Flora Soulier
M1 GAED EnviTerr

SORBONNE UNIVERSITÉ

Rapport d'activité d'analyse de données

Séance 2

Questions :

Dans le domaine de la géographie, les statistiques sont inclues dans les « informations géographiques ». Elles servent alors à mesurer, comparer et modéliser des phénomènes spatiaux mais aussi à rester objectif dans l'analyse de données. Cela permet de donner une rigueur scientifique à la géographie. Lors des phénomènes dits « aléatoires » par l'irrégularité de leur fréquence et le fait que certains facteurs ne soient pas prévisibles, le hasard peut être évoqué. Cependant, les causes peuvent être éclairées par les statistiques. Le caractère aléatoire peut-être prouvé et à l'inverse, les caractères réguliers le sont également. Les statistiques réduisent ainsi le hasard et l'expliquent.
Les types d'information géographique sont les suivants :
- Données quantitatives
- Données qualitatives
- Données spatiales
- Données temporelles
- Données relationnelles
- Données perceptuelles (=non objective)
- Données cartographiques

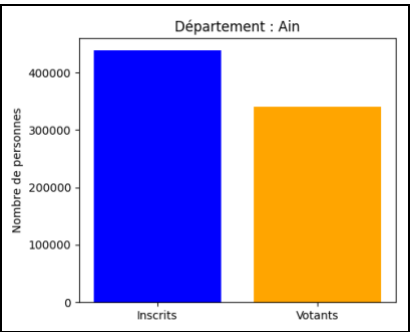
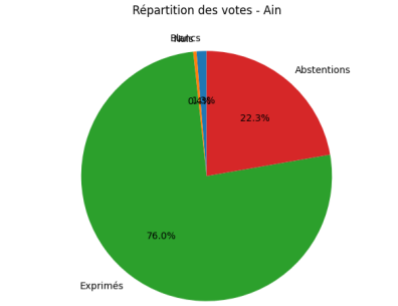
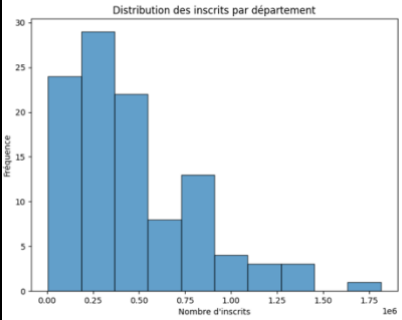
En géographie, les besoins liés à l'analyse de données sont variés : Décrire les territoires de manière objective pour connaître les réalités spatiales, montrer les régularités spatiales en repérant des motifs spatiaux comme des zones de concentration par exemple. Prévoir en modélisant les évolutions spatiales à l'aide de modèles statistiques et spatiaux, expliquer les variables et leurs liens, représenter les résultats (cartographie, graphiques…) et enfin, croiser les disciplines (données variées).
Les statistiques descriptives résument et décrivent les données de manière simple. Ces statistiques ne permettent pas d'émettre d'hypothèse. Les statistiques explicatives cherchent à comprendre en mettant en lien les variables entre elles. Pour résumer, les statistiques descriptives répondent à la question « Quoi ? » quand les statistiques explicatives répondent à la question « Comment ? ».
Il existe 4 types de visualisation de données en géographie. Tout d'abord, Les cartes thématiques pour représenter un phénomène à partir des données ; les graphiques statistiques qui constituent un résumé des données (histogramme, diagramme) ; les visualisations spatiales avancées pour des réalisations en 3D ; les visualisations non cartographiques pour les données qui ont un caractère spatial mais qui ne nécessitent pas de carte. Le choix entre ces différents types de visualisation s'opère en fonction de ce que l'on veut représenter, du type de données mais aussi de comment le phénomène est réparti. Sans oublier de prendre en compte le public visé.
Il existe plusieurs méthodes d'analyse de données, à choisir selon l'usage que l'on veut en faire. Les méthodes descriptives pour résumer, les méthodes explicatives pour expliquer les relations entre les variables, les méthodes multivariées pour analyser plusieurs données en même temps. Mais aussi les méthodes d'analyse spatiale pour identifier la localisation et les méthodes de modélisation pour simuler et donc prévoir.
Une population statistique est l'ensemble observé, est l'élément de cet ensemble. Et l'on mesure le caractère sur une population. Le caractère peut être quantitatifs ou qualitatifs.

