IZERN Diego

21204059

Compte rendu du cours d'analyse de données en géographie

Maxime Forriez

Sommaire

Je m'inscris dans le cadre du parcours débutant.

Séance 2 : Les principes généraux de la statistique.

2.1 Questions de cours

La géographie entretient deux dynamiques avec les statistiques : elle est productrice de données mais ne les analyse pas nécessairement de la même manière. La géographie accorde une place importante aux statistiques dans le but de comprendre l'information géographique et ce à toutes les échelles sans être descriptif.

Pour répondre à la question, nous devons faire intervenir la philosophie. Pour les déterministes, le hasard n'existe pas et n'a donc pas sa place en géographie. Cependant, en géographie, le hasard est admis, on peut identifier des schémas généraux au moyen de la statistique mais pas prévoir exactement comment l'espace évolue. Par exemple, pour l'étude des séismes, les géographes peuvent prévoir où un séisme a le plus de possibilités de survenir en fonction de la fréquence dans le temps et l'espace de l'activité sismique.

En géographie, il existe 2 types d'informations géographiques. La première correspond à l'information attributaire, donc les données descriptives d'un territoire comme le climat ou l'économie. La seconde est l'information géométrique, donc les formes et structures spatiales des objets géographiques.

La géographie a 4 besoins au niveau de l'analyse de données : la production et la collecte de données fiables, la disposition de nomenclatures et métadonnées cohérentes, le fait de pouvoir décrire, comparer et expliquer des phénomènes spatiaux à partir des données et la maîtrise des outils statistiques et informatiques.

Les statistiques descriptives permettent de résumer, organiser et visualiser les données afin de dégager des propriétés. Les statistiques explicatives au contraire, permettent d'expliquer ou prédire une variable via d'autres variables (analyse discriminante…).

Il existe 4 types de variables : des variables qualitatives nominales (représentées avec des diagrammes en secteurs), les variables qualitatives ordinales, les variables quantitatives continues (boîtes à moustache) et les variables quantitatives discrètes (diagrammes en bâtons).

Pour savoir à quelle variable correspond une donnée, cela dépend du type de variable (qualitative ou quantitative) et du but de l'analyse (comparaison ou description).

Il existe 3 grandes classes de méthodes d'analyse de données. La première est la méthode descriptive : CAH, AFM, AFDM, ACP, AFC. La seconde est la méthode explicative : régression logistique, analyse discriminante, régression, analyse de variance. Enfin, la dernière méthode est prédictive : analyse de séries temporelles.

La population statistique désigne l'ensemble des éléments étudiés (ex : tous les adultes de plus de 35 ans dans une ville). L'Individu statistique correspond à l'élément de la population (ex : un

adulte de plus de 35 ans). Le caractère est la propriété mesurée chez chaque individu (ex : l'âge). Enfin, la modalité est la valeur prise par un caractère (ex : 35 ans). Le caractère peut être de 4 types : qualitatif nominal, qualitatif ordinal (possibilité d'effectuer une hiérarchie), quantitatif discret, quantitatif continu.
Voici les 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}.

Afin de 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 et donc la répartition des effectifs en fonction des modalités d'un caractère.

2.2 Analyse des résultats

L'objectif de cette séance est d'utiliser un fichier CSV relatif aux résultats du premier tour de l'élection présidentielle 2022 à l'échelon départemental. Nous devons utiliser plusieurs bibliothèques que nous importons : Pandas et Matplotlib. Cette séance nous mène également à identifier les caractéristiques du jeu de données, calculer des indicateurs statistiques et produire des représentations graphiques (histogrammes, camembert).

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

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

Dimensions du DataFrame

Indicateur	**Valeur**
Nombre de lignes	107
Nombre de colonnes	56

Nous avons produit le tableau suivant, celui du total des inscrits, qui correspond à l'ensemble des électeurs inscrits au 1er tour sur l'ensemble du territoire français. Ce chiffre s'élève à 48.747.876 individus.

Total des inscrits

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

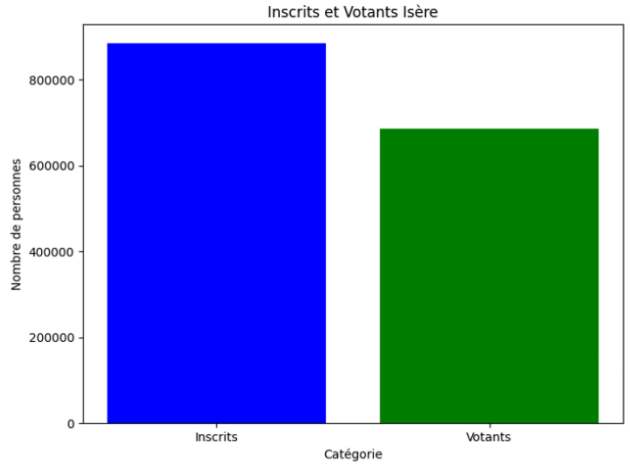
Au moyen d'une boucle conditionnelle, nous avons calculé les effectifs uniquement sur les colonnes numériques. Voici le tableau représentant les sommes des variables quantitatives.

Nous pouvons constater que les valeurs d'*Inscrits*, *Votants* et *Exprimés* ne sont pas aberrantes. En effet, il y a une participation nationale importante (36 millions de votants). Cependant, on constate qu'il y a une abstention assez forte (12 millions d'inscrits).

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

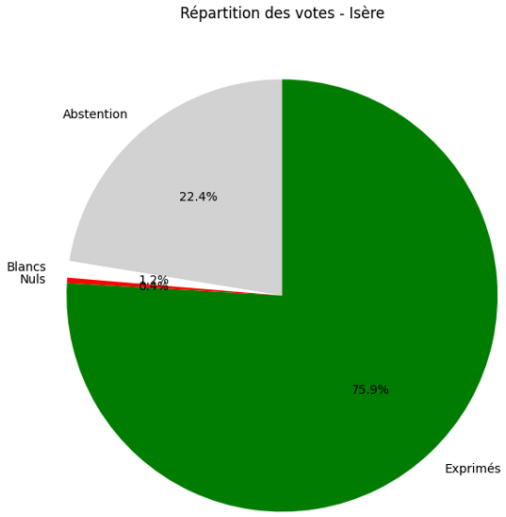
Ensuite, nous avons créé pour chacun des départements un diagramme en barres avec comme ordonnée le nombre de personnes et en abscisse le nombre d'inscrits et celui des votants.

Ces diagrammes nous donnent la possibilité de visualiser l'écart directement entre le nombre d'inscrits et celui de votants. Dans le cas du département de l'Isère. On constate un écart important. En effet, le nombre de votants est inférieur au nombre d'inscrits, seuls ⅘ des inscrits ont voté.



