Rapport d’activité

Analyse de données - Parcours débutant

Remy LAVENIER

Master 1 GEAD, parcours EnviTerr, Sorbonne Université

Table des matières

Séance 2 : principes généraux de la statistique

En géographie, la perception des données statistiques est complexe, elles sont à la fois perçues comme un outil utile pour traiter les nombreuses données créées par la géographie et comme un champ externe à la géographie car trop proche des mathématiques. Cette situation conduit donc les géographes à sous-utiliser les données et analyses statistiques. Néanmoins, l'utilisation des données statistiques est aujourd'hui en essor avec un décloisonnement des disciplines en géographie. On peut tout de même noter que la géographie française est bien moins orientée vers la statistique que les autres géographies du monde qui sont souvent « classées » dans les sciences de la nature.

Le positionnement des géographes par rapport au hasard est assez binaire. Dans la géographie sans statistiques, le hasard serait la source de ce qui peut être observé dans l'espace. Au contraire, la géographie qui utilise la statistique considère le hasard comme mineur voire négligeable en prenant appui sur les sciences dites dures telles que la physique. C'est le cas, notamment, de l'école de l'Analyse spatiale qui utilise une démarche nomothétique, c'est-à-dire qui cherche à établir des lois universelles et des modèles. Cependant, il est assez bien établi que le hasard en géographie physique ou humaine existe par contingence (possibilité qu'un événement se produise ou pas, au contraire de la nécessité). On peut, dans cette ligne de pensée, dégager une certaine certitude du résultat d'un événement.

Les statistiques en géographie, souvent appelées informations géographiques se divisent en plusieurs types : des données attributaires et des données spatiales (aussi appelées géométriques). D'une part, les données attributaires correspondent à des données aspatiales, elles sont souvent représentées en CSV, tableur, etc. D'autre part, les données géométriques sont des données spatialisées mais elles ne comportent pas d'informations autres que leurs données spatiales.

La géographie a de nombreux besoins au niveau de l'analyse de données. Elle a besoin, tout d'abord de données, elle peut les faire elle-même ou en utiliser d'autres déjà faites. Ensuite, les géographes ont besoin d'outils pour traiter les données, notamment des logiciels (SIG, cartographie, etc) et des langages de programmation (R, Python, etc). Pour traiter ces données, elle nécessite aussi des méthodes pour établir des relations et les expliquer, ces méthodes peuvent notamment prendre la forme d'équations mathématiques (autocorrélation spatiale, méthode des plus proches voisins (fonction k de Ripley), etc). Ces analyses permettent ensuite de créer des modèles, des simulations et des projections qui sont utiles pour prédire des changements et des évolutions. Enfin, les géographes ont besoin de communiquer leurs données, leurs résultats et leurs méthodes avec d'autres géographes mais aussi avec des acteurs non-initiés pour aider à la décision.

Dans les statistiques, on peut différencier deux branches : la statistique descriptive et la statistique explicative. La statistique descriptive permet, comme son nom l'indique, de décrire une situation de manière simplifiée en comparant cette situation à des situations théoriques (basées sur les modèles et des lois). Elle permet, par la suite de dégager des prédictions : elle est, en quelque sorte, la base des statistiques. A la suite, la statistique explicative permet d'expliquer des situations et d'analyser des relations et pas seulement de les observer.

En géographie, les données peuvent être visualisées selon différents types en fonction des variables étudiées : quantitatives, qualitatives. La visualisation de variables quantitatives se base sur des variables qui ne sont pas uniques par individu. Par exemple, le nombre d'habitants dans une commune est une variable quantitative car deux communes peuvent avoir le même nombre d'habitants. On peut traiter ces données directement avec des opérations mathématiques (moyennes, médianes, etc). Pour les représenter, on peut utiliser un graphique circulaire, en barre, un histogramme, etc. La visualisation de variables qualitatives se base sur des variables uniques par individu. Par exemple, le nom d'une commune (il n'y a pas deux fois la même). On ne peut pas faire directement des calculs mathématiques, même si ce sont des nombres (comme le code postal). On ne peut pas toujours les représenter graphiquement (faire un graphique du nom de communes aurait peu de sens). Or, on peut en représenter de certaines manières grâce à leur occurrence notamment

Pour analyser les statistiques, on peut utiliser trois groupes de méthodes : les méthodes descriptives, les méthodes explicatives ou les méthodes de prévision. Les méthodes descriptives sont utilisées dans l'analyse d'une quantité d'individus caractérisés selon des variables à valeur identique. On cherche à décrire un phénomène en croisant plusieurs variables notamment. Les méthodes explicatives sont utilisées pour expliquer une variable précise selon d'autres variable dans une quantité d'individus. Les méthodes de prévision sont, quant à elles, utilisées pour chercher à relier le passé au futur pour en tirer des prévisions ; on s'appuie sur le passé et les données que nous avons déjà pour essayer de prédire l'avenir selon un contexte.

Pour pouvoir faire de l'analyse de données en statistique, il est nécessaire de connaître le vocabulaire de base. Une population statistique est simplement un ensemble d'individus. Un individu statistique correspond à une entité unique définie par des caractères statistiques qui le rendent particulier dans la population statistique, il n'a de sens que dans une population (une ville parmi toutes les villes ou un habitant parmi tous les habitants d'une ville). En géographie, cet individu est une unité spatiale s'il est localisable et cartographiable.
Les caractères statistiques qui définissent un individu statistique sont eux-mêmes définis par des modalités statistiques qui sont les valeurs prises par un caractère (elles sont qualitatives ou quantitatives)
On remarque donc qu'il y a une hiérarchie entre ces termes comme le montre ce schéma de synthèse :

Population statistique individu caractères un caractère modalités

Pour mesurer une amplitude, c'est-à-dire l'écart entre la valeur la plus faible et la plus élevée, on fait la différence entre b (la valeur la plus forte) a (la valeur la plus faible) : $b - a$ (a et b doivent appartenir à la même classe).

Pour mesurer une densité, on fait le rapport entre un effectif n, choisi selon une modalité i, et l'amplitude de la classe décrivant i : $d = \frac{n_i}{b-a}$.

Les formules de Sturges et Yule permettent de savoir le nombre de classes idéal qu'on peut faire pour un caractère statistique quantitatif. Cela permet d'éviter une perte d'information en créant des classes trop larges (généralisation, à l'extrême, tous les individus sont dans la même classe) ou trop fines (individualisation, à l'extrême, tous les individus ont chacun une classe ou ils sont seuls).

L'effectif (ou fréquence absolue) correspond au nombre d'individus d'une population qui possèdent un critère X dont la modalité est x_i. Il est noté n_i.

