Juline VERHAEGHE
M1 ACA
Place n°6 dans le groupe des GEOINT

Rapport Analyse de données en géographie

LE FOND

Séance 2 :

La géographie entretient un rapport ambivalent avec les statistiques. Longtemps méfiante, elle produit pourtant aujourd'hui des données massives dont l'étude repose nécessairement sur les outils statistiques. Cette relation conduit à sous-estimer leur apport, alors même que les statistiques sont devenues essentielles au développement de la discipline.
La question du hasard en géographie donne lieu à plusieurs positions. Le déterminisme nie son existence, tandis qu'une approche intermédiaire le considère comme une cause cachée. En géographie, s'il est impossible de prévoir les comportements individuels, il reste possible d'en dégager des tendances générales. Le hasard est ainsi intégré dans une démarche probabiliste.

L'information géographique se décline en deux types principaux : des entrées territoriales claires et précises, et une morphologie d'ensemble délimitée.
Face à la massification des données, la géographie a besoin de l'analyse statistique pour étudier leur structure interne, traiter l'information, en évaluer la fiabilité et produire des applications opérationnelles. L'analyse de données est donc un outil central du travail géographique.
La statistique descriptive permet de résumer et de simplifier la réalité en dégageant des propriétés, tandis que la statistique explicative cherche à établir des liens entre les variables afin de comprendre les phénomènes observés.
Les données peuvent être représentées par différents types de visualisations : histogrammes pour les variables continues, diagrammes sectoriels pour les variables qualitatives, ainsi que polygones de fréquence, courbes cumulatives ou rectangles horizontaux pour les données discrétisées. Le choix de la représentation dépend du type de variable étudiée.
Trois méthodes d'analyse de données sont mobilisables : descriptive, explicative et de prévision.
La population statistique correspond à l'ensemble des individus étudiés, l'individu statistique est une unité de la population. Le caractère statistique est une propriété observable, et les modalités sont les différentes valeurs possibles de ce caractère, incompatibles et exhaustives, constituant une « partition du caractère ».
L'amplitude se calcule par la différence entre la valeur maximale et la valeur

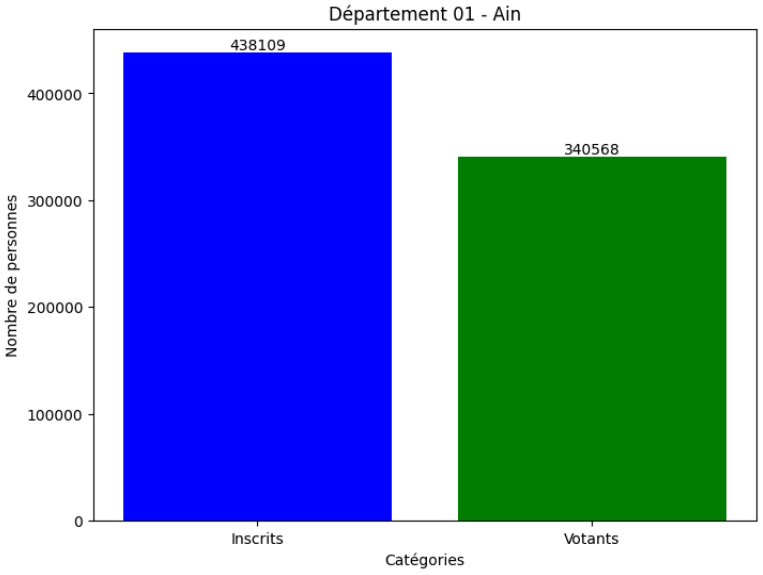
minimale des données relevées, tandis que la densité correspond au rapport entre l'effectif des données et l'amplitude.
Les formules de Sturges et de Yule permettent de déterminer la valeur du nombre de classes, c'est-à-dire en intervalles de valeurs.
L'effectif désigne le nombre d'occurrences d'une valeur. La fréquence est le rapport entre l'effectif d'une classe et l'effectif total, et la fréquence cumulée correspond à la somme des fréquences jusqu'à une valeur donnée. Une distribution statistique associe des classes de valeurs à leurs fréquences.

Résultats de séance:
Question 6 : 107 lignes et 56 colonnes

Question 7 : float64(17) pour les comptabilisation des votes, int64(1) correspondant à la colonne du nombre d'inscrits, object(38) pour les noms et prénoms des candidats, et le sexe des votants.

Question 10 : Exprimés : 35132947. Voix : 197094. Voix.1 : 802422. Voix.2 : 9783058. Voix.3 : 1101387. Voix.4 : 8133828. Voix.5 : 2485226. Voix.6 : 7712520. Voix.7 : 616478. Voix.8 : 1627853. Voix.9 : 1679001. Voix.10 : 268904. Voix.11 : 725176.

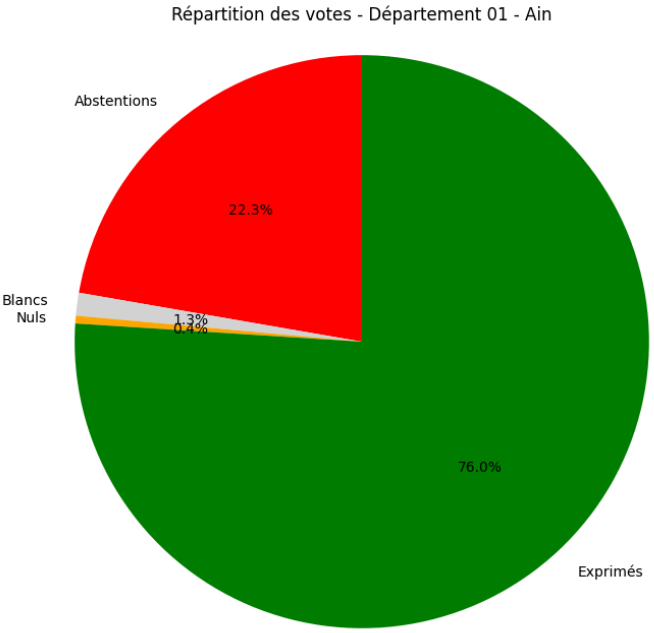
Question 11: diagramme en barres pour chaque département:



Ce diagramme en barres illustre la participation électorale dans le département de l'Ain (01) lors du premier tour de l'élection présidentielle de 2022. On observe un nombre d'inscrits de 438 109 personnes pour 340 568 votants effectifs. Cet écart de près de 97 541 personnes entre inscrits et votants révèle une abstention significative d'environ 22,3%, un indicateur important de l'engagement citoyen dans ce territoire. La visualisation permet d'appréhender rapidement le phénomène d'abstention électorale, qui constitue un enjeu démocratique majeur. Ce type de représentation graphique simple et efficace facilite la comparaison entre le corps électoral potentiel (inscrits) et la participation réelle

