Rapport final des séances - analyse de données

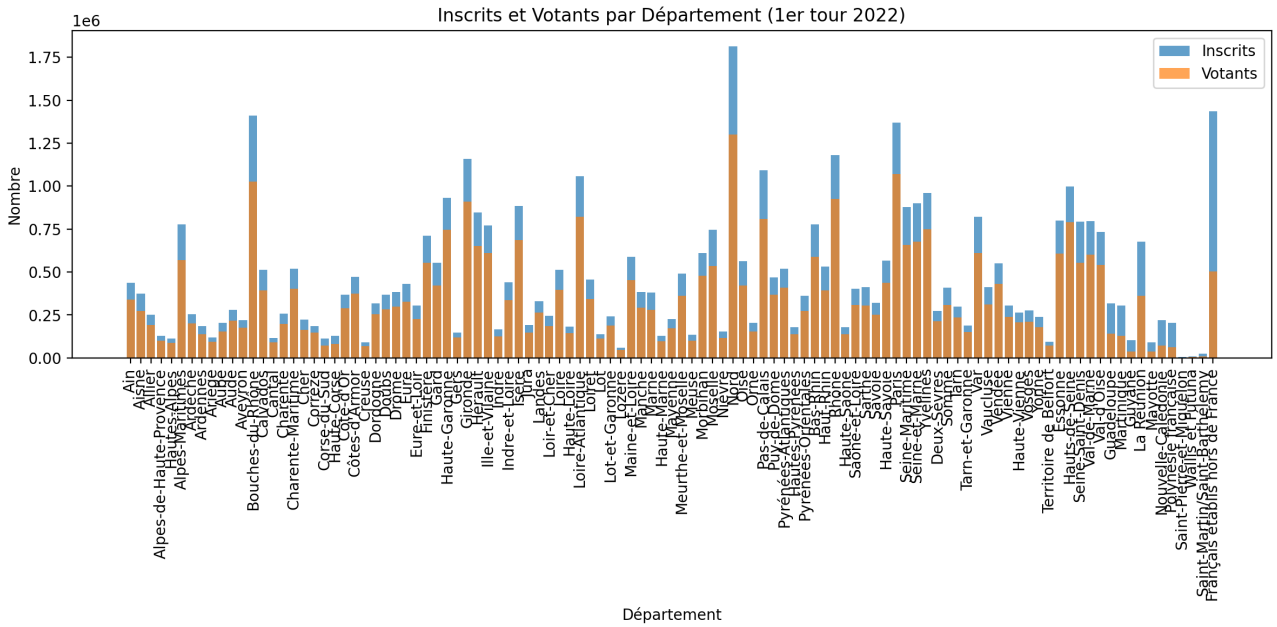
Le fond

Séance 2 :

Effectifs nationaux pour le 1er tour des élections présidentielles 2022 :

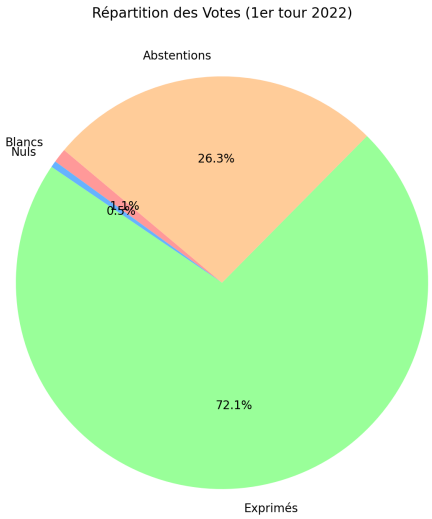
Catégorie	Effectif
Inscrits	48 747 876
Votants	35 923 707
Blancs	543 609
Nuls	247 151
Exprimés	35 132 947
Abstentions	12 824 169

Diagramme en barres comparant le nombre d'inscrits et de votants par département pour le 1er tour des élections présidentielles 2022 :