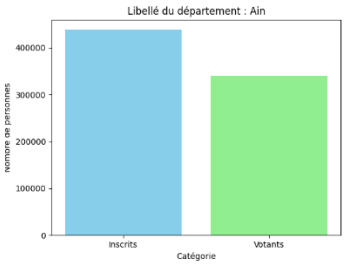
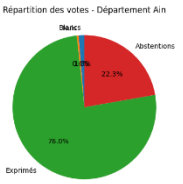
La fréquence relative est le rapport entre l'effectif selon la modalité i et l'effectif total (noté n). Elle est notée f_i. $f_i = \frac{n_i}{n}$ Soit : $f_i = \frac{n_i}{\sum_{i=1}^{i} n_i}$

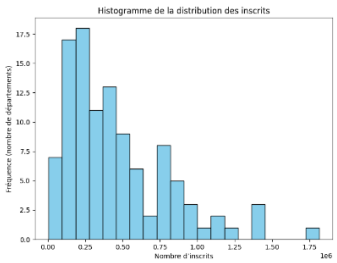
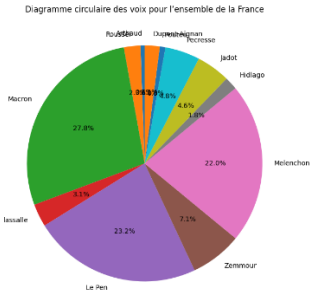
La fréquence cumulée absolue correspond à la somme des effectifs des modalités dont la valeur de i va de 1 à k : $f_i = \sum_{i=1}^{k} n_i$ avec $k \in \mathbb{R}_+^*$

La fréquence cumulée relative correspond à la somme des fréquences relatives f_i des effectifs n_i où i varie de 1 à k. $F_i = \sum_{i=1}^{k} f_i = \sum_{i=1}^{k} \frac{n_i}{n} = \sum_{i=1}^{k} \left[\frac{n_i}{\sum_{i=1}^{i} n_i}\right]$

La fréquence permet de créer une distribution statistique empirique. Une distribution statistique est une fonction qui associe la fréquence d'apparition d'une modalité ou d'une classe.

Mise en pratique, séance 2 :

QUESTION N°	RESULTAT(S)	GRAPHIQUES (EXEMPLES)
QUESTION 5	Affichage sommaire du contenu du CSV	
QUESTION 6	Affichage du nombre de lignes et colonnes du CSV	
QUESTION 7	Affichage par colonne du type de données (str (texte), float64 ou int64)	
QUESTION 8	Affichage du nom des colonnes	
QUESTION 10	Affichage de la somme des colonnes (si c'est en str, affichage de « non quantifiable »)	
QUESTION 11	Création de diagrammes en barre des inscrits et des votants dans chaque département. Enregistrement de ces diagrammes dans le fichier "graph_departements"	
QUESTION 12	Création de diagrammes circulaire des votes blanc, des votes nuls, de l'abstention et des exprimés par département. Enregistrement dans le fichier "graph_rond_departements"	

QUESTION 13	Création d'un histogramme de la distribution des inscrits dans toute la France. Enregistrement de l'histogramme.	
QUESTION BONUS	Création de diagrammes circulaires par département de la répartition des voix en fonctions des candidats à la présidentielle de 2022. Enregistrement des diagrammes dans "graph_rond_departements_voix". Création d'un diagramme circulaire national de la répartition des voix par candidat. Enregistrement du diagramme.	

Les résultats obtenus sont assez satisfaisants. On remarque qu'il y a eu peu d'erreurs sur les données et celles qui ont été observées ont été corrigées. On peut prendre l'exemple du Libellé du département de "Saint-Martin/Saint-Barthélemy" ou « / » posait problème dans l'enregistrement des graphiques, il a donc fallu changer « / » en « _ ».
Les résultats présentés sur le terminal manquent un peu de clarté visuelle. Les questions 7 à 10, amènent à des affichages longs et n'exploitent pas la largeur du terminal, je n'ai pas trouvé de solution efficace que je comprenais pour résoudre ce petit problème. De plus, j'ai fait le choix de ne pas afficher un message dès qu'un graphique pour un département était réalisé pour éviter de surcharger visuellement le terminal.

Séance 3 : Les paramètres statistiques élémentaires

Le plus souvent, on trouve dans les statistiques des variables à caractère quantitatif par rapport aux variables ayant un caractère qualitatif. En effet, les variables quantitatives sont des données qui viennent des sciences "dures" et des sciences humaines et sociales ce qui englobe donc toutes les sources potentielles de données. Au contraire, les variables qualitatives sont plus rares car plus généralement issues des sciences humaines et sociales. En étant vues comme plus objectives, les variables au caractère quantitatif sont majoritairement préférées pour décrire et expliquer des phénomènes avec un point de vue objectif.

Les caractères discrets sont à différencier des caractères continus. En effet, une variable discrète est très différente d'une variable continue. Une variable discrète n'existe que pour des points qu'on pourrait qualifier de "séparées". Par exemple, les années de naissance, on ne peut pas dire qu'un individu est né en 2012,3, ça n'a pas de sens. Ces variables sont généralement définies dans $\mathbb{Z}$ c'est-à-dire dans les entiers relatifs (positifs et négatifs). Au contraire, une variable continue ne présente pas de "séparation" des points. Par exemple la taille des individus, c'est une variable continue car ces individus peuvent théoriquement prendre toutes les tailles (même si elles sont comprises dans un intervalle pour des raisons biologiques). Un individu peut donc faire, par exemple 156,27 (avec une infinité de nombres après la virgule), ce sont des nombres, en général, inclus dans $\mathbb{R}$ (les réels). Ainsi, on ne traite pas exactement de la même manière les caractères quantitatifs continus et les caractères quantitatifs discrets, pour les discrets, on utilise le symbole de somme $\sum$ et pour les continus le symbole de la somme continue (intégrale) $\int$.

La moyenne est un calcul basique en statistique. Bien que la moyenne la plus utilisée et la plus connue soit la moyenne arithmétique (somme des valeurs ÷ nombre de valeurs) il existe de très nombreuses moyennes en fonction de ce qu'on souhaite étudier. L'utilisation de plusieurs sortes de moyennes permet un large panel d'information en fonction de leur nature. La moyenne quadratique, bien que peu utilisée en statistiques, permet de moyenner des surfaces, la moyenne harmonique est une moyenne plus utilisée en statistique notamment avec des liens de proportionnalités, comme par exemple, avec la vitesse moyenne sur un trajet aller-retour, elle correspond à l'inverse de la moyenne arithmétique. La moyenne géométrique, quant à elle, permet de calculer des taux moyens comme des taux de croissance moyens ou des taux de rendement moyens. La moyenne glissante, ou moyenne mobile, permet de calculer une moyenne arithmétique sur un échantillon n de l'ensemble des valeurs N. La moyenne glissante est très utile dans l'utilisation de séries temporelles en supprimant des variations mineures afin de souligner la tendance générale. Enfin, la moyenne fonctionnelle permet de calculer la moyenne des valeurs prises par $f(x)$ sur un intervalle $[a, b]$, elle est utilisée uniquement pour des variables continues. Pour pouvoir avoir des résultats les plus pertinents, il faut bien choisir le type de moyenne en fonction de ce qu'on étudie.

Calculer une médiane, quand on a une moyenne semble inutile. Or, en réalité, la médiane est très différente de la moyenne. En effet, la moyenne est fortement influencée par les valeurs extrêmes alors que la médiane considère ces valeurs au même titre que les valeurs non-extrêmes. Elle permet de trouver le "point" qui sépare la population étudiée en deux parts égales, elle est donc souvent plus pertinente que la moyenne.