Dans le premier cas ils peuvent être discrets, c'est-à-dire que les valeurs entières uniquement sont prises en compte. Ils peuvent aussi être continu et les valeurs sont donc toutes considérées. Dans le cas où il s'agit de caractères qualitatifs ils peuvent être nominaux lorsqu'il n'y a pas d'ordres logiques. Ils sont ordinaux si un ordre logique existe.
Il existe une hiérarchie entre ces caractères. En effet, le caractère qualitatif nominal est le plus petit niveau d'informations. Nous retrouvons ensuite le caractère qualitatif ordinal puis quantitatif discret et enfin continu qui permet une mesure complète.
Enfin, les modalités statistiques sont les variantes des éléments d'une population.

Une amplitude est la différence entre la valeur maximale et la valeur minimale.
A = valeur max – valeur min
Une densité se calcule en divisant l'effectif total par la surface du territoire.
La formule de Sturges est le calcul du nombre de classes et permet de regrouper les données quantitatives continues. Elle est utile pour déterminer des classes de valeur et à préparer les données pour des calculs. Quant à la formule de Yule mesure la concentration d'un phénomène c'est-à-dire à quel point un ensemble est dominé par d'autres.
Un effectif est le nombre d'individus qui ont la même valeur du caractère étudié. À partir de cet effectif on peut calculer la fréquence qui est la proportion de l'effectif d'une modalité par rapport à l'effectif total. Elle mesure le poids d'une catégorie dans l'ensemble.
F = effectif de la modalité/ effectif total
Une fréquence cumulée est la somme des fréquences. Cela permet de connaître le nombre d'individus qui ont une valeur inférieure ou égale à un seuil donné. Enfin, une distribution statistique est une vue d'ensemble de la répartition des données étudiées.

Exercices :

Question 3	Ouverture du main.py dans VS code	
Question 6	Nombre de lignes Nombre de colonnes	
Question 7	Affichage du type de colonnes	
Question 8	Affichage du nom des colonnes	
Question 9	Affichage du nombre d'inscrits	
Question 10	Affichage de la somme des colonnes quantitatives	

Question 11	Création de diagrammes par département	
Question 12	Création de diagrammes circulaires par département	
Question 13	Création de l'histogramme de la distribution des inscrits	

La bibliothèque Pandas permet de manipuler des données pour les analyser et y voir plus clair. En effet, Pandas permet de visualiser les données sous forme de tableau (DataFrame) et facilite alors la lisibilité des données. De plus, il est plus simple de filtrer les données et de les trier.
La bibliothèque Matplotlib permet la création de graphiques. Cela rend plus simple l'analyse de données puisque le code permet de générer immédiatement un certain nombre de graphiques.

Séance 3

Le paramètre statistique est très général pour l'analyse de données et son caractère quantitatif est le plus utilisé puisque dans un premier temps, la plupart des outils de calculs tels que la moyenne, l'écart-type ou la médiane ne s'applique qu'aux caractères quantitatifs alors que les caractères qualitatifs permettent uniquement d'établir un mode ou des fréquences. Les caractères quantitatifs permettent alors une analyse plus élargie. Cependant, il existe deux types de variables quantitatives et elles ne sont pas utilisées de la même manière et dans le même but. En effet, les variables discrètes sont des nombres fini, dénombrable de valeurs, c'est-à-dire que les valeurs sont séparées et souvent entières. Il s'agit de valeurs isolées pour lesquelles on utilise des sommes dans les méthodes de calcul. Quant aux caractères quantitatifs continus, leurs différents calculs font appel à des intégrales. Il s'agit de valeurs comprises dans un intervalle et elles sont mesurables.

Il existe justement plusieurs types de moyennes pour effectuer des calculs sur les différents types de caractères quantitatifs. Les variables discrètes se calculent par des sommes (moyenne géométrique par exemple) et les variables continues se calculent par des intégrales (moyenne géométrique continue par exemple). Ainsi chaque moyenne répond à une situation et une valeur différente, c'est pour cela qu'il en existe plusieurs. Cependant, une moyenne peut être plus ou moins sensible et pour contourner cela et éviter un calcul trop biaisé, on calcul une médiane qui est moins sensible aux valeurs extrêmes puisqu'elle sépare la population étudiée en deux parties égales. Enfin, le mode n'existe pas toujours mais il permet d'observer la variable quantitative maximale, cela admet qu'il existe bien une variable qui revient fréquemment.

