurs numériques et d'effectuer des calculs statistiques.

Les caractères quantitatifs discrets sont des valeurs numériques finies alors que les caractères quantitatifs continus peuvent prendre tous les nombres compris dans un intervalle. Il est nécessaire de les distinguer car ce ne sont pas les mêmes méthodes de calcul et d'interprétations en fonction des caractères.

Il existe plusieurs types de moyenne car chaque type de moyenne à des usages particuliers, certaines moyennes peuvent être tronquées si elles sont sensibles aux valeurs extrêmes et certaines moyennes sont plus adaptées à un type de variable. La médiane est utile à déterminer car elle permet de connaître la valeur située au centre d'une série de données. Le mode se calcule aussi bien avec les variables discrètes, qu'avec les variables continues.

L'intérêt de la médiale est de partager en deux parties égales la totalité des valeurs. L'intérêt de l'indice de Gini est de mesurer la concentration d'une série de données en comparant la médiale à la médiane.

Il est utile de calculer la variance à la place de l'écart à la moyenne car elle quantifie la dispersion des valeurs sans que les écarts positifs et négatifs s'annulent. On peut la remplacer par l'écart type, car elle ramène l'unité à celle des données. Il est intéressant de mesurer l'étendue car cela met en évidence la différence entre la valeur la plus élevée et la valeur la plus basse d'une série et permet d'obtenir un indicateur de dispersion. Les quartiles divisent une série en des catégories avec le même nombre de données et résument la répartition des valeurs. Ceux qui sont les plus utilisés sont les Q1, Q2 (médiane) et Q3. La boîte à moustache permet de visualiser la distribution (médiane, quartiles, valeurs extrêmes) afin de faire des comparaisons.

Les moments centrés permettent de mesurer la dispersion autour de la moyenne, tandis que les moments absolus permettent de déterminer la distance à une valeur donnée. Ils sont utiles car ils permettent de donner la nature d'une forme d'une distribution (symétrique ou dissymétrique). Il est intéressant de vérifier la symétrie d'une distribution pour savoir si les mesures comme la moyenne, la médiane et le mode sont représentatives de la série de données.

On utilise alors le coefficient de dissymétrie : $\beta 1$

→Asymétrie positive : $\beta 1 > 0$

→Asymétrie négative : $\beta 1 < 0$

→Distribution symétrique : $\beta 1 = 0$

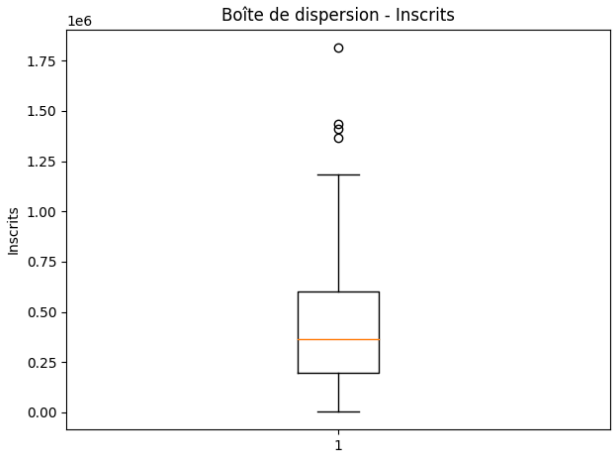
Analyse des résultats

Dans une première partie, on analyse les résultats du premier tour de l'élection présidentielle française de 2022, ensuite on analyse des surfaces d'îles pour les catégoriser en fonction de leur surface.

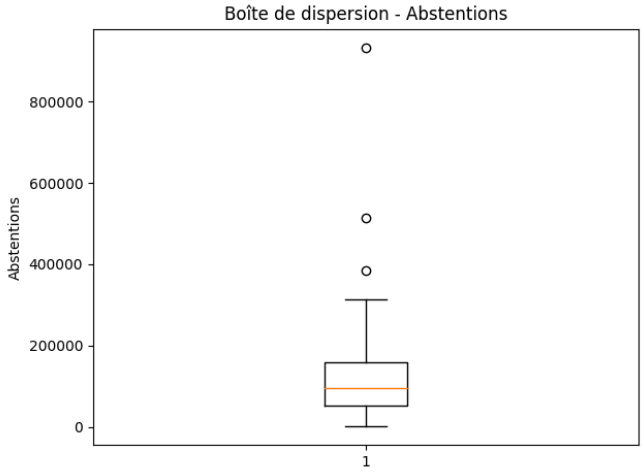
Tout d'abord, pour analyser le premier tour de l'élection présidentielle française de 2022, on va calculer une série de paramètre statistique dont des paramètres de position, tels que la moyenne, la médiane et le mode, des paramètres de dispersion, tels que l'écart-type, l'écart absolu à la moyenne et l'étendue et enfin des paramètres de distribution tels que les quartiles, les déciles, la distance interquartile et la distance interdécile. L'analyse de nos données via ces paramètres nous permet ainsi de mieux comprendre la structure des distributions électorales. Les résultats de notre analyse montrent alors une forte hétérogénéité entre les variables quantitatives étudiées. Les paramètres statistiques calculés traduisent une grande dispersion des résultats selon les communes, notamment en ce qui concerne les inscrits, les votants et les suffrages exprimés.

La représentation en boîte à moustaches permet alors de représenter graphiquement nos résultats. Nous pouvons ainsi visualiser simultanément la médiane, la dispersion, l'étendue des valeurs et la présence éventuelle de valeurs atypiques.