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

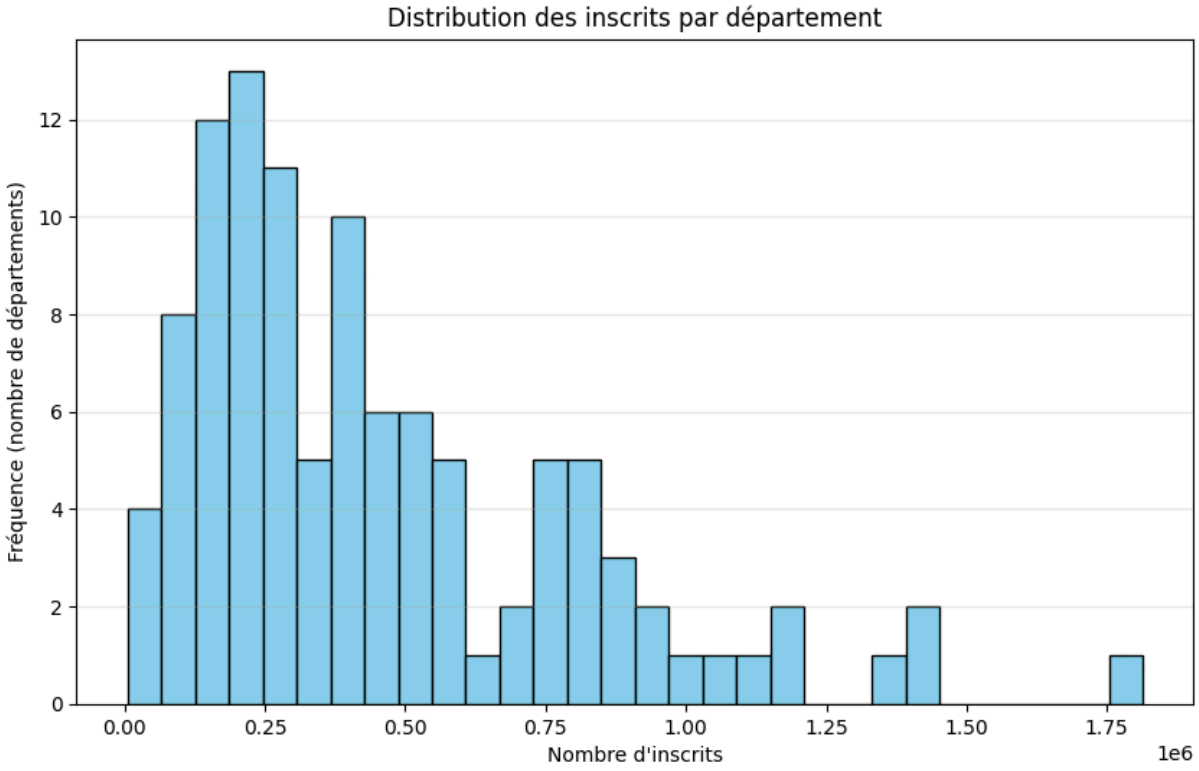
On constate que l'abstention en Isère est proche du quart des inscrits. Les votes blancs et nuls sont beaucoup plus faibles mais sont les témoins de comportements électoraux différenciés.



Enfin, nous avons créé un histogramme de la distribution des inscrits.

Nous constatons une distribution asymétrique du nombre d'inscrits par département. En effet, la majorité des départements ont entre 100.000 et 400.000 inscrits, sauf pour quelques départements (Nord, Paris, Bouches-du-Rhône). Il y a une distribution asymétrique vers la droite, une queue à droite avec beaucoup de départements de taille moyenne et quelques départements très peuplés.

L'ordonnée représente le nombre d'inscrits dans chaque département et l'abscisse la fréquence du nombre de départements qui se trouvent dans un intervalle de valeurs. Les barres correspondent à l'intervalle de nombre d'inscrits et la hauteur de la barre le nombre de départements qui ont une quantité d'inscrits dans cet intervalle. Alors, les grands départements ont une grande part d'inscrits mais ils sont numériquement minoritaires et la plupart des départements ont un nombre d'inscrits modéré.


Distribution des inscrits par département
Fréquence (nombre de départements)
0
2
4
6
8
10
12
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Nombre d'inscrits
1e6

Séance 3 : Les paramètres statistiques élémentaires.

3.1 Questions de cours

Le caractère le plus général est le caractère quantitatif. En effet, ils permettent d'effectuer des calculs statistiques et de mesurer des valeurs numériques

Les caractères quantitatifs discrets sont des valeurs numériques finies (exemple : nombre de français à Genève). Alors que les caractères quantitatifs continus peuvent prendre tous les nombres dans un intervalle (exemple : poids). Il est nécessaire de les distinguer car on n'utilise pas les mêmes méthodes de calcul et d'interprétation.

Il existe plusieurs types de moyenne car chaque type de moyenne a des usages particuliers, certaines moyennes peuvent être tronquées si elles sont sensibles aux valeurs extrêmes, certaines moyennes sont plus adaptées à un type de variable. La médiane permet de calculer la valeur située au centre d'une série de données, elle n'est pas influencée par la distribution des valeurs (si elle est asymétrique). Le mode se calcule par les variables discrètes (valeur avec l'effectif le plus élevé), avec les variables continues (la valeur correspondant à la densité maximale).

L'intérêt de la médiane est de partager en 2 parties égales la masse totale des valeurs afin de déterminer la concentration des valeurs. L'intérêt de l'indice de C. Gini est de mesurer la concentration ou l'inégalité 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 exergue la différence entre la plus élevée et la plus basse valeur d'une série et permet d'obtenir un indicateur de dispersion. Les quantiles 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. Les quantiles divisent une série en parties égales et permettent de résumer la répartition des valeurs. La boîte à moustache permet de visualiser la distribution (médiane, quartiles, valeurs extrêmes) afin de comparer plusieurs distributions

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

On utilise pour cela le coefficient de dissymétrie : $\beta 1$
$\beta 1 > 0 \longrightarrow$ l'asymétrie est positive

$\beta 1 < 0 \longrightarrow$ l'asymétrie est négative

$\beta 1 = 0 \longrightarrow$ la distribution est symétrique

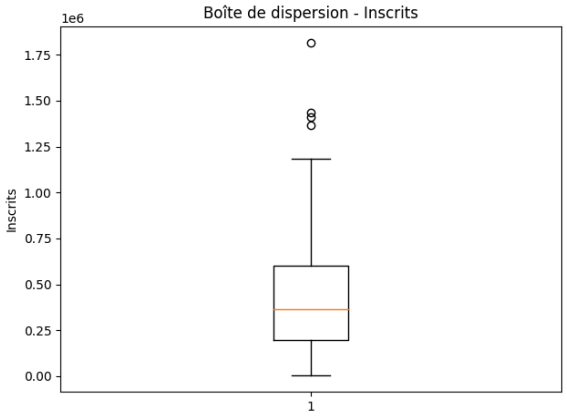
3.2 Analyse des résultats

Le but de cette séance est d'analyser les résultats du premier tour de l'élection présidentielle française de 2022. Pour cela, nous utilisons plusieurs bibliothèques : Pandas pour traiter les données, NumPy pour traiter les calculs numériques avancés et enfin Matplotlib pour créer des boîtes à moustaches (*boxplot*). Les résultats sont présentés sous la forme de boîtes à moustache.

