CHASSAING Eva
Sorbonne Université – M1 Géopolitique-Geoint

Compte-rendu des séances du cours d'analyse de données
Parcours débutant (séances 2 à 6)

Maxime Forriez

Séance 2 : Les principes généraux de la statistique

I) Questions de cours

1. La géographie a souvent méprisé et sous-estimé l'apport des analyses statistiques car cela n'entre pas dans son champ disciplinaire. Pourtant, la géographie produit des données massives qui nécessitent des outils statistiques. Les relations entre les deux disciplines sont donc souvent tendues et complexes. Cependant, certains géographes sont aussi des statisticiens comme Marchand, Béguin ou encore Dumolard. Aujourd'hui, l'analyse statistique est indispensable pour traiter l'information géographique et faire de la géographie une science à part entière, notamment grâce à l'analyse spatiale et aux modèles socio-spatiaux.

2. Beaucoup de géographes affirmaient que le hasard était à l'origine de toute chose, ce qui faisait que la géographie ne pouvait pas être une science proprement dite. Cependant, les spatialises de l'école de l'analyse spatiale s'opposaient à cette thèse. Ils ont ainsi construit des modèles spatiaux ou socio-spatiaux. Ainsi, en géographie, même si le hasard est admis, on peut tout de même dégager des tendances globales grâce aux statistiques.

3. L'information géographique se décompose en deux séries statistiques possibles. Tout d'abord, les attributs qui correspondent aux caractéristiques des territoires (population humaine, volume des précipitations...). Ensuite, les données géométriques qui correspondent aux structures spatiales des ensembles géographiques délimités, comme par exemple la forme des territoires.

4. Les besoins de la géographie au niveau de l'analyse de données sont premièrement la production et la collecte de données fiables, c'est-à-dire l'utilisation de nomenclatures et de métadonnées pour structurer l'information. Le deuxième besoin est l'analyse des données qui permet d'étudier les données en les confrontant aux connaissances du phénomène. Enfin le troisième besoin de la géographie au niveau de l'analyse des données est la visualisation et la classification de celles-ci via des outils statistiques ou informatiques.

5. La statistique descriptive permet de résumer et de visualiser les données, comme la moyenne ou l'écart type. La statistique explicative permet d'expliquer une variable via des variables dites « explicatives ». Elle permet ainsi de modéliser des relations entre les variables.

6. Il existe plusieurs types de visualisation de données en géographie qui correspondent aux différents types de variables. Les variables quantitatives sont souvent représentées par des histogrammes, des boîtes à moustache ou des diagrammes en bâton tandis que les variables qualitatives sont souvent représentées par des diagrammes en secteurs ou par des diagrammes à rectangles.

Pour choisir les types de visualisation de données en géographie, il faut donc tout d'abord les choisir en fonction de la variable (si elle est quantitative ou qualitative) et du but de l'analyse (comparaison, distribution, relation…). Par exemple, une boîte à moustache est idéale pour analyser les distributions.

7. Il y a trois méthodes d'analyse de données possibles. La première méthode est la méthode descriptive (CAH, AFM, AFDM, ACP, AFC). La deuxième méthode est la méthode explicative qui

est utilisée pour la régression linéaire et logistique, l'analyse discriminante et l'analyse de variance. La dernière méthode est la méthode prédictive qui sert principalement à analyser des séries temporelles.

8. La population statistique peut se définir comme l'ensemble des individus étudiés, comme par exemple tous les élèves du M1 de Géopolitique-Geoint à Sorbonne Université. L'individu statistique désigne un élément de la population, si on reprend notre exemple, un élève au sein de la classe du M1 de Géopolitique-Geoint. Les caractères statistiques correspondent aux propriétés des individus, par exemple l'âge ou l'attitude d'un élève. Enfin, les modalités statistiques sont les valeurs qui sont prises par un caractère, par exemple « homme » ou « femme » pour le caractère « sexe ».

Le caractère peut être de quatre types : qualitatif nominal (catégorie sans ordre), qualitatif ordinal (possibilité d'effectuer une hiérarchie), quantitatif discret (valeurs isolées), quantitatif continu (valeurs au sein d'un intervalle donné). Il n'existe pas de hiérarchie à proprement parler entre les différents caractères, mais on peut mentionner que les variables quantitatives sont plus faciles à traiter directement que les variables qualitatives, qu'il faut souvent convertir en données quantitatives pour pouvoir les analyser efficacement.

9. Si l'on définit une classe [a, b], l'amplitude se mesure en effectuant le calcul A= b -a. Si l'effectif de la classe est ni, alors on mesure la densité en effectuant le rapport entre l'effectif de la classe et son amplitude, c'est-à-dire en faisant le calcul d = ni / (b-a).

10. Les formules de Sturges et de Yule servent toutes les deux à choisir le bon nombre de classes pour regrouper des données, par exemple dans un histogramme ou un tableau statistique. Cela permet d'éviter de créer trop de classes, car les données seraient alors trop dispersées, ne permettant pas de dégager les grandes tendances, ou au contraire de ne pas en créer assez, car les données seraient alors trop regroupées et l'on perdrait des détails importants pour l'analyse. Les deux formules sont cependant un peu différente l'une de l'autre puisque la formule de Struges donne un nombre de classes souvent un peu plus petit que la formule de Yule qui est très utile quand les données à étudier sont très variées. Ainsi, ces formules permettent de découper adéquatement les données pour qu'elles soient compréhensibles.

11. Un effectif peut être défini comme le nombre de fois où une catégorie apparaît dans les données, par exemple le nombre de chemises rouges dans une armoire. La fréquence correspond à la proportion d'une valeur par rapport au total. Elle se calcule en divisant l'effectif total de l'effectif de la valeur. Si l'on reprend notre exemple, le nombre de chemises rouges par rapport au nombre total de chemises. La fréquence cumulée indique la somme des fréquences au fur et à mesure qu'on parcourt des valeurs jusqu'à une certaine modalité. Enfin, la distribution statistique correspond à la répartition de toutes les valeurs dans les données, en prenant en compte à la fois leur effectif et leur fréquence.

II) Mise en œuvre avec Python

Analyse du code

Le code analyse les résultats du premier tour de l'élection présidentielle française de 2022 à partir d'un fichier CSV qui contient les résultats de ces élections département par département. Le code

permet dans un premier temps d'explorer les données, d'abord en affichant les 5 premières lignes du tableau et en comptant le nombre de lignes (107 qui correspondent aux départements) et de colonnes (56 qui correspondent aux variables) puis en identifiant les types de données (float, int, object) et en indiquant le nombre de personnes inscrites pour voter, au nombre de 48 747 876.

Le code calcule ensuite la somme des valeurs pour chaque colonne numérique (nombre total de votants, de blancs, de nuls…). Enfin, il génère trois types de graphiques pour chaque département : un graphique en barre comparant le nombre d'inscrits et le nombre de votant, des diagrammes circulaires pour les votes blancs, les votes nuls, les votes exprimés et les absentions pour chaque département et enfin un histogramme qui montre la distribution du nombre d'inscrits dans tous les départements.

Analyse des résultats

Analysons désormais les graphiques obtenus. Tout d'abord, le code a permis de générer des digrammes en barre. Pour faciliter l'analyse, et parce que les dynamiques sont globalement similaires sur tout le territoire français, nous ne prendrons l'exemple que d'un seul département, le Puy-de-Dôme (63).

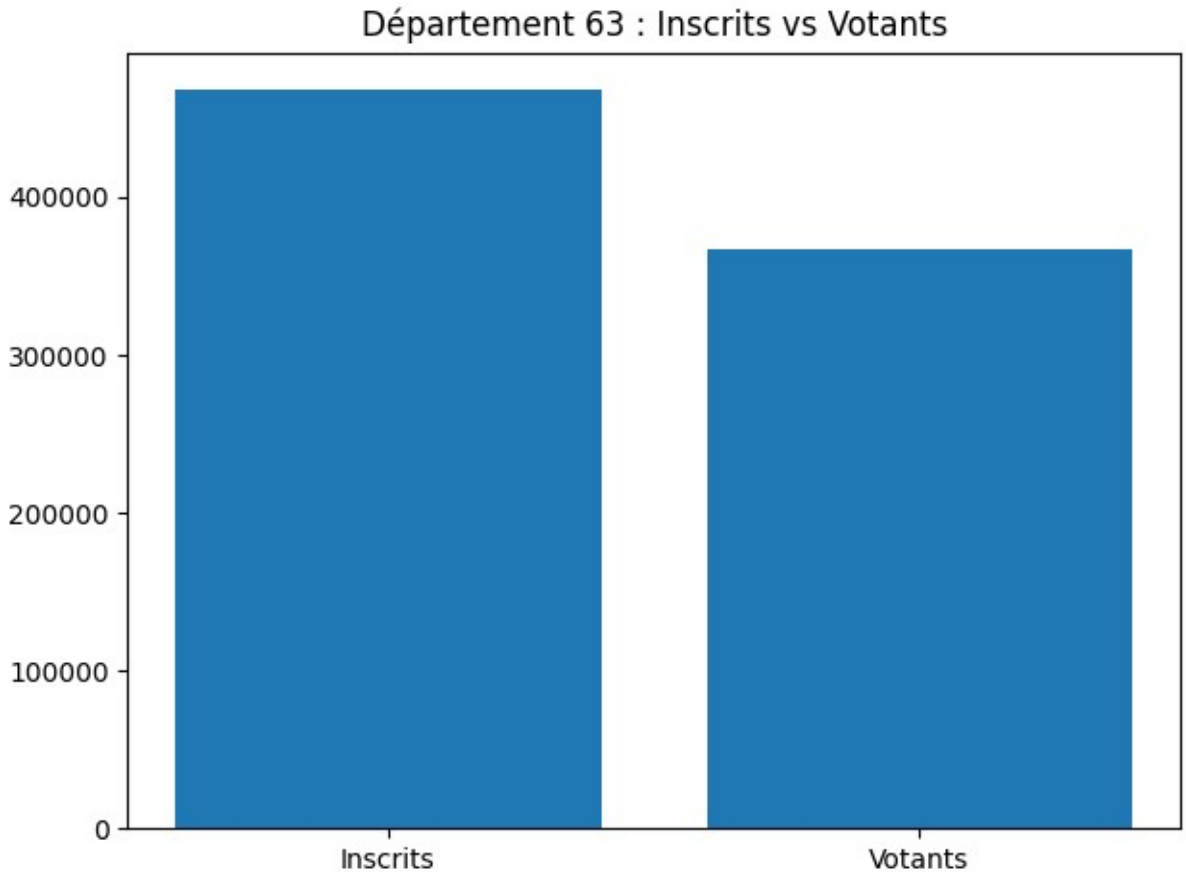


FIGURE 1 – Nombre d'inscrits VS nombre de votants dans le Puy-de-Dôme

Le digramme ci-dessus (figure 1) permet de visualiser l'écart entre le nombre d'inscrits au premier tour de l'élection présidentielle de 2022 et le nombre de votants. Dans le cas du département du Puy-de-Dôme, on peut constater un écart entre le nombre d'inscrits (autour de 50 000 personnes) et le nombre de votants (autour de 36 000 personnes). Ce diagramme montre qu'il y a un écart

important entre le nombre d'inscrits et le nombre de votants, qui correspond à une dynamique d'abstention qu'on retrouve sur tout le territoire. Cela est révélateur d'un certain désabusement de la population française, pour qui le vote a perdu son sens car il paraît ne pas avoir de réelles répercussions sur leur quotidien.