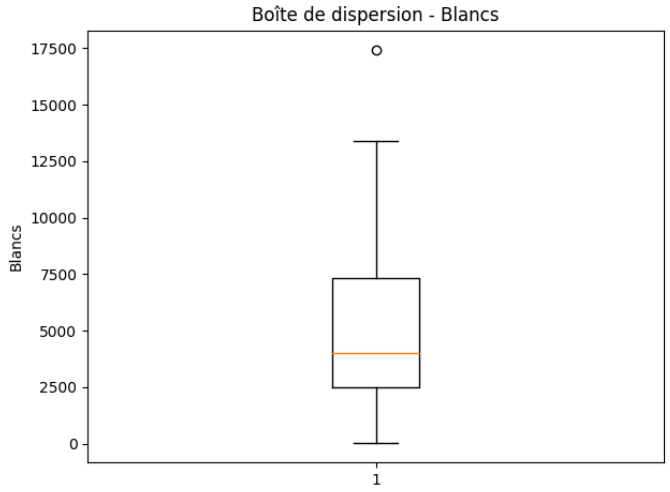
La boîte à moustaches des inscrits révèle une distribution fortement asymétrique. En effet, la majorité des communes ont peu d'inscrits, tandis que quelques grandes communes ont des effectifs très élevés. Cela permet d'exposer les grandes différences entre les tailles de populations des différentes communes.



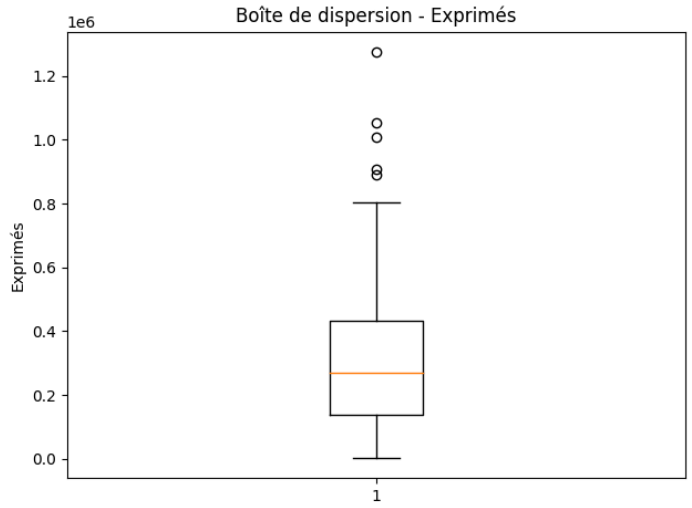
La boîte à moustaches des abstentions montre également une importante dispersion. La médiane est relativement basse, ce qui permet de comprendre que la plupart des communes ont peu d'abstentionnistes. Les valeurs élevées correspondent aux grandes communes qui mécaniquement du fait de leur taille, ont plus d'abstentionnistes.



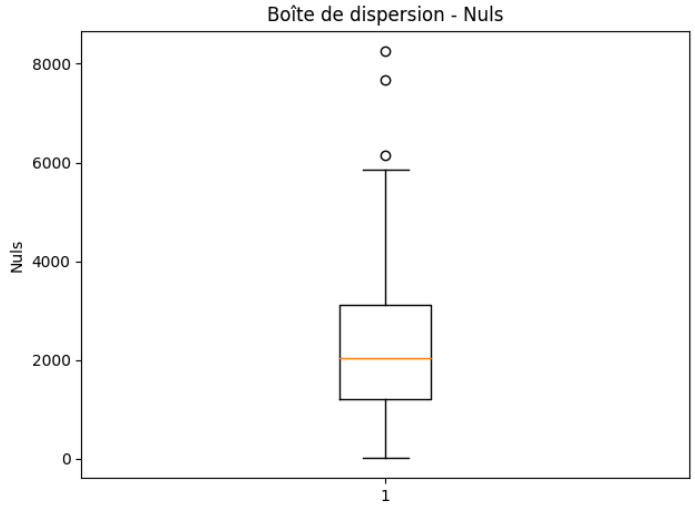
La boîte à moustaches des votes blancs montre une médiane faible, ce qui indique que les votes blancs sont plutôt marginaux dans la plupart des communes.



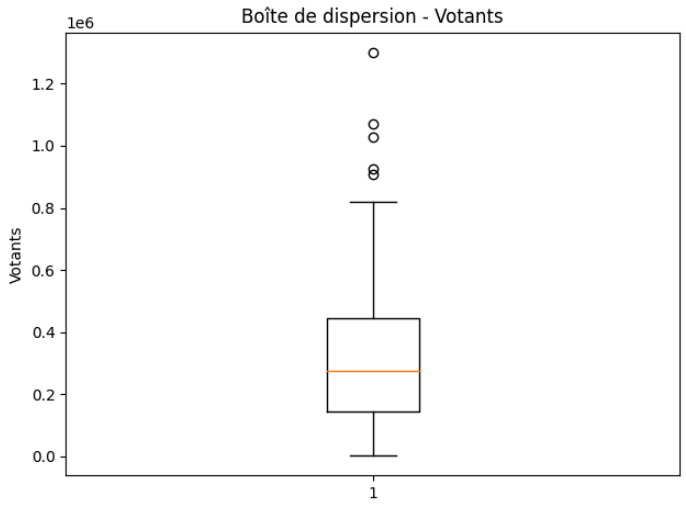
La boîte à moustaches des suffrages exprimés montre une distribution fortement asymétrique. En effet, la plupart des communes ont un nombre limité de votes exprimés tandis que les grandes villes en ont un nombre élevé. On voit ainsi bien qu'il y a une forte hétérogénéité entre les communes.