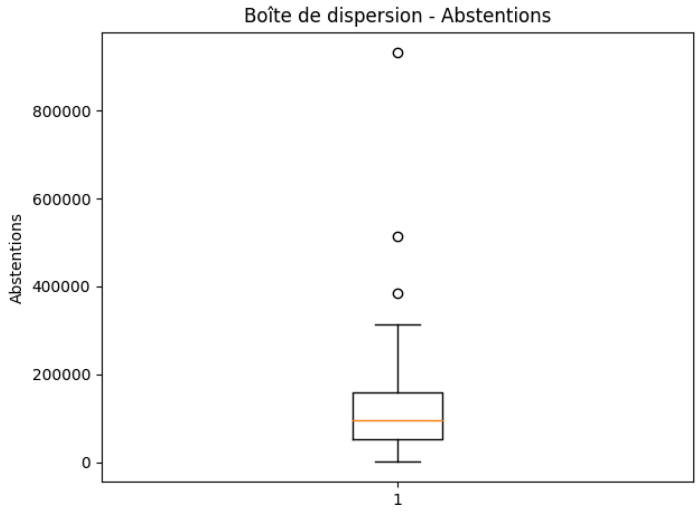
Grâce au code, nous avons pu catégoriser les variables quantitatives, cela étant utile car seules ces variables peuvent être utilisées pour calculer des paramètres de position et de dispersion.

Le code permet de calculer pour chaque colonne quantitative une *boxplot* : paramètres de position (moyenne et médiane), paramètres de dispersion (écart-type, écart absolu à la moyenne, étendue) et des paramètres de distribution (quartiles, déciles, distance interquartile et distance interdécile). Cela permet de mieux comprendre les distributions des données du premier tour de l'élection présidentielle de 2022. Par la suite, le code permet de créer visuellement des *boxplot* qui peuvent être analysées.

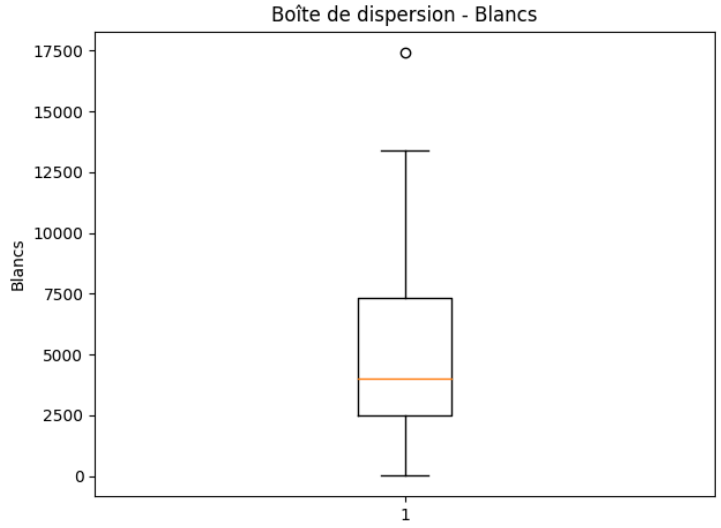
La *boxplot* des inscrits révèle une distribution fortement asymétrique. En effet, la grande partie des communes ont peu d'inscrits, alors que quelques grandes villes ont des effectifs très élevés. Cela permet de mettre en relief les fortes différences entre la taille de la population et celle des communes.



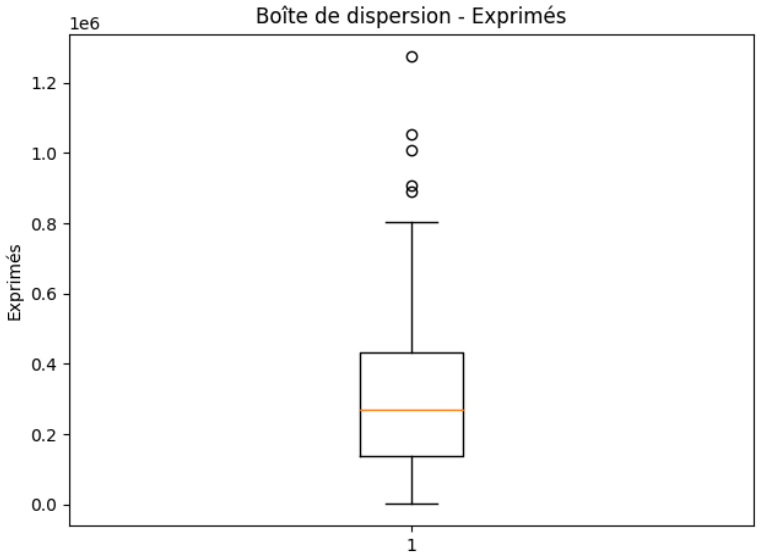
La *boxplot* des abstentions montre également une importante dispersion. La médiane est basse, donc la plupart des communes ont peu d'abstentionnistes. Les valeurs élevées correspondent aux grandes communes qui numériquement ont plus d'abstentionnistes.



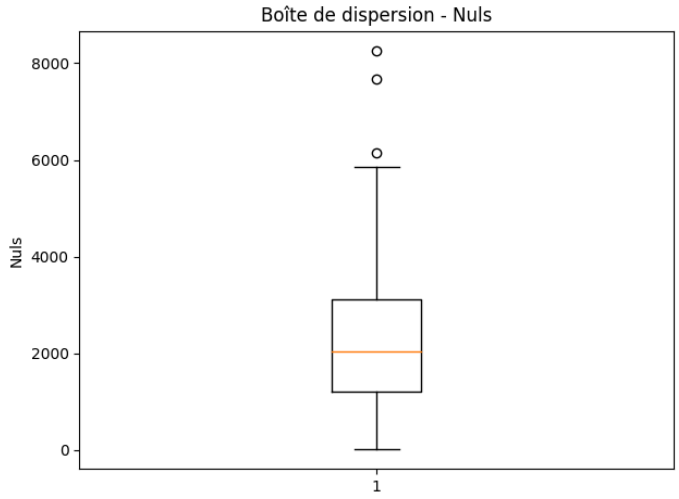
La *boxplot* des votes blancs révèle qu'il sont peu nombreux dans la plupart des communes, la médiane est donc faible.