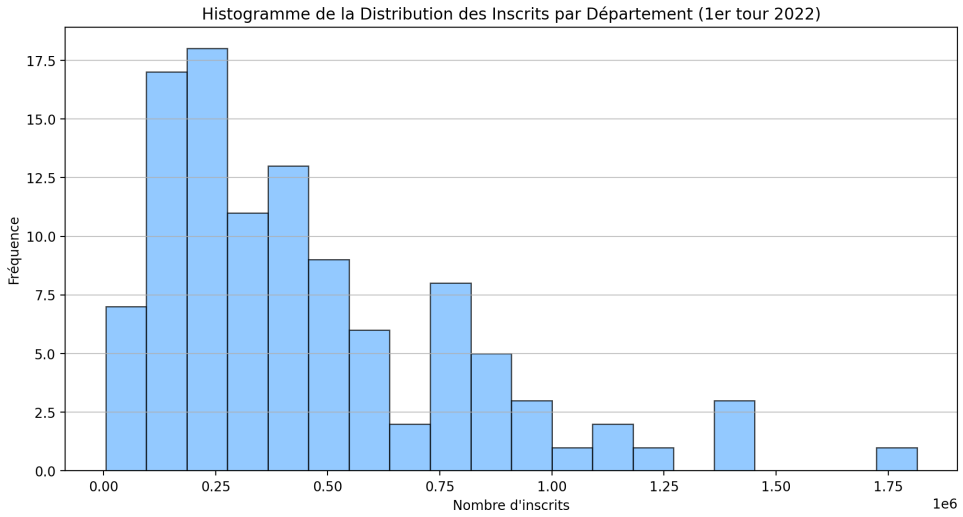
Commentaire technique :
Ce graphique met en évidence les disparités entre les départements en termes de participation électorale. Certains départements, comme Paris, montrent un taux de votants élevé par rapport aux inscrits, tandis que d'autres, souvent ruraux, affichent une participation plus faible. Ces différences pourraient refléter des dynamiques socio-économiques ou culturelles spécifiques, qu'une analyse plus approfondie pourrait éclairer.

Diagramme circulaire illustrant la répartition des votes (blancs, nuls, exprimés) et des abstentions pour le 1er tour des élections présidentielles 2022 :



Commentaire technique :
La majorité des électeurs ont exprimé un vote (97,8 %), tandis que les votes blancs (1,5 %) et nuls (0,7 %) restent marginaux. Les abstentions représentent 26,3 % des inscrits, ce qui souligne un enjeu majeur de participation citoyenne. Ce résultat pourrait refléter un désintérêt ou une défiance envers le processus électoral, ou encore des contraintes pratiques (ex. : mobilité, accès aux bureaux de vote).

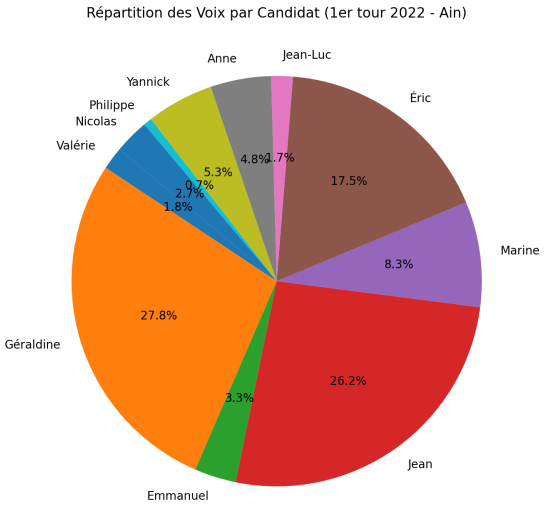
Histogramme de la distribution des inscrits par département pour le 1er tour des élections présidentielles 2022 :