La boîte à moustache des votes nuls montre une dispersion assez importante avec un large écart interquartile. On y voit alors des valeurs aberrantes élevées, du fait de la présence de grandes villes.

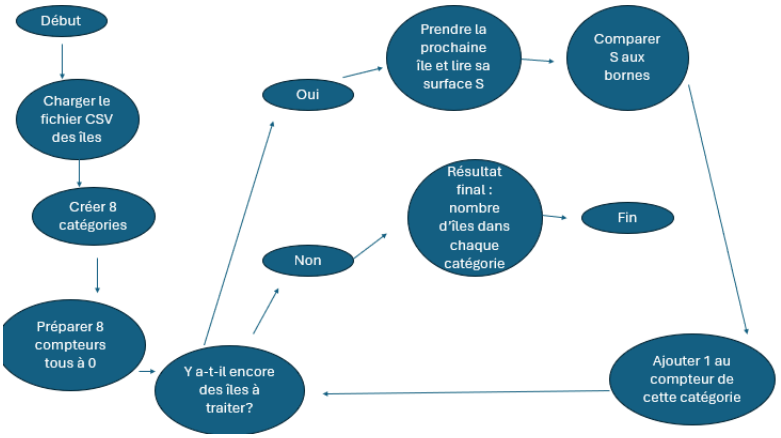


Enfin, la boîte à moustaches des votants révèle une forte hétérogénéité avec une grande dispersion liée à la grande variabilité de la taille des communes.



Dans un second temps, les résultats de notre analyse sur les surfaces d'îles mettent en évidence une forte concentration d'îles de petite taille, tandis que les îles de très grande superficie sont rares. La catégorisation que l'on obtient alors apparaît comme particulièrement utile pour nous en tant que géographe car elle nous permet par exemple d'effectuer par la suite plus simplement un travail de cartographie.

Voici ci-dessous un organigramme qui permet de comprendre la méthode pour catégoriser et dénombrer le nombre d'îles selon leur surface.



On crée dans un premier temps huit catégories pour définir nos boîtes : petites îles de 0-10km² ; moyennes 10-25, 25-50, 50-100 km² ; grandes 100-2500, 2500-5000, 5000-10000 km² ; très grandes : > 10000 km². Ensuite on crée huit compteurs et on met en place une boucle pour laquelle chaque surface d'île est comparée en fonction des catégories, une fois que cette dernière est identifiée on ajoute 1 au compteur de la catégorie. Enfin, une fois que toutes les surfaces d'îles ont été catégorisées, on affiche le nombre d'îles dans chaque catégorie.

Ainsi, les paramètres calculés montrent que les données géographiques et électorales sont rarement symétriques et homogènes, ce qui démontre l'intérêt de l'usage combiné de plusieurs indicateurs de position et de dispersion pour les classer et les analyser.

<u>Séance 4 :</u> Les distributions statistiques

<u>Questions de cours</u>

Pour choisir entre une distribution statistique à variable discrète et une distribution statistique à variable continue, il est nécessaire de s'interroger sur la nature du phénomène étudié. En effet, on choisit une loi discrète seulement quand la variable prend un nombre fini ou dénombrable de valeurs alors qu'on choisit une loi continue quand la variable peut prendre une infinité de valeurs dans un intervalle. Il faut de plus s'interroger sur la forme de la distribution empirique, on regarde comment se distribuent les données observées pour décider de ce qui est le plus adapté entre une distribution statistique à variable discrète et une distribution statistique à variable continue. Les caractéristiques statistiques des données sont également nécessaires à prendre en compte. En effet, les moments mesurés peuvent orienter vers certains types de lois, par exemple une loi symétrique ou asymétrique. Enfin, le nombre de paramètres de la loi est également à prendre en compte. Certaines lois discrètes ou continues ont peu de paramètres, d'autres davantage, ce qui influence le choix du modèle le plus pertinent.

Les lois les plus utilisées en géographie sont selon moi, la loi uniforme discrète car elle est utilisée pour les sondages, ce qui est régulièrement effectué par des travaux en géographie humaine.

La loi de Zipf-Mandelbrot car elle est utile pour les lois rang-taille confrontant, au sein d'un territoire, le nombre d'habitants d'une ville avec son rang. C'est une loi qui apparaît comme essentielle pour les travaux en géographie urbaine.