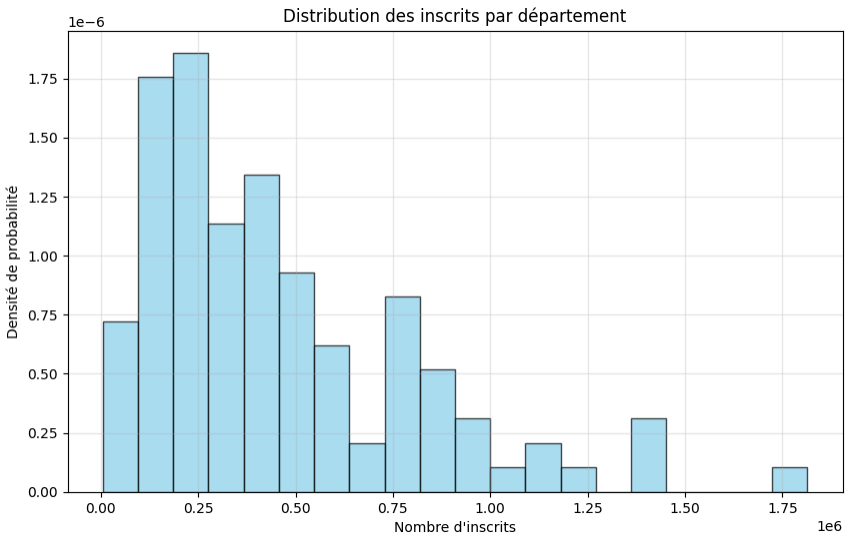
(votants), mettant en évidence le différentiel entre capacité de vote et exercice effectif du droit de vote.

Question 12: diagramme circulaire:



Ce diagramme circulaire représente la répartition des votes dans le département de l'Ain lors du premier tour de la présidentielle 2022. Les votes exprimés dominent largement avec 76,0% du total des inscrits, témoignant d'une participation effective significative. L'abstention représente 22,3% des inscrits, soit près d'un quart de l'électorat. Les votes blancs et nuls demeurent marginaux avec seulement 1,3%, ce qui indique une faible proportion d'électeurs ayant choisi de voter sans se prononcer pour un candidat. Cette visualisation permet d'apprécier la qualité de la participation électorale en distinguant clairement les votes valides des autres comportements électoraux.

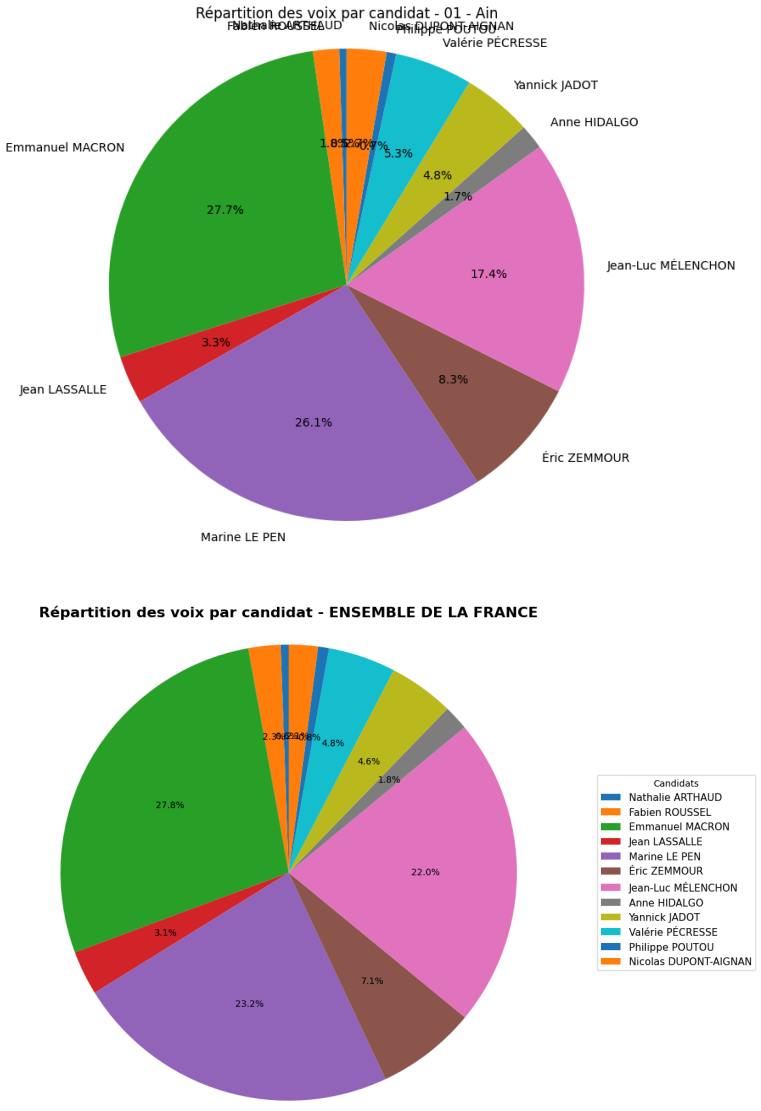
Question 13: histogramme



Cet histogramme représente la distribution du nombre d'inscrits par département en France lors du premier tour de la présidentielle 2022. On observe une distribution asymétrique avec une forte concentration des départements dans les classes inférieures (entre 100 000 et 500 000 inscrits), ce qui correspond aux départements à population moyenne ou faible. La majorité des départements français possèdent

moins de 500 000 inscrits, avec un pic de densité autour de 200 000-250 000 inscrits. On remarque quelques départements isolés avec des effectifs beaucoup plus élevés (au-delà de 1,5 million), correspondant aux départements très peuplés comme le Nord ou les Bouches-du-Rhône. Cette distribution révèle l'hétérogénéité démographique du territoire français et la prédominance des départements de taille petite à moyenne.

Bonus:



Ces deux diagrammes circulaires permettent de comparer la répartition des voix par candidat dans le département de l'Ain (en haut) et au niveau national (en bas) lors du premier tour de la présidentielle 2022. On observe une forte similarité entre les deux distributions, avec Emmanuel Macron en tête (27,7% dans l'Ain vs 27,8% au national), suivi de près par Marine Le Pen (26,1% vs 23,2%) et Jean-Luc Mélenchon (17,4% vs 22,0%). Toutefois, l'Ain se distingue par un score légèrement supérieur pour Marine Le Pen et inférieur pour Jean-Luc Mélenchon par rapport à la moyenne nationale, ce qui peut refléter les spécificités sociologiques et économiques de ce territoire rural et périurbain. Éric Zemmour obtient des scores comparables (8,3% vs 7,1%), tandis que les candidats mineurs restent marginaux dans les deux cas. Cette visualisation comparative illustre à la fois la cohérence du paysage électoral français et les nuances territoriales qui le caractérisent.

Pour plus de clarté, dans le terminal j'ai affiché les résultats nationaux :
Nathalie ARTHAUD : 197094 voix (0.56%)

Fabien ROUSSEL : 802422 voix (2.28%)
Emmanuel MACRON : 9783058 voix (27.85%)
Jean LASSALLE : 1101387 voix (3.13%)
Marine LE PEN : 8133828 voix (23.15%)
Éric ZEMMOUR : 2485226 voix (7.07%)
Jean-Luc MÉLENCHON : 7712520 voix (21.95%)
Anne HIDALGO : 616478 voix (1.75%)
Yannick JADOT : 1627853 voix (4.63%)
Valérie PÉCRESSE : 1679001 voix (4.78%)
Philippe POUTOU : 268904 voix (0.77%)
Nicolas DUPONT-AIGNAN : 725176 voix (2.06%)

Séance 3 :

Les paramètres statistiques s'appliquent principalement aux caractères quantitatifs, et plus marginalement aux caractères qualitatifs. Les caractères quantitatifs peuvent en effet être mesurés et faire l'objet de calculs, ce qui les rend plus généraux dans l'analyse statistique que les caractères qualitatifs, qui décrivent avant tout des catégories.