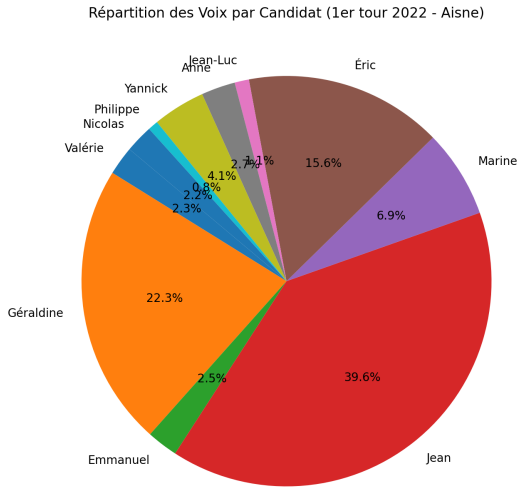


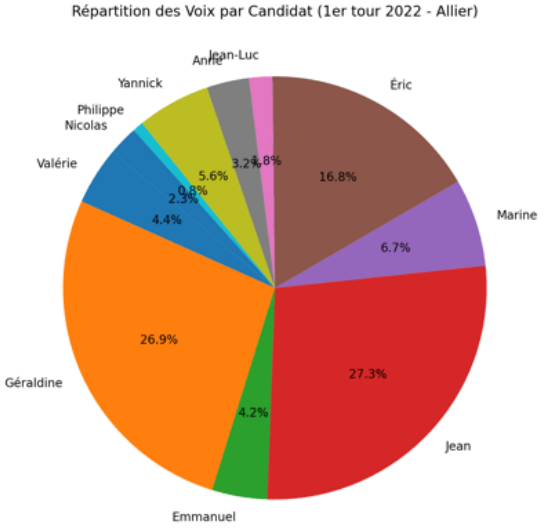
Commentaire technique :
L'histogramme montre une distribution inégale du nombre d'inscrits par département. La majorité des départements ont entre 200 000 et 600 000 inscrits, mais certains départements très peuplés (comme le Nord ou Paris) se distinguent par un nombre d'inscrits bien supérieur.

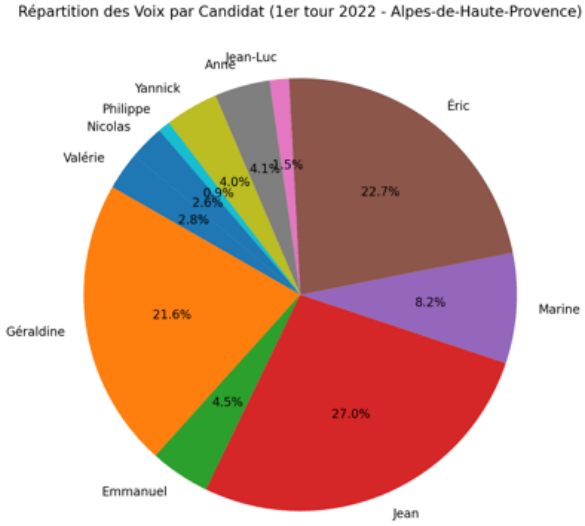
Cette disparité reflète les inégalités démographiques entre les territoires, avec des zones urbaines densément peuplées et des zones rurales moins peuplées.

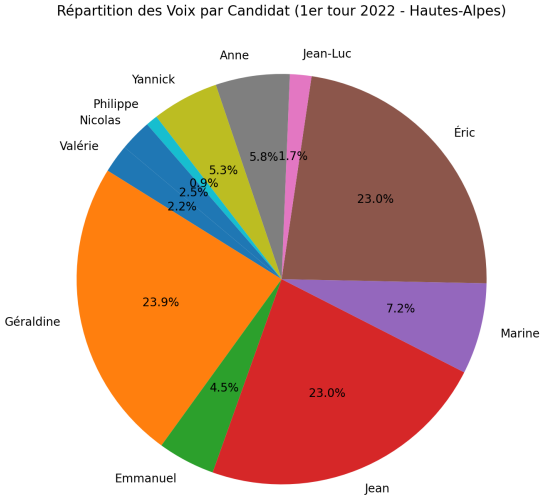
Bonus : diagrammes circulaires montrant la répartition des voix par candidat pour 5 départements (exemples) lors du 1er tour des élections présidentielles 2022 :