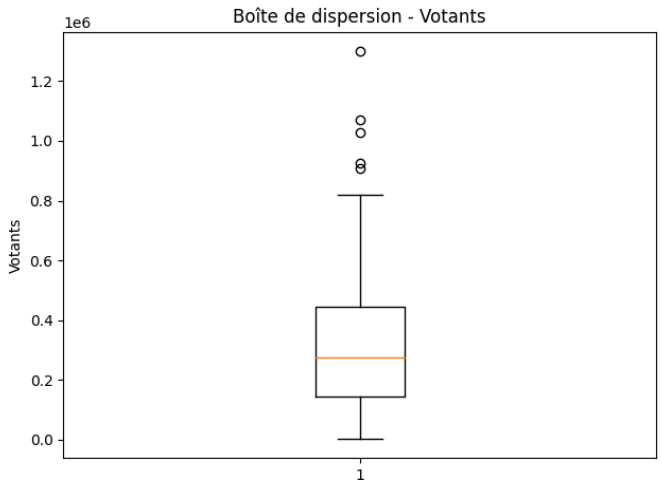
La *boxplot* des expressions montre une distribution asymétrique. En effet, la plupart des communes ont un nombre limité de votes exprimés alors que les grandes villes en ont un nombre élevé.



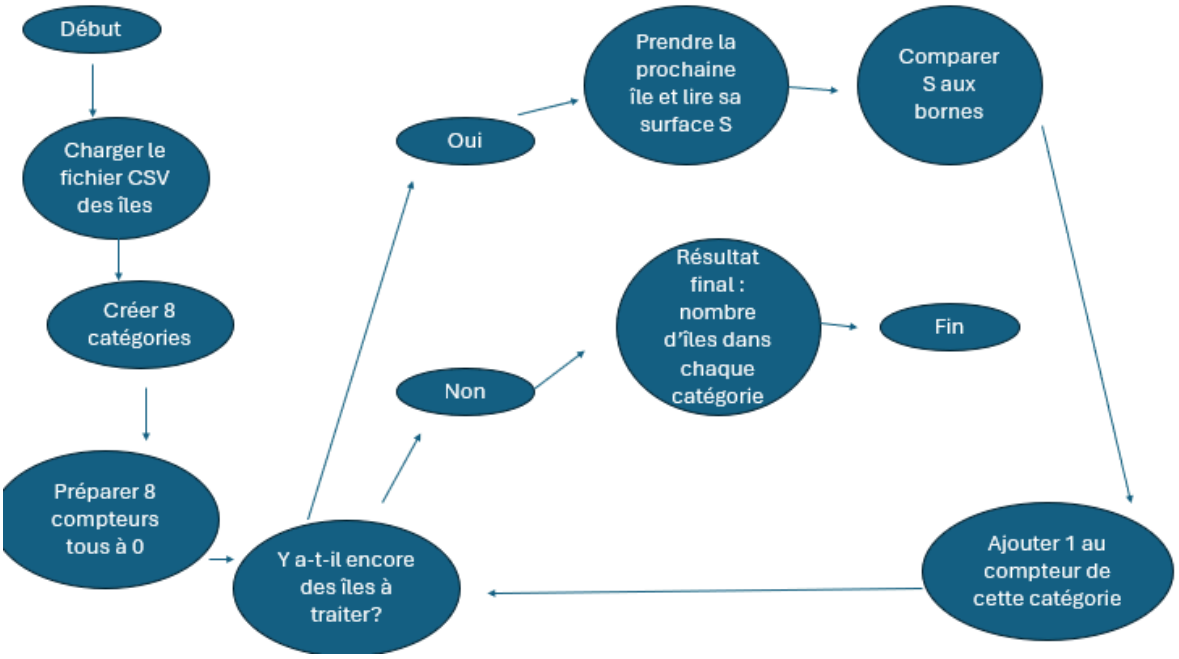
La *boxplot* des votes nuls montre une médiane située autour de 2.000, alors la moitié des communes ont un nombre de votes nuls inférieur à cette valeur. La dispersion est assez importante avec l'écart interquartile qui est large. Il existe également des valeurs aberrantes élevées, qui correspondent à des valeurs numériques importantes fondées sur la taille plus importante des grandes villes.



Enfin, la *boxplot* des votants révèle une médiane élevée de 250.000, il y a donc une forte hétérogénéité de la taille des communes, les données ont une grande dispersion.



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



La première étape est de charger le fichier CSV qui contient les îles. Ensuite, nous devons créer 8 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 très grandes : > 10000 km². Ensuite nous préparons 8 compteurs et nous intégrons une boucle pour laquelle chaque île a sa surface qui est lue. Elle 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.

Séance 4 : Les distributions statistiques.

4.1 Questions de cours

Les critères à mettre en avant pour choisir entre variable discrète et variable continue sont :

La nature du phénomène étudié : on choisit une loi discrète quand la variable prend un nombre fini ou dénombrable de valeurs (exemple : nombre de succès, nombre d'individus, tirages dans une urne). On choisit une loi continue quand la variable peut prendre une infinité de valeurs dans un intervalle (exemple : mesures continues).

La forme de la distribution empirique : on regarde comment se distribuent les données observées pour décider si un modèle discret ou continu est adapté.

Les caractéristiques statistiques des données : les moments mesurés peuvent orienter vers certains types de lois, par exemple une loi symétrique (normale) ou asymétrique (géométrique, loi Poisson etc.).

Le nombre de paramètres de la loi : certaines lois discrètes ou continues ont peu de paramètres (Bernoulli, uniforme), d'autres davantage (Zipf-Mandelbrot), influençant le choix du modèle le plus pertinent.

Plusieurs lois sont régulièrement utilisées en géographie :

La loi uniforme discrète est utilisée pour les sondages, qui sont très utiles en géographie humaine.

La loi binomiale 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 (succès/échec)

La loi de Poisson est la loi des événements rares. 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 (comme pour le trafic, les accidents, les occurrences spatiales) et pour la distribution dans l'espace de phénomènes rares.

La loi hypergéométrique est utilisée pour des contrôles de qualité dans lesquels on retire de la population étudiée les éléments défaillants. Cela est particulièrement utile en géographie dès qu'on échantillonne sans remise (enquêtes, populations finies)

La loi multinomiale permet de modéliser des catégories multiples, très communes en géographie et donne la probabilité d'observer plusieurs catégories simultanément. Cela correspond exactement aux distributions de populations en classes, catégories socio-professionnelles, type d'occupation du sol etc.