On distingue deux types de caractères quantitatifs. Les caractères quantitatifs discrets prennent des valeurs précises et isolées, souvent entières, tandis que les caractères quantitatifs continus peuvent prendre l'ensemble des valeurs possibles sur un intervalle. Cette distinction est nécessaire car elle conditionne les méthodes de traitement et de représentation statistique, notamment pour le calcul de certains indicateurs.

Il existe plusieurs types de moyennes car aucune ne résume parfaitement toutes les situations statistiques : le choix dépend de la nature des données et de l'objectif de l'analyse. La médiane est utile car elle permet d'identifier le caractère symétrique ou asymétrique d'une distribution et de mesurer la tendance centrale sans être influencée par les valeurs extrêmes. Le mode, quant à lui, ne peut être calculé que lorsque des valeurs se répètent, ce qui est surtout le cas pour les variables quantitatives discrètes.

La médiale permet de repérer le poids des valeurs élevées dans la somme totale et d'évaluer le degré de concentration d'une variable. L'indice de Gini mesure le niveau d'égalité ou d'inégalité dans la répartition d'une variable, en indiquant si les valeurs sont réparties de manière homogène ou concentrée.

La variance permet d'évaluer la dispersion des valeurs autour de la moyenne, mais son unité rend l'interprétation moins intuitive. C'est pourquoi elle est souvent remplacée par l'écart type, qui s'exprime dans la même unité que la variable étudiée

et facilite l'analyse de l'homogénéité d'une population. L'étendue permet d'observer l'écart entre la valeur minimale et la valeur maximale, donnant une première idée de l'amplitude des données. Les quantiles servent à découper une distribution afin d'analyser la dispersion et la répartition des valeurs, les plus utilisés étant ceux qui partagent la série en parts égales. La boîte de dispersion permet de visualiser ces informations de manière synthétique et d'identifier la dispersion, la médiane et les éventuelles valeurs extrêmes.

Les moments centrés mesurent les écarts autour de la moyenne et servent notamment à analyser la dispersion, l'asymétrie ou l'aplatissement. Les moments absolus permettent de mesurer l'écart en ne tenant pas compte du signe, cela évite les compensations entre valeurs positives et négatives. Vérifier la symétrie d'une distribution permet de savoir si les valeurs sont équilibrées ou déformées vers la droite ou la gauche. Cette symétrie peut être évaluée à l'aide d'un coefficient d'asymétrie, indiquant le sens et l'intensité de la dissymétrie.

Question 5 :

	Moyenne	**Médiane**	**Mode**	**Écart type**	**Écart absolu**	**Étendue**
Inscrits	455587.63	366859.0	5045	351003.78	272240.72	1808861
Abstentions	119852.05	95369.0	2272.0	117017.8	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	328345.3	268568.0	2701.0	253758.58	197762.2	1272080.0
Voix	1842.0	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	60278.1	42514.72	372668.0
Voix 5	23226.41	16885.0	9657.0	20760.6	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	15691.6	11918.0	51.0	13027.13	9432.01	69513.0
Voix 10	2513.12	2118.0	3663.0	1781.41	1404.5	8686.0
Voix 11	6777.35	6152.0	7271.0	4636.02	3689.5	20535.0

Ce tableau présente les statistiques descriptives des données électorales du premier tour de la présidentielle 2022 par département. La médiane des inscrits (366 859) est inférieure à la moyenne (455 587), révélant une distribution asymétrique avec quelques départements très peuplés. Les écarts types élevés (351 003 pour les inscrits) traduisent une forte hétérogénéité territoriale, confirmée par l'étendue considérable de 1 808 861 inscrits. Les candidats "Voix 2" et "Voix 3" dominent avec les moyennes les plus élevées, tandis que leurs écarts types importants reflètent des implantations géographiques variables selon les territoires.

Question 7 : Distance interquartile et interdécile

	Q1	**Q3**	**Distance IQ**
Inscrits	198067.50	599117.50	401050.0
Abstentions	51659.50	158148.50	106489.0
Votants	142568.50	444339.00	301770.5
Blancs	2484.00	7336.50	4852.5
Nuls	1202.50	3119.50	1917.0
Exprimés	137370.00	434240.50	296870.5
Voix	1012.50	2530.00	1517.5
Voix.1	3263.50	9528.00	6264.5
Voix.2	32077.00	133394.00	101317.0
Voix.3	5594.50	13594.00	7999.5
Voix.4	34552.00	97894.00	63342.0
Voix.5	8856.50	29495.00	20638.5
Voix.6	25770.00	86513.50	60743.5
Voix.7	2610.00	7389.00	4779.0

Voix.8	4853.50	19687.00	14833.5
Voix.9	6870.50	20136.00	13265.5
Voix.10	1133.50	3599.50	2466.0
Voix.11	3177.50	9324.00	6146.5

Ce tableau présente les mesures de dispersion (distance interquartile) des données électorales par département. La distance interquartile (IQ), calculée comme la différence entre Q3 et Q1, mesure l'étendue des 50% centraux de la distribution et constitue un indicateur robuste de dispersion, moins sensible aux valeurs extrêmes que l'écart type. On observe des distances IQ importantes pour les inscrits (401 050), les votants (301 770) et les exprimés (296 870), confirmant la forte hétérogénéité démographique entre départements. Pour les résultats par candidat, "Voix.2" présente la distance IQ la plus élevée (101 317), indiquant une forte variabilité territoriale de ce vote, tandis que "Voix.10" affiche la plus faible (2 466), suggérant une implantation plus homogène sur le territoire national.

	D1	**D9**	**Distance ID**
Inscrits	117721.40	911710.20	793988.8
Abstentions	28880.80	222557.00	193676.2
Votants	77697.00	680384.20	602687.2
Blancs	1472.00	10317.80	8845.8
Nuls	762.20	4002.80	3240.6
Exprimés	75306.60	665475.80	590169.2
Voix	499.00	3514.60	3015.6
Voix.1	1693.80	14798.00	13104.2
Voix.2	18110.60	195450.80	177340.2
Voix.3	3430.80	17243.80	13813.0
Voix.4	19442.00	149536.60	130094.6
Voix.5	5312.20	48981.00	43668.8
Voix.6	13426.60	172847.80	159421.2
Voix.7	1203.80	11916.00	10712.2
Voix.8	2224.00	40414.80	38190.8

Voix.9	4078.00	31764.80	27686.8
Voix.10	659.00	4925.60	4266.6
Voix.11	1741.00	14052.00	12311.0

Ce tableau présente la distance interdécile (ID) des données électorales, calculée comme la différence entre le 9e décile (D9) et le 1er décile (D1). Cette mesure capture l'étendue des 80% centraux de la distribution, offrant une vision encore plus robuste de la dispersion que la distance interquartile. Les distances ID considérables pour les inscrits (793 988), les votants (602 687) et les exprimés (590 169) confirment l'importante hétérogénéité démographique entre départements. Parmi les candidats, "Voix.2" présente la distance ID la plus élevée (177 340), révélant une très forte variabilité territoriale de son électorat, tandis que "Voix" et "Voix.10" affichent les distances les plus faibles (respectivement 3 015 et 4 266), témoignant d'une répartition plus uniforme sur le territoire national.