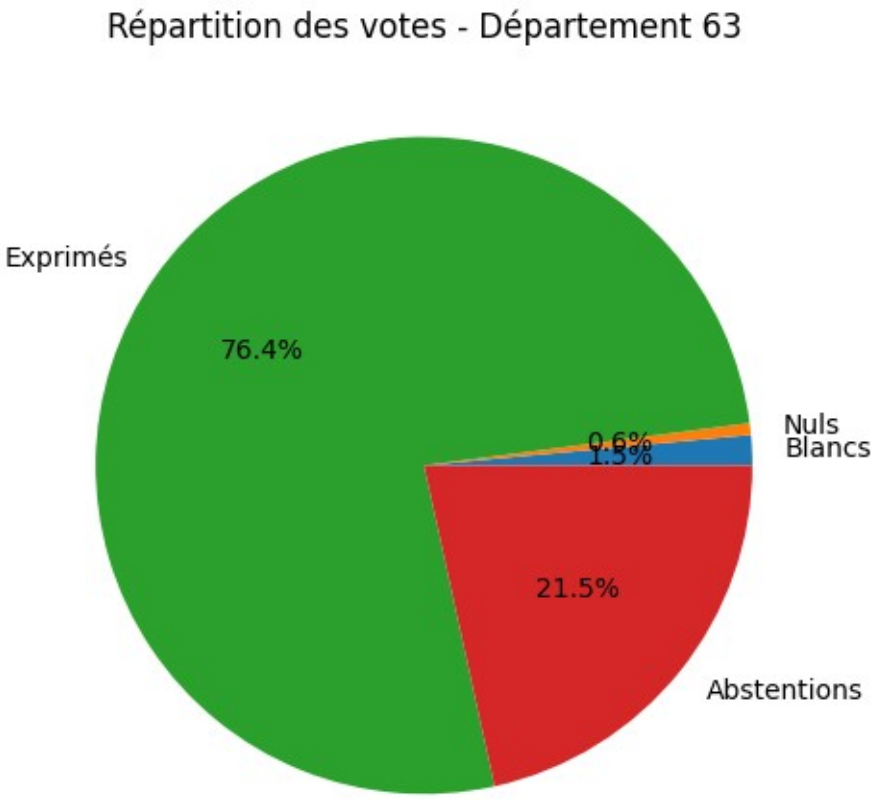


FIGURE 2 – Répartition des votes dans le département du Puy-de-Dôme

Ensuite, le code a généré des digrammes en barre. Comme au-dessus, pour faciliter l'analyse, et parce que les dynamiques sont globalement similaires sur tout le territoire français, nous ne prendrons l'exemple que d'un seul département, le Puy-de-Dôme (63).

Le diagramme circulaire ci-dessus (figure 2) permet de visualiser très facilement la répartition des votes : votes exprimés, abstentions, votes blancs, votes nuls. On peut constater que le nombre de votes exprimés est de 76,4 % contre 21,5 % d'abstentions, 1,5 % de votes blancs et 0,6 % de votes nuls. Si les votes blancs et nuls, votes contestataires, restent à la marge, l'abstention est très forte, puisqu'elle est proche du quart des inscrits. Ce diagramme circulaire possède une dimension visuelle très forte qui est accentué par les couleurs choisies : vert pour les votes exprimés et rouge pour les abstentions, faisant particulièrement ressortir l'importance de ce phénomène.

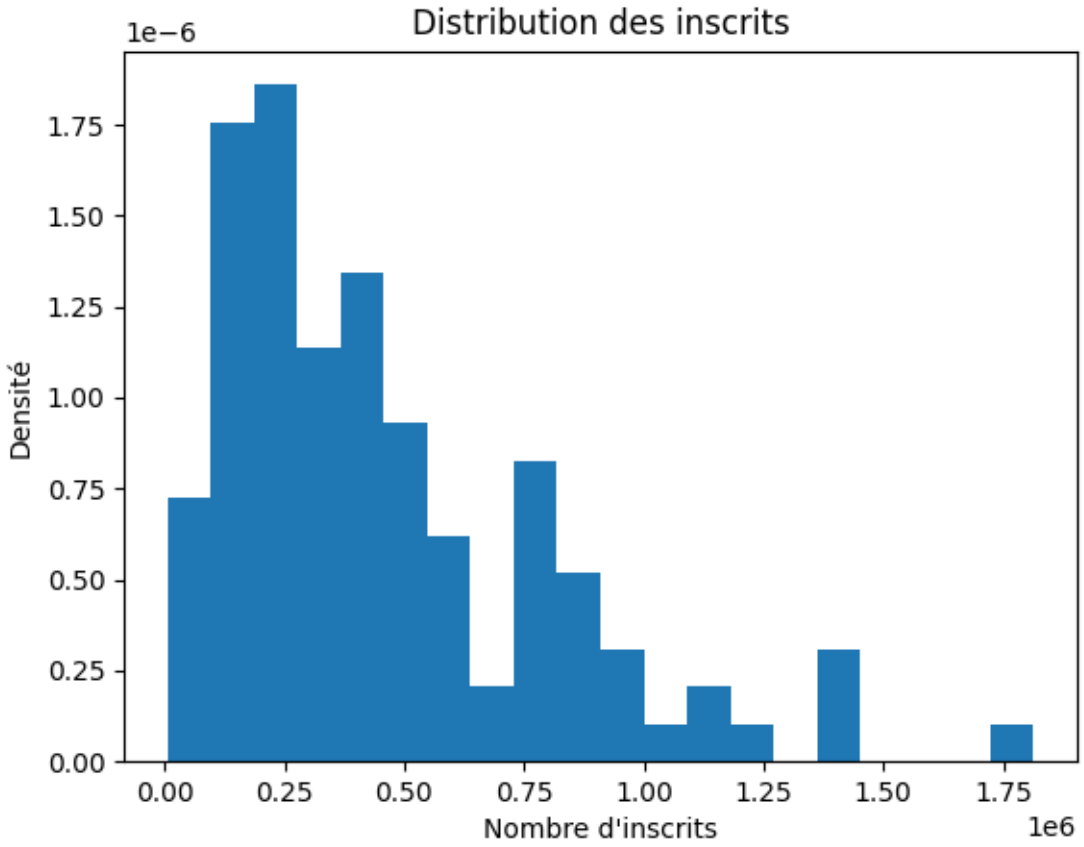


FIGURE 3 – Distribution des inscrits

Enfin, le code a généré un histogramme de densité représentant la distribution des inscrits au premier tour des élections présidentielles françaises de 2022. Il indique comment se répartissent les départements français selon leur nombre d'inscrits. On peut constater une distribution asymétrique. En effet, il y a une forte concentration à gauche, ce qui signifie que la majorité des département ont entre 0 et 500 000 inscrits avec un pic de 200 000 à 300 000 inscrits. Au contraire, à droite, quelques départements isolés ont un nombre d'inscrits beaucoup plus important, avec des pics à 1,5-1,8 millions d'inscrits. Ainsi, l'histogramme révèle une majorité de petits départements et quelques département très peuplés.

Séance 3 : Les paramètres statistiques élémentaires

I) Question de cours

1. Le caractère le plus général est le caractère quantitatif. En effet, les paramètres statistiques concernent principalement les variables quantitatives, et ponctuellement qualitatives.

2. Les caractères quantitatifs discrets prennent des valeurs isolées et qui sont dénombrables, comme par exemple le nombre d'enfants ou les notes obtenues. Les caractères quantitatifs continus prennent toutes les valeurs possibles au sein d'un intervalle. C'est le cas par exemple pour la taille ou le poids. Il est important de les distinguer pour plusieurs raisons. D'abord, les méthodes de calculs différent : sommes pour les discrets et intégrales pour les continues. De plus, la médiane n'existe pas toujours pour une variable discrète, d'où l'intérêt d'utiliser alors une variable continue.

3. Il existe plusieurs types de moyennes car chaque type de moyenne répond à un usage spécifique. En effet, certaines moyennes sont plus adaptées à un type de variable. Ainsi, la moyenne arithmétique est la moyenne la plus souvent utilisée, mais les valeurs extrêmes peuvent fortement influencer la moyenne, en la tirant vers le haut ou vers le bas. La moyenne harmonique est utilisée pour les vitesses moyennes et les taux. La moyenne géométrique est utilisée pour les taux de croissance et les proportions. La moyenne quadratique est utilisée pour les phénomènes physiques tels que les aires et les énergies.

La médiane est particulièrement utile à calculer, car elle permet de contourner le problème de la moyenne arithmétique, puisque la médiane n'est pas influencée par les valeurs extrêmes et est donc aussi moins affectée par les valeurs aberrantes. Elle permet de partager la population en deux groupes de taille égale, et résume bien les distributions asymétriques.

Le mode peut être calculé pour les variables discrètes (valeur la plus fréquente) et pour les variables continues (la valeur correspond à la densité maximale).

4. La médiale permet de partager la masse totale de la variable en deux parties égales, ce qui la diffère de la médiane qui elle partage les effectifs.
L'indice de C. Gini permet de mesurer la concentration. Plus la courbe de Gini s'éloigne de la diagonale, plus la concentration est forte. En géographie, cet indice est très souvent utilisé pour visualiser les inégalités de distribution, comme pour les salaires.

5. Si on calcule l'écart à la moyenne, certains écarts sont positifs tandis que d'autres sont négatifs. Or si on fait la moyenne de ces écarts, ils s'annulent et l'on obtient alors 0, même si les données sont dispersées. Donc l'écart à la moyenne ne permet pas de mesurer la dispersion. C'est pourquoi on utilise la variance, qui permet d'éviter cette annulation en mettant les écarts au carré, rendant tous les écarts positifs. A partir de là, on peut mesurer à quel point les valeurs sont éloignées de la moyenne. Le problème est que la variance est exprimée en unités au carré, or cela est difficile à interpréter.

C’est là qu’intervient l’écart type, qui consiste à calculer la racine carré de la variance. Les résultats sont alors de la même unité que les données initiales étudiées, ce qui permet de les interpréter beaucoup plus facilement.

Calculer l’étendue est très utile car elle permet d’obtenir un indicateur de dispersion, puisqu’elle se calcule en soustrayant la valeur minimale à la valeur maximale.

Créer un quantile partage une série en un certain nombre de parties égales, permettant de répartir les valeurs. Les quantiles les plus utilisés sont les quartiles, qui partagent la série en 4 parties égales. Le Q2 correspond à la médiane et l’écart interquartile (Q3-Q1) permet de facilement connaître les données centrales. Les quartiles sont par exemple fréquemment utilisés en statistiques, dans les boxplot notamment. Les centiles peuvent aussi être utile quand la population est importante.

Il est très intéressant de construire une boîte de dispersion car elle permet de visualiser rapidement la répartition des données et donne des indicateurs clés (quartiles, valeurs extrêmes, médiane…). L’interprétation est assez simple et intuitive. Le rectangle (Q1 à Q3) représente 50 % des données centrales et le trait à l’intérieur de la boîte correspond à la médiane. Les « moustaches » correspondent aux valeurs minimales et maximales. Plus la boîte est large, plus la dispersion est grande.