Commentaire technique :

Ces diagrammes révèlent des disparités régionales dans les préférences électorales. Par exemple, certains candidats obtiennent des scores plus élevés dans des départements spécifiques, reflétant des dynamiques locales (ex. : influence des enjeux socio-économiques ou culturels). Ces variations pourraient être analysées plus en détail pour comprendre les déterminants géographiques des choix électoraux.

Questions de cours

1. Positionnement de la géographie par rapport aux statistiques
La géographie utilise les statistiques comme un outil essentiel pour analyser des données massives. Historiquement, cette discipline a parfois sous-estimé leur importance, mais aujourd'hui, les statistiques permettent de réduire l'incertitude et de structurer l'analyse des phénomènes géographiques. Elles transforment la géographie en une science basée sur l'analyse rigoureuse des données.

2. Le hasard existe-t-il en géographie ?
Deux visions coexistent : le déterminisme, où tout a une cause, et le hasard comme cause cachée, explorable grâce aux progrès des connaissances. En géographie, le hasard existe localement sous forme de variabilité, mais des tendances globales sont identifiables. Les statistiques permettent de modéliser cette variabilité (loi normale, loi de Pareto, etc.).

3. Types d'information géographique
Il existe deux grands types d'informations géographiques : les données attributaires (population, variables sociales, économiques, climatiques) et les données géométriques (formes, contours, surfaces, réseaux). Ces deux éléments sont constitutifs des Systèmes d'Information Géographique (SIG).

4. Besoins de la géographie en analyse de données
La géographie nécessite la production et la collecte de données, ainsi que l'étude de leur structure interne (matrice individus-variables). Les outils statistiques permettent de résumer, modéliser, comparer et visualiser les phénomènes complexes, afin d'extraire des connaissances et de confronter théorie et réalité.

5. Différences entre statistique descriptive et explicative
La statistique descriptive résume et ordonne les données (moyennes, quantiles, histogrammes, ACP, AFC). La statistique explicative cherche à comprendre une variable Y à partir de variables explicatives X (régression, analyse discriminante, ANOVA), avec pour objectif d'expliquer ou de prédire.

6. Types de visualisation de données en géographie
Les diagrammes sectoriels conviennent aux variables qualitatives. Pour les variables quantitatives, on utilise des histogrammes, boîtes à moustaches, courbes cumulatives ou polygones de fréquences. Pour les données multivariées, on recourt aux cartes de proximité, cartes thématiques, ACP, AFC, ou ACM.

7. Méthodes d'analyse de données possibles
Les méthodes descriptives incluent l'ACP, l'AFC, l'ACM, et les classifications (CAH, nuées dynamiques). Les méthodes explicatives comprennent la régression, l'analyse discriminante, et l'ANOVA. Les méthodes de prévision analysent les séries chronologiques et les modèles prédictifs.

8. Définitions : population, individu, caractères, modalités statistiques
Une population statistique est un ensemble d'unités étudiées (ex. : villes d'une région). Un individu statistique est une unité spatiale (ex. : commune). Les caractères statistiques sont des

caractéristiques mesurées (ex. : âge, superficie). Les modalités sont les valeurs possibles d'un caractère. Les caractères peuvent être qualitatifs (nominal/ordinal) ou quantitatifs (discret/ continu), avec une hiérarchie selon les opérations possibles (ex. : tests paramétriques pour les variables quantitatives).

9. Mesure d'une amplitude et d'une densité
L'amplitude d'une classe est calculée par $A = b - a$, où b et a sont les bornes de la classe. La densité d'une classe est $d = ni / b - a$, où ni est l'effectif de la classe.