Question 8 : Boîte à moustache

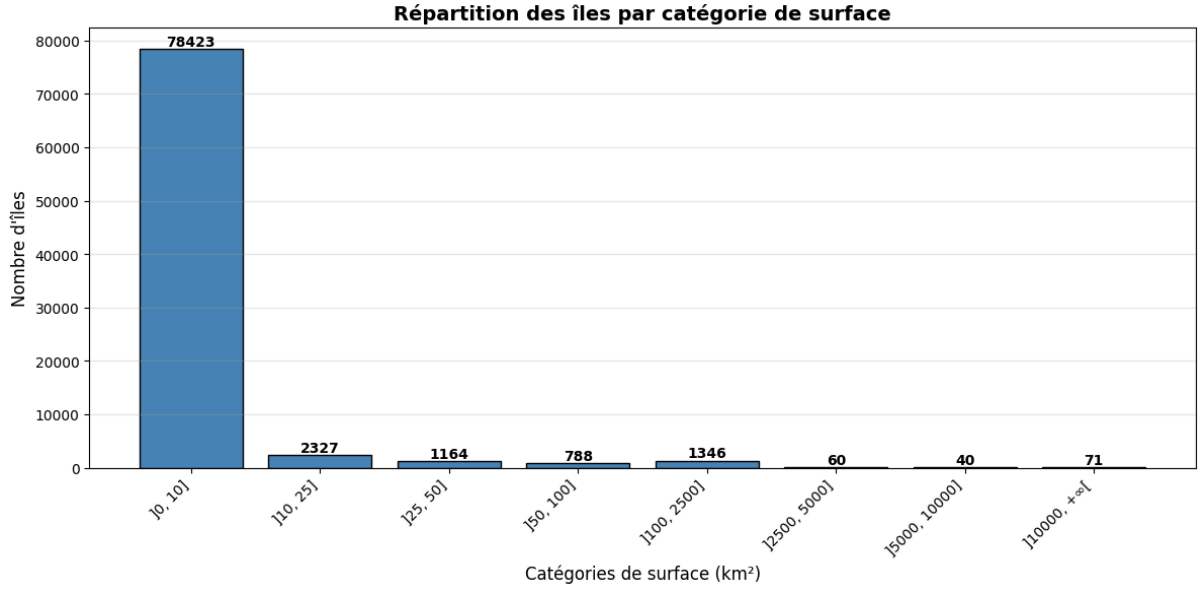
Cette boîte à moustache représente la distribution du nombre d'inscrits par département. La visualisation révèle une distribution fortement asymétrique avec une médiane (ligne orange) autour de 350 000 inscrits, positionnée dans la partie inférieure de la boîte, indiquant que la majorité des départements ont des effectifs modestes. L'écart interquartile (hauteur de la boîte) s'étend approximativement de 200 000 à 600 000 inscrits. On observe trois valeurs aberrantes (outliers) au-dessus de 1,2 million d'inscrits, correspondant aux départements les plus peuplés comme Paris, le Nord ou les Bouches-du-Rhône. Ces valeurs extrêmes confirment l'hétérogénéité démographique du territoire français et justifient l'utilisation de la médiane plutôt que de la moyenne comme indicateur de tendance centrale pour ce type de données.

Question 9 : Catégoriser et dénombrer le nombre d'îles

Catégorie (km²)	Nombre d'îles
]0, 10]	78423

]10, 25]	2327
]25, 50]	1164
]50, 100]	788
]100, 2500]	1346
]2500, 5000]	60
]5000, 10000]	40
]10000, +∞[	71
TOTAL	84219

Ce tableau et ce graphique en barres illustrent la répartition des 84 219 îles mondiales selon leur superficie. La visualisation met en évidence une distribution extrêmement asymétrique avec une domination écrasante des petites îles : 78 423 îles (93,1%) mesurent moins de 10 km², créant une première barre qui écrase visuellement toutes les autres. Le nombre d'îles décroît exponentiellement avec l'augmentation de la superficie, passant de 2 327 îles dans la catégorie]10, 25] km² à seulement 71 îles au-delà de 10 000 km².



Question Bonus :

Catégorie (km²)	Nombre d'îles
]0, 10]	78423
]10, 25]	2327
]25, 50]	1164

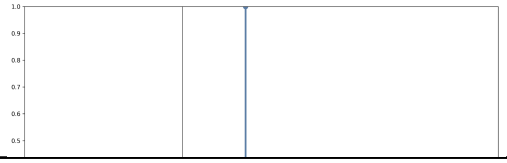
]50, 100]	788
]100, 2500]	1346
]2500, 5000]	60
]5000, 10000]	40
]10000, +∞[	71
TOTAL	84219

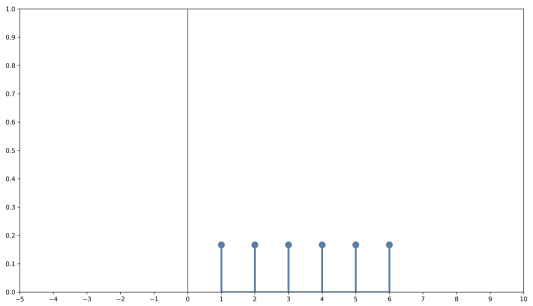
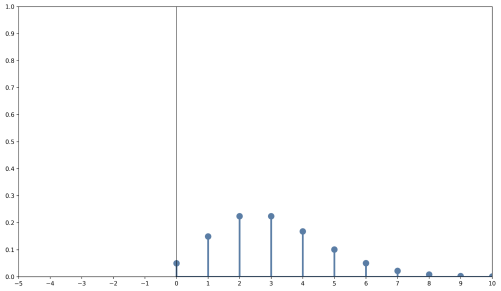
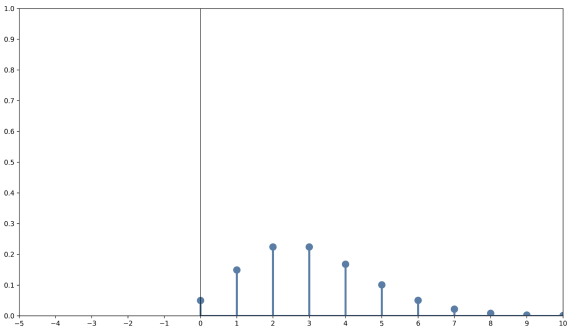
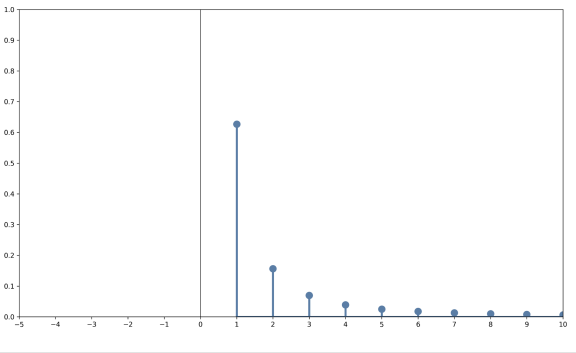
Séance 4 :