La médiale permet de mesurer une concentration en partageant « en deux parties égales la masse totale du caractère. » et ainsi d'avoir un autre point de vue sur la distribution étudiée puisqu'elle met en valeur le partage de la quantité totale et non des individus. Puisque ce calcul permet de mesurer la concentration, on peut mesurer en la comparant avec la médiane ; l'inégalité. On l'observe lorsque la médiale est plus élevée que la médiane. L'inégalité de la distribution se mesure aussi avec l'indice de Gini. Cet indice est matérialisé par une courbe et une diagonale qui représente l'égalité parfaite, plus la courbe s'éloigne de la diagonale, plus l'inégalité est forte.

Les écarts à la moyenne sont inutilisables pour mesurer la dispersion puisque la somme des écarts est toujours de 0. Or, si on utilise la variance, on prend en compte le carré des écarts et ils ne s'annulent pas, la somme n'est donc plus de 0. Cependant, la variance s'exprime en carré, l'écart-type permet alors de faire revenir la variance à la même unité que la moyenne par exemple. L'étendue permet de calculer l'amplitude des valeurs extrêmes, l'écart-type et la variance permettent de prendre en compte la dispersion moyenne et proposent ainsi une vision plus précise que l'étendue sur la distribution. La dispersion centrale se mesure par la création de quantiles dans un premier temps. Le découpage par quantile permet en effet d'observer et décrire la position d'une valeur dans la distribution. Ce découpage sert ensuite à calculer l'écart interquartile. Cet écart peut être matérialisé par la taille d'une boîte à moustache dont le début et la fin de la boîte correspondent aux différents quartiles précédemment séparés (Q1, Q3). Cette boîte permet d'avoir l'écart-interquartile, la médiane et les valeurs minimales et maximales pour « représenter visuellement une série statistique ».

La différence entre un moment centré et un moment absolu est que le moment absolu mesure toujours la dispersion de manière positive contrairement au moment centré. Il mesure

la dispersion en évitant que les écarts négatifs s'annulent avec les positifs. Le moment centré permet lui de caractériser la moyenne de la distribution.

Si une distribution est symétrique, cela indique que les valeurs extrêmes sont équilibrées mais l'asymétrie montre aussi des tendances naturelles dans les distributions. Il est alors utile de vérifier si la distribution est asymétrique ou non car si c'est le cas il peut être nécessaire de modifier des données. Cette symétrie peut se mesurer par un graphique histogramme, une boîte à moustache ou des graphiques quantile-quantile.

Exercices :

Question 4	Ouverture du fichier main.py dans VS code	
Question 5	Affichage des colonnes quantitatives et calcul de leur moyenne ; médiane ; écart-type ; écart absolu à la moyenne ; étendue	
Question 6	Affichage en liste des paramètres calculés	
Question 7	Calcul de la distance interquartile et interdécile de chaque colonne avec Pandas	
Question 8	Affichage et création de boîtes à moustache	
Question 9 & 10	Ouverture du second fichier et catégorisation puis dénombrement des « îles »	
	Création d'un organigramme	