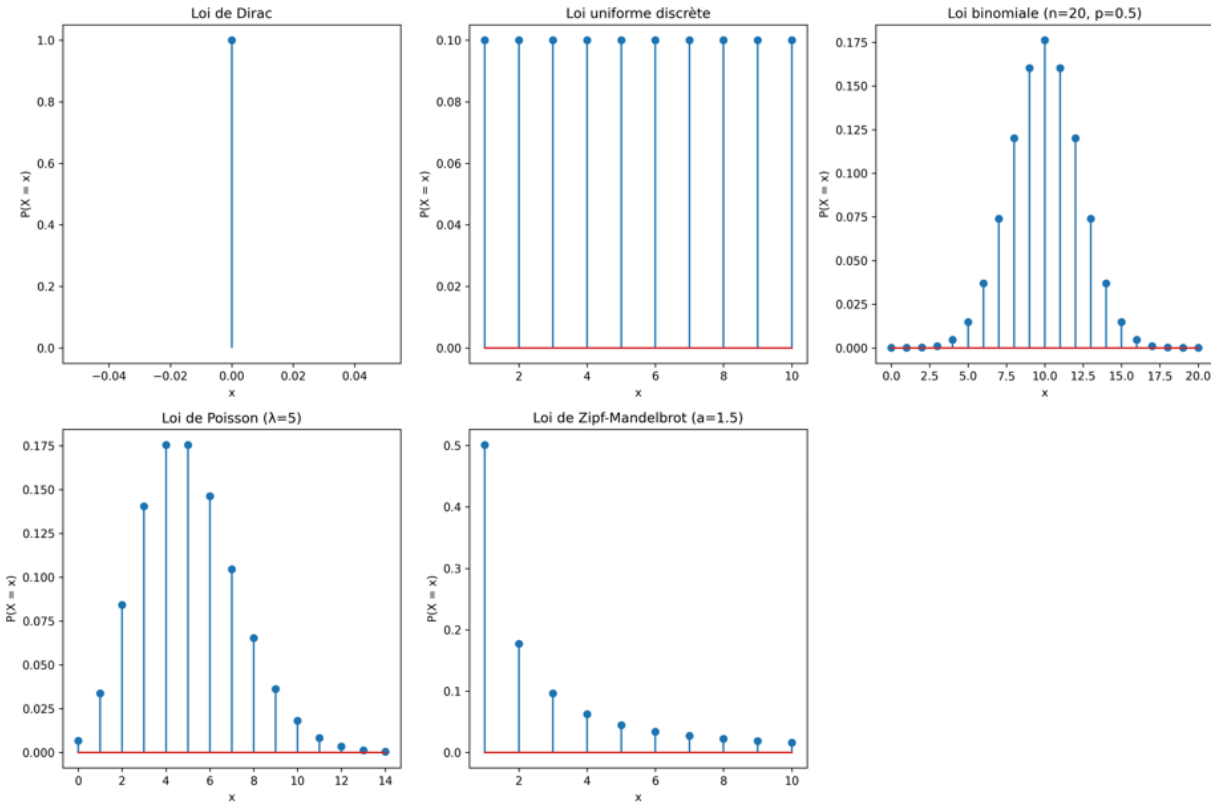
La loi binomiale me semble également souvent utilisée car elle s'applique aux phénomènes ne pouvant prendre que deux états s'excluant mutuellement. Un certain nombre de phénomènes prennent cette forme en géographie, notamment les phénomènes environnementaux ou sociaux.

La loi de Poisson également car elle est utilisée lorsque les événements se produisent dans une succession d'épreuves très nombreuses. Cette loi peut être utile en géographie pour les comptages ou pour la distribution dans l'espace de phénomènes rares.

Enfin, la loi multinomiale me semble également souvent utilisée en géographie car elle permet de modéliser des catégories multiples comme la distribution de populations en classes, catégories socioprofessionnelles, type d'occupation du sol etc.

Analyse des résultats

L'objectif de cette séance est de savoir afficher une distribution statistique afin de comparer une distribution théorique avec une distribution observée. Avec le code, on obtient une série de graphiques que l'on peut alors analyser.



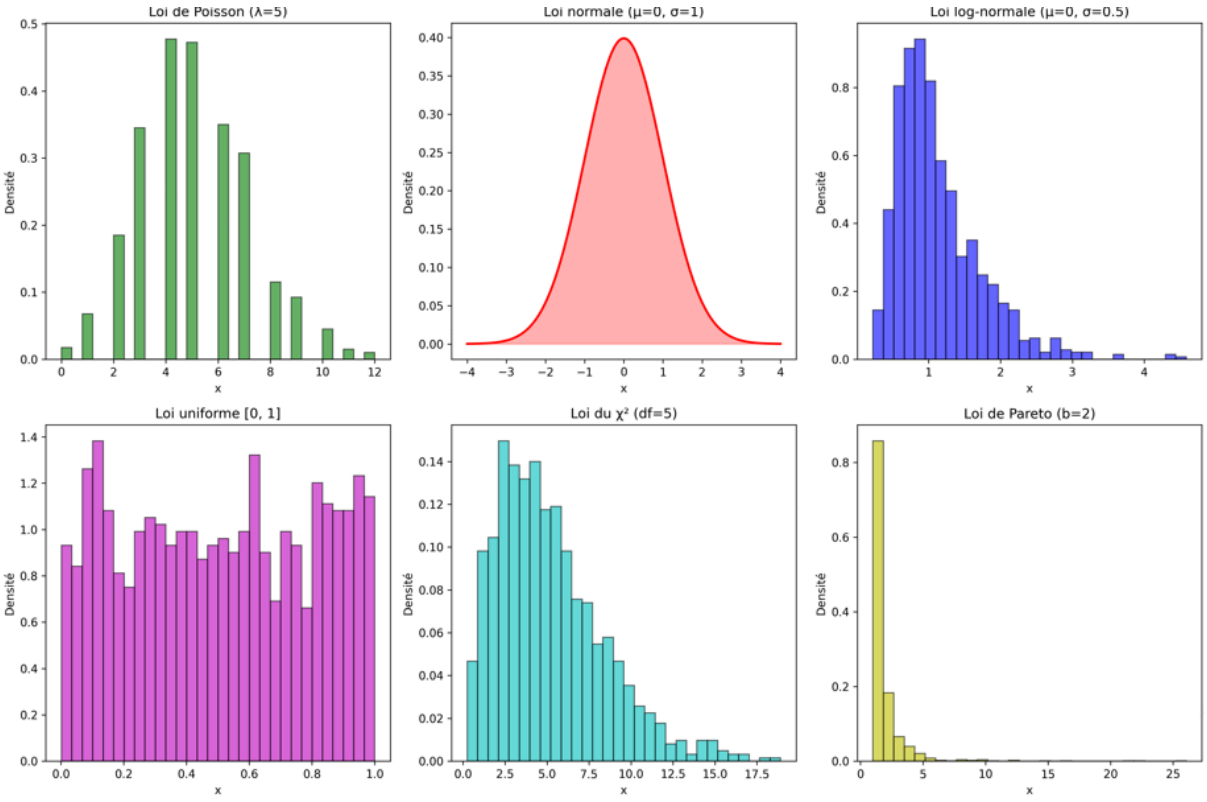
Le premier graphique représente la loi de Dirac, une loi dite "dégénérée", où toute la masse de probabilité est concentrée en un point unique, ce qui est bien visible avec le pic de probabilité sur le graphique qui est égal à 1.