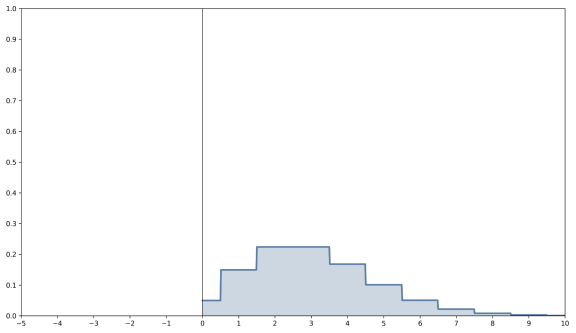
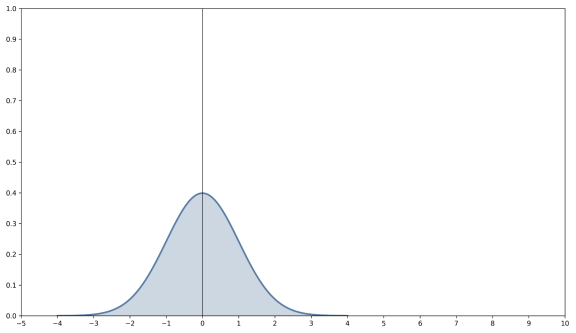
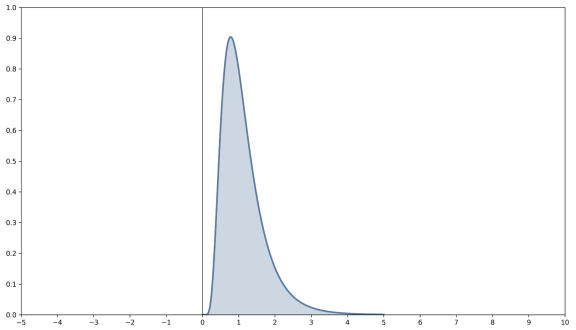
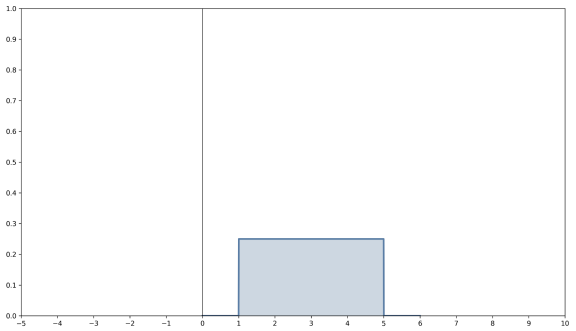
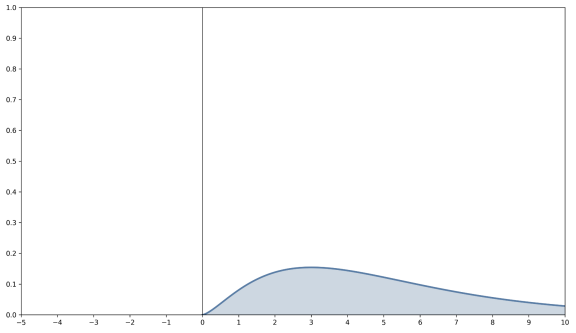
Le choix entre une distribution statistique à variables discrètes et une distribution à variables continues repose avant tout sur la nature du phénomène étudié. Lorsque la variable observée prend un nombre fini ou dénombrable de valeurs, comme dans le cas d'un comptage d'individus, d'événements ou de résultats, il est pertinent d'utiliser une distribution à variables discrètes. À l'inverse, lorsque la variable peut prendre l'ensemble des valeurs possibles au sein d'un intervalle continu, comme une température, une durée, une distance ou un revenu, une distribution à variables continues est plus adaptée.
Ce choix dépend également de la forme de la distribution, c'est-à-dire de la manière dont les données se répartissent. Les caractéristiques statistiques de la série, telles que l'espérance, la médiane, la variance, l'écart type ou encore l'asymétrie, constituent aussi des critères importants pour orienter la sélection de la distribution la plus appropriée. Enfin, le nombre de paramètres propres à chaque loi statistique peut jouer un rôle, certaines lois s'ajustant mieux que d'autres selon la complexité des données à modéliser.

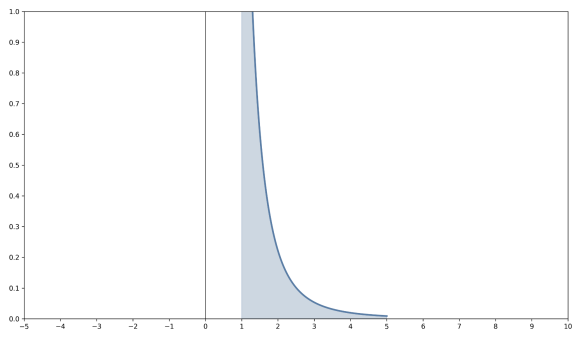
Parmi les lois statistiques fréquemment mobilisées en géographie, la loi de Zipf occupe une place centrale. Elle établit une relation entre le rang d'un élément et sa taille, en montrant que celle-ci est inversement proportionnelle à son rang. Cette loi est particulièrement utilisée pour analyser les hiérarchies urbaines ou les distributions de population, car elle met en évidence des régularités dans la répartition inégale des villes au sein d'un territoire.

<u>Question 1</u> :

Loi de Dirac: Distribution dégénérée concentrant toute la masse probabiliste en deux points	

spécifiques (1 et 2). Modélise une variable aléatoire prenant des valeurs constantes avec certitude, utilisée pour représenter des phénomènes déterministes.	
<u>Loi uniforme discrète</u>: Chaque valeur possible (0 à 5) possède exactement la même probabilité d'occurrence (~0,167). Caractéristique des phénomènes équiprobables comme le lancer d'un dé équilibré où aucune modalité n'est favorisée.	
<u>Loi binomiale</u> Distribution en cloche asymétrique avec un maximum autour de 2, modélisant le nombre de succès dans une série d'essais indépendants. La concentration des probabilités autour des valeurs centrales reflète le comportement typique des expériences de type succès/échec répétées. .	
<u>Loi de Poisson</u> Distribution asymétrique modélisant le nombre d'événements rares survenant dans un intervalle fixe. Les probabilités décroissent progressivement après le pic, caractéristique des phénomènes aléatoires discrets peu fréquents.	
Loi de Zipf-Mandelbrot Distribution fortement asymétrique avec un pic maximal proche de zéro puis une décroissance rapide. Modélise des phénomènes de rang comme la fréquence des mots dans un texte ou la taille des villes.	

Loi de Poisson Représentation par classes de la loi de Poisson montrant une distribution approximativement symétrique centrée autour des valeurs 2-3. La forme en cloche traduit la concentration des probabilités autour de la moyenne.	
Loi normale Distribution symétrique en forme de cloche centrée sur zéro, caractéristique de nombreux phénomènes naturels. La courbe illustre la concentration des valeurs autour de la moyenne avec des probabilités décroissantes vers les extrêmes.	
Loi log-normale Distribution asymétrique positive caractérisée par une queue étendue vers la droite. Modélise des variables dont le logarithme suit une loi normale, typique de phénomènes multiplicatifs comme les revenus ou les tailles de particules.	
Loi uniforme Distribution rectangulaire où toutes les valeurs d'un intervalle continu ont la même densité de probabilité. Représente l'absence totale d'information préférentielle sur la position d'une variable dans un intervalle donné.	
Loi du χ2 Distribution asymétrique positive utilisée principalement dans les tests statistiques d'indépendance et d'adéquation. La forme varie selon le nombre de degrés de liberté, avec une concentration vers les valeurs positives.	