Le mode correspond à la valeur qui a le plus d'occurrence (pour les variables discrètes) ou qui a le plus de probabilité (pour les variables continues). On peut calculer un mode quand la population étudiée est assez importante pour que le mode soit pertinent. Il ne faut évidemment pas que ce soit avec des variables qualitatives.

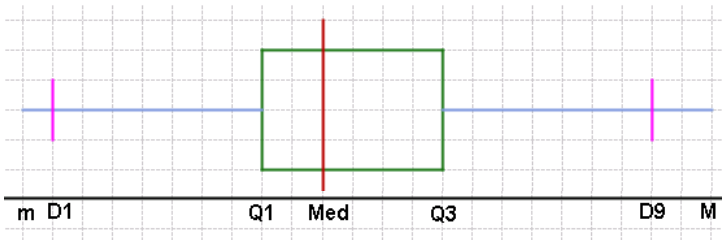
La médiale (médiane de la somme des valeurs) est utile pour observer la concentration des valeurs quand on la compare à la variable. Plus l'écart est grand entre les deux, plus il y a une concentration. La courbe de Gini est très utile pour étudier la concentration, notamment en sciences sociales avec l'indice de Gini de la concentration des richesses par exemple. La courbe de Gini varie de 0 à 1 pour x (abscisse) variant de 0 à 1, en général. On peut comparer la courbe de Gini avec la courbe de Lorenz (distribution théorique parfaite entre tous les individus ou $x = y$). Plus la courbe de Gini s'écarte de la courbe de Lorenz (ou diagonale du carré de Lorenz), plus la concentration est importante.

La variance est un paramètre de dispersion à privilégier par rapport à l'écart à la moyenne car elle est toujours positive (avec les carrés). En effet, on peut décrire la variance comme étant la moyenne des écarts à la moyenne au carré, ce qui donne : $V = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. La moyenne des écarts à la moyenne, sans carré, amène des imprécisions, notamment car les nombres positifs et négatifs risquent de s'annuler en les sommant : $M = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$. Cependant, on utilise beaucoup les écarts types qui reprennent la variance et permettent de déterminer un écart à la moyenne (absolu, donc positif) standard, il est calculé comme ceci : $\sigma = \sqrt{V}$. C'est un très bon indice de dispersion de données. L'écart type est donc un outil plus simple à manier que la variance, notamment pour communiquer des résultats compréhensibles par un public non connaisseur des statistiques.

Calculer une étendue d'une série statistique est facile : $E = x_{\max} - x_{min}$, cependant, cette étendue a peu d'utilité seule. Elle ne prend en compte que les individus les plus extrêmes et n'est pas représentative de la distribution. Néanmoins, elle est utile dans d'autres calculs en statistique, notamment la médiale.

Les quantiles sont un découpage de la série statistique étudiée en x parties égales (même nombre d'individus), c'est la valeur qui sépare deux parties. On utilise le plus souvent les quartiles (découpage en 4, donc 3 quantiles) qui représentent chacun 25% de la population, les quintiles (découpage en 5, donc 4 quantiles) qui représentent chacun 20% de la population, les déciles (découpage en 10 donc 9 quantiles) qui représentent chacun 10% de la population, les centiles (découpage en 100, donc 99 quantiles) qui représentent chacun 1% de la population et enfin la médiane qui est le nom du quantile qui sépare la population en 2 parties égales.

On utilise ces quantiles pour créer une boîte de dispersion (aussi appelé boxplot ou boîte à moustaches). Une boîte de dispersion représente, souvent schématiquement, les principales caractéristiques d'une distribution : médiane, quartiles, minimum, maximum et parfois aussi déciles. On peut prendre ce schéma comme exemple :



Ici, tous les éléments les plus représentés y sont : le minimum (m) et le maximum (M) qui correspondent aux extrémités des moustaches, D1 (décile 1) et D9 (décile 9) qui montrent que les 10% les plus faibles de la population se trouvent entre m et D1 et les 10% les plus forts entre D9 et M. On y trouve aussi les quartiles (Q1 et Q3, Q2 étant égal à la médiane) qui montrent, dans la boîte, la concentration de 50% de la population 25% à gauche de la médiane et 25% à droite. Ces boîtes de dispersion sont très utiles pour avoir une représentation graphique des données que nous avons et que cette représentation soit simple et visuelle.

Les paramètres de forme caractérisent la forme prise par la distribution (symétrie, etc). Pour ce faire on peut utiliser les moments statistiques et notamment les moments centrés et les moments absolus. Les moments centrés sont utiles pour caractériser la forme de la distribution de manière générale. Par exemple, l'espérance et la variance sont des moments centrés, respectivement, d'ordre 1 et d'ordre 2.
Les moments absolus sont, eux, centrés sur un point α. L'utilisation des moments dits absolus est plus courante dans des contextes spécifiques comme pour éviter l'annulation des écarts comme avec le moment absolu d'ordre 1 centré sur la moyenne aussi appelé écart moyen absolu.

Vérifier la symétrie d'une distribution peut être utile pour connaître sa loi et identifier une répartition inégale des valeurs (plus à gauche ou plus à droite. Une distribution est symétrique si la moyenne, la médiane et le mode sont égaux. On peut mesurer l'asymétrie avec, par exemple les coefficients d'asymétrie de Pearson :

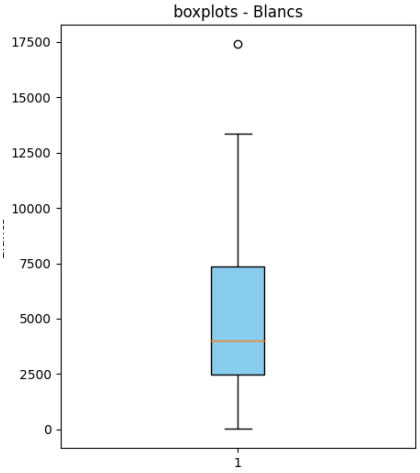
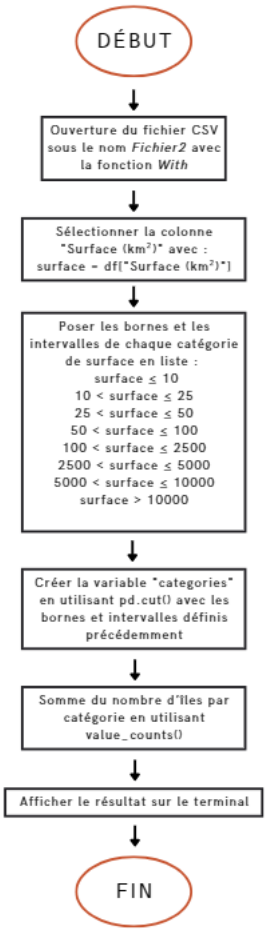
Le coefficient d'asymétrie de médiane $\beta_1 = \frac{3(moyenne - médiane)}{\sigma}$

Le coefficient d'asymétrie de mode $\beta_1 = \frac{moyenne - mode}{\sigma}$

Il existe bien d'autres méthodes pour mesurer l'asymétrie d'une distribution mais ce sont les plus simples.