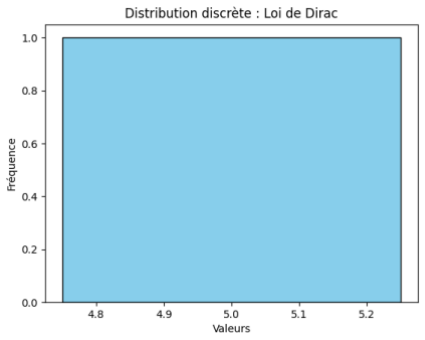
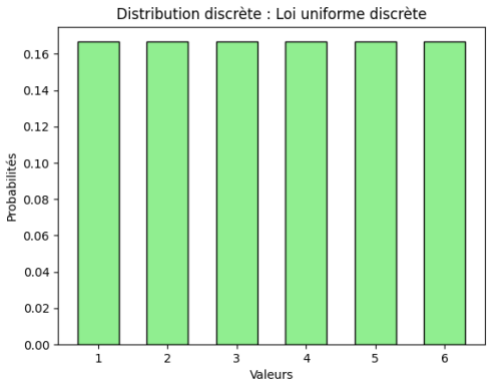
Cette séance permet d'obtenir des calculs en grand nombre pour optimiser l'analyse de données (moyenne, médiane…). Il se concentre davantage sur les valeurs interquartiles pour créer des boîtes à moustaches. Je suis plutôt satisfaite de cette première partie d'exercice puisque les boxplot sont lisibles et parlantes.
Cependant, pour la seconde partie de l'exercice, j'ai eu des difficultés à ce que Python reconnaisse mes colonnes comme des valeurs numériques. J'ai en effet, fait l'erreur d'ouvrir le CSV avec Excel, ce qui a détérioré mon CSV en incluant des virgules dans toutes les colonnes. Ainsi, les valeurs étaient reconnues comme du texte. Après rectification et création d'un Dataframe, l'intitulé de la colonne voulu est apparu et la catégorisation a pu se faire.

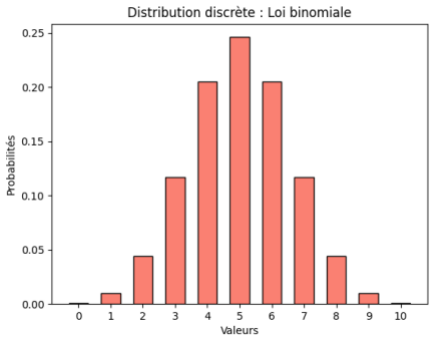
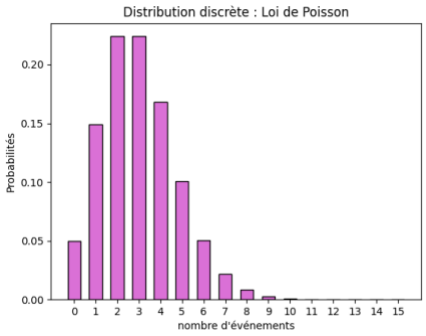
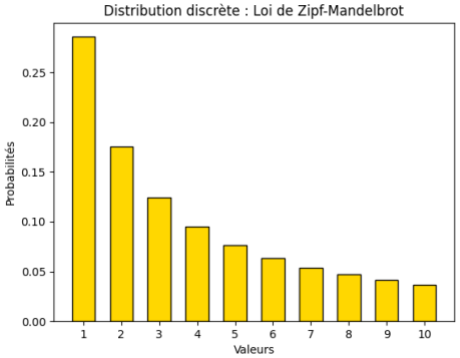
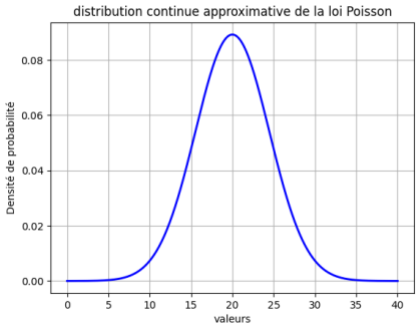
Séance 4 :

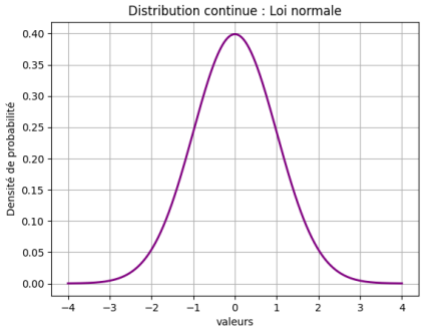
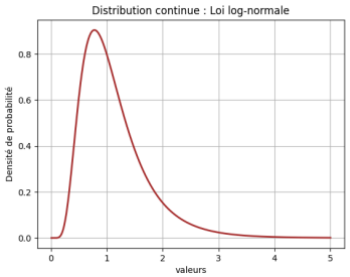
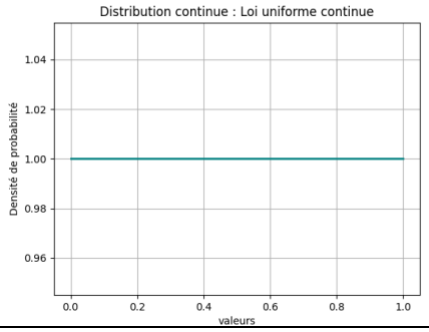
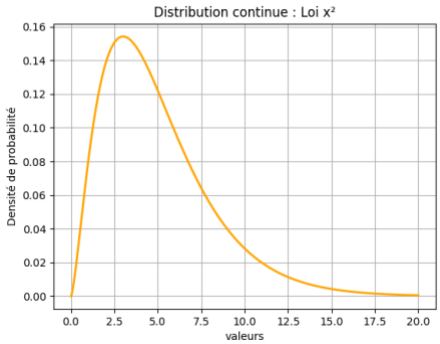
Questions :