Loi de Pareto Distribution fortement asymétrique avec un pic prononcé suivi d'une longue queue décroissante. Modélise les phénomènes de type "80-20" comme la répartition des richesses ou la fréquence des événements rares extrêmes.	

Question 2 :

	Moyenne	Écart-type
UNIFORME_DISCRETE	3.5000	1.7078
BINOMIALE	5.0000	1.5811
POISSON	3.0000	1.7320
ZIPF	2.2540	2.7294
POISSON_CONTINUE	3.0069	1.6784
NORMALE	0.0000	0.9994
LOG NORMALE	1.1294	0.5920
UNIFORME_CONTINUE	3.0000	1.1547
CHI2	4.9722	3.1016
PARETO	1.5176	0.6478

Ce tableau présente les moyennes et écarts types des différentes lois de probabilité étudiées. On observe une forte diversité dans les valeurs centrales : la loi normale est centrée sur 0 par construction, tandis que la binomiale présente la moyenne la plus élevée (5,0). Les écarts types révèlent des dispersions très variables selon les distributions. La loi de Zipf affiche l'écart type le plus important (2,73), témoignant d'une forte variabilité inhérente aux distributions de rang. À l'inverse, la loi log-normale (0,59) et la loi de Pareto (0,65) présentent les dispersions les plus faibles. La loi normale, avec un écart type proche de 1, illustre une dispersion standard typique. Ces indicateurs permettent de quantifier et comparer les caractéristiques de position et de dispersion propres à chaque modèle probabiliste.

Séance 5 :

L'échantillonnage consiste à sélectionner aléatoirement une partie d'une population mère afin d'en tirer des informations sur l'ensemble de celle-ci. Cette démarche relève de la statistique inférentielle. L'étude exhaustive d'une population entière est généralement impossible ou trop coûteuse sur le plan logistique, notamment lorsque les effectifs sont très élevés. De plus, il est difficile de choisir un modèle statistique pertinent lorsque l'effectif exact de la population mère est inconnu. Il existe principalement deux méthodes d'échantillonnage : l'échantillonnage non biaisé, fondé sur le hasard et l'équiprobabilité, et l'échantillonnage biaisé. Le choix de la méthode dépend des contraintes de l'étude et des objectifs de l'analyse.

Un estimateur est une fonction appliquée à un échantillon permettant d'approcher au mieux la valeur d'un paramètre inconnu de la population mère. L'estimation correspond à la valeur numérique obtenue en appliquant cet estimateur à un échantillon concret, fournissant ainsi un résultat chiffré fondé sur les données observées.

Il convient de distinguer l'intervalle de fluctuation de l'intervalle de confiance. L'intervalle de fluctuation indique la plage de valeurs dans laquelle une fréquence observée a une forte probabilité de se situer, à partir d'un paramètre connu de la population on peut regarder jusqu'où les résultats d'un échantillon peuvent varier.
L'intervalle de confiance, en revanche, permet d'estimer un paramètre de la population mère à partir d'un échantillon, en déterminant une plage de valeurs dans laquelle ce paramètre est susceptible de se trouver.

Un biais correspond à l'écart entre l'espérance mathématique d'un estimateur et la valeur réelle du paramètre estimé. Lorsque cet écart est nul, l'estimateur est dit sans biais ; dans le cas contraire, il est biaisé.

Une statistique fondée sur l'étude de la population totale est qualifiée de statistique exhaustive. Elle mobilise l'ensemble des informations disponibles sur le paramètre étudié pour conserver toutes les informations utiles. Les données massives correspondent à des données très volumineuses, et s'inscrivent dans cette logique d'exhaustivité, car elles couvrent l'ensemble de la population connue et nécessitent des méthodes spécifiques de traitement en raison de leur volume.

Le choix d'un estimateur est un enjeu central de l'analyse statistique. Un bon estimateur doit fournir une estimation ponctuelle la plus proche possible du paramètre inconnu, indépendamment de l'échantillon considéré. Les critères principaux sont l'absence de biais, une variance faible traduisant une bonne précision, ainsi qu'une robustesse face aux valeurs aberrantes.

Plusieurs méthodes permettent d'estimer un paramètre. La méthode des moindres carrés est utilisée lorsque les quantités à estimer correspondent à des espérances.

La méthode du maximum de vraisemblance et la méthode des moments constituent également des outils couramment mobilisés. Le choix entre ces méthodes repose sur les propriétés attendues de l'estimateur, notamment sa convergence, son efficacité et sa robustesse.

Les tests statistiques permettent de vérifier des hypothèses à partir de données observées. On distingue les tests paramétriques portant sur la moyenne, l'écart type ou la forme de la distribution, comme le test de Student ; et les tests non paramétriques, qui reposent sur des statistiques telles que les effectifs ou la médiane, comme le test de Mann-Whitney. Il existe différents types de tests, notamment les tests de conformité, d'homogénéité, d'indépendance, d'adéquation à une loi de probabilité ou encore les tests sur séries appariées. Ces tests servent à établir des liens entre plusieurs échantillons, selon un événement, phénomène ou une action afin de chercher des liens de cause à effet.
Créer un test d'hypothèse consiste d'abord à formuler une hypothèse de travail, et également une hypothèse nulle et une hypothèse alternative. Après avoir fixé le seuil d'erreur et choisi la loi de probabilité, on détermine les paramètres du test, notamment l'échantillon et sa variance. La collecte de données permet ensuite de sélectionner et de calculer la statistique de test. En comparant cette valeur à la région critique préalablement établie, on peut finalement rejeter ou accepter l'hypothèse nulle, concluant ainsi le test.

Les critiques adressées à la statistique inférentielle soulignent sa complexité et interrogent parfois sa fiabilité. Je pense que si ces critiques peuvent fragiliser la confiance accordée à ses résultats, elles ne remettent pas en cause son utilité dans l'analyse des phénomènes, à condition que ses méthodes soient appliquées de manière rigoureuse.