6. Les moments centrés permettent de mesurer la dispersion autour de la moyenne tandis que les moments absolus utilisent la valeur absolue et permettent de déterminer la distance à une valeur donnée. Ils peuvent être utiles pour caractériser une distribution et calculer les paramètres de forme (asymétrie, aplatissement).
Il est intéressant de vérifier la symétrie d’une distribution pour savoir si les mesures comme la moyenne, la médiane ou le mode sont représentatives de la série de données. Pour vérifier la symétrie d’une distribution, il faut utiliser les coefficients de Pearson et Fisher. Pour le coefficient d’asymétrie (β_1), si $\beta_1 > 0$, la distribution est étalée à droite (asymétrie positive), si $\beta_1 < 0$, la distribution est étalée à gauche (asymétrie négative) et si $\beta_1 = 0$, la distribution est symétrique. Pour le coefficient d’aplatissement (β_2), si $\beta_2 > 0$, la distribution est platicurtique (aplatie), si $\beta_2 < 0$, la distribution est leptocurtique (pointue), si $\beta_2 = 0$, la distribution est mésocurtique (comme la loi normale).

II) Mise en œuvre avec Python

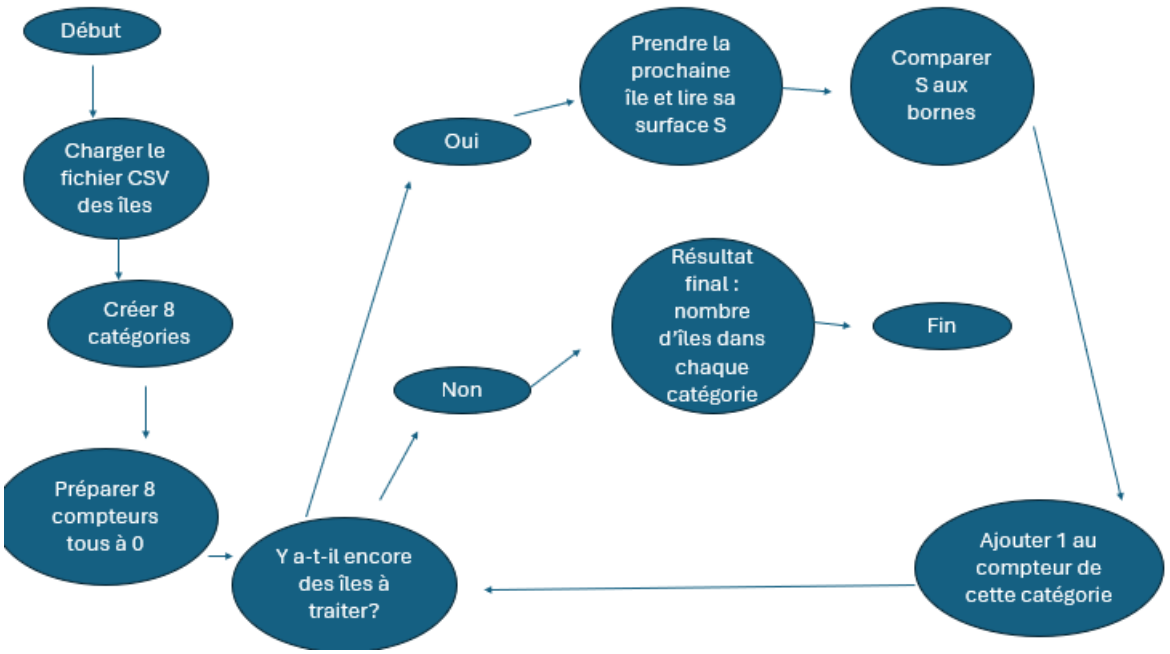
Analyse du code

Le code est composé de deux parties principales. Dans un premier temps, le code permet d’analyser les résultats du premier tour de l’élection présidentielle française de 2022 à partir d’un fichier CSV qui contient les données. Le code permet de calculer les moyennes, les médianes, les écart-types et les quartiles, puis de les visualiser grâce à la création de boites à moustache qui rendent compte de la dispersion des données.

Dans un deuxième temps, le code permet de classer les îles en fonction de leur surface. On peut constater une forte concentration d’îles de petite taille, tandis que les îles de grandes superficies sont

plus rares. Cette catégorisation peut s'avérer très utile pour des travaux cartographiques par exemple.

L'organigramme ci-dessous s'attache à illustrer visuellement la solution proposée par le code. On commence par charger le fichier CSV qui contient les données des îles. On créé ensuite 8 catégories qui correspondent aux petites îles (0-10 km²), aux îles de taille moyenne (10-25, 25-50, 50-100 km²), aux grandes îles (100-2500, 2500-5000, 5000-10000 km²) et aux très grandes îles (supérieures à 10000 km²). A partir de là, on prépare 8 compteurs où l'on va pouvoir compter combien d'îles vont dans chaque catégorie. La boucle permet de regarder chaque île une par une. Elle lit d'abord la surface de l'île, par exemple 5350 km², puis compare avec les catégories créées pour la ranger au bon endroit, ici la catégorie 6. L'île est alors ajoutée au compteur de la catégorie 6, puis on passe à l'île suivante et l'opération recommence jusqu'à ce que toutes les îles soient traitées. Une fois qu'elles ont toutes été passées en revue par la boucle, on affiche combien d'îles il y a dans chaque catégorie.



Analyse des boites de dispersion

Les boîtes de dispersion générées par le code permettent de dégager de manière visuelle les grandes tendances lors du premier tour de l'élection présidentielle française de 2022.