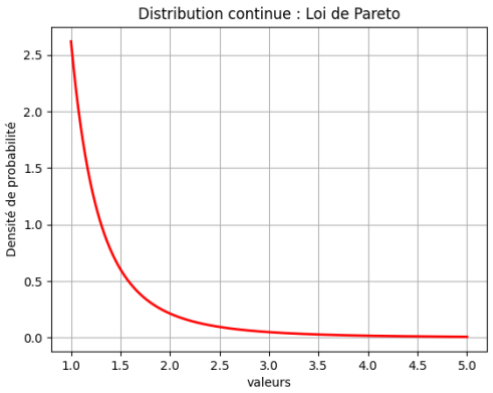
Il existe les variables discrètes et les variables continues. La nature des variables permet de choisir quelle distribution est la plus adaptée à la situation. En effet, si la variable est un nombre fini, alors on utilise la distribution discrète mais on utilise la distribution continue lorsque la variable est un chiffre compris dans un intervalle. La représentation de la distribution permet également de choisir. Si le but est de représenter une courbe, alors on préfère les lois qui s'appliquent aux distributions continues et non aux distributions discrètes qui elles, représentent les variables sous la forme d'histogrammes. Enfin, le choix dépend aussi du contexte puisque certaines lois sont adaptées à la mesure de phénomènes réels. Par exemple, le comptage d'évènements rares se fait avec la loi Poisson puisque les variables sont indépendantes et sur de petits intervalles temporels. En géographie, selon moi, les lois les plus utilisées sont les lois « rang-taille », par conséquent la loi de Zipf. Une loi rang-taille compare le rang d'une ville avec sa taille et établit alors la relation suivante entre les deux variables : Plus une ville est petite (en taille), plus sa population diminue. Cette loi permet d'analyser des rapport « rang-taille » donc mais aussi d'évaluer la hiérarchie, d'une série de villes par exemple.

Exercices :

Question 1	Calcul de la Loi Dirac	
	Calcul de la loi Uniforme discrète	

	Calcul de la loi Binomiale	
	Calcul de la loi de Poisson	
	Calcul de la loi de Zipf	
	Calcul distribution continue loi Poisson	

	Calcul distribution continue loi normale	
	Calcul distribution continue loi log-normale	
	Calcul loi uniforme continue	
	Calcul loi de x2	

	Calcul loi de Pareto	
Question 2	Calcul de la moyenne et de l'écart-type des variables discrètes	
	Calcul de la moyenne et de l'écart-type des variables continues	

Cette séance fait appel à la théorie et est composée de code simple pour chaque calcul. Cela permet de garder en mémoire ces formules.
J'ai cependant rencontré un problème dans la création des graphiques. Au fur et à mesure des formules entrées, le graphique précédent s'ajoutait au nouveau. Il faut simplement écrire « plt.clf » la fin de la création de chaque graphique pour repartir de 0.

Séance 5 :

Questions :