Question 1 : Théorie de l'échantillonnage
Moyennes arrondies pour chaque opinion
Pour : 391
Contre : 416
Sans opinion : 193
Somme des moyennes : 1000

Fréquences de l'échantillon
Pour : 0.39
Contre : 0.42
Sans opinion : 0.19

Fréquences de la population mère
Total population mère : 2185
Pour : 0.39
Contre : 0.42
Sans opinion : 0.19

Catégorie	Échantillon	Population	Différence
Pour	0.39	0.39	0.0
Contre	0.42	0.42	0.0
Sans opinion	0.19	0.19	0.0

Taille de l'échantillon : 1000.0
Valeur critique z : 1.96

Calcul des intervalles de fluctuation

Catégorie	Fréquence	Borne inf	Borne sup	Largeur
Pour	0.39	0.36	0.42	0.06
Contre	0.42	0.39	0.45	0.06
Sans opinion	0.19	0.17	0.21	0.05

L'analyse révèle une parfaite cohérence entre les fréquences de l'échantillon et celles de la population mère (différence nulle pour les trois catégories). Les intervalles de fluctuation calculés au seuil de confiance de 95% (z = 1,96) encadrent correctement les fréquences observées : pour "Pour" [0,36 ; 0,42], pour "Contre" [0,39 ; 0,45], et pour "Sans opinion" [0,17 ; 0,21]. Les valeurs réelles de la population (0,39, 0,42 et 0,19) se situent toutes à l'intérieur de leurs intervalles respectifs, confirmant la représentativité de l'échantillon de 1000 individus. La largeur des intervalles (0,05 à 0,06) témoigne d'une précision satisfaisante. On peut conclure que l'échantillon utilisé est statistiquement représentatif de la population mère, validant ainsi la théorie de l'échantillonnage : un échantillon aléatoire de taille suffisante permet d'estimer fidèlement les paramètres de la population avec une marge d'erreur contrôlée.

Question 2 : Théorie de l’estimation

Somme de la ligne : 1000
Fréquence
Pour : 0.3950
Contre : 0.3960
Sans opinion : 0.2090

Intervalles de confiance

Catégorie	Freq obs	Borne inf	Borne sup
Pour	0.3950	0.3647	0.4253
Contre	0.3960	0.3657	0.4263
Sans opinion	0.2090	0.1838	0.2342

Vérification

Catégorie	Freq population	Dans IC?
Pour	0.3900	OUI ✓
Contre	0.4200	OUI ✓
Sans opinion	0.1900	OUI ✓

L'estimation par intervalles de confiance à 95% confirme la représentativité de l'échantillon observé. Les fréquences observées (Pour : 0,395, Contre : 0,396, Sans opinion : 0,209) génèrent des intervalles qui capturent correctement les fréquences réelles de la population mère (0,39, 0,42 et 0,19 respectivement). Cette approche diffère de la question précédente : au lieu de partir des paramètres connus de la population pour prédire la variabilité d'échantillons futurs (intervalle de fluctuation), on part ici d'un échantillon observé pour estimer les paramètres inconnus de la population (intervalle de confiance). Les deux approches sont complémentaires et aboutissent à des conclusions convergentes : l'échantillon de 1000 individus permet une estimation fiable avec une marge d'erreur d'environ ±3%. La vérification confirme que les trois paramètres populationnels tombent bien dans leurs intervalles respectifs, validant la méthode d'estimation statistique.

Question 3 : Théorie de la décision

Test 1 - Statistique W : 0.9639, p-value : 0.000000
Test 2 - Statistique W : 0.2609, p-value : 0.000000

Séance 6 :

La statistique ordinale correspond à des données organisées selon un ordre. Elle se distingue de la statistique nominale, qui classe les données sans hiérarchie. La statistique ordinale repose sur des variables quantitatives et permet d'établir un classement des entités à partir d'un critère commun. En géographie, ce type de statistique peut matérialiser une hiérarchie spatiale en ordonnant les territoires ou les unités spatiales selon leur position relative.

Dans les classifications, il est préférable d'adopter un ordre croissant, considéré comme l'ordre naturel, car il facilite la lecture et l'interprétation des résultats.

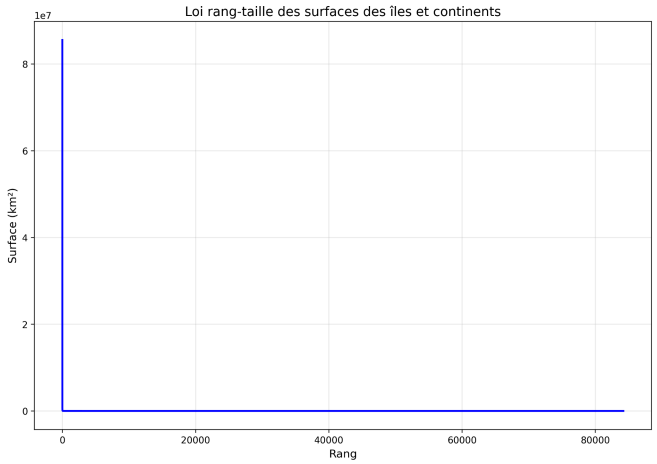
La corrélation des rangs vise à vérifier si deux séries de données sont classées de manière similaire, en mesurant la relation entre deux listes de rangs. La concordance de classements s'intéresse davantage à la cohérence des critères utilisés pour ordonner les données au sein d'un groupe.

Les tests de Spearman et de Kendall permettent tous deux d'évaluer une corrélation de rangs, mais reposent sur des méthodes de calcul différentes. Le coefficient de

Spearman est calculé en se basant sur les différences entre les rangs de chaque paire d'éléments: il mesure à quel point les distances entre les positions dans la première liste sont similaires aux distances dans la deuxième liste.
Le test de Kendall, quant à lui, examine l'ensemble des paires d'éléments possibles et compare le nombre de paires dont l'ordre est respecté (paires concordantes) ou inversé (paires discordantes)entre deux classements.

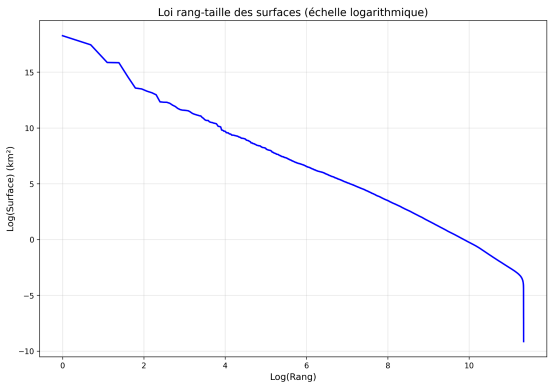
Les coefficients de Goodman-Kruskal et de Yule servent à mesurer la concordance entre des classements. Le coefficient de Goodman-Kruskal évalue l'association entre variables ordinales en comparant le nombre de paires concordantes et discordantes, avec une valeur comprise entre −1 et +1 et son interprétation est similaire à celle des coefficients de Spearman et de Kendall.
Le coefficient de Yule constitue un cas particulier de Goodman-Kruskal, applicable lorsque les données sont organisées dans une matrice avec tableau de contingence de type 2×2. Dans ce cas, une valeur négative indique une association inverse, une valeur positive une association parfaite, et une valeur nulle l'absence d'association.

Question 5 : loi rang-taille



Ce graphique illustre la loi rang-taille des surfaces d'îles et continents en échelle linéaire. On observe une chute brutale de la surface dès les premiers rangs, suivie d'une asymptote quasi-horizontale proche de zéro pour la majorité des îles. Cette représentation met en évidence la domination écrasante de quelques grandes masses terrestres (les continents et grandes îles) face à l'immense majorité des petites îles. Cependant, cette visualisation est peu exploitable car l'écrasement des valeurs faibles rend impossible la distinction entre les rangs élevés, justifiant le recours à une échelle logarithmique.

Question 6 : loi rang-taille log



La relation apparaît désormais linéaire sur la majeure partie du graphique, caractéristique d'une distribution en loi de puissance. La pente négative régulière indique que log(Surface) décroît proportionnellement avec log(Rang), traduisant mathématiquement que la

surface d'une île est inversement proportionnelle à son rang. Cette linéarité confirme que la distribution des surfaces insulaires suit une loi de Zipf, phénomène universel observé dans de nombreux systèmes naturels et sociaux.