Mise en pratique, séance 3 :

QUESTION N°	RESULTATS	GRAPHIQUES/AFFICHAGE
QUESTION 4	Ouverture du document en tant que *fichier*. Affichage sommaire du contenu du CSV.	
QUESTION 5 & 6	Calcul et affichage de la moyenne, de la médiane, du mode, de l'écart type, de l'écart absolu et de l'étendue de chaque colonne quantitative et affichage de ces données sous forme d'une liste.	Liste des paramètres statistiques par colonne : Moyenne Médiane Mode Écart-type Écart absolu moyen Étendue Inscrits 455587.63 366859.0 5045.0 351003.78 272240.72 1808861.0 Abstentions 119852.05 95369.0 2272.0 117017.88 74959.07 929183.0 Votants 335735.58 274372.0 2773.0 258393.81 201517.17 1297100.0 Blancs 5080.46 4001.0 4577.0 3492.52 2817.95 17389.0 Nuls 2309.82 2039.0 17.0 1501.38 1131.99 8236.0 Exprimés 328345.30 268568.0 2701.0 253758.58 197762.20 1272080.0 Voix 1842.00 1627.0 1203.0 1268.37 977.36 7651.0 Voix.1 7499.27 5968.0 19.0 6501.29 4474.96 45883.0 Voix.2 91430.45 67831.0 534.0 77226.14 59929.14 372286.0 Voix.3 10293.34 8944.0 17010.0 7464.32 5140.37 48168.0 Voix.4 76017.08 64543.0 459.0 60278.10 42514.72 372668.0 Voix.5 23226.41 16885.0 9657.0 20760.60 15278.36 108537.0 Voix.6 72079.63 51556.0 501.0 66210.68 49157.01 316871.0 Voix.7 5761.48 4881.0 75.0 4581.79 3333.34 22826.0 Voix.8 15213.58 9561.0 72.0 14807.62 11136.57 80196.0 Voix.9 15691.60 11918.0 51.0 13027.13 9432.01 69513.0 Voix.10 2513.12 2118.0 3663.0 1781.41 1404.50 8686.0 Voix.11 6777.35 6152.0 7271.0 4636.02 3689.50 20535.0
QUESTION 7	Calcul et affichage de la distance interquartile $(Q3 - Q1)$ et interdécile $(D9 - D1)$ avec la méthode *quantile()* de chaque colonne quantitative. Affichage sous forme de liste	distance interquartile et interdécile de chaque colonne : Distance interquartile Distance interdécile Inscrits 401050.0 793988.8 Abstentions 106489.0 193676.2 Votants 301770.5 602687.2 Blancs 4852.5 8845.8 Nuls 1917.0 3240.6 Exprimés 296870.5 590169.2 Voix 1517.5 3015.6 Voix.1 6264.5 13104.2 Voix.2 101317.0 177340.2 Voix.3 7999.5 13813.0 Voix.4 63342.0 130094.6 Voix.5 20638.5 43668.8 Voix.6 60743.5 159421.2 Voix.7 4779.0 10712.2 Voix.8 14833.5 38190.8 Voix.9 13265.5 27686.8 Voix.10 2466.0 4266.6 Voix.11 6146.5 12311.0

QUESTION 8	Création d'une boîte à moustache de chaque colonne quantitative et sauvegarde dans le dossier boxplots. Ci-contre, la boîte à moustaches des votes blancs par départements. On remarque que 50% des départements ont entre environ 2500 et 7500 votes blanc, avec une médiane vers 3700. Seul un département à un nombre de votes blancs extrême avec 17500 votes, c'est le maximum de la distribution des votes blancs.	
QUESTION 10	Catégorisation d'intervalles et dénombrement des îles ayant une surface en km^2 comprise dans chaque intervalle. Ouverture du nouveau fichier sous le nom *fichier2*. Création de bornes (bins) et d'intervalles (labels). Dénombrement avec la formule *value_counts()*.	Nombre d'îles par catégorie de surface 0-10 78423 10-25 2327 25-50 1164 50-100 788 100-2500 1346 2500-5000 60 5000-10000 40 10000+ 71
QUESTION 10	Création d'un organigramme de la question n°10 (en commentaire sur le main.py) ou en image ci-contre (fait hors python).	

QUESTION BONUS	Création d'un CSV sur les élections avec les calculs faits précédemment (questions 5 et 7). Création d'un CSV sur le dénombrement des îles par catégorie de surface (question 10). Création d'un document excel qui combine sur une feuille les données sur les élections et sur une autre les données calculées sur les îles.	

Les résultats obtenus lors de la séance 3 sont satisfaisants et correspondent à ce qui était attendu. L'organisation en liste, qui s'affiche en tableau sur le terminal, est particulièrement agréable visuellement puisqu'elle évite la saturation du terminal. Il y avait un problème lors de la création des documents CSV et Excel pour les îles puisqu'il n'y avait pas de nom de colonne. J'ai donc décider de rajouter des titres aux colonnes avec la ligne 118 : *compte_categories.index.name = "Intervalles (km²)"*.

Séance 4 : Les distributions statistiques

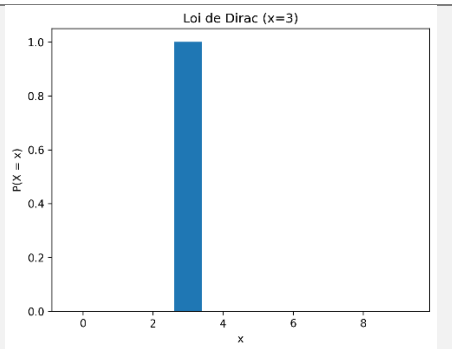
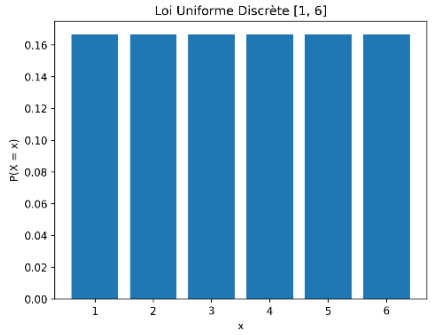
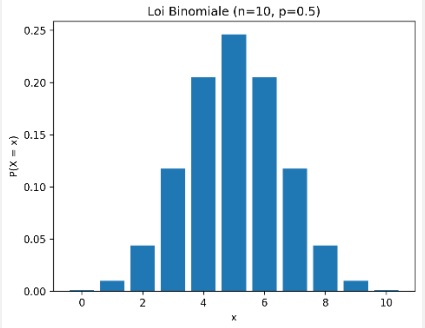
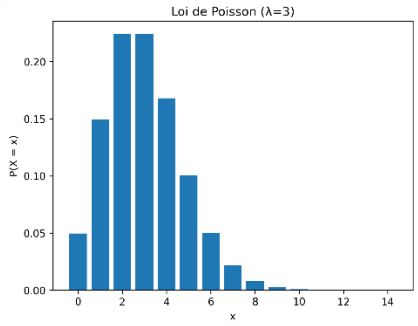
Les distributions statistiques sont divisées en deux grandes catégories : les distributions statistiques de variables discrètes et les distributions statistiques de variables continues. Les variables discrètes sont caractérisées par le fait qu'elles prennent des valeurs dénombrables et distinctes. Par exemple, un lancer de dé. C'est une variable discrète qui peut prendre les valeurs $\{1, 2, 3, 4, 5, 6\}$ elle est dite discrète car elle ne peut pas prendre n'importe quelle valeur dans $[1{,}6]$, ce serait absurde qu'un dé puisse tomber sur 1,5 par exemple. Une variable discrète est le plus souvent limitée à des entiers naturels ($\mathbb{N}$), par exemple, la population, elle ne peut pas être négative ni être décimale. Lorsqu'on somme les variables discrètes, pour une moyenne par exemple, on utilise le symbole Σ.