L'échantillonnage consiste à prélever un sous ensemble d'individus à partir d'une population mère pour évincer certains éléments dans l'analyse de données. L'étude se base sur une part limitée de la population car l'analyse de l'entièreté de cette dernière serait impossible. Il existe 5 méthodes d'échantillonnage. Premièrement, les méthodes dites aléatoires sont les suivantes : échantillonnage aléatoire simple par tirage avec ou sans remise. Ensuite, les méthodes Monte-Carlo établies sur des tirages aléatoires répétés. La méthode des quotas est une méthode non aléatoire qui consiste en la reproduction des proportions connues de la population dans l'échantillon. L'échantillonnage systématique par la sélection des individus dans une population (nombre à définir, fixer l'intervalle systématique, établissement aléatoire d'un point de départ). Le choix d'une méthode d'échantillonnage dépend tout d'abord de l'existence ou non d'une base de sondage. Ensuite, le niveau de précision souhaité ne fait pas appel aux mêmes méthodes, les méthodes les plus précises sont l'échantillonnage aléatoire simple et l'échantillonnage systématique ou encore la méthode des quotas si menée correctement. La structure de la population influence également le choix de la méthode puisque cette dernière est différente suivant de s'il s'agit d'une population homogène (aléatoire simple), ordonnée, hétérogène (quotas). Enfin, le temps disponible à l'analyse détermine aussi la méthode. Par exemple, la méthode des quotas ou systématique est plus rapide que les méthodes Monte-Carlo.
Un estimateur détermine une valeur inconnue d'un paramètre à partir d'un échantillon. Il s'agit d'une variable aléatoire puisqu'il dépend de l'échantillon choisit. Suite à l'application de l'estimateur à l'échantillon, on obtient l'estimation sous forme de valeur numérique.
Un intervalle de fluctuation permet de prévoir la variabilité d'une statistique comme la variation de l'estimateur par exemple. Quant à l'intervalle de confiance, il estime un paramètre inconnu. L'intervalle de confiance se construit donc autour de l'estimateur puisqu'il intervient pour proposer un certain nombre de valeurs probables pour identifier le paramètre inconnu. L'estimateur comprend un biais dans la théorie de l'estimation. Ce biais mesure l'écart moyen entre l'estimateur et le paramètre réel que l'on veut estimer. Ainsi, si le biais est égal à 0 alors l'estimateur n'est pas biaisé et donne la vraie valeur du paramètre recherché. Dans le cas contraire, il surestime ou sous-estime le paramètre.
Dans le cas où l'on ne parle pas d'échantillon et donc pas d'estimateur mais d'une population totale, on parle de paramètre. Ce dernier représente la valeur réelle de toute la population. Le paramètre intervient à la place de l'estimateur lorsqu'on a la possibilité de calculer toutes les variables à partir des Big Data.
Le choix d'un estimateur est important puisqu'il est nécessaire qu'il donne l'estimation ponctuelle la plus proche possible du paramètre. Il faut donc qu'il soit cohérent et robuste pour garantir une certaine fiabilité lors du calcul. Le choix du meilleur estimateur doit prendre en compte des statistiques exhaustives pour éliminer le biais de variance minimale ou en tenant compte de la quantité d'information de Fisher pour avoir davantage d'informations sur la précision. L'estimation d'un paramètre se fait selon différentes méthodes, pour obtenir des caractéristiques fiables d'une population. Premièrement, la méthode du bootstrap qui utilise les méthodes de Monte-Carlo. Cette méthode procède à des tirages au sort et est efficace pour les variables aberrantes. La méthode par intervalle de pari permet de calculer un intervalle plus fiable que l'intervalle de confiance. La méthode des moindres carrés est plus puissante

pour calculer plusieurs variables aléatoires, qui sont des espérances. L'estimation par la méthode du maximum de vraisemblance augmente la probabilité des données puisque la fonction de vraisemblance trie les données selon leur probabilité. Un test statistique permet de décider de la compatibilité d'une hypothèse sur un paramètre avec les données. Dans le but de créer un test on formule donc dans un premier temps une hypothèse, on choisit la statistique du test puis on calcule la statistique avec les données en fixant le niveau de signification. Enfin, on compare la valeur puis on interprète le résultat. Il permet de comparer des moyennes, de valider une hypothèse, un modèle, ou encore de contrôler la variance. Une statistique inférentielle permet d'estimer des paramètres inconnus et de tester des hypothèses à partir d'un échantillon. Même si elle est utile, elle ne remplace pas le jugement et la prise en compte du contexte. En effet, il est nécessaire d'intervenir pour tenir compte des hypothèses et de la pertinence ou non des données.