Le deuxième graphique représente la loi uniforme discrète, une loi qui attribue la même probabilité à chaque valeur possible au sein d'un ensemble fini. Nous pouvons bien le voir sur le graphique, avec les barres qui sont de même hauteur, ainsi les résultats ont la même probabilité de se réaliser.

Le troisième graphique représente la loi binomiale. Cette loi modélise le nombre de succès obtenus au sein d'un nombre fini d'essais indépendants, sachant que chaque essai a la même probabilité de succès. Ainsi, chaque barre du graphique représente la probabilité d'obtenir un certain nombre de succès. Les valeurs qui se trouvent au centre du graphique sont les plus élevées car le résultat le plus probable est d'obtenir un nombre de succès proche de la moyenne, tandis que les valeurs aux aux extrémités du graphique sont faibles car la probabilité d'obtenir très peu ou beaucoup de succès est bien plus rare.

Le quatrième graphique représente la loi de Poisson qui décrit le nombre d'événements qui se produisent au sein d'un intervalle donné. Les barres sont plus hautes pour les petites valeurs, ce qui signifie que le nombre d'événements le plus probable est faible.

Le cinquième graphique représente la loi de Zipf-Mandelbrot. Cette loi décrit la distribution de fréquence d'entités dans un ensemble de données. Nous pouvons alors voir une forte inégalité entre les occurrences avec notamment la première valeur qui a une probabilité très élevée tandis que la deuxième a une probabilité beaucoup plus faible, alors même que la probabilité des autres valeurs décroît plus lentement. C'est ainsi le phénomène « de longue traîne », une minorité domine tandis que la majorité est peu représentée.



Le premier graphique représente ici la loi de Poisson continue. Cette loi reprend la loi de Poisson mais les événements se produisent cette fois de manière continue et non pas discrète. Le graphique modélise le nombre d'événements rares survenant dans un intervalle de temps ou d'espace fixe. Ici, la distribution présente un pic autour de la moyenne.

Le deuxième graphique représente la loi normale. Le graphique est en forme de cloche symétrique, centrée autour de la moyenne et elle décroît de manière symétrique de part et d'autre de la courbe. La valeur centrale est la plus fréquente, c'est-à-dire que plus on s'éloigne de cette valeur centrale, plus les valeurs sont rares. Ainsi, la plupart des observations sont proches de la moyenne.

Le troisième graphique représente la loi log-normale, avec le logarithme qui suit une loi normale, comme sur le graphique précédent. On peut ainsi voir que le graphique est asymétrique et étalé vers la droite. En effet, les petites valeurs sont plus fréquentes que les grandes valeurs. Cette loi est utile pour des phénomènes où les valeurs varient sur plusieurs ordres de grandeur.

Le quatrième graphique représente la loi uniforme. Sur le graphique on voit alors bien que chaque valeur au sein d'un intervalle donné a la même probabilité de se produire. La courbe plate indique bien que la probabilité est constante dans l'intervalle donné.

Le cinquième graphique illustre la loi du Khi2 qui est très utile dans les tests statistiques pour comparer les distributions observées et les distributions théoriques. Le test du Khi2 permet alors de tester l'indépendance entre deux variables ou l'adéquation d'un modèle. Sur le graphique la distribution est asymétrique puisqu'elle est centrée autour de la moyenne, tandis que sa queue est étendue vers la droite. Cette loi est très utile en géographie, elle permet par exemple d'établir si la distribution des risques naturels diffère selon les régions.

Le sixième graphique illustre la loi de Pareto. Cette loi modélise des phénomènes où une petite partie de la population possède la majorité des ressources. Cela est bien visible sur le graphique puisqu'il y a une majorité de valeurs élevées concentrées sur un tout petit intervalle. Cette loi est essentielle en géographie humaine pour décrire des inégalités spatiales, notamment pour montrer la répartition des richesses entre les régions par exemple.

Séance 5 : Les statistiques inférentielles.

Questions de cours

L'échantillonnage peut se définir comme le prélèvement dans une population dite mère d'une partie de cette dernière. L'échantillon correspond à une quantité restreinte, un sous-ensemble de la population mère. Nous prenons seulement un échantillon car utiliser l'ensemble des données de la population mère serait trop lourd en termes de traitement de données, cela ne serait pas efficace et coûteux en ressources d'un point de vue matériel.