10. Formules de Sturges et de Yule
Les formules de Sturges $(k \approx 1+3,2222 \times log10(n))$ et de Yule $(k \approx 2,5 \times \sqrt[4]{n})$ servent à déterminer le nombre optimal de classes lors de la discrétisation d'un caractère quantitatif, évitant un découpage trop fin ou trop grossier.

11. Effectif, fréquence, fréquence cumulée, distribution statistique
Un effectif est le nombre d'apparitions d'une modalité dans une population. La fréquence $Fi = ni / n$, où n est l'effectif total. La fréquence cumulée Fk est la somme des fréquences des modalités inférieures à une valeur donnée. La distribution statistique est l'ensemble des fréquences observées pour les différentes modalités, permettant d'identifier la loi de probabilité sous-jacente.

Séance 3 :

Questions de cours

Synthèse des concepts statistiques en analyse de données
En statistique, le caractère quantitatif se distingue par sa généralité, car il permet de calculer une gamme complète de paramètres — moyenne, variance, moments — et de réaliser l'ensemble des opérations statistiques, qu'il s'agisse d'analyser la dispersion, la forme ou la position des données. À l'inverse, les caractères qualitatifs se limitent à des descripteurs plus simples, comme les effectifs ou les fréquences, ce qui en fait des outils moins polyvalents. Cette distinction est fondamentale, car elle détermine les possibilités d'analyse : les variables quantitatives, qu'elles soient discrètes (valeurs dénombrables, comme le nombre d'enfants) ou continues (valeurs sur un intervalle, comme un revenu), nécessitent des approches différentes. Les premières utilisent des sommes pour calculer leurs paramètres, tandis que les secondes reposent sur des intégrales. Cette différence influence non seulement les formules employées, mais aussi les représentations graphiques (histogrammes continus vs. discontinus) et les méthodes de calcul des indicateurs comme la médiane ou les quantiles.

Paramètres de position : moyennes, médianes et modes
L'existence de plusieurs types de moyennes — arithmétique, géométrique, quadratique, harmonique, ou mobile — répond à des besoins spécifiques. Par exemple, la moyenne géométrique est adaptée aux produits (comme les taux de croissance), tandis que la moyenne harmonique est utilisée pour les vitesses. La moyenne mobile, quant à elle, permet de lisser les séries temporelles. Cependant, la moyenne n'est pas toujours le meilleur indicateur de position centrale, surtout en présence de valeurs extrêmes. Dans ces cas, la médiane s'avère plus robuste, car elle n'est pas influencée par les outliers et résume efficacement la position centrale, même pour des séries très dissymétriques. Le mode, quant à lui, n'est calculable que lorsque la

distribution présente une valeur dominante claire. Il peut être absent (distribution uniforme) ou multiple (séries plurimodales), et son calcul dépend du regroupement en classes pour les variables continues.

Concentration et inégalités : médiale et indice de Gini
Pour évaluer la concentration d'un caractère dans une population, deux outils sont particulièrement utiles : la médiale et l'indice de Gini. La médiale, qui partage la masse totale en deux parties égales (50 %–50 %), est toujours supérieure à la médiane et permet d'évaluer les inégalités de distribution (par exemple, dans la répartition des revenus ou des surfaces). L'indice de Gini, quant à lui, mesure le degré de concentration : un indice élevé indique qu'une petite proportion d'individus concentre une grande part de la masse totale, révélant ainsi des inégalités marquées.

Dispersion et variabilité : variance, écart-type et quantiles
La variance et l'écart-type sont des paramètres clés pour mesurer la dispersion des données. La variance, calculée à partir des carrés des écarts à la moyenne, offre des propriétés mathématiques utiles, mais son unité (au carré) la rend moins intuitive. L'écart-type, en revanche, ramène cette mesure à l'unité d'origine, ce qui le rend plus interprétable. L'étendue (différence entre la valeur maximale et minimale) donne une première idée de la dispersion, mais sa fiabilité est limitée, car elle ne dépend que des valeurs extrêmes. Pour une analyse plus fine, les quantiles (quartiles, déciles) divisent la série en parties égales, permettant d'étudier la répartition interne des valeurs et de construire des indicateurs robustes. Enfin, la boîte de dispersion (ou boxplot) est un outil visuel puissant : elle résume à la fois la position centrale (médiane), la dispersion (quartiles, étendue), et l'asymétrie, tout en mettant en évidence les valeurs atypiques.