Exercices :

Question 1	Ouverture du fichier par la fonction « ouvrirUnFichier() »	
Question 2	Calcul de la moyenne de chaque colonne et introduction de la fonction round() pour arrondir le nombre obtenu	
Question 3	Étapes du calcul des fréquences du premier fichier	
Question 3b	Étapes du calcul des fréquences du second fichier	
Question 4	Calcul de l'intervalle de fluctuation L'intervalle de fluctuation ne porte pas sur les valeurs réelles de la population mère mais quantifie la variabilité des échantillons. Cependant, l'intervalle permet de savoir ou non la nature des estimateurs. Si un échantillon se situe dans l'intervalle, il est compatible avec la population supposée. À l'inverse, s'il se situe hors intervalle, cela contribue à contredire l'hypothèse.	
Question 5	Étapes de la théorie de l'estimation par la construction d'intervalles de confiance. Il faut calculer les fréquences puis l'intervalle de chaque opinion du fichier 1.	
Question 6	L'intervalle de confiance est moins large que l'intervalle de fluctuation. En effet, prenons l'opinion « pour ». Pour l'intervalle de fluctuation, les valeurs de la moyenne de la population qui sont compatibles avec l'échantillon se situent entre 0,29 et 0,48. Or, pour l'intervalle de confiance, ces valeurs se situent entre 0,38 et 0,41. On prend en compte alors plus de valeurs en utilisant l'échantillon de fluctuation.	

Question 7	Étapes de la théorie de la décision par le test de Shapiro selon deux fichiers représentant chacun une distribution	
Question 8	Aucune de ces distributions ne suit une loi normale. En effet, le test révèle qu'il est peu probable que ces valeurs suivent une loi normale. D'après moi c'est parce que la distribution est trop plate ou trop concentrée autour de la médiane. Ce qui est incompatible avec la loi normale.	

Cette séance permet l'apprentissage de nouvelles fonctions locales.

Séance 6

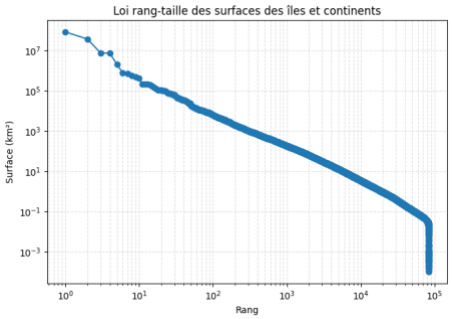
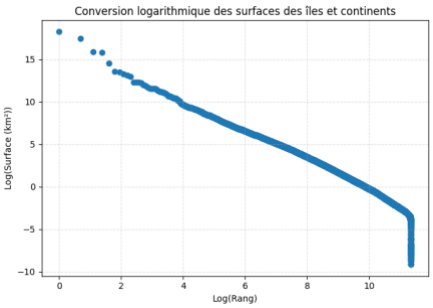
Questions :

Les statistiques ordinales utilisent des variables qualitatives qui possèdent un ordre logique (naturel ou croissant). Le fait que les données suivent un ordre permet de les trier et de chercher des variables aberrantes par exemple. Ces statistiques s'opposent aux statistiques nominales qui ne suivent pas d'ordre. Les statistiques suivant un ordre permettent de matérialiser une hiérarchie spatiale puisque cela permet de classer des objets géographiques tels que des villes par exemple, à l'aide de la loi rang-taille. Les classements peuvent être matérialisés par des tableaux.
L'ordre croissant est à privilégier pour établir des classifications. En effet, les valeurs étant classées par ordre de grandeur, il est plus facile de repérer les variables trop grandes, trop petites ou bien aberrantes.
La corrélation des rangs permet de mesurer l'association entre deux variables ordinales. En effet, en comparant leur rang, on mesure la force et le sens de la relation entre les deux variables. La corrélation des rangs est composée de deux tests. Dans un premier temps, le test de Spearmanr est utilisé pour évaluer des relations entre variables ordinales. Quant au test de Kendall, il permet de savoir si les variables peuvent être considérées comme statistiquement indépendantes. Le test de Spearmanr est à favoriser si l'on dispose d'une variable qualitative ordinale et d'une variable quantitative alors que le test de Kendall est souvent utilisé pour des variables qualitatives uniquement.

La concordance des classements se différencie de la corrélation des rangs puisqu'elle constitue une généralisation du coefficient de cette dernière. Le coefficient de Goodman-Krusdal permet de mesurer une proportion en calculant « le surplus de paires concordantes par rapport aux paires discordantes ». Quant au coefficient de Yule, il est associé au coefficient de Goodman-Krusdal et permet de mesurer l'association entre deux variables qualitatives dans le cas des matrices 2x2 (binaires). Il permet donc de mesurer la force de la relation de deux variables dites « dichotomiques ».