Question 7 : Il est possible de réaliser des tests statistiques sur les rangs. Les tests non-paramétriques sont spécifiquement conçus pour analyser des données ordinales ou de rang, sans nécessiter les hypothèses strictes des tests paramétriques.Dans le contexte de la loi rang-taille, un test de corrélation de Spearman entre le logarithme du rang et le logarithme de la surface permettrait de quantifier statistiquement la force de la relation linéaire observée et de tester sa significativité.

Question 14 :

Coefficient de Spearman : 0.9862 (p-value: 1.0697e-134)
Tau de Kendall : 0.9043 (p-value: 2.0819e-69)

Coefficient de Spearman : 0.9673 (p-value: 3.6881e-103)
Tau de Kendall : 0.8588 (p-value: 8.6261e-63)

Les tests de corrélation de rangs révèlent une très forte concordance entre les classements. Le premier couple de variables présente un coefficient de Spearman exceptionnel de 0,986 et un tau de Kendall de 0,904, tous deux avec des p-values extrêmement faibles, indiquant une relation hautement significative.
Le second couple montre également une forte corrélation (Spearman : 0,967 ; Kendall : 0,859) avec des p-values tout aussi significatives. Ces résultats confirment que les deux classements sont remarquablement cohérents : les départements les plus peuplés tendent également à être les plus denses. La différence entre les coefficients de Spearman et de Kendall est normale, ce dernier étant généralement plus conservateur car il compte les paires concordantes différemment.
Les p-values quasi-nulles permettent de rejeter catégoriquement l'hypothèse d'indépendance entre les rangs, validant statistiquement l'existence d'une structure commune dans l'organisation démographique des départements français.

Question Bonus :

1. Spearman : 0.1446 (p=0.0000e+00)

   Kendall : 0.0965 (p=0.0000e+00)

L'analyse de corrélation entre le classement par surface et le classement par longueur de trait de côte révèle une relation faible mais statistiquement significative (Spearman : 0,145 ; Kendall : 0,097, $p < 0,001$). Ces coefficients bas indiquent que la taille d'une île n'est pas un bon prédicteur de la longueur de son littoral. La

significativité statistique malgré la faiblesse de la corrélation confirme qu'il existe bien une tendance générale.

2. algorithme & code (réalisé sur un échantillon de 3000 îles pour des soucis de performance).

Résultats POPULATION
2007-2008: Spearman=0.9996, Kendall=0.9887
2008-2009: Spearman=0.9996, Kendall=0.9910
2009-2010: Spearman=0.9845, Kendall=0.9672
2010-2011: Spearman=0.9855, Kendall=0.9674
2011-2012: Spearman=0.9989, Kendall=0.9846
2012-2013: Spearman=0.9933, Kendall=0.9816
2013-2014: Spearman=0.9928, Kendall=0.9713
2014-2015: Spearman=1.0000, Kendall=0.9990
2015-2016: Spearman=0.9904, Kendall=0.9670
2016-2017: Spearman=0.9993, Kendall=0.9850
2017-2018: Spearman=0.9999, Kendall=0.9960
2018-2019: Spearman=0.9974, Kendall=0.9839
2019-2020: Spearman=0.9968, Kendall=0.9801
2020-2021: Spearman=0.9994, Kendall=0.9931
2021-2022: Spearman=0.9999, Kendall=0.9956
2022-2023: Spearman=0.9995, Kendall=0.9863
2023-2024: Spearman=0.9871, Kendall=0.9808
2024-2025: Spearman=0.9870, Kendall=0.9712

Résultats DENSITÉ
2007-2008: Spearman=0.9992, Kendall=0.9819
2008-2009: Spearman=0.9992, Kendall=0.9857
2009-2010: Spearman=0.9781, Kendall=0.9562
2010-2011: Spearman=0.9790, Kendall=0.9576
2011-2012: Spearman=0.9978, Kendall=0.9775
2012-2013: Spearman=0.9915, Kendall=0.9730
2013-2014: Spearman=0.9903, Kendall=0.9587
2014-2015: Spearman=1.0000, Kendall=0.9990
2015-2016: Spearman=0.9975, Kendall=0.9702
2016-2017: Spearman=0.9984, Kendall=0.9783
2017-2018: Spearman=0.9998, Kendall=0.9935
2018-2019: Spearman=0.9970, Kendall=0.9763
2019-2020: Spearman=0.9929, Kendall=0.9717
2020-2021: Spearman=0.9953, Kendall=0.9852
2021-2022: Spearman=0.9997, Kendall=0.9905
2022-2023: Spearman=0.9982, Kendall=0.9730
2023-2024: Spearman=0.9995, Kendall=0.9894
2024-2025: Spearman=0.9990, Kendall=0.9786

L'analyse réalisée sur un échantillon de 3000 îles pour des raisons de performance, révèle une stabilité remarquable entre 2007 et 2025. Les coefficients de Spearman (> 0,98) et de Kendall (> 0,96) indiquent que l'ordre relatif des pays reste quasi-constant d'une année à l'autre, tant pour la population que pour la densité. Cette forte concordance temporelle s'explique par l'inertie démographique : les populations évoluent lentement et les écarts entre pays maintiennent les positions relatives. Les valeurs légèrement plus faibles pour certaines années (2009-2010, 2023-2025) peuvent signaler des changements ponctuels dus à des événements exceptionnels.

RÉFLEXION PERSONNELLE :

Sujet : les sciences des données et les humanités numériques.

Ce parcours m'a permis de mieux comprendre le rôle des sciences des données dans les sciences humaines et sociales, et en particulier en géographie. Les exercices montrent que les outils statistiques servent à structurer et analyser une réalité complexe, en faisant émerger des tendances générales à partir de données souvent nombreuses.
Les humanités numériques mettent en évidence le fait que le traitement des données repose sur des choix méthodologiques qui ne sont jamais neutres. Le choix des variables, des indicateurs, des tests ou des visualisations influence directement l'interprétation des phénomènes étudiés et la représentation des dynamiques spatiales.
Ce travail souligne également les limites des méthodes quantitatives. Les résultats dépendent fortement des choix effectués en amont et ne peuvent se substituer à une analyse qualitative. Les sciences des données apparaissent ainsi comme un outil complémentaire essentiel, à condition d'être mobilisées avec rigueur

OPTIONS :

Dans le cadre de ce cours, utilisant un Mac, j'ai dû adapter légèrement l'environnement de travail. J'ai privilégié Visual Studio Code et son extension Python, à la place de Docker, tout en conservant les fichiers Docker au cas où ils seraient nécessaires pour l'évaluation.
À chaque séance, j'ai créé un environnement virtuel dédié (venv), ce qui m'a permis de revenir facilement sur mes travaux précédents en cas de besoin.
Par ailleurs, l'installation des requirements était bloquée par SpaCy sur mon ordinateur, j'ai donc commenté cette dépendance afin de pouvoir poursuivre l'installation et le travail demandé.
Pour déposer l'ensemble de mes dossiers sur GitHub, j'ai utilisé l'IA Claude afin d'être guidée dans les démarches.