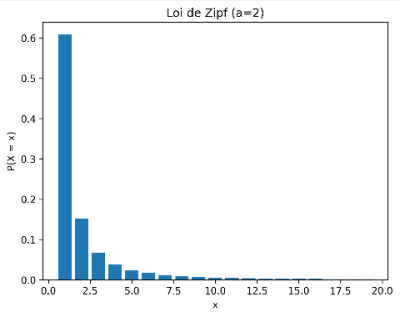
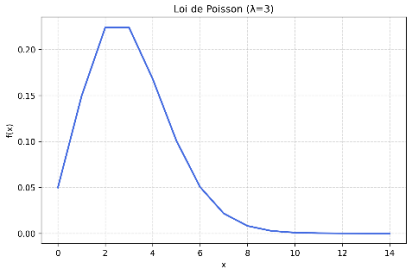
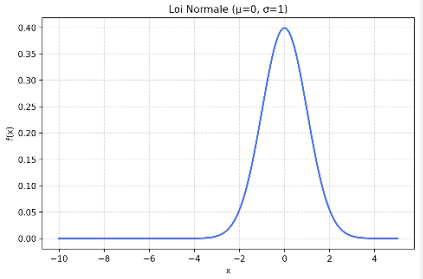
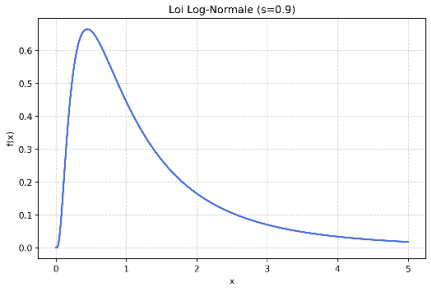
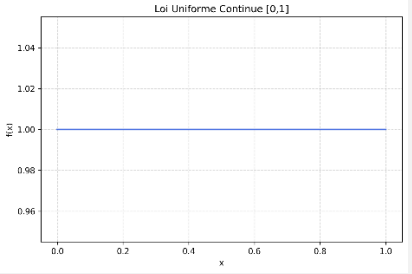
Au contraire, les variables continues sont définies sur un intervalle où elles peuvent prendre toutes les valeurs incluses dans cet intervalle. Par exemple la superficie d'une ville peut prendre n'importe quelle valeur entre la superficie de la ville la plus petite et la superficie de la ville la plus grande. En général, les variables continues le sont une partie de $\mathbb{R}$ ou de $\mathbb{R}^+$ (les valeurs sont positives), en géographie, c'est surtout sur $\mathbb{R}^+$, tout simplement car on se place dans l'espace et donc les nombres négatifs sont rares. Lorsqu'on somme les variables continues, pour une espérance (c'est le nom de la moyenne pour les variables continues) par exemple, on utilise le symbole $\int$ dans l'intervalle $[a, b]$ défini.

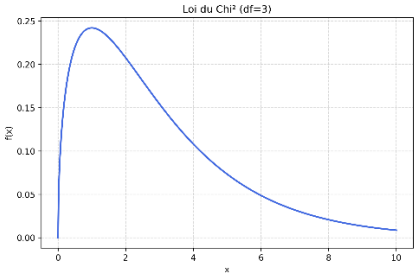
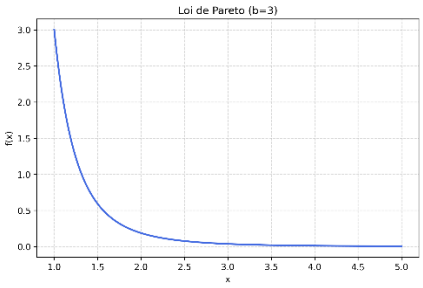
Pour choisir entre une distribution statistique avec des variables discrètes et une distribution statistique avec des variables continues, on aussi peut se référer à ce qu'on souhaite comme résultats graphiques (histogramme pour les variables discrètes et courbe pour les variables continues) et aux lois que l'on souhaite utiliser. Par exemple, la Loi de Poisson permet de compter le nombre d'événements exceptionnels sur un intervalle de temps.

En géographie, on utilise de nombreuses lois pour décrire une distribution statistique. Pour décrire une hiérarchie ou un système, comme la hiérarchie urbaine, on peut utiliser la loi de Zipf (loi rang/taille) ou encore la loi de Pareto, qui permet de décrire une distribution inégalitaire en puissance comme, par exemple une distribution où une petite partie de la population détient une grande partie de la richesse. On utilise aussi des distributions plus générales comme avec la distribution de la loi normale (aussi appelée loi de Gauss) où les valeurs se concentrent sur une moyenne. La loi de Poisson, qui est une loi applicable aux variables discrètes, est utilisée dans la modélisation du nombre d'événements rares dans un intervalle de temps. On peut aussi parler de la loi exponentielle qui permet, par exemple, de modéliser la densité de population qui décroît avec la distance par rapport au centre d'une ville. Evidemment, ce n'est pas une liste exhaustive mais ce sont ici les lois les plus utilisées et les plus connues en géographie.

Mise en pratique, séance 4 :

LOI STATISTIQUE ETUDIEE	RESULTATS	GRAPHIQUES
Loi de dirac	La distribution discrète de la loi de Dirac est caractérisée par le fait que pour une valeur (ici $x = 3$), tous les individus se concentrent sur cette valeur.	
Loi uniforme discrète	La distribution de la loi uniforme discrète montre une probabilité de réalisation identique pour chaque modalité (ici 6 modalités). On appelle ça l'équirépartition la probabilité de "tomber" sur une modalité est ici de $\frac{1}{6}$.	
Loi binomiale	La distribution de la loi binomiale est concentrée en un point, ici $x = 5$. La probabilité d'obtenir une valeur dépend de sa proximité avec 5, plus on s'en éloigne, plus la probabilité est faible. On utilise l'épreuve de Bernoulli pour réaliser cette distribution (résultat de 0 (échec) ou de 1 (succès)).	
Loi de poisson	Cette distribution se concentre sur un point λ (ici $\lambda = 3$). Ce point λ correspond à moyenne d'apparition d'un événement rare. La moitié des valeurs est comprise entre 0 et λ et l'autre moitié entre λ et $+\infty$ avec un maximum en λ.	