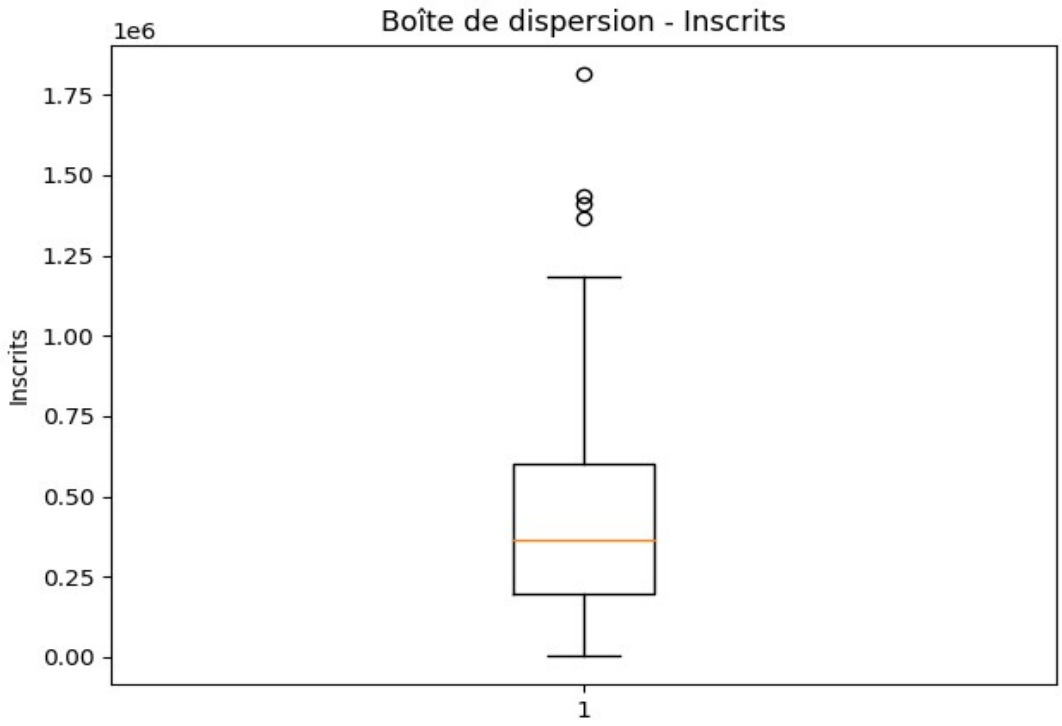


FIGURE 1 – Nombre d'inscrits par circonscription

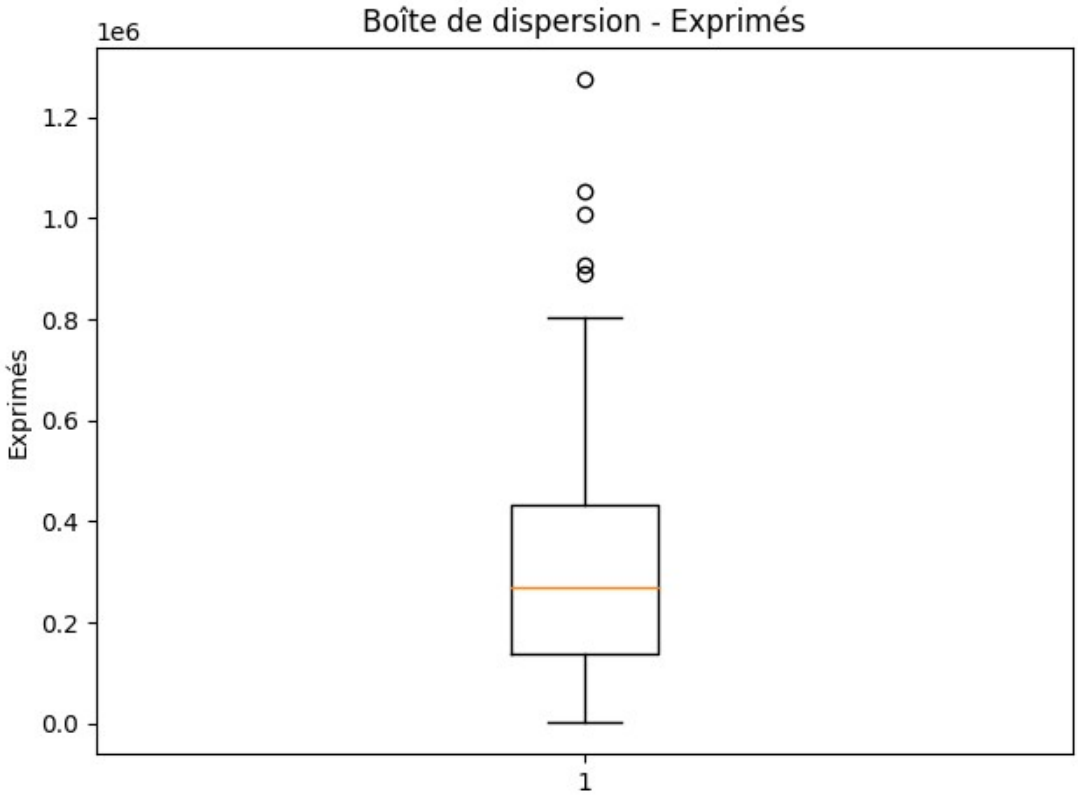


FIGURE 2 – Nombre de voix exprimées

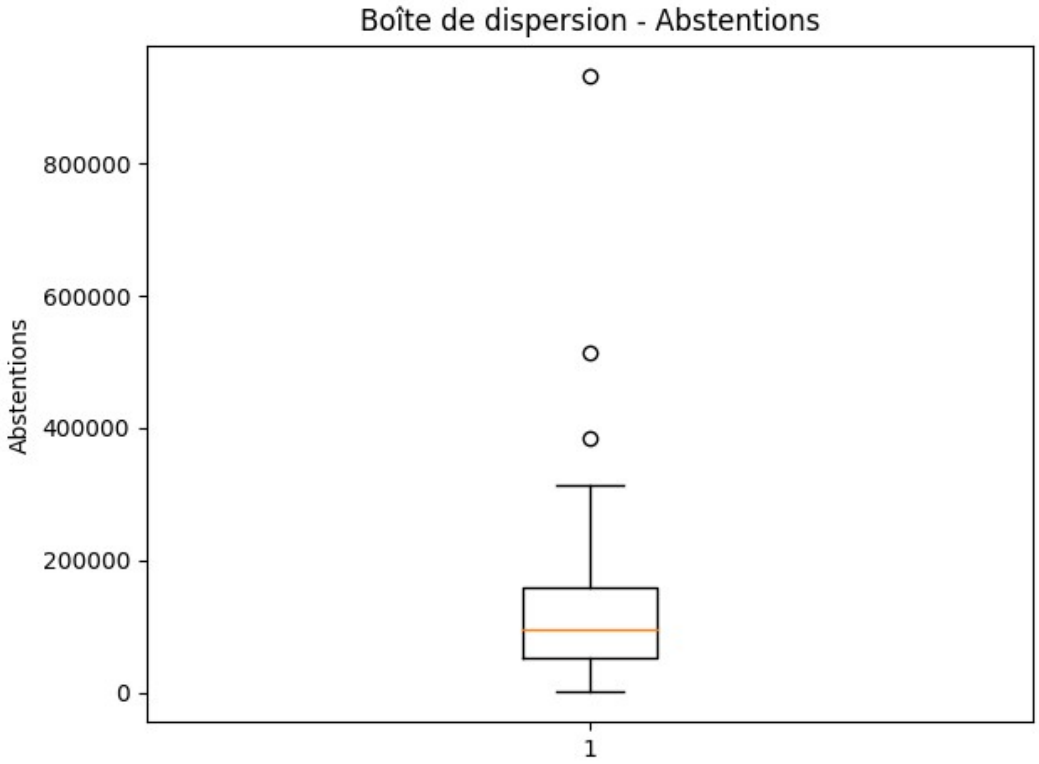


FIGURE 3 – Absentions

En analysant les boîtes de dispersion des inscrits et des voix exprimées (figure 1 et figure 2), on peut constater une importante dispersion qui reflète une forte hétérogénéité territoriale entre les départements peu peuplés du territoire français et les départements qui sont au contraire densément peuplés, comme l'est par exemple la ville de Paris. De la même manière, la boîte à dispersion représentant l'abstention (figure 3) montre une dispersion importante avec une médiane autour de 100 000 voix mais avec la présence de valeurs extrêmes allant jusqu'à 950 000 voix qui révèle des différences importantes de mobilisation suivant les territoires.

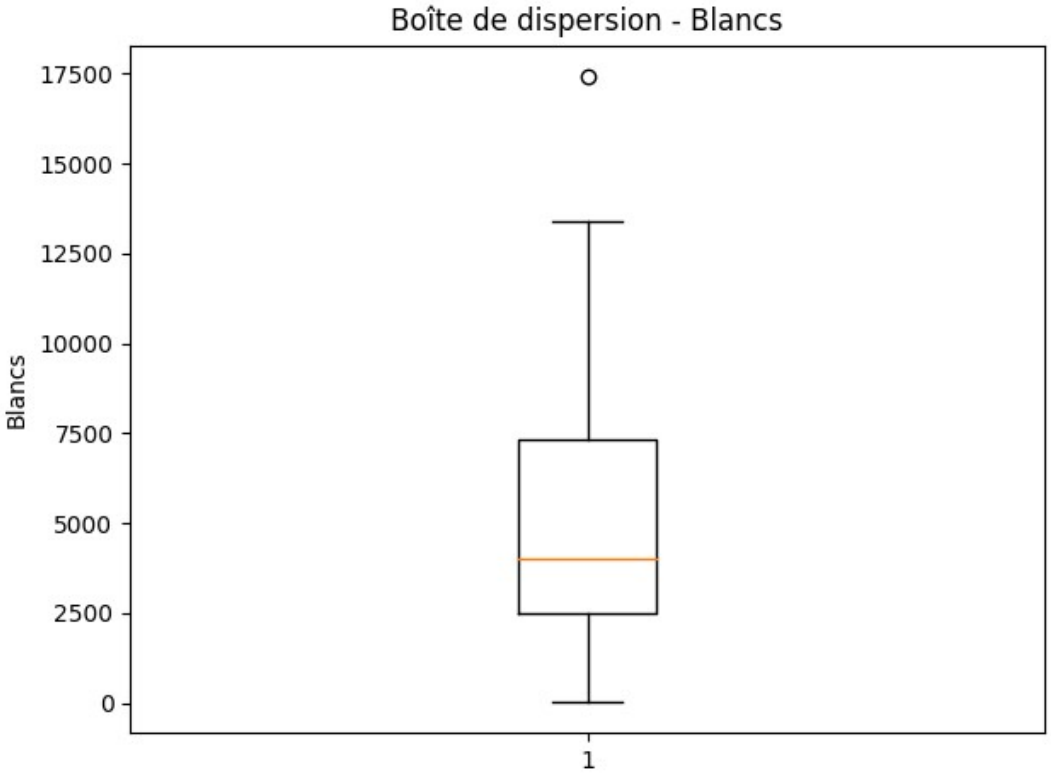


FIGURE 4 – Votes blancs

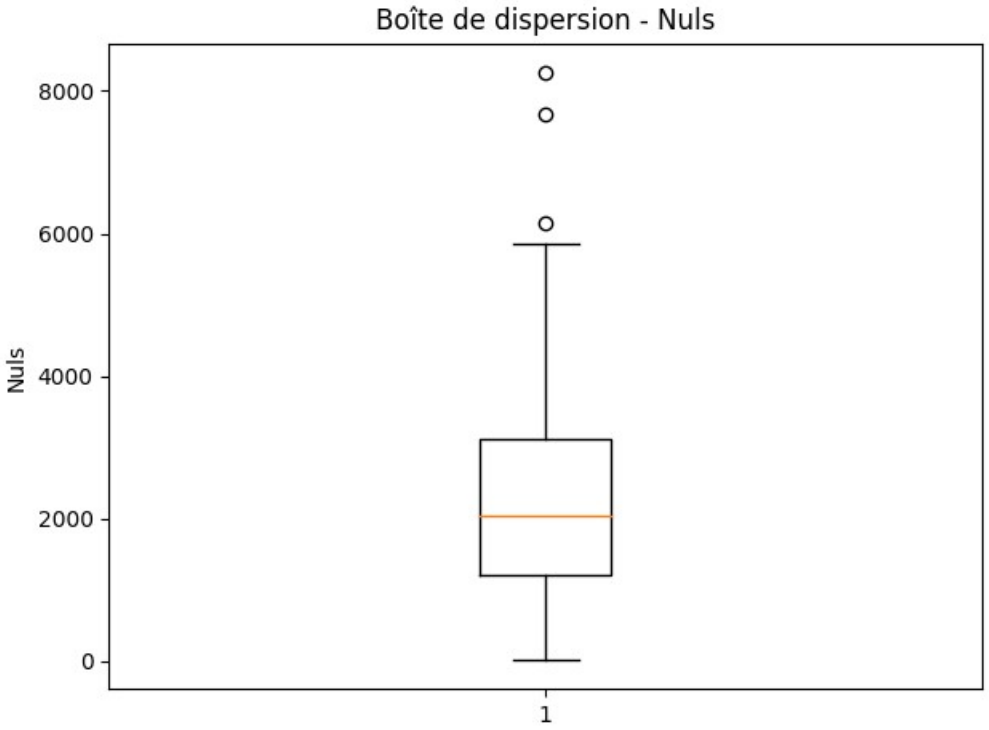


FIGURE 5 – Votes nuls

Les boites de dispersion représentant les votes blancs (figure 4) et les votes nuls (figure 5) révèlent que ces choix de vote, souvent contestataires et révélateurs d'un mécontentement, restent marginaux, puisque la médiane pour les votes blancs se situe à 4000 voix et la médiane pour les votes nuls se situe à 2000 voix. De plus, la dispersion est faible, ce qui permet de conclure que ces comportements restent à la marge dans toutes les circonscriptions de France.