Exercices :

Question 2	Ouverture du fichier par la fonction « ouvrirUnFichier() »	
Question 3	Isolement de la colonne « Surface (Km2) » et ajout d'une liste supplémentaire	
Question 4	Ordonner de manière décroissante la liste par la fonction « ordreDecroissantt() »	

Question 5	Visualisation de la loi rang-taille	
Question 6	Conversion des axes en logarithme si l'image est illisible	
Question 9	Ouverture du fichier par la fonction « ouvrirUnFichier() »	
Question 10	Isolement des colonnes sur lesquelles on veut analyser les données	
Question 11	Ordonner de manière décroissante la liste par une fonction différente de la précédente « ordrePopulation() »	
Question 12	Préparation de la comparaison des listes avec fonction locale « (classementPays() » - Rien ne s'affiche après ce code ; la liste s'affiche à la question 14	
Question 13	Isolement des colonnes crées par boucle	
Question 14	Calcul de la corrélation de Spearmanr et la concordance de Kendalltau. Affichage de l'extrait des rangs puis des calculs	

	Le fait que les valeurs soient proches de 0 (Spearmanr = 0,122 & Kendall = 0,089) admet que la corrélation entre la population et la densité est assez faible et surtout que l'une des variables ne permet pas de prédire l'autre. Cependant, ces variables n'admettent pas non plus aucune relation entre elles.	

Cette séance est pour moi une des plus utiles pour l'analyse de données géographiques. À l'issue de la question 5, ne voyant pas la courbe s'afficher j'ai trouvé une autre solution avant d'utiliser la conversion des axes en logarithme.

*Texte en rouge : réponses aux questions posées dans les consignes des manipulations de code.

Remarques générales :

J'ai souvent rencontré des difficultés d'encodage et donc des difficultés à ouvrir les fichiers csv. Mais également des difficultés quant à l'exécution de certains codes dans le terminal (VS code). En effet, à chaque ouverture de chaque main.py, peu importe si la séance avait été entamée ou non, l'exécutant ne reconnaissait pas le dossier dans lequel je travaillais. En utilisant « cd « nom_dossier » », j'ai pu facilement retrouver chaque dossier. Enfin, j'ai trouvé que ces séances permettaient un réel gain de temps dans l'analyse de tableaux statistiques (surtout séance 2 & 6), malgré que le langage python m'était inconnu.

Réflexion sur les humanités numériques

Les humanités numériques désignent un domaine interdisciplinaire associé à la recherche scientifique. En effet, ce domaine regroupe les sciences humaines et les outils numériques. Dans le domaine des sciences humaines et plus particulièrement dans le domaine de la géographie, les outils numériques permettent l'analyse de données en grandes quantités et surtout rapidement. De plus, cette analyse est complétée par des représentations graphiques possibles (cartes, histogrammes, diagrammes, boxplot…). Enfin, certains outils permettent aussi la sauvegarde intellectuelle par la numérisation d'archives par exemple.

Les SHS étudient les relations sociales, comportements ; plus généralement, le fonctionnement des sociétés. Ainsi, il est utile pour les chercheurs d'identifier des tendances dans ces comportements sociétaux. C'est pourquoi des outils numériques permettent la

quantification de ces tendances par le biais d'une analyse au sein d'un ensemble de textes par exemple. De plus, les relations de corrélation, concurrence, concordance entre différents éléments relatifs aux SHS sont également identifiables plus facilement par les méthodes d'analyses de données.

Les outils numériques jumelés aux SHS sont donc un vrai atout pour expliquer certains phénomènes. Cela permet une analyse plus rapide et en grande quantité. Cependant, dans certains contextes ou pour certains phénomènes, les outils numériques ne peuvent pas remplacer l'expertise des chercheurs. Notamment pour certaines données moralement sensibles… À l'inverse, pour certaines situations qui réclament une analyse objective, alors les outils numériques sont idéaux pour quantifier ces phénomènes.

En conclusion, les études en SHS ne seraient pas aussi précises et objectives sans les humanités numériques puisque cela remplace le facteur temps mais aussi l'abondance de données qui ralentissent les recherches scientifiques.