La loi normale est omniprésente dans les phénomènes géographiques. Cette loi régit sous des conditions très générales beaucoup de phénomènes aléatoires. Elle sert pour analyser des mesures continues, construire des intervalles de confiance, modéliser des distributions symétriques dans la nature ou la société.

La loi de Zipf (et Zipf-Mandelbrot) qui est la seule explicitement liée à un domaine géographique. En géographie, cette loi se rencontre dans les lois rang-taille confrontant, au sein d'un territoire, le nombre d'habitants d'une ville avec son rang. C'est une loi majeure en géographie urbaine.

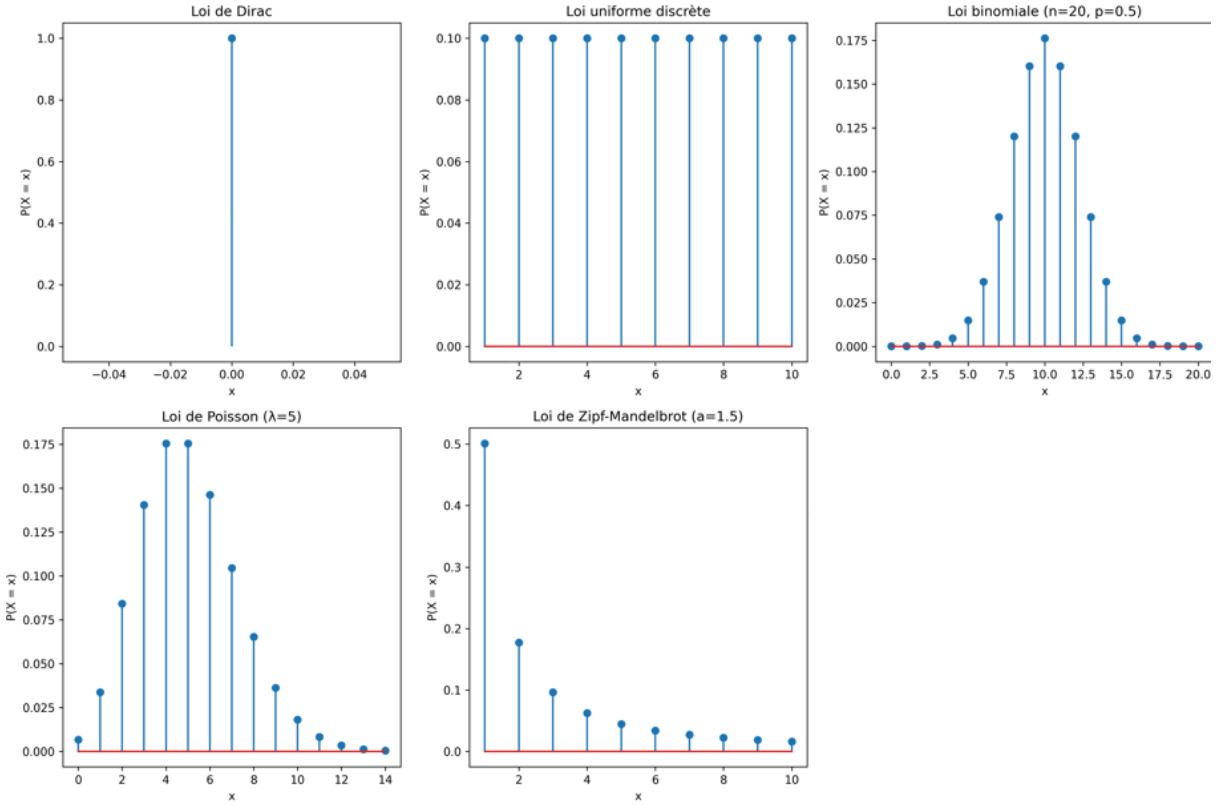
4.2 Analyse des résultats

Cette séance nous invite à nous pencher sur les distributions statistiques. Le but étant de savoir afficher une distribution statistique afin de comparer une distribution théorique avec une distribution observée.

Voici ci-dessous les analyses des graphiques obtenus.

Analyse des graphiques obtenus

Première série de graphiques obtenus



La loi de Dirac est illustrée par le premier graphique, elle est dite "dégénérée". Toute la masse de probabilité est concentrée en un point unique, cela est le cas sur notre exemple, le pic de probabilité est égal à 1.

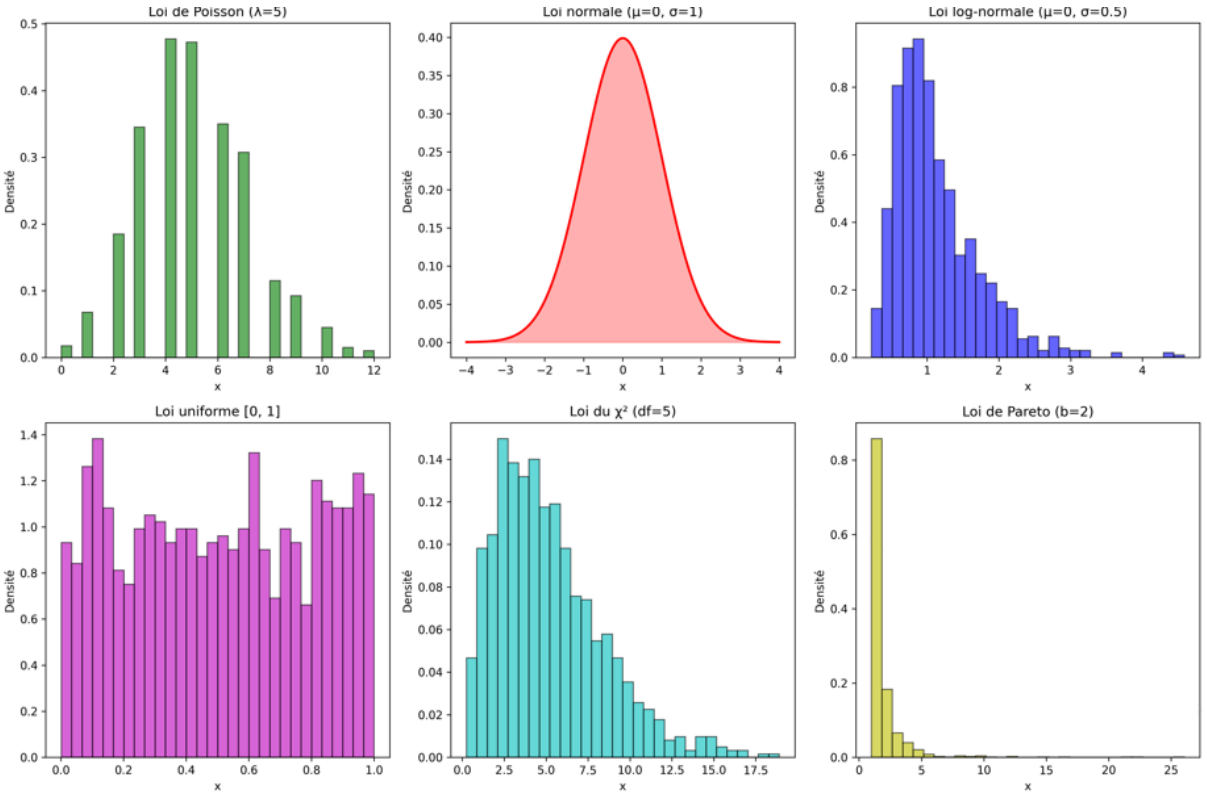
La loi uniforme discrète est illustrée par le second graphique. Cette dernière permet d'attribuer la même probabilité à chaque valeur possible au sein d'un ensemble fini. Nous pouvons constater cela sur le graphique, en effet les barres ont la même hauteur, ainsi les résultats ont la même probabilité d'advenir.