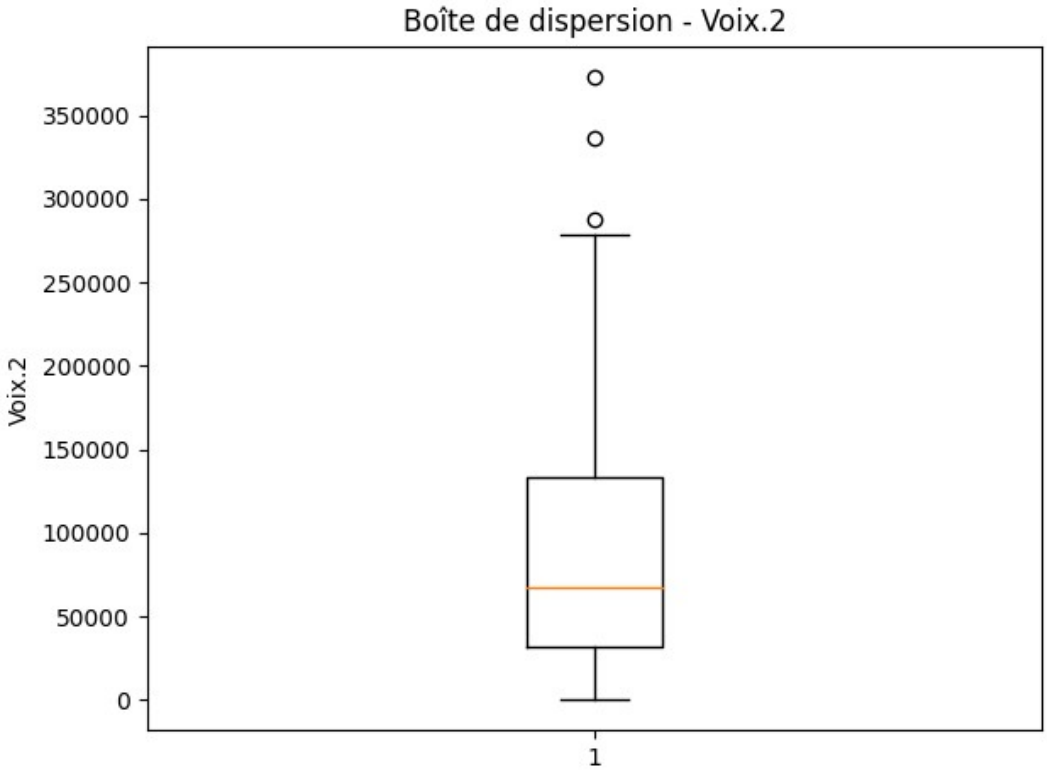


FIGURE 6 – Voix 2

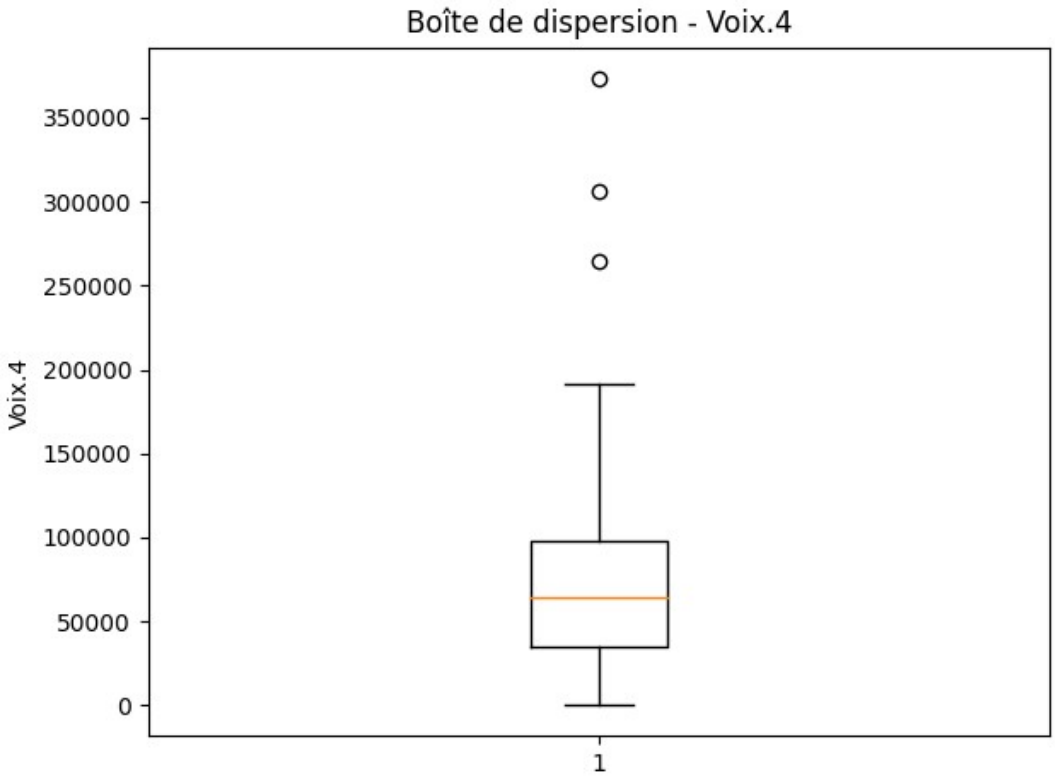


FIGURE 7 – Voix 4

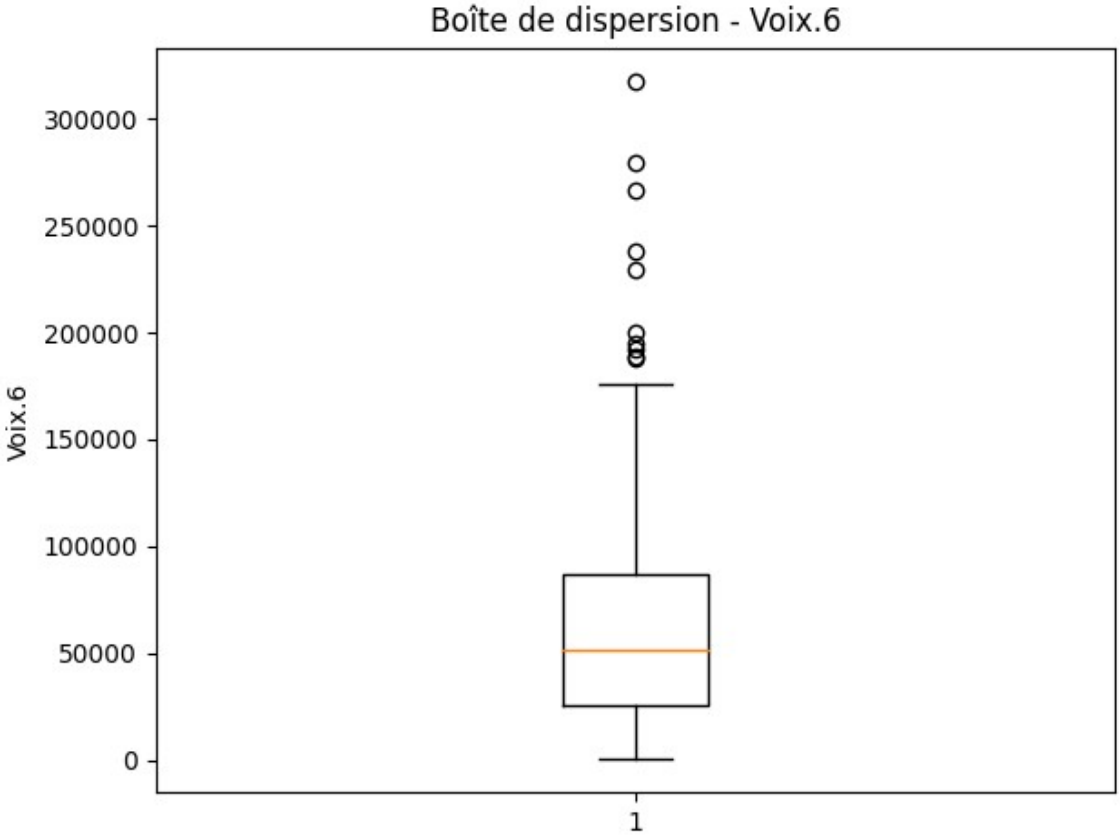


FIGURE 8 – Voix 6

Les boites de dispersion représentant les voix obtenues par les 11 candidats indiquent une hiérarchie très claire : la voix 2 et la voix 4, correspondant à Emmanuel Macron et Marine le Pen, se détachent largement avec des médianes à environ 65 000/70 000 voix. La voix 6, correspondant à Jean-Luc Mélanchon, se détache aussi avec une médiane à 50 000 voix. Tous les autres candidats ont des scores beaucoup plus faibles, avec des médianes oscillant entre 5000 et 20 000 voix. La majorité des voix est donc concentrée sur 3 candidats principaux, tandis que les petites candidats peinent à émerger. Ces boîtes de dispersion permettent de brosser le paysage politique français lors du premier tour des élections présidentielles en France en 2022 avec une France coupée entre l'extrême gauche, l'extrême droite, et un candidat plus centriste.

Séance 4 : Les distributions statistiques

I) Questions de cours

1. Les critères à mettre en avant pour choisir entre une distribution statistique avec des variables discrètes et une distribution statistique avec des variables continues sont multiples. Il est plus pertinent de choisir une loi discrète quand on compte des éléments distincts, comme le nombre de succès ou le nombre d'événement et quand la variable ne peut prendre que des valeurs entières séparées. Une loi discrète convient parfaitement à des sondages d'opinion par exemple. Au contraire, il est plus pertinent de choisir une loi continue quand la variable peut prendre n'importe quelle valeur dans un intervalle donné et quand on mesure des quantités comme le temps, la température ou la distance. Une loi continue peut être particulièrement utile pour calculer le temps d'attente par exemple. Ainsi, quand l'on peut quantifier, les lois discrètes sont à privilégier tandis que lorsqu'on peut mesurer, les lois continues sont à privilégier.

2. Selon moi, parmi les lois discrètes et continues, plusieurs lois peuvent être utilisées en géographie. Les lois continues sont souvent les plus utilisées en géographie car beaucoup de mesures géographiques sont continues (surfaces, distances, flux…). Cependant, les lois discrètes sont particulièrement utiles pour les événements ponctuels et rares et pour les sondages.

Plusieurs lois continues sont fondamentales à la géographie. La loi de Zipf est une loi très fréquemment utilisée en géographie pour les lois rang-taille. En effet, elle sert à analyser la relation entre le nombre d'habitants d'une ville et son rang dans un territoire, ce qui en fait une loi clé pour la géographie urbaine. La loi normale et la loi log-normale sont aussi souvent utilisées en géographie. La loi normale est utilisée pour les phénomènes où il y a une somme d'effets indépendants. Elle sert notamment à modéliser des distributions symétriques. La loi log-normale est utile dès lors que les processus sont multiplicatifs, par exemple pour analyser la croissance des villes (rapport entre le taux de croissance et la population). Enfin, la loi de Pareto est régulièrement utilisée en géographie pour étudier les inégalités spatiales, mais aussi pour comparer les distributions de richesse et de population.

Des lois discrètes peuvent aussi être utiles en géographie. La loi binomiale s'applique aux phénomènes qui ne peuvent prendre que deux états s'excluant mutuellement. Elle peut ainsi être utile dans le cadre de la géographie sociale pour analyser des phénomènes binaires de type succès/échec. La loi hypergéométrique peut aussi être utile en géographie dès que l'on échantillonne sans remise. C'est le cas dans le cadre des contrôles de qualité où on retire les éléments défaillants de la population. Cela peut aussi être utile pour des sondages territoriaux où l'enquête est réalisée sur une population finie comme un village ou un quartier et où chaque individu ne peut être interrogé qu'une fois. Enfin, la loi de Poisson est aussi souvent utilisée en géographie car c'est la loi des événements rares. C'est donc un outil pertinent pour analyser la distribution dans l'espace de phénomènes rares, notamment lorsque surviennent des catastrophes naturelles (tremblements de terre, inondations…).

II) Mise en œuvre avec Python

Analyse du code

Ce code permet d'analyser et de visualiser des distributions statistiques discrètes et continues. Il permet de transformer les données en informations exploitables, ce qui facilite l'analyse géographique.

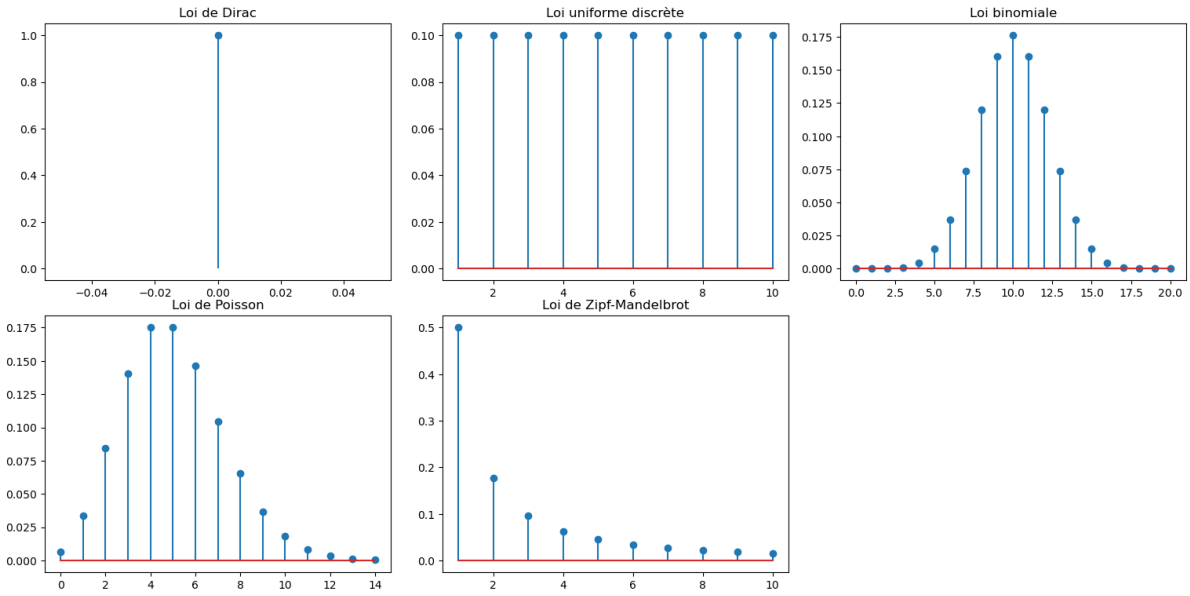
J'ai commencé par importer les bibliothèques nécessaires : numpy qui sert aux calculs numériques, matplotlib.pyplot qui permet de créer des graphiques et d'ainsi visualiser les données et spicy.stats qui sert aux fonctions statistiques avancées.

J'ai créé trois fonctions. J'ai commencé par créer la fonction plot_discrete_distributions () qui permet de générer des graphiques pour les distributions discrètes : loi de Dirac, loi uniforme discrète, loi binomiale, loi de Poisson et loi de Zipf-Mandelbrot. J'ai ensuite crée la fonction plot_continuous_distributions () qui permet de générer des graphiques pour des distributions continues : loi normale, loi log-normale, loi uniforme, loi du Khi2, loi de Pareto. J'ai enfin créé la fonction calculate_mean_std qui permet de calculer la moyenne et l'écart-type pour chaque distribution.