Il existe deux grandes catégories d'échantillonnage. La première est la méthode aléatoire avec des tirages qui peuvent être sans remise ou avec avec remise, avec des échantillonnage systématique et des sondages aléatoires simples avec un tirage au sort. La seconde est la méthode non aléatoire avec la méthode des quotas, la méthode d'échantillonnage "Monte Carlo".

On choisit notre méthode en fonction de la disponibilité de sondés, du coût des tirages, de la possibilité d'avoir plusieurs échantillons, de la taille de la population ou encore du niveau de précision qui est souhaité.

Un estimateur est relatif à la variable aléatoire, il est construit pour que sa valeur soit la plus proche de la vraie valeur du paramètre. Une estimation correspond à la valeur numérique concrète qui est obtenue via un échantillon observé.

L'intervalle de fluctuation suppose que la proportion théorique (p) est connue, c'est un échantillonnage et non une estimation. La formule est : $[p - z\_C \times \sqrt{(p(1-p)/n)}, p + z\_C \times \sqrt{(p(1-p)/n)}]$. On estime un intervalle de fluctuation asymptomatique avec un seuil de 95% avec comme effectif n. Le but est de pouvoir prendre une décision avec un risque d'erreur relativement à une appartenance ou non de la fréquence observée à l'intervalle. Pour l'intervalle de confiance, la valeur exacte du paramètre dans la population n'est pas connue, car c'est une estimation du paramètre qui est basée sur l'échantillon observé. L'intervalle de confiance a comme point de départ les observations afin d'estimer le paramètre inconnu.

Un biais, ou l'erreur d'estimation, dans la théorie de l'estimation désigne la différence entre l'espérance de l'estimateur (θ') et la valeur à estimer (θ). L'estimateur peut être considéré comme sans biais, si la moyenne des erreurs s'harmonise, ce qui procure la valeur juste. L'estimateur peut être asymptotiquement sans biais si la limite de sa valeur n tend vers l'infini : $\lim(n \to +\infty)\ E(\theta') - \theta = 0$. L' estimateur peut être considéré comme biaisé si l'estimation varie autour de son espérance mathématique et pas autour de la valeur à estimer du paramètre.

Une statistique travaillant sur la population totale est un recensement. Avec l'arrivée de données massives appelé le "big data", et la multiplication d'entreprises visant à le maîtriser et à l'exploiter comme par

exemple la société Palantir, la question d'étudier sur une population quasiment complète peut se poser. Toutefois, le travail à l'échelle de l'ensemble de la population ne peut pas encore se généraliser aujourd'hui en raison de nos nombreuses limites techniques actuelles pour parvenir à gérer autant d'informations afin de proposer des résultats vraiment fiables.

Les enjeux autour du choix d'un estimateur sont d'abord l'absence de biais afin d'obtenir la valeur juste du paramètre. De plus, l'estimateur doit être assez précis pour produire des résultats qui sont stables et qui ne varient pas entre les différents échantillons. Ensuite, l'estimateur doit être efficace en respectant deux conditions qui sont de ne pas avoir de biais et ne pas présenter de grande dispersion. Enfin, l'estimateur doit être exhaustif pour assurer une réelle fiabilité des résultats.

Il existe cinq méthodes d'estimation d'un paramètre. La première est la méthode des moindres carrés qui permet de déterminer des grandeurs pour représenter des valeurs moyennes. La méthode est basée sur la mesure de l'écart entre ce que le modèle prédit et ce qui est observé. La seconde méthode est celle du maximum de vraisemblance. La valeur du paramètre qui donne les données les plus probables est choisie. La troisième méthode est celle de l'intervalle de confiance, elle procure un encadrement avec une probabilité de contenir la vraie valeur. La quatrième méthode de l'intervalle de pari, qui est fondée sur les caractéristiques de la population. Enfin, la cinquième méthode est celle du bootstrap qui vise à créer des faux échantillons en tirant au hasard avec remise.

Pour choisir une méthode, cela dépend de la taille de l'échantillon, de la disponibilité d'informations, de la présence de données aberrantes, de la recherche de robustesse, de la nature des données, ou encore de la connaissance ou non de la loi suivie par la variable.

Il existe quatre types de tests. Les premiers sont les tests spécifiques, les seconds sont les tests paramétriques et les troisièmes sont les tests non paramétriques. Les tests statistiques sont utiles pour comparer des groupes, tester l'indépendance des variables, l'évaluation des hypothèses et enfin pour vérifier si des données correspondent à un modèle théorique.

Pour créer un test, il faut :

1. Formuler problématique
2. Trouver l'hypothèse à tester pour la recherche
3. Définir une hypothèse de travail et une hypothèse neutre
4. Identifier la loi suivie par la statistique si l'hypothèse neutre était vraie
5. Choisir un niveau de rigueur
6. Vérifier les conditions d'application du test
7. Recueillir ou préparer les données

8. Choisir une statistique de test
9. Déterminer les conditions pour que l'hypothèse neutre soit rejetée
10. Calculer la valeur obtenue à partir de l'échantillon
11. Conclure

La statistique inférentielle a en effet plusieurs limites car les tests réagissent à la taille de l'échantillon, les méta-analyses sont ainsi difficiles car les tests sont liés au contexte, les hypothèses neutres sont en grande partie fausses sur le terrain et les p-valeurs sont souvent mal comprises.