La loi binomiale est illustrée par le troisième graphique. Elle permet la modélisation du nombre de succès obtenus au sein d'un nombre fini d'essais indépendants, sachant que chaque essai a la même probabilité de succès. Alors, 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.

La loi de poisson est illustrée par ce quatrième graphique. 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.

La loi de Zipf-Mandelbrot est illustrée par le cinquième graphique. Elle permet de décrire la distribution de la fréquence des mots, des noms et d'autres entités dans un ensemble de données. Ce graphique révèle une forte inégalité entre les occurrences: la première valeur a une probabilité très élevée alors que la deuxième a une probabilité plus faible. C'est le phénomène de "longue traîne", la minorité domine largement alors que la majorité est peu représentée.

Deuxième série de graphiques obtenus



La loi de poisson en continue est représentée par le premier graphique. Les événements se produisent cette fois de manière continue et non pas discrète. Ce 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.

La loi normale est illustrée par le deuxième graphique. 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.

La loi log-normale est illustrée par le troisième graphique. Cette loi montre que le logarithme suit une loi normale, comme vue 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.

La loi uniforme est illustrée par le quatrième graphique. 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.

La loi du Khi-2 est illustrée par le cinquième graphique. Elle permet de 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.

La loi de Pareto est illustrée par le sixième graphique. Elle permet la modélisation 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.

5.1 Questions de cours

L'échantillonnage correspond au prélèvement dans une population dite mère d'une partie (n) de cette dernière. L'échantillon correspond à une quantité restreinte, un sous-ensemble de la population mère.

L'utilisation de la population entière n'est pas efficace car l'étude sur l'ensemble d'une population n'est pas réalisable d'un point de vue pratique, cela est trop cher.

Il existe 2 grandes catégories d'échantillonnage. La première est la méthode aléatoire avec des tirages sans remise, avec remise, des échantillonnages systématiques et des sondages aléatoires simples avec un tirage au sort. La seconde est la méthode non aléatoire avec la méthode des quotas, et la méthode d'échantillonnage "Monte Carlo".

Pour les choisir, cela dépend de la disponibilité de sondés, le coût des tirages, la possibilité d'avoir plusieurs échantillons, la taille de la population et le niveau de précision.

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. Il est estimé 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. Voici la formule : $[p - z\_C \times \sqrt{(p(1-p)/n)}, p + z\_C \times \sqrt{(p(1-p)/n)}]$

De son côté, 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 ($\hat{\theta}$) et la valeur à estimer (θ).

L'estimateur peut être sans biais, c'est à dire qu'en moyenne les erreurs s'harmonisent, l'estimateur procure la valeur juste.

L'estimateur peut être asymptotiquement sans biais si la limite de sa valeur n tend vers l'infini : $\lim(n \to +\infty)\ E(\hat{\theta}) - \theta = 0$

Enfin, l'estimateur peut être biaisé si l'estimateur 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. Elle est associée aux enquêtes, aux sondages et aux recensements.

Le lien avec la notion de données massives est que la population étudiée est un échantillon, or avec l'arrivée de données massives, on peut se poser la question d'étudier sur une population quasiment complète. Nonobstant, il ne faut pas oublier que l'on ne peut pas travailler souvent avec la totalité d'une population, alors si tous les individus ne sont pas représentés, les données sont biaisées.

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 convergent en probabilité vers la valeur à estimer quand n tend vers l'infini. De plus, l'estimateur doit être précis pour produire des résultats qui sont stables et ne varient pas entre les différents échantillons. Pour la mesurer il faut analyser l'erreur globale (erreur de la dispersion des résultats et erreur relative à l'éloignement de la vraie valeur). Ensuite, l'estimateur doit être efficace, pour cela il doit respecter 2 conditions : ne pas avoir de biais et ne pas présenter de grande dispersion. De plus, l'estimateur doit être robuste donc ne pas être influencé par les données extrêmes. Enfin, l'estimateur doit être exhaustif et être basé sur toutes les informations disponibles au sein des données.

Il existe 5 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 puis choisir la valeur du paramètre qui rend cet écart le plus petit possible. La seconde méthode est celle du maximum de vraisemblance. Il faut choisir la valeur du paramètre qui donne les données les plus probables. La troisième méthode est celle de l'intervalle de confiance, elle procure un encadrement qui a une probabilité d'être composée par 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. Cette méthode vise à créer des faux échantillons en tirant au hasard avec remise.

Pour choisir une méthode, nous pouvons activer plusieurs critères : la taille de l'échantillon, la disponibilité d'informations, la présence de données aberrantes, la recherche de robustesse, la nature des données, la connaissance ou non de la loi suivie par la variable.

Il existe 4 types de tests. Les premiers sont les spécifiques (tests d'ajustement, tests de comparaison, tests d'indépendance, ANOVA), les seconds sont les tests paramétriques (Student, Fisher-Snedecor, Khi-deux, test de normalité), les troisièmes sont les tests non paramétriques (Mann-Whitney, Wilcoxon, Kruskal-Wallis, Kolmogorov-Smirnov, Cramér-von Mises, Smirnov…).

Les tests statistiques sont utiles pour comparer des groupes, tester l'indépendance des variables, l'évaluation des hypothèses et enfin vérifier si des données correspondent à un modèle théorique.