Le code permet donc dans un premier temps de générer des graphiques qui permettent de visualiser les distributions et donc d'analyser plus facilement les données et, dans un second temps, il permet de calculer les moyennes et les écarts-types, ce qui permet une analyse quantitative des phénomènes étudiés.

Analyse des graphiques obtenus

Première série de graphiques obtenus



Le premier graphique correspond à la loi de Dirac, cette loi est une loi dite « dégénérée », c'est-à-dire que toute la masse de probabilité est concentrée en un point unique. C'est ce qu'on peut constater sur le graphique, puisqu'il n'y a qu'un seul pic de probabilité qui est égal à 1.

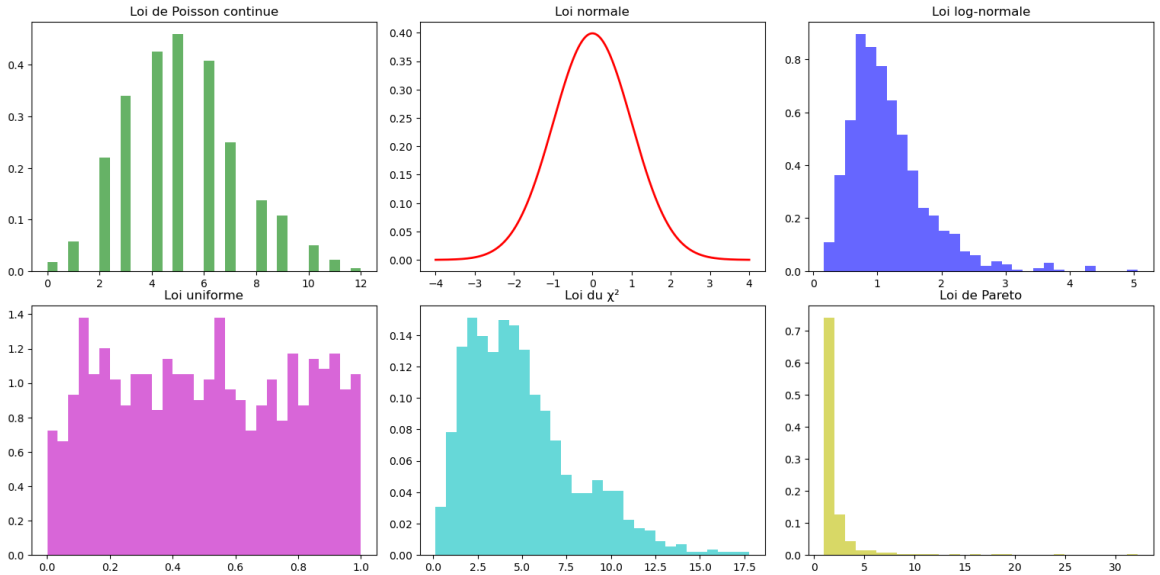
Le deuxième graphique illustre la loi uniforme discrète. Cette loi attribue la même probabilité à chaque valeur possible au sein d'un ensemble fini. Cela est bien visible sur le graphique : les barres sont de hauteur identique, ce qui montre bien que les résultats ont la même probabilité d'advenir.

Le troisième graphique représente la loi binomiale : elle 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. Au contraire, les valeurs aux extrémités du graphique sont faibles car la probabilité d'obtenir très peu ou beaucoup de succès est bien plus rare.

Le quatrième graphique illustre la loi de Poisson. Cette loi décrit le nombre d'événements qui se produisent au sein d'un intervalle donné. Sur le graphique, les barres sont plus hautes pour les petites valeurs, ce qui signifie que le nombre d'événements le plus probable est faible. Il est donc courant d'observer peu d'événements mais il est au contraire rare d'en observer beaucoup.

Le cinquième et dernier graphique correspond à la loi de Zipf-Mandelbrot qui décrit la distribution de fréquence des mots, des noms ou d'autres entités dans un ensemble de données. Le graphique révèle une forte inégalité entre les occurrences : en effet, la première valeur a une probabilité très élevée mais la deuxième a une probabilité beaucoup plus faible, tandis que la probabilité des autres valeurs décroît plus lentement. Ce phénomène est nommé phénomène « de longue traîne », c'est-à-dire que qu'une minorité domine largement tandis que la majorité est peu représentée.

<u>Deuxième série de graphiques obtenus</u>



Le premier graphique illustre la loi de Poisson continue. Elle reprend la loi de Poisson vue précédemment, 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. On peut constater que la distribution présente un pic autour de la moyenne.

Le deuxième graphique représente la loi normale. Le graphique est en forme de cloche symétrique. On peut constater qu'elle est centrée autour de la moyenne et qu'elle décroît de manière symétrique de part et d'autre de la courbe. La valeur centrale est la plus fréquente : plus on s'éloigne de cette valeur centrale, plus les valeurs sont rares. Cela montre que la plupart des observations sont proches de la moyenne.

Le troisième graphique s'intéresse à la loi log-normale. Cette loi montre que le logarithme suit une loi normale, comme vu sur le graphique précédent. On peut voir que le graphique est asymétrique et étalé vers la droite : les petites valeurs sont plus fréquentes que les grandes valeurs. Cette loi est souvent utilisée pour des phénomènes où les valeurs varient sur plusieurs ordres de grandeur, par exemple pour les revenus au sein d'une population.

Le quatrième graphique correspond à la loi uniforme. Cette loi montre que chaque valeur au sein d'un intervalle donné a la même probabilité de se produire. Cela est bien visible sur le graphique. En effet, la courbe est plate, indiquant que la probabilité est constante dans l'intervalle donné.

Le cinquième graphique illustre la loi du Khi-2 qui est très fréquemment utilisée 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. On peut constater sur le graphique que la distribution est asymétrique puisqu'elle est centrée autour de la moyenne, tandis que sa queue est étendue vers la droite. Cette loi est particulièrement utile en géographie. Elle peut par exemple servir à établir si la distribution des risques naturels diffère selon les régions.

Le sixième graphique illustre la loi de Pareto qui modélise des phénomènes où une petite partie de la population possède la majorité des ressources. C'est ce qu'on peut constater sur le graphique puisqu'il y beaucoup de valeurs élevées concentrées sur un tout petit intervalle. Cette loi est fréquemment utilisée en géographie pour décrire des inégalités spatiales, par exemple pour montrer la répartition des richesses entre les régions ou encore la concentration des flux touristiques sur quelques destinations majeures.

Séance 5 : Les statistiques inférentielles

I) Questions de cours

1. L'échantillonnage peut se définir comme le fait de prélever une partie de la population, qu'on appelle alors l'échantillon, pour en étudier les caractéristiques. La population d'où est tiré l'échantillon est appelée population mère. On n'utilise pas la population en entière car cela ferait beaucoup trop d'individus à étudier et l'étude sur toute la population serait beaucoup trop onéreuse à réaliser.

Les méthodes d'échantillonnage sont de trois types. La première est la méthode aléatoire, la base de sondage est alors fiable et précise puisqu'on choisit en avance les individus à interroger. Dans les méthodes aléatoires, on trouve le tirage avec remise, c'est-à-dire que l'individu peut être tiré plusieurs fois et le tirage sans remise (chaque individu n'est tiré qu'une fois). La deuxième méthode est la méthode non aléatoire, qui cherche à fabriquer des « modèles réduits » d'une population mère en utilisant d'autres procédés que le tirage au sort. Dans la méthode aléatoire, on trouve la méthode des quotas qui respecte les proportions de la population mère et l'échantillonnage systématique, c'est-à-dire que tous les n personnes on choisit une personne. Enfin, la troisième méthode d'échantillonnage est la méthode Monte-Carlo. Cette méthode diffère des autres méthodes car c'est une méthode d'échantillonnage par simulation, les populations étudiées sont donc des populations virtuelles, et non pas réelles. Cette méthode utilise le hasard de manière répétée et massive pour résoudre des problèmes qui sont très difficiles à traiter via les deux autres méthodes.

Pour choisir quelle méthode est la plus efficace, il faut s'intéresser à la disponibilité des sondés, au coût des tirages, à la possibilité d'avoir plusieurs échantillons, à la taille de la population et au niveau de précision souhaité. Dans tous les cas, l'échantillon doit être représentatif et suffisamment grand.

2. Un estimateur est une variable aléatoire, une fonction mathématique tandis que l'estimation est la valeur numérique concrète qui est obtenue via un échantillon observé. L'estimateur est donc la formule tandis que l'estimation est le résultat obtenu grâce à l'estimateur.

3. L'intervalle de fluctuation suppose que le paramètre théorique de la population (p) est connu. La question est de savoir où vont se situer les échantillons. L'intervalle de fluctuation permet de prendre une décision en se demandant si l'échantillon observé est compatible avec p.

Au contraire pour l'intervalle de confiance, on ne connaît pas le paramètre de la population donc on l'estime à partir d'un échantillon. La question est cette fois de savoir de où se situe probablement le vrai paramètre de la population.

Ainsi, l'intervalle du fluctuation se fonde sur le théorique pour aller vers l'observé tandis que l'intervalle de confiance se fonde sur l'observé pour aller vers le théorique.

4. Dans la théorie de l'estimation, le biais peut se définir comme la différence entre l'espérance de l'estimateur ($\hat{\theta}$) et la valeur à estimer (θ). C'est une erreur systématique qui fait qu'un estimateur se trompe toujours dans la même direction. Au contraire, un estimateur est sans biais si $E(\hat{\theta}) - \theta = 0$, c'est-à-dire si, en moyenne, sur tous les échantillons possibles, l'estimateur donne la bonne valeur.

5. Une statistique travaillant sur la population totale se nomme recensement. On peut faire un lien avec les données massives (*big data*) car à l'ère des *big data*, on a parfois accès à des populations entières, par exemple tous les utilisateurs d'une plateforme. On peut donc questionner la pertinence des méthodes statistiques et de l'échantillonnage si désormais nous avons accès directement à la population mère. Cependant, les méthodes statistiques restent très utiles même à l'ère des données massives car les données recueillies peuvent être biaisées.

6. Les enjeux autour du choix d'un estimateur sont grands car le choix d'un estimateur se répercute sur la qualité des conclusions tirées dans le cadre d'analyses statistiques. Tout d'abord, il faut veiller à choisir un estimateur sans biais, c'est-à-dire qui vise juste en moyenne. Sous l'influence du hasard, l'estimateur peut parfois donner des valeurs estimées trop grandes ou trop petites, mais ces erreurs doivent se balancer de telle sorte que l'estimateur donne en moyenne la valeur juste. Deuxièmement, l'estimateur doit être convergent, c'est-à-dire que plus on collecte de données, plus l'estimateur doit se rapprocher de la vraie valeur. Troisièmement, l'estimateur doit être efficace c'est-à-dire qu'il doit avoir la plus petite variance possible parmi tous les estimateurs sans biais. Enfin, l'estimateur ne doit pas être influencé par les valeurs extrêmes ou erronées.

7. Les méthodes d'estimation d'un paramètre sont au nombre de 5. Tout d'abord, la méthode des moments consiste à égaliser les moments théoriques et les moments empiriques. La méthode des moindres carrées permet de minimiser la somme des carrés des écarts. La méthode du maximum de vraisemblance permet de choisir la valeur qui rend les observations les plus probables. La méthode de l'intervalle de confiance procure un encadrement qui a une probabilité d'être composée par la vraie valeur. La méthode de l'intervalle de pari est elle fondée sur les caractéristiques de la population. Enfin, le bootstrap permet de rééchantillonner massivement pour estimer la distribution.
Pour sélectionner adéquatement la bonne méthode d'estimation d'un paramètre, il faut veiller à connaître les avantages de chaque méthode. Par exemple, la méthode des moindres carrées est simple et est souvent utilisée en régression. De plus, il faut choisir la méthode par rapport à la taille de l'échantillon, la disponibilité des informations, la nature des données et la connaissance ou non de la loi suivie par la variable.