Loi de zipf-mandelbrot	La distribution de la loi Zipf-Mandelbrot montre une forte probabilité sur un point, donc une forte concentration des valeurs sur ce point. Plus l'indice a est élevé, plus cette concentration est importante. On l'utilise notamment dans les lois rang/taille.	
Loi de poisson continue	Présentation de la loi de poisson de manière continue même si elle ne peut pas être correctement représentée de manière continue via python.	
Loi normale	Loi très utilisée, aussi appelée loi de Gauss. Sa distribution est centrée en un point x, ici en 0. C'est une distribution symétrique par rapport au point x.	
Loi log-normale	C'est une distribution asymétrique. Elle suit une loi normale avec l'utilisation d'un logarithme.	
Loi uniforme continue	C'est une distribution simple où sur un intervalle donné, ici [0,1], la probabilité est équivalent et vaut suit la fonction suivante : $f(x) = \begin{cases} \frac{1}{b-a} & pour\ a < x < b \\ 0\ sinon \end{cases}$	

Loi du chi^2	Distribution asymétrique avec une concentration à gauche et un étalement à droite pour $df = 3$.	
Loi de pareto	Distribution décroissante quand x croit. Elle est asymétrique et très étendue à gauche dans cette configuration.	
Moyennes et ecarts-type	Réalisation de la moyenne et de l'écart type de la loi demandé avec les variables indiquées dans le programme. Utilisation de scipy.	

La séance 4 a été assez compliquée du fait du manque d'informations et de précision dans l'énoncé des questions. C'est la plus grosse difficulté rencontrée pour cette séance.

La mise en place du programme a été compliquée au début, notamment la prise en main de Scipy. Je n'ai pas pu réaliser les représentations graphiques théoriques avec des variables non définies (laisser x, λ, etc au lieu de choisir un chiffre) car il faut définir des variables pour pouvoir créer ces distributions. Le programme python est visuellement assez répétitif mais je n'ai pas connaissance d'une solution sur ce point.

Séance 5 : Les statistiques inférentielles

Un échantillon est un sous-ensemble d'une population mère. C'est un groupe d'individus issus de la population mère selon une variable aléatoire X. La raison pour laquelle on choisit de travailler sur un échantillon plutôt que sur la population mère en entier est simple : on ne peut pas recueillir les informations nécessaires sur une population mère de grande taille. Par exemple, si on cherche à déduire les intentions de vote de la population d'un pays pour une élection, on choisira de prendre un échantillon représentatif de cette population car il y aurait un besoin énorme de moyens pour recueillir les intentions de vote de plusieurs millions de personnes plutôt que de quelques milliers.

Pour avoir un échantillon représentatif, on utilise différentes méthodes d'échantillonnage : aléatoire, non-aléatoire.

Les méthodes aléatoires font appel à un tirage au sort. Pour ce faire, on doit disposer d'une base de sondage (c'est une liste qui répertorie tous les individus d'une population), ce qui n'est pas toujours le cas. On tire au sort les individus de la population mère pour constituer l'échantillon garce au numéro qui les identifie. Il existe, théoriquement, deux façons de faire un tirage au sort : avec remise ou sans remise. Sans remise, l'individu tiré au sort lors du tirage n n'est pas inclus dans la population soumise au tirage suivant, le tirage $n+1$; c'est un échantillonnage dit exhaustif. Au contraire, dans un tirage avec remise, l'individu tiré au sort lors du tirage n est toujours présent dans la population soumise au tirage suivant $n+1$, il peut donc être tiré de nouveau ; c'est un échantillonnage dit non-exhaustif. En réalité, il existe une infinité de façons de tirer au sort des individus dans une population mère.

Les méthodes non-aléatoires permettent de créer des "modèles réduits" d'une population mère. On peut utiliser l'échantillonnage systématique : le premier individu est choisi aléatoirement, puis les suivants sont déterminés à intervalle régulier par rapport à ce premier individu. Autre méthode non-aléatoire : l'échantillonnage par quotas. Cet échantillon respecte la proportion d'éléments distinctifs de sa population totale. C'est une méthode très utilisée en sciences sociales, notamment en sociologie puisque si elle est bien faite, les résultats sont plus représentatifs de la population mère que les autres méthodes.

On choisit la méthode d'échantillonnage en fonction de ce qu'on souhaite étudier et de quelle manière mais aussi des données que nous disposons. Si nous avons une base de sondage, on peut très bien utiliser les méthodes aléatoires qui sont assez simples mais qui ne représentent pas le mieux la population mère. Si, nous n'en avons pas, il est plus difficile de mettre en place un échantillon. L'échantillonnage par quotas (méthode non-aléatoire) est une méthode très efficace pour représenter une population mère mais cet échantillonnage demande beaucoup de travail et de moyens pour identifier les individus selon les critères que nous prenons en compte (qui peuvent être très nombreux).

Une estimation, en statistique, est le processus par lequel on cherche à estimer un paramètre, d'en donner une valeur la plus proche possible, à partir des résultats obtenus sur un échantillon issu de la population mère. Pour ce faire, on va utiliser des estimateurs. Un estimateur c'est une fonction des données permettant d'estimer un paramètre et caractéristiques d'une loi de probabilité. Le but de la théorie de l'estimation est de trouver quel estimateur est le plus proche du paramètre, quel que soit l'échantillon issu de la même population mère.

En statistiques inférentielles, on utilise souvent des intervalles comme l'intervalle de fluctuation ou l'intervalle de confiance.
L'intervalle de fluctuation peut être calculé quand on connaît p (la proportion théorique d'un type d'individu dans une population). L'intervalle de fluctuation est écrit comme ceci : $[a, b]$ et est appliqué à un échantillon. En général, l'intervalle de fluctuation dit que dans 95% des cas (la précision la plus utilisée), la fréquence observée du type cible dans l'échantillon est comprise entre a et b.
Au contraire, l'intervalle de confiance peut être calculé sans connaître la proportion théorique p. L'objectif de l'intervalle de confiance est de donner une estimation de p. Pour estimer p, on utilise la fréquence observée dans un échantillon. L'intervalle obtenu est une estimation, en général à 95%, de la proportion théorique du type ciblé dans la population mère. Plus l'échantillon utilisé est grand, plus l'intervalle de confiance est restreint et donc précis.

Dans la théorie de l'estimation, un biais correspond à la différence entre l'espérance de l'estimateur $\hat{\theta}$ et la valeur à estimer θ dans la population. L'estimateur est dit sans biais si $E(\hat{\theta}) - \theta = 0$. C'est une erreur d'estimation.

Lorsque l'on travaille sur la population entière, la statistique est appelée : paramètre ou statistique exhaustive lorsqu'elle résume toute l'information pertinente. Une statistique est exhaustive lorsqu'elle contient toute l'information possible sur le paramètre de la population, c'est-à-dire qu'une fois la statistique connue, l'échantillon n'apporte rien de plus. Quand la population entière est disponible, tous les paramètres sont connus, donc aucune estimation n'est nécessaire.
Dans les situations de big data ou données massives, les bases de données se rapprochent d'une population totale. On peut alors travailler sans échantillonnage, avec des paramètres calculés directement sur toutes les données disponibles. Cependant, même avec les big data, il se peut qu'on ne travaille que sur une partie de la population totale, ces données peuvent aussi être biaisées ou de mauvaise qualité.