Voici les 11 étapes pour créer un test :

1. Formuler une question scientifique / problématique
2. En déduire l'hypothèse à tester
3. Définir 2 hypothèses : une de travail et une 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 la zone où l'hypothèse neutre sera rejetée
10. Calculer la valeur obtenue à partir de l'échantillon
11. Conclure

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

5.2 Analyse des résultats

L'objectif de cette séance sur les statistiques inférentielles est d'utiliser les fonctions natives avec les méthodes pandas et de comprendre les trois théories qui permettent de valider un résultat en analyse de données. Les 3 théories sont les suivantes : la théorie de l'échantillonnage, celle de l'estimation et celle de la décision.

La première théorie est celle de l'échantillonnage donc des intervalles de fluctuation. Nous nous sommes posé la question suivante : Obtient-on toujours les mêmes résultats si on réalise 100 sondages différents sur une même population?

La population mère est de 2.185 personnes, avec 852 « Pour » (39%) ; 911 « Contre » (42%) ; 422 « Sans opinion » (19%). Nous générons 100 échantillons au hasard dans cette population. Nous calculons donc les moyennes sur les 100 échantillons, effectuons une conversion en fréquences et calculons l'intervalle de fluctuation à 95%. Cet intervalle permet de comprendre l'intervalle de variation des résultats en cas de répétition du sondage plusieurs fois.

La fréquence moyenne est de 0,39 et est donc inclus dans l'intervalle de fluctuation [0,36 ; 0,42], alors les échantillons sont représentatifs.

Ensuite, nous allons utiliser la théorie de l'estimation basée sur les intervalles de confiance. Le but étant de confirmer la fiabilité d'un échantillon, de savoir où se situe la vraie valeur dans la population

Le code permet d'isoler le premier échantillon, de calculer les fréquences observées et de calculer l'intervalle de confiance à 95%. Nous pouvons comparer l'intervalle de confiance avec la vraie valeur de la population mère, 0,39 étant incluse dans [0,38 ; 0,46], l'échantillon est considéré comme fiable.

In fine, nous mobilisons la théorie de la décision fondée sur les tests d'hypothèse. La problématique est de savoir si les données ont une distribution particulière, dans notre cadre nous devons connaître si parmi 2 séries de nombres, laquelle a une loi normale donc une courbe en cloche.

Le code permet de réaliser le test de Shapiro Wilk (normalité) et la règle de décision avec la p-value qui correspond à la probabilité que les données observées soient dues au hasard. Si p-value $\geq 0,05$, on ne peut pas rejeter la normalité donc la distribution est probablement normale, si p-value $< 0,05$, on doit rejeter la normalité, la distribution est non normale.

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

6.1 Questions de cours

Une statistique ordinale est une statistique qui a été appliquée à des variables qualitatives ordinales, on peut donc en faire un ordre.

Elles s'opposent aux statistiques nominales qui ne peuvent être ordonnées (exemple : occupation du sol : tourisme, commerce, nature, transport…).

Elle utilise des variables qualitatives ordonnées.

Cela se matérialise avec une hiérarchie spatiale, comme par exemple des indicateurs pour classifier l'attractivité des lieux touristiques.

L'ordre croissant est le plus simple afin d'identifier les valeurs extrêmes, même si la loi rang-taille est utilisée en géographie.

Une corrélation des rangs permet de calculer le degré de ressemblance entre 2 séries de rangs. Alors que la concordance de classement permet de connaître si plusieurs classements (plus de 2) 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 permet de calculer le surplus de paires concordantes relativement aux paires discordantes. Il varie entre -1 et +1.

De son côté le coefficient de Yule permet d'évaluer la force de l'association entre 2 modalités binaires, il varie également entre -1 et +1.

6.2 Analyse des résultats

L'objectif de cette séance est d'aborder la statistique d'ordre des variables qualitatives. Nous allons manipuler des fonctions locales et comprendre la nécessité de factoriser son code en une liste de fonctions ou de procédures exécutant une tâche unique. De plus, nous allons créer des fonctions locales spécifiques au traitement d'un problème et comprendre l'analyse de variables qualitatives ordinales.

Nous avons commencé par l'analyse des surfaces des îles et continents via la loi rang-taille. Après avoir isolé la variable quantitative Surface (km2), les surfaces des îles ont été complétées par celles des grands ensembles continentaux. Nous avons ordonné de manière croissante la liste

obtenue. La représentation de la relation rang-taille en échelle linéaire révélant une distribution déséquilibrée, ces quelques entités concentrent l'essentiel de la surface, alors que la majorité des îles ont de très faibles superficies. Pour rendre plus lisible le graphique, nous avons converti 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 relief 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 en l'espèce que d'une seule variable qualitative qu'est la surface.

Ensuite nous avons effectué le classement des États selon la population et la densité (2007-2025). Nous avons extrait les variables de densité en 2007, en 2025, de population en 2007 et en 2025 et les Etats.

Nous avons pu ordonner les États par ordre décroissant de la population et de la densité pour chaque année, comparer le classement entre 2007 et en 2025 et isoler les rangs.

Finalement, nous allons appliquer 2 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 élevés et positifs, cela permet d'affirmer la forte stabilité du classement des Etats selon leur population. Alors, 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. Alors, cela signifie la stabilité partielle des classements. Les variations de densité sont influencées par des dynamiques plus complexes comme par exemple les migrations et les politiques territoriales.

Humanités numériques : Réflexion sur les sciences des données et les humanités numériques.

Au terme de ce parcours débutant en analyse de données avec Python, il apparaît clairement que les sciences des données et les humanités numériques constituent aujourd'hui un champ en pleine expansion, marqué par une hybridation croissante entre technologie, traitement automatisé de l'information et traditions intellectuelles issues des sciences humaines. Mais que sont les humanités numériques ? Les humanités numériques se définissent comme l'ensemble des connaissances textuelles ayant subi une numérisation, et plus largement, l'intégration du numérique dans les pratiques de recherche en lettres, sciences humaines et sociales. Cette perspective s'inscrit dans une continuité historique : dès le XIX^{e} siècle, l'automatisation du traitement textuel accompagnait le progrès industriel, réintroduisant les lettres au sein du domaine des sciences sous une forme renouvelée.

Aujourd'hui, le numérique n'est plus seulement un outil technique, mais un véritable objet de recherche et un instrument de communication, capable de rapprocher les connaissances scientifiques et de recomposer les projets des humanités du XVI^{e} au XX^{e} siècle. Il s'impose comme un facteur commun aux sciences et aux lettres, ouvrant de nouvelles perspectives interdisciplinaires. Les humanités numériques poursuivent ainsi plusieurs objectifs : revitaliser les filières des humanités et l'écriture scientifique, transformer les pratiques de recherche, encourager les collaborations entre disciplines et articuler approches qualitatives et quantitatives. Elles s'organisent également autour d'enjeux majeurs, tels que l'inscription des humanités dans la modernité technologique, la diversité des formes d'édition ou encore l'analyse des représentations et des usages du numérique.