8. Il y a deux grands types de tests statistiques existants : premièrement les tests paramétriques qui comprennent le test de Student (comparaison de moyennes), le test de Fisher (comparaison de variances) et le test du Khi-2 (ajustement et indépendance), deuxièmement les tests non paramétriques qui comprennent le test de Mann-Whitney, le test de Wilcoxon et le test de Kruskal-Wallis.
Ces tests ont en commun de servir à prendre une décision c'est-à-dire à accepter ou rejeter une hypothèse précédemment formulée, dans le cadre d'une enquête par exemple. En effet, ils permettent de vérifier si une différence observée est significative ou si au contraire elle est due au hasard.
Pour créer un test, il y a onze étapes différentes à respecter. D'abord il faut émettre une hypothèse, ensuite il faut formuler H_0 (hypothèse nulle) et H_1 (hypothèse alternative), choisir un risque d'erreur

α, choisir la statistique de test appropriée, déterminer sa loi sous H_0, définir la région critique (de rejet), collecter les données, calculer la valeur de la statistique, calculer la p-value, comparer à α et enfin conclure en rejetant ou non H_0.

9. Plusieurs critiques sont émises à l'encontre de la statistique inférentielle. Tout d'abord, H_0 est souvent fausse dès le départ car l'hypothèse "aucun effet" est irréaliste. La taille de l'échantillon joue aussi un grand rôle puisque avec un n très grand, tout devient significatif tandis qu'avec n très petit, rien ne l'est. De plus, ne pas rejeter H_0 ne revient pas à confirmer H_0. De même la p-value ne revient pas à dire que la probabilité H_0 soit vraie. Ensuite, les données récoltées par la statistique inférentielle sont souvent difficiles à analyser. Lors de l'interprétation des résultats, on cherche aussi souvent à vouloir à tout prix trouver un sens aux données recueillies, même quand elles n'en ont pas. Enfin, les intervalles de confiance sont parfois plus informatifs.
Ces critiques paraissent justifiées puisque les tests sont en effet souvent mal utilisés et mal compris. Cependant, bien employés, ces tests statistiques restent utiles. Donc ce n'est pas tant l'outil qui pose problème que son utilisation et son interprétation.

II) Mise en œuvre avec Python

Analyse du code et de ses résultats

1. On observe que les fréquences réelles de la population mère sont systématiquement incluses dans les intervalles de fluctuation calculés au seuil de 95%. Cela confirme que nos 100 échantillons sont représentatifs de la population mère. L'opération de moyennage sur 100 tirages a permis de réduire la variance et d'obtenir une estimation des opinions quasi-identique à la réalité.

Il convient de noter que le calcul des intervalles de confiance joue un rôle majeur et bien souvent négligé en science, en particulier dans les sciences sociales. Si notre simulation a démontré la robustesse mathématique de l'intervalle de confiance à 95% (toutes les valeurs réelles y sont incluses), la pratique humaine est souvent bien moins rigoureuse. Comme l'a mis en évidence le psychologue et prix Nobel d'économie Daniel Kahneman.

Dans ses travaux (notamment présentés dans *Système 1 / Système 2*), Kahneman montre que lorsqu'on demande à des individus d'estimer une quantité incertaine (comme une fréquence ou une valeur future) en donnant un intervalle de confiance à 90%, ils ont tendance à fournir une fourchette beaucoup trop étroite. Ils sous-estiment systématiquement la variance réelle du monde. Ainsi, être capable de calculer des intervalles de confiance est fondamental pour prétendre faire de la recherche sérieuse.

2. La deuxième partie du code s'intéresse à la théorie de l'estimation. On n'a qu'un seul échantillon et on ne connaît pas la population mère. La question est donc de savoir si l'échantillon est fiable ou non. A partir de l'échantillon unique, on calcule l'effectif total et les fréquences. Ces valeurs correspondent aux valeurs observées dans l'échantillon, mais il faut déterminer ce qu'elles valent si on les applique à la population entière. Pour cela, on calcule l'intervalle de confiance à 95 %. Cet intervalle signifie que si je répétais cette enquête 100 fois avec 100 échantillons différents de 1 000 personnes, dans 95 cas sur 100, la vraie proportion dans la population totale se trouverait dans mon

intervalle.

3. Analysons à présent les résultats du test de Shapiro-Wilk sur nos deux échantillons. Commençons par le deuxième cas : on trouve W = 0.2608882349902276 pour une p-value = 7.04938990116743*10^(-67). A la fois la W << 1 et p << 0.05, donc on peut en conclure que les données du premier échantillon ne suivent pas une loi normale. Pour le deuxième jeu de données, on trouve W = 0.9639482021309311 très proche de 1; cependant on a une p-value = 7.04938990116743*10^(-67) qui est extrêmement faible, ce qui laisse penser que la distribution ne suit pas une loi normale non plus. Toutefois, le fait que W soit si proche de 1 intrigue. Après examen des données, on remarque que l'échantillon de données est très large : plus de 2000 valeurs. Le test de Shapiro-Wilk est très sensible quand l'échantillon est de grande taille (ce qui est le cas ici), ainsi il suffit qu'une faible proportion des valeurs infirment la possibilité d'une loi normale pour obtenir une p-value extrêmement faible et que le test réponde à la négative. Étant donné qu'on nous assure qu'un des deux échantillons suit une loi normale, on peut en déduire qu'on est sûrement tombé dans ce dernier cas de figure et que c'est le premier échantillon qui suit une loi normale.

Séance 6 : La statistique d'ordre des variables qualitatives

I) Questions de cours

1. Une statistique ordinale est une branche de la statistique qui classe les observations selon un ordre, sans que l'écart entre ces valeurs soit quantifié. Concrètement, l'information réside uniquement dans le rang relatif (X(i) < X(j)), et non dans la magnitude de la différence (X(i) − X(j) = …).
Elle s'oppose à la statistique nominale. Dans une variable nominale, les catégories sont des étiquettes pures sans hiérarchie logique (ex : climat océanique / continental ; on ne peut pas dire que « climat océanique < climat continental »).
Une statistique ordinale utilise des variables ordinales : ce sont des variables pour lesquelles il existe une hiérarchie stricte (ex : faible / moyen / fort avec faible < moyen < fort). On peut comparer les positions, mais pas mesurer précisément les écarts.
Une statistique ordinale peut matérialiser une hiérarchie spatiale. Par exemple, si l'on classe les villes d'un pays selon leur influence (métropole mondiale > métropole régionale > ville moyenne), on crée une structure ordinale.

2. L'ordre à privilégier dans les classifications est l'ordre croissant, également appelé ordre naturel.
Cependant, il existe une exception majeure en géographie : la loi rang-taille, où l'on classe souvent les villes de la plus grande à la plus petite (ordre décroissant) afin d'étudier la distribution de leur taille.

3. La corrélation des rangs sert à comparer deux classements : si l'on a deux juges (ou deux critères), comment savoir mathématiquement s'ils disent la même chose ?
La concordance des classements s'utilise lorsqu'on n'a pas deux mais p > 2 classements. Elle permet de répondre à la question suivante : parmi ces p classements, y a-t-il un consensus ou bien est-ce le chaos ?

4. Le test de Spearman regarde l'écart numérique entre les rangs. Si un élément est 1er dans la liste A et 10e dans la liste B, l'écart est grand.
Le test de Kendall regarde la cohérence des paires. Si A est mieux classé que B dans la liste 1, est-ce que A est toujours mieux classé que B dans la liste 2 ? C'est une approche plus « structurelle ».
Dans les deux cas, les tests renvoient un coefficient compris entre −1 et +1 qui quantifie à quel point les classements sont d'accord :
+1 indique des classements identiques,
−1 des classements parfaitement inverses,
0 indique que les rangs sont indépendants.

5. Les coefficients de Goodman-Kruskal et de Yule permettent de répondre quantitativement à la question suivante : quand une variable augmente (ou change de modalité), l'autre a-t-elle tendance à évoluer dans le même sens, dans le sens inverse, ou indépendamment ?
C'est à rapprocher du concept d'élasticité en économie (variation d'une grandeur provoquée par la variation d'une autre).

Supposons que l'on souhaite vérifier l'hypothèse : « plus on s'éloigne du centre (densité faible), plus les ménages sont des familles avec enfants ». On dispose de deux variables ordinales : la densité du lieu (3 niveaux : centre-ville (dense) > banlieue (moyen) > périurbain (faible)), le type de ménage (3 niveaux : célibataire > couple sans enfant > famille avec enfants). En calculant le coefficient Γ de Goodman-Kruskal, on peut confirmer ou infirmer mathématiquement cette hypothèse.
Le coefficient de Yule est le cas particulier où les deux variables ordinales ont seulement 2 niveaux chacune.

II) Mise en œuvre avec Python

1) Partie obligatoire

Voici les modifications apportées aux fonctions pré-écrites :

- Ajout du paramètre encoding="utf8" à la fonction open dans la fonction ouvrirUnFichier, sans quoi Python utilise un encodage par défaut incompatible avec les fichiers CSV.

- Modification du test isnan(...) pour vérifier si un float est NaN de manière plus native en Python (si x vaut NaN, alors x == x renvoie False).

- Suppression de la conversion float(pop[element]) puisque pop[element] est déjà un float.

- Correction d'une faute de frappe dans l'avant-dernière ligne de classementPays (utilisation d'une variable inexistante).

- Correction des bornes des boucles dans classementPays, qui oubliaient le dernier élément des listes.

- Remarque intéressante : pour classementPays, on peut supposer sans perte de généralité que len(ordre1) ≤ len(ordre2) et inverser les arguments si ce n'est pas le cas. Cela évite de dupliquer du code et rend la fonction plus lisible.