Un bon estimateur est caractérisé par différentes propriétés : il doit être sans biais ($E(\hat{\theta}) - \theta = 0$), il doit être convergent, il doit être robuste – c'est-à-dire qu'il "résiste" aux valeurs extrêmes –, il doit être précis – sa variance est faible afin d'avoir des estimations plus fiables – et, enfin, doit être le plus exhaustif possible.
Le choix d'un estimateur est important pour maximiser la qualité de l'information et minimiser le risque d'erreur. Au contraire, choisir un mauvais estimateur peut amener un biais important, augmenter la marge d'erreur et induire des mauvaises/fausses conclusions.

Il existe différentes méthodes d'estimation d'un paramètre : estimation ponctuelle, intervalle de confiance, méthode des moindres carrés, méthode du maximum de vraisemblance et la méthode du Bootstrap. La méthode de l'estimation ponctuelle renvoie une seule valeur, si l'estimation renvoie un intervalle, on parle de la méthode par intervalle de confiance. La méthode des moindres carrés est utilisée lorsque la quantité à estimer est une espérance, souvent en lien avec les modèles de régression. Avec la méthode du maximum de vraisemblance, on maximise la vraisemblance de l'observation. Enfin, la méthode du Bootstrap est un ré-échantillonnage aléatoire de l'échantillon pour construire un intervalle ou un estimateur.

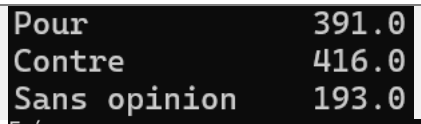
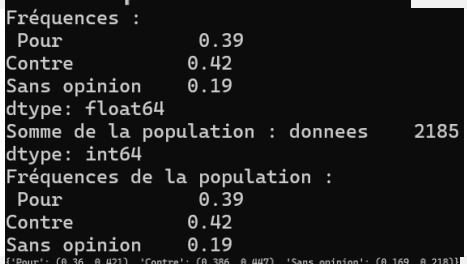
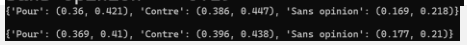
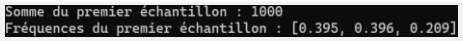
On choisit la méthode la plus adéquate en fonction de plusieurs paramètres : la loi supposée des donnée, la taille de l'échantillon, le paramètre à estimer, le niveau voulu de précision et les hypothèses possibles (comme la normalité de la distribution de l'échantillon pour l'intervalle de confiance)

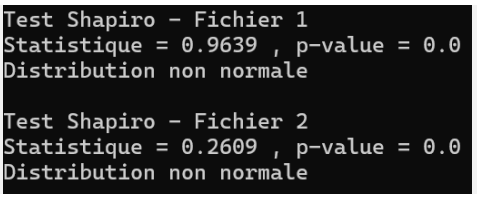
Il existe de nombreux tests statistiques différents, ils peuvent être répartis en deux groupes : les tests paramétriques et les tests non-paramétriques. Dans les tests paramétriques, on peut par exemple trouver le test de Student, celui de Fisher-Snedecor. Dans les tests non-paramétriques, on peut, par exemple, citer le test de Mann-Whitney, de Wilcoxon, de Fisher exact ou encore celui du χ^2.

Les tests statistiques servent à décider de valider ou pas l'hypothèse H_0 , à mesurer la significativité d'un effet, à comparer des estimateurs, à tester l'indépendance des variables, etc. Pour créer un test, il faut d'abord définir l'hypothèse supposée de départ (H_0) puis choisir un type de test : test de Student, test du χ^2, etc. Il faut ensuite déterminer un seuil d'erreur/certitude, généralement, on choisit une marge d'erreur de 5% (certitude de 95%). Ensuite, on calcule la statistique observée avec le test et on la compare à la valeur critique ou on calcule la p-value. Enfin, on conclut en rejetant ou pas l'hypothèse H_0.

Les critiques adressées aux statistiques inférentielles sont, en général, dues à des erreurs d'interprétation, de présentation ou de simplification des résultats (ne pas rejeter H_0 ne veut pas dire que H_0 est vraie). Il est donc nécessaire de bien connaître les statistiques inférentielles et leur fonctionnement pour bien les comprendre et les interpréter, il faut être initié à ces statistiques pour pouvoir les lire et les utiliser, et donc, en parler.

Mise en pratique, séance 5 :

QUESTION	RESULTAT(S)	IMAGES
1. Calcul de la moyenne de l'echantillon	Affichage de la moyenne des "Pour", "Contre" et "Sans opinion"	
1. Calcul des frequences	Calcul et affichage des fréquences de l'échantillon et de la population mère	
1. Calcul de l'intervalle de fluctuation	Calcul et affichage des intervalles de fluctuation de l'échantillon et de la population mère	
2. Prise du premier echantillon dans la liste	Définition et affichage des données du premier échantillon (Pour, contre, sans opinion)	
2. Calculer la somme de la ligne et les frequences	Calcul et affichage de la somme du 1er échantillon et des fréquences associées	
2. Calcul de l'intervalle de confiance	Calcul et affichage de l'intervalle de confiance à 95%	

3. Test de shapiro-wilks	Calcul et affichage des résultats. Les deux distributions ne sont pas normales.	

La séance 5 a été la plus compliquée jusque-là. Les questions de cours ainsi que le document d'explication étaient tous particulièrement difficiles à comprendre. Le document était assez difficile à comprendre et à assimiler.

Le programme n'a pas été le plus difficile. Au contraire, la vérification des résultats savoir si le programme est correct a été plus difficile. Je n'ai pas trouvé laquelle des distributions était normale (partie 3), ces résultats ont été confirmés par d'autres méthodes hors-python et par des comparaisons avec mes camarades.

Séance 6 : Statistique d'ordre des variables qualitatives

Une statistique ordinale regroupe l'ensemble des méthodes reposant sur le classement d'objets ou d'individus, c'est-à-dire sur l'ordre des observations plutôt que sur leurs valeurs numériques absolues. Elle s'appuie sur les rangs, notés $X(1) \leq \cdots \leq X(n)$, obtenus après ordonnancement d'une série d'observations. Elle se distingue ainsi des statistiques nominales, qui se limitent à répartir les individus en catégories sans relation d'ordre entre elles. La statistique ordinale mobilise des variables éponymes, également qualifiées d'« ordinales ». Ce sont des variables qualitatives pour lesquelles un ordre naturel peut être défini, qu'il soit croissant ou décroissant. Dans la majorité des situations, l'ordre croissant est privilégié, à l'exception de certains cas particuliers, comme la loi rang-taille. Ce type de statistique permet de visualiser aisément une hiérarchie spatiale, car de nombreux phénomènes géographiques produisent des classements : taille des villes, intensité de phénomènes physiques (crues, etc), etc. Le classement permet alors d'identifier les entités « en tête », « moyennes » ou « en queue », rendant visible l'organisation hiérarchique d'un ou des territoires étudiés.

Dans les démarches de classification, l'ordre généralement retenu est l'ordre croissant, aussi appelé ordre naturel. Celui-ci facilite l'analyse des rangs, la détection des valeurs aberrantes et l'étude de certaines distributions, comme l'identification de la valeur maximale d'une série.