Analyse des résultats

L'objectif de notre code est de montrer comment valider ou non, un résultat statistique en tenant compte de l'incertitude liée aux données observées.

Dans un premier temps, le code utilise les données du fichier qui contient les résultats de 100 échantillons simulés issus d'une population mère de 2 185 individus. Sur chaque échantillon a été réalisée une enquête d'opinion avec trois modalités : "Pour", "Contre" et "Sans opinion". Les résultats du calcul des fréquences comparé aux fréquences de la population mère mettent en évidence un écart entre les fréquences réelles et celles issues des échantillons. Toutefois, cet écart est attendu car un échantillon ne représente jamais parfaitement une population mère. L'enjeu est alors surtout de savoir si cet écart est suffisamment petit pour pouvoir considérer l'échantillon comme représentatif. C'est pourquoi on calcule des intervalles de fluctuation à 95% pour chaque fréquence observée. Les résultats sont alors compris dans les intervalles de fluctuations, ce qui permet alors de conclure que les différences observées ne sont pas aberrantes. Il s'agit alors de la variabilité normale liée à cette méthode statistique.

Dans un second temps, on se place dans une situation plus réaliste en sciences humaines et sociales, où la population mère est inconnue et où l'on ne dispose que d'un seul échantillon. À partir de cet échantillon, les fréquences associées sont calculées. Des intervalles de confiance à 95 % sont ensuite construits. Contrairement aux intervalles de fluctuation, ces intervalles ne sont pas utilisés pour comparer un échantillon à une population connue, mais pour estimer un paramètre inconnu de la population mère et déterminer si l'échantillon est alors fiable. Les résultats démontrent que l'échantillon peut être considéré comme fiable même si les fréquences observées ne sont que des estimations des proportions réelles de la population et qu'elles doivent donc toujours être interprétées avec une marge d'erreur.

Enfin, la troisième partie illustre la théorie de la décision. Nos données sont analysées à l'aide du test de Shapiro-Wilk. Ce test permet de déterminer si une distribution suit une loi normale. Les résultats montrent que, pour les deux fichiers, la p-value est inférieure au seuil critique. Ainsi, l'hypothèse de normalité est rejetée dans les deux cas. Les distributions statistiques ne suivent donc pas une loi normale.

Séance 6 : La statistique des variables d'ordres qualitatives.

Questions de cours

Une statistique ordinale est une statistique qui a été appliquée à des variables qualitatives ordinales, des variables qui peuvent ainsi être ordonnées. Elle s'oppose alors à la statistique nominale qui ne peut être ordonnée.

L'ordre croissant est le plus simple afin d'identifier les valeurs extrêmes. Une corrélation des rangs permet de calculer le degré de ressemblance entre deux séries de rangs, tandis que la concordance de classement permet de connaître si plusieurs classements sont cohérents entre eux.

Le test de Spearman est fondé sur la corrélation des rangs en utilisant la différence entre ces derniers. Le test de Kendall est lui basé sur le comptage des paires concordantes et discordantes.

Le coefficient de Goodman-Kruskal varie entre -1 et +1 et permet de calculer le surplus de paires concordantes relativement aux paires discordantes.

Le coefficient de Yule varie également entre -1 et +1 et permet d'évaluer la force de l'association entre 2 modalités binaires.

Analyse des résultats

Le code nous permet de travailler sur les statistiques ordinales avec des variables qualitatives. Le but est de comprendre la nécessité de factoriser son code en une liste de fonctions ou de procédures exécutant une tâche unique.

On commence par l'analyse des surfaces des îles et des continents via la loi rang-taille. On ordonne ainsi de manière croissante la liste.

Pour rendre plus lisible le graphique, on convertit en échelle logarithmique pour mieux visualiser la distribution. La relation entre le rang et la surface tend vers une forme quasi linéaire, caractéristique de la loi rang-taille et mettant en évidence une hiérarchisation des espaces insulaires et continentaux.

Il n'est pas possible d'effectuer de test sur les rangs car les tests de corrélation ou de concordance des rangs sont liés à 2 variables qualitatives ordonnées, nous disposons alors que d'une seule variable qualitative qu'est la surface.

Ensuite on effectue le classement des États selon la population et la densité. On extrait les variables de densité et de population en 2007 et en 2025. On ordonne alors les États par ordre décroissant de la population et de la densité pour chaque année, et on compare le classement entre 2007 et 2025.

Finalement, on applique deux tests de corrélation (coefficient de Spearman p) et un de concordance des rangs (coefficient de Kendall τ). Pour la population de 2007 et celle de 2025, les coefficients obtenus sont alors élevés et positifs, cela permet d'affirmer la forte stabilité du classement des Etats selon leur population. Ainsi, les pays les plus peuplés en 2007 sont majoritairement les plus peuplés en 2025.

Pour la densité de population en 2007 et en 2025, les coefficients sont également positifs mais plus faibles que pour la population. Cela signifie ainsi la stabilité partielle des classements. Les variations de densité sont influencées par des dynamiques plus complexes, ce qui explique ces résultats.

<u>Humanités numériques :</u> Réflexion sur les sciences des données et les humanités numériques.