Cependant, malgré leurs apports, les humanités numériques comportent certains risques : celui de privilégier les méthodes quantitatives au détriment de l'interprétation, de sous-estimer le contexte social de production des données, d'accentuer l'hégémonie de l'informatique dans les sciences humaines et sociales ou encore, de confondre le traitement des données avec la compréhension du sens. Ces limites mettent en évidence la nécessité d'une vigilance éthique et méthodologique : il revient au chercheur de connaître les outils, d'en maîtriser les limites et de maintenir un regard critique sur les pratiques employées.

De plus, l'évolution des liens entre humanités et informatique, des premières métadonnées des années 1970 jusqu'à l'avènement des humanités numériques, témoigne d'un processus structurant. L'usage croissant des bases de données a permis de cataloguer et relier les documents ; la numérisation du patrimoine a rendu accessibles des corpus volumineux ; le Web sémantique a offert aux machines la possibilité de comprendre et relier les données selon leur sens. Ces transformations ont largement contribué à la structuration des humanités numériques et à leurs enjeux contemporains.

Les cinq étapes de leur méthodologie (trouver l'information, modéliser les données, numériser les sources, analyser le contenu et valoriser les résultats), trouvent un écho direct dans ma pratique débutante de Python. Qu'il s'agisse de nettoyer un jeu de données, d'écrire des scripts d'automatisation ou de produire des visualisations (tableaux, graphiques, boîtes à moustaches, Khi², etc.), chaque étape rappelle que le numérique n'est pertinent que s'il est articulé autour d'une réflexion sur la structure, le sens et les usages des données.

Ainsi, ce parcours débutant trouve également des points d'ancrage concrets avec d'autres domaines de mon Master 1 GAED Géopolitique-GEOINT, notamment les Systèmes d'Information Géographique (SIG) et le Geospatial Intelligence (GEOINT). Dans le cours de SIG dispensé par M. de Matos-Machado, je manipule quotidiennement de vastes ensembles de données spatialisées (rasters, vecteurs, données attributaires), accompagnées de métadonnées fournies par des institutions comme l'Institut National de l'information Géographique et Forestière (IGN). Leur prise en compte est indispensable pour garantir la validité des analyses spatiales. L'apprentissage de Python m'a ainsi permis de mieux comprendre ces enjeux, notamment en automatisant le nettoyage, en détectant les incohérences ainsi que les données aberrantes et en structurant des bases complexes, ce qui rejoint directement les problématiques de rigueur et de normalisation propres aux SIG.

En GEOINT, champ dédié à la fusion de données géolocalisées multi-sources et multi-capteurs, la gestion de données hétérogènes soulève des défis supplémentaires : compatibilité, interopérabilité, sécurité, protection de l'information. La difficulté à faire dialoguer des bases issues de services différents illustre parfaitement les enjeux des humanités numériques : comment croiser et interpréter des données produites dans des contextes variés ? La fusion de données satellitaires, terrestres, institutionnelles ou ouvertes met en lumière la nécessité d'une méthodologie rigoureuse et d'une compréhension fine des contraintes techniques et sémantiques.

Même des outils plus simples comme Excel (lien avec le cours "Méthodes quantitatives" de Madame Huguenin-Richard) montrent combien la manipulation des données repose sur une compréhension fine de leur structure : trier, filtrer, structurer un tableau ou vérifier la cohérence des valeurs constituent une première approche de la logique de la donnée. La transition entre Excel et Python n'est donc pas une rupture mais plutôt une montée en complexité, permettant une automatisation poussée, une reproductibilité accrue et une capacité à traiter des volumes bien plus importants. Ces compétences se retrouvent au cœur de la *cultural analytics*, des *digital methods* ou encore de la *distant reading*, qui renouvellent l'étude des textes, des images, des discours et des comportements numériques. Elles montrent que la maîtrise des outils ne peut être dissociée d'une réflexion critique sur la nature des données et leurs analyses.

Au moyen de cette initiation à l'analyse de données de Python, j'ai pu développer des compétences techniques essentielles pour les sciences humaines et sociales. Elle m'a également fait prendre conscience de la place désormais centrale du numérique dans la production de connaissances, en particulier dans la pratique de la géographie. Les disciplines des sciences humaines et sociales mobilisent aujourd'hui une grande diversité de jeux de données (spatiales, textuelles, statistiques ou multimédias). La maîtrise d'outils tels que Python devient alors indispensable pour les structurer, les analyser et en tirer des interprétations pertinentes.

Ce cours m'a également permis de comprendre que les humanités numériques ne constituent pas une rupture avec les traditions intellectuelles, mais bien un prolongement des méthodes d'analyse et un accélérateur des traitements et des opérations complexes. Elles permettent à la fois de revisiter des questionnements anciens grâce à de nouvelles capacités techniques, tout en renforçant la complémentarité entre rigueur scientifique, interprétation qualitative et analyse quantitative.

Cependant, j'ai pris conscience que cette évolution rapide des outils impose une vigilance éthique. À l'avenir, les humanités numériques et l'analyse de données seront donc confrontées à des défis majeurs qui redéfiniront profondément nos manières de produire, d'organiser et d'interpréter le savoir. L'augmentation continue des volumes d'information, leur diversification et leur circulation mondiale imposeront des capacités techniques et théoriques toujours plus pointues, notamment en matière de gestion des big data, de standardisation, de sécurité et de transparence des méthodes. Les frontières entre disciplines continueront de s'estomper, tandis que les questions de souveraineté numérique, de protection des données, de biais algorithmiques et de traçabilité des modèles deviendront centrales dans les sciences humaines et sociales. L'arrivée d'IA génératives, capables d'interagir avec les corpus, les images ou les données spatiales, posera également de nouveaux dilemmes : Comment garantir la fiabilité des analyses ? Comment préserver l'intégrité du sens lorsque les traitements sont automatisés ? Quelle place accorder à l'interprétation humaine dans des environnements où l'automatisation devient la norme ? Fort de ce constat, je dois ainsi veiller à prendre du recul sur les technologies utilisées, en particulier sur les intelligences artificielles qui s'intègrent progressivement aux *workflows* d'analyse de données. Leur développement transforme en profondeur les pratiques : automatisation accrue, assistance aux traitements, génération de contenus. Plus que jamais, l'avenir des humanités numériques dépendra de notre capacité à concilier innovation technologique, exigence scientifique et responsabilité éthique, afin de construire une recherche qui demeure à la fois rigoureuse, critique et profondément humaine.