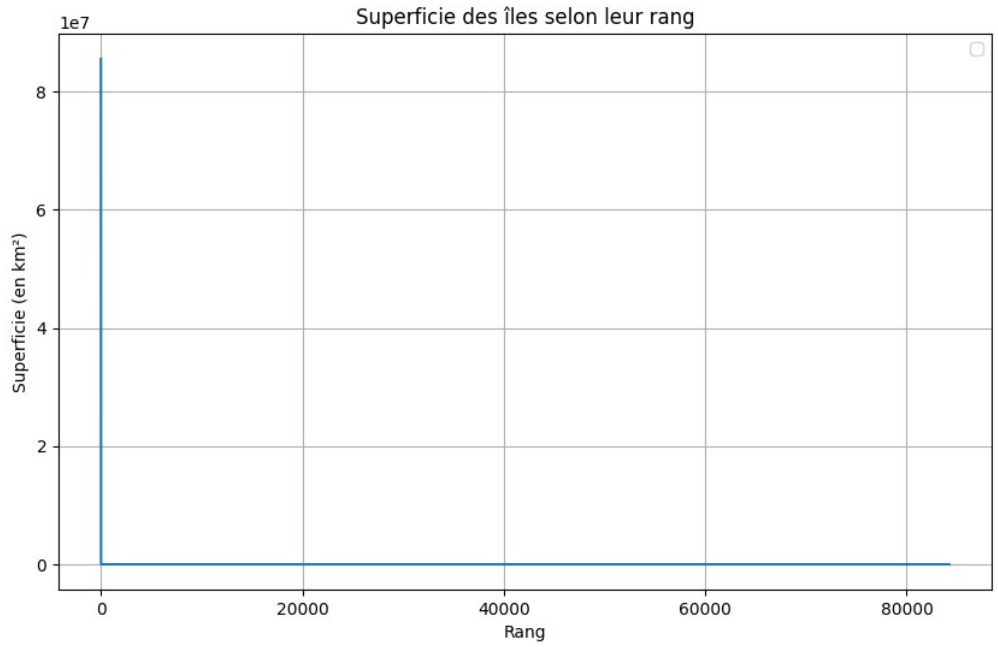


FIGURE 1 – Question 5

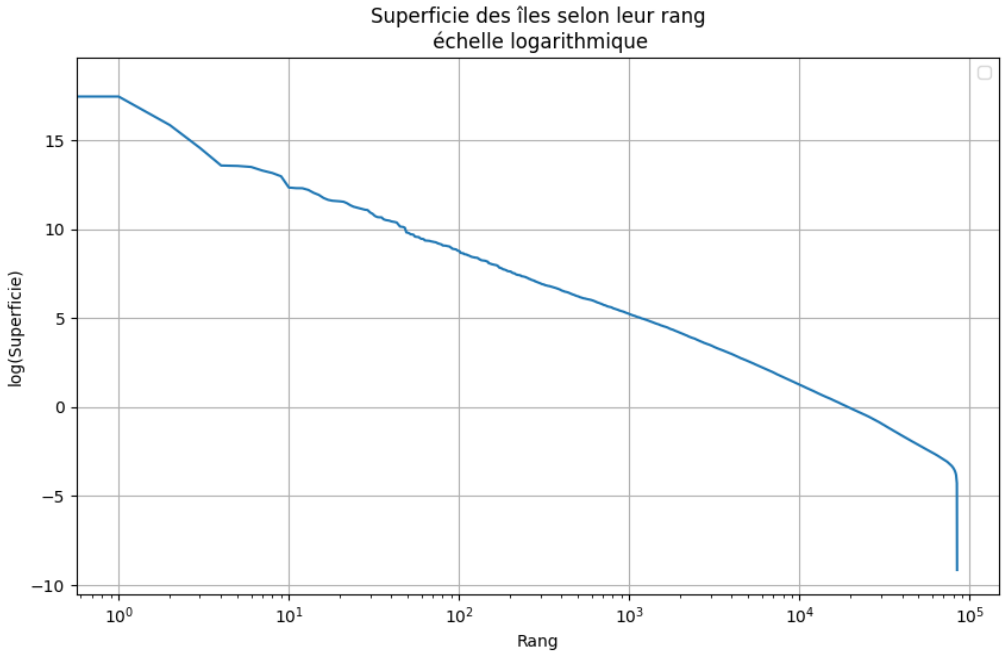


FIGURE 2 – Question 6

Question 14 : Dans l'ensemble, les coefficients obtenus sont très proches de +1, ce qui confirme une forte inertie démographique sur la période 2007–2025. L'absence de bouleversements majeurs (tels que des conflits mondiaux ou des catastrophes démographiques massives) explique que ni la hiérarchie des pays les plus peuplés, ni celle des densités, n'aient été fondamentalement remises en cause.

Cependant, une analyse comparée des coefficients de Spearman (r_s) et de Kendall (τ) révèle une nuance intéressante : r_s est très légèrement supérieur à τ (on a $r_s - \tau \approx 0.11$, que ce soit pour la

population ou pour la densité). Il serait tentant d'interpréter cet écart géographiquement (en supposant par exemple que de nombreuses petites inversions locales de positions pèsent sur le Kendall sans affecter le Spearman). Il convient toutefois de rester prudent quant à la comparaison directe des magnitudes entre ces deux indicateurs. On ne sait pas si la relation entre r_s et τ est linéaire (pour reprendre le langage des économistes, on ne connaît pas l'élasticité). Par conséquent, l'écart observé n'est pas nécessairement le signe d'un phénomène démographique particulier (comme des flux migratoires spécifiques), mais pourrait simplement refléter la nature mathématique différente des deux métriques.

III) Bonus

1) Première partie

Note préliminaire : Les fonctions du module scipy.stats que l'on utilise, à savoir scipy.stats.spearmanr et scipy.stats.kendalltau, sont conçues pour prendre en entrée des données brutes (non triées). Les questions 8 à 13 de la partie obligatoire visent à traiter les données pour ensuite les passer en argument à scipy.stats.spearmanr et scipy.stats.kendalltau ; en réalité, un traitement aussi abouti n'était pas nécessaire puisque les fonctions acceptent des données brutes. Toutefois, supprimer les données NaN (Not a Number) est bel et bien nécessaire pour obtenir le bon résultat.

Pour cette partie bonus, j'ai suivi la même trame ; cependant, les données des îles ne contiennent pas de NaN et ce traitement est donc parfaitement inutile (bien qu'intéressant pédagogiquement). Ainsi, il existe une réponse en une ligne à la première partie de la question bonus :

```
print("Coefficient de Spearman : ",
        scipy.stats.spearmanr(
                iles["Trait de cote (km)"].tolist(),
                iles["Surface (km^2)"].tolist()
        )
)
```

On obtient les résultats suivants : $r_s \approx 0.9712866815534842$ et $\tau \approx 0.8539337169239855$, tous deux proches de 1, ce à quoi on pouvait s'attendre. En effet, on s'attend à avoir des côtes assez longues pour une île très grande en superficie.

2) Deuxième partie

Comme il n'existe pas de fonction permettant de calculer W dans scipy.stats, j'ai implémenté un algorithme qui calcule W « à la main ». Le problème est qu'il faut gérer les NaN présents dans les données de 2007 à 2014. Pour éviter de devoir supprimer une partie des pays de la table de données (qui pourtant ne présentent pas de données manquantes sur la période 2015–2025), j'ai choisi de travailler uniquement sur la période 2015–2025.

Après exécution de l'algorithme, on trouve $W \approx 0.994838198352725$, ce qui est cohérent avec les coefficients de Kendall calculés pour la population à la question 14. L'interprétation est la même.

Conclusion : Réflexion personnelle sur les sciences des données et les humanités numériques en fonction des exercices du parcours débutant

Dans le cadre du cours « Analyse de données », j'ai choisi le parcours débutant qui m'a permis, tout au long du semestre, de développer une réflexion sur les humanités numériques, qui désignent la transformation des sciences humaines et sociales par le numérique ainsi que l'ensemble des pratiques de recherche à l'intersection entre les technologies numériques et les disciplines des sciences humaines comme la géographie.

Intéressée depuis toujours par les domaines aussi bien scientifiques que littéraires, j'ai justement choisi d'entrer dans le master Géopolitique-Geoint car il permet de faire le lien entre numérique et humanités, qui sont deux disciplines qui me passionnent depuis que je suis petite. Les enseignements proposés par le master mêlent des cours liés aux humanités comme les cours de géopolitique ou d'environnement numérique et des cours liés au numérique comme le cours d'analyse de donnée, le cours de Système d'Information Géographique (SIG), et le cours de méthodes d'enquête.

Ces cours sont intéressants car ils permettent de traiter les données géographiques de manière différente mais complémentaire : programmation en Python pour automatiser le traitement de données, création de cartes sur ArcGIS Pro à partir de bases de données en ligne (IGN, Sentinel, données du gouvernement...) et regroupement de données sur Excel. Les trois compétences requises pour les humanités numériques sont donc présentes au sein du master Géopoltique-Geoint : les compétences techniques avec le cours d'analyse de données et de SIG, les compétences analytiques avec le cours de méthodes d'enquête et d'analyse de données et les compétences critiques avec les cours de géopolitique.

En effet, les cours de géopolitique, alliés aux séminaires organisés par le master où interviennent des professionnels du monde militaire, me permettent d'avoir particulièrement conscience de la masse de données qu'il faut aujourd'hui être capable de traiter si l'on veut prendre les bonnes décisions. Dans un monde où les données prolifèrent, il faut savoir faire le tri et savoir quelles sont les informations utiles à un moment T.

Le parcours débutant du cours d'Analyse de données m'a ainsi permis d'étayer ces réflexions sur les humanités numériques de manière concrète. En effet, le langage informatique Python permet de traiter des données, comme dans la séance 2 et la séance 3 où nous avons utilisé une base de données sur le premier tour des élections présidentielles de 2022 qui était directement disponible sur le site du gouvernement. Les algorithmes permettent de traiter des volumes de données qui seraient beaucoup plus longues, voire parfois impossible, à analyser manuellement.

La plupart des codes permettaient de créer des graphiques ou des boites à moustache qui sont très utiles pour visualiser facilement les données. De grandes tendances peuvent ainsi être dégagées, tandis qu'elles restent généralement invisibles dans des tableaux de chiffres. Si cela permet de générer très rapidement des visualisations, l'analyse, elle, reste à la charge de celui qui produit l'étude. Il faut donc se souvenir que si le numérique permet de traiter plus rapidement les données, cela ne remplace pas les connaissances humaines qui sont les seules capables d'interpréter

adéquatement les résultats. J'ai donc trouvé intéressant de ne pas seulement avoir à rédiger les codes, mais de devoir ensuite les « faire parler » et d'analyser les résultats produits par le code.

En effet, il ne faut pas limiter les humanités numériques aux équations ou aux algorithmes. On peut trouver un exemple parlant dans la séance 4 du parcours débutant, où la loi normale, qui est très fréquemment utilisée en statistiques, ne correspond pourtant pas toujours aux phénomènes sociaux car elle suppose une symétrie et une concentration autour de la moyenne qui ne reflète pas forcément les ruptures et asymétries de ces phénomènes.

Si les termes d'« humanités » et de « sciences » sont souvent perçus comme des termes antithétiques, ils sont pourtant liés depuis toujours. Par exemple la géométrie, aujourd'hui considérée comme une « science » et la géographie, aujourd'hui considérée comme une « humanité », ont pourtant la même origine. Les humanités numériques permettent finalement de revenir à ce syncrétisme qui était présent dans l'Antiquité. Le lien entre humanités et numérique se retrouve d'ailleurs dans certaines lois, comme par exemple la loi de Zipf-Mandelbrot. En effet, cette loi mathématique s'applique aussi aux humanités, notamment la littérature, puisqu'elle décrit la fréquence des mots dans un texte, mais aussi la géographie, puisqu'elle permet la distribution des villes par taille.

Il faut cependant veiller à ce que le côté scientifique et numérique ne prenne pas le dessus sur les humanités. En effet, on a tendance à attribuer une supériorité aux sciences par rapport aux humanités. C'est pourtant oublier que l'esprit critique est indispensable, et encore plus à l'ère du *big data* où les données doivent être traitées avec prudence car elles sont la proie de manipulation, comme par exemple lors d'ingérences numériques étrangères.

Dès lors, pour que les humanités numériques soient pérennes et que le dialogue entre humanités et numérique soit fructueux, il faut avoir une bonne maîtrise des outils informatiques mais aussi comprendre les concepts mathématiques et statistiques qui vont avec. Mais la priorité est bien de garder un esprit critique sur les données, seul garant de la fiabilité des résultats.

Ainsi, ce cours d'analyse de données a confirmé mon attrait pour le numérique et les humanités, et m'a permis de réaliser concrètement, en manipulant et en analysant des données, que numérique et humanités étaient intrinsèquement liés et qu'il fallait entretenir leur dialogue pour faire face aux données massives.