Forme et symétrie : moments et coefficients
Les moments centrés (calculés par rapport à la moyenne) et les moments absolus (basés sur la valeur absolue) servent à caractériser la forme d'une distribution. Les premiers mesurent la variance, l'asymétrie (skewness) et l'aplatissement (kurtosis), tandis que les seconds sont moins sensibles aux valeurs extrêmes. Vérifier la symétrie d'une distribution est crucial, car une distribution symétrique simplifie l'analyse : moyenne, médiane et mode coïncident, et les tests statistiques supposent souvent cette symétrie. Pour l'évaluer, on peut utiliser le coefficient d'asymétrie (β_1) :

$\beta_1 > 0$: distribution étalée à droite.
$\beta_1 < 0$: distribution étalée à gauche.
$\beta_1 = 0$: distribution symétrique.
Une comparaison des paramètres (mode ≈ médiane ≈ moyenne) confirme cette symétrie.

Question 5 :

colonne	moyenne	mediane	mode	écart type	écart absolu	etendue
Inscrits	455587,6	366859	5045	351003,8	272240,7	1808861
Abstentions	119852,1	95369	2272	117017,8	74959,07	929183
Votants	335735,6	274372	2773	258393,8	201517,2	1297100

Blancs	5080,46	4001	4577	3492,52	2817,95	17389
Nuls	2309,82	2039	17	1501,38	1131,99	8236
Exprimés	328345,3	268568	2701	253758,6	197762,2	1272080
Voix	1842	1627	1203	1268,37	977,36	7651
Voix.1	7499,27	5968	19	6501,29	4474,96	45883
Voix.2	91430,45	67831	534	77226,14	59929,14	372286
Voix.3	10293,34	8944	17010	7464,32	5140,37	48168
Voix.4	76017,08	64543	459	60278,1	42514,72	372668
Voix.5	23226,41	16885	9657	20760,6	15278,36	108537
Voix.6	72079,63	51556	501	66210,68	49157,01	316871
Voix.7	5761,48	4881	75	4581,79	3333,34	22826
Voix.8	15213,58	9561	72	14807,62	11136,57	80196
Voix.9	15691,6	11918	51	13027,13	9432,01	69513
Voix.10	2513,12	2118	3663	1781,41	1404,5	8686
Voix.11	6777,35	6152	7271	4636,02	3689,5	20535

Ce tableau représente les statistiques des données électorales par département du premier tour de la présidentielle 2022. Il révèle des disparités marquées dans les comportements électoraux selon les circonscriptions. Les inscrits et les votants affichent des moyennes élevées, mais avec des écarts-types et des étendues très larges, ce qui suggère que certaines circonscriptions (probablement urbaines ou très peuplées) pèsent lourdement sur les résultats globaux, tandis que d'autres, plus petites, restent marginales. Les abstentions, bien que moins nombreuses en moyenne, montrent aussi une forte variabilité, reflétant des dynamiques locales très contrastées.

Côté voix, quelques candidats (comme Voix.2, Voix.4, et Voix.6) se distinguent par des scores moyens élevés et une grande dispersion, indiquant une popularité inégale selon les territoires. À l'inverse, d'autres (comme Voix.10 ou Voix.11) obtiennent des résultats modestes et plus homogènes, signe d'un ancrage limité. Enfin, la présence de modes souvent éloignés des moyennes confirme que certaines valeurs reviennent fréquemment, probablement dans des circonscriptions de taille similaire.