La corrélation des rangs a pour objectif d'évaluer la ressemblance entre deux séries ordonnées, en comparant les rangs attribués à chaque individu. Deux outils principaux peuvent être mobilisés : le coefficient de Spearman et le coefficient τ de Kendall. Ils permettent de déterminer si les classements sont similaires, inversés ou indépendants. La concordance de classements, quant à elle, s'attache à mesurer le nombre de paires concordantes et discordantes entre deux ordres. Elle repose donc sur l'examen du respect ou non de l'ordre naturel pour chaque couple de rangs. La concordance est dite complète lorsque toutes les paires évoluent dans le même sens, et nulle lorsque concordances et discordances se compensent. Ainsi, la corrélation renseigne sur une proximité globale entre deux classements, tandis que la concordance évalue leur cohérence paire par paire.

Bien que ces deux tests soient employés pour comparer des classements, leurs logiques diffèrent. Le test de Spearman s'appuie directement sur les rangs et calcule une corrélation à partir de la différence entre deux séries, selon le terme $(u_i - v_i)^2$. Sa formulation finale est la suivante : $r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (u_i - v_i)^2$ Ce test est sensible à la présence d'ex æquo et, pour des effectifs supérieurs à 30, sa distribution peut être assimilée à une loi normale pour $n > 30$. Le test de Kendall, en revanche, repose sur le comptage des paires concordantes et discordantes, et son coefficient s'exprime de la manière suivante : $\tau = \frac{2S_C}{n(n-1)}$ Il est souvent considéré comme plus simple sur le plan conceptuel, car il compare directement l'ordre de chaque paire d'individus. Il présente également l'avantage de pouvoir être généralisé à plusieurs classements. De plus générale, Spearman mesure quantitativement la proximité des rangs, tandis que Kendall évalue qualitativement la cohérence de l'ordre. Ces deux approches sont complémentaires et particulièrement pertinentes pour l'analyse des hiérarchies spatiales lorsque plusieurs classements sont disponibles.

Le coefficient de Goodman-Kruskal permet de mesurer la force de l'association d'ordre entre deux variables ordinales en comparant le nombre de paires concordantes (N_a) et

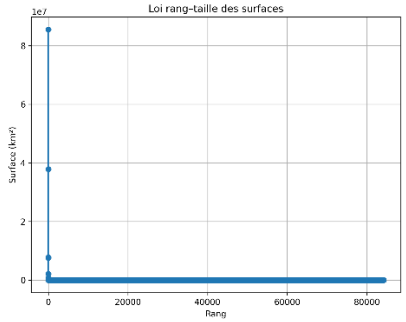
discordantes (N_d). Il est noté : $\Gamma = \frac{N_a - N_d}{N_a + N_d}$ Pour son interprétation, il convient de rappeler qu'il varie entre −1 et +1 et qu'il s'analyse comme un indice de concordance : $\Gamma = 1$ correspond à une concordance parfaite, $\Gamma = -1$ à une inversion complète, et $\Gamma = 0$ à une absence d'association détectable. Ce coefficient est conceptuellement proche du τ de Kendall.

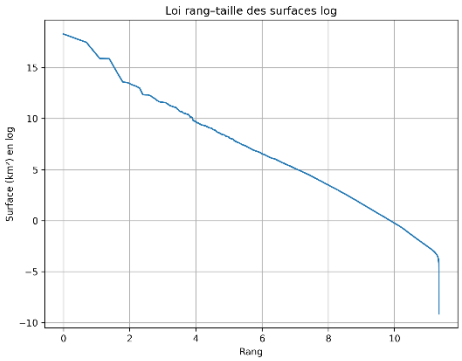
Le coefficient de Yule, ou Q de Yule, constitue un cas particulier du Γ, réservé aux tableaux de contingence (2 × 2). Il s'écrit sous la forme suivante : $Q = \frac{ad - bc}{ad + bc}$ Il sert à mesurer l'association entre deux variables dichotomiques, telles que oui/non ou présent/absent. À l'instar de Γ, il prend des valeurs comprises entre −1 et 1, indiquant respectivement une association positive parfaite, une association négative parfaite ou une absence d'association. En définitive, le coefficient de Goodman-Kruskal offre une mesure générale de l'association d'ordre fondée sur les couples classés, tandis que le Q de Yule fournit un outil plus spécifique pour les relations binaires. Leur utilisation permet ainsi de quantifier rigoureusement la force des relations entre variables catégorielles et d'interpréter les hiérarchies ou dépendances observées dans les données géographiques.

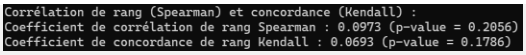
Mise en pratique, séance 6 :

QUESTION	RESULTAT(S)	EXEMPLE(S)
Isoler la colonne « surface (km2) » et ajouter les continents et leur surface	Calcul et affichage d'un aperçu des données sur les îles et sur les continents.	
Loi rang taille	Calcul et création d'un graphique sur la loi rang taille des îles.	
Loi rang taille logarithmique	Calcul et création d'un graphique sur la loi rang taille des îles en logarithme.	

	Surface (km²)	Continent
0	1.819384e+02	NaN
1	8.857096e+01	NaN
2	2.679248e+01	NaN
3	2.988454e+00	NaN
4	9.018293e-02	NaN
...	...	...
84218	9.213744e+01	NaN
84219	8.554532e+07	Asie/Afrique/Europe
84220	3.785684e+07	Amérique
84221	7.768030e+06	Antarctique
84222	7.605049e+06	Australie





Ordonner de manière décroissante les listes	Classement par ordre décroissant des listes « Pop 2007 », « Pop 2025 », « Densité 2007 » et « Densité 2025 » avec la fonction locale ordrePopulation(). Affichage d'un extrait du résultat.	
Coefficient de corrélation des rangs et la concordance des rangs	Calcul et affichage des méthodes spearmanr() et kendalltau() pour calculer le coefficient de corrélation des rangs et la concordance des rangs.	

Remarque pour le coefficient de corrélation des rangs et la concordance des rangs : ces coefficients indiquent le degré de similarité entre le classement par population et le classement par densité. Valeurs proches de 1 => forte concordance ; proches de 0 => classement indépendant ; valeurs négatives => classement inverse. Les résultats obtenus sont assez proches de 0, respectivement 0,2 et 0,17. Les données sont donc assez indépendantes les unes des autres.

Réflexion sur les sciences des données et les humanités numériques

Les exercices menés tout au long de ce parcours mettent en évidence l'importance croissante des sciences des données et des humanités numériques dans les disciplines des sciences humaines, et plus particulièrement en géographie. La manipulation de jeux de données variés, l'automatisation des traitements statistiques et la production de visualisations montrent que ces outils ne se limitent pas à une approche technique, mais participent pleinement à la construction de l'analyse scientifique. L'usage de langages de programmation comme Python permet de gagner en rigueur, en reproductibilité et en efficacité, tout en obligeant à une réflexion critique sur la qualité des données, les choix méthodologiques et l'interprétation des résultats. Les humanités numériques apparaissent ainsi comme un espace de dialogue entre méthodes quantitatives et questionnements disciplinaires, offrant aux géographes de nouvelles manières d'explorer, de comprendre et de représenter les phénomènes spatiaux, sans se substituer à l'analyse théorique mais en la renforçant.