Les humanités numériques peuvent se comprendre comme l'intégration du numérique dans les sciences humaines et sociales. Le numérique est aujourd'hui incontournable dans notre vie quotidienne. Le numérique s'impose désormais partout peu importe le champ de recherche. Ainsi, même les humanités, c'est-à-dire l'ensemble des disciplines qui ont pour objet principal l'Homme et la société, ne peuvent aujourd'hui plus être pensées comme à part du numérique, tant par la place que prend aujourd'hui le numérique dans nos vies, que par également l'apport que le numérique procure dans la pratique de ces disciplines. En effet, la numérisation des sciences humaines et sociales marque une réelle révolution. Les échanges entre chercheurs sont facilités, encourageant ainsi à la collaboration pour une production de savoirs plus efficace et plus robuste. De plus, le développement de l'automatisation de certaines tâches annonce des gains de temps phénoménaux dans la pratique de ces disciplines. Enfin, la diffusion des savoirs est démultipliée via le numérique ce qui annonce également une potentielle opportunité de rayonnement bien plus grande pour ces disciplines. Ainsi, l'arrivée du numérique marque bien selon moi un réel tournant pour les sciences humaines et sociales.

L'essor du numérique entraîne alors une multiplication des données sans précédent. On assiste alors à un véritable "déluge de données". Des données qui sont alors à haute valeur stratégique, créant ainsi le marché du "big data" évalué à 245,9 milliards de dollars en 2023 et devrait enregistrer un taux de croissance annuel de 15% entre 2024 et 2032. Face à ce déluge de données, le numérique apparaît comme une opportunité majeure, notamment via le développement de l'intelligence artificielle qui permet déjà des capacités de traitements de données de masses en des temps extrêmement réduits. Ainsi, les humanités doivent selon moi se saisir pleinement du tournant apporté par le numérique et l'explosion actuelle de données pour continuer à exercer un rôle majeur dans la production de données scientifiques.

Pour moi, le développement des pratiques numériques dans mon champ d'étude m'apparaît comme évident. Aujourd'hui on ne peut plus faire de la géopolitique et de l'analyse géospatiale sans utiliser des outils de traitement numérique. Il est donc essentiel pour moi de toujours lier mes connaissances en sciences humaines et sociales à des connaissances dans la sphère numérique. Aujourd'hui rien que pour faire des cartes, il est nécessaire de maîtriser des outils de SIG (Systèmes d'Information Géographique) tels que ArcGIS Pro. C'est pourquoi, dès ma première année de Master, je suis formé à l'utilisation de tels outils numériques. Ma spécialisation en GEOINT, qui se définit comme de la fusion de données géolocalisées multi-sources et multi-capteurs, démontre qu'une analyse géopolitique performante de nos jours, se base sur une synthétisation d'une grande quantité de données très diverses qui doivent être fusionnées et synthétisées sur des supports

cartographiques afin de pouvoir être exploitées pour de l'aide à la décision. Ainsi, le numérique est un outil central dans toutes les étapes de ce processus. Il faut ainsi constamment manipuler de vastes ensembles de données, les traiter puis les exploiter, ce qui est grandement facilité par les outils numériques. Les bases de données sont nettoyées et structurées de manière bien plus efficace via un traitement automatisé par ordinateur. La fusion de ces données est alors bien plus rapide en passant par des logiciels de SIG. Enfin, l'exploitation d'un grand nombre d'informations se développe de plus en plus avec l'arrivée de l'IA qui vient désormais assister l'Homme dans sa prise de décision. Ainsi, je vois donc bien tout l'intérêt d'inclure le numérique dans mes études en sciences humaines et sociales. La pratique de Python me permet déjà d'être bien plus à l'aise dans la gestion de grandes bases de données à traiter.

Enfin, par mon projet professionnel qui est de m'engager dans les forces armées en tant qu'officier spécialiste expert géographe, la pratique des humanités numériques est alors au cœur du métier. En effet, la fonction consiste à exploiter et analyser des données issues de capteurs divers pour produire le renseignement nécessaire au bon déroulement des opérations. Ainsi, le numérique est l'outil central dans l'analyse et la donnée peut alors être perçue comme une véritable arme déterminante dans l'issue d'un affrontement. Sur le champ de bataille où chaque seconde compte, l'apport du numérique et notamment de l'IA permet de raccourcir la boucle décisionnelle en démultipliant l'efficacité de traitement et d'exploitation des données dans le but d'agir plus vite que l'adversaire. C'est pourquoi la pratique de disciplines comme la géographie et la géopolitique est désormais indissociable de l'utilisation du numérique. Les sciences humaines et sociales comme toutes les autres sciences sont désormais des sciences qui sont également numériques.

Toutefois, cette évolution rapide impose tout de même une certaine vigilance. Des problèmes éthiques apparaissent et nécessitent d'être pris en compte. La place de l'automatisation avec le risque de sortir l'Homme de la boucle doit être un véritable sujet de réflexion pour déterminer jusqu'où l'automatisation de tâches peut aller. L'augmentation continue du volume de données nécessite également de revoir en profondeur nos méthodes de traitement afin de ne pas être submergé par toutes ces données et continuer à exploiter efficacement les données. Ainsi l'intégration du numérique dans les humanités nécessite tout de même une certaine vigilance afin que le numérique reste un outil d'appui en démultipliant nos capacités de traitement pour une meilleure exploitation dans nos analyses.

C'est pourquoi il m'apparaît aujourd'hui véritablement essentiel de me former dès à présent à la science des données et aux outils numériques afin de pouvoir continuer à faire pleinement, aujourd'hui et demain, des sciences humaines et sociales qui soient utiles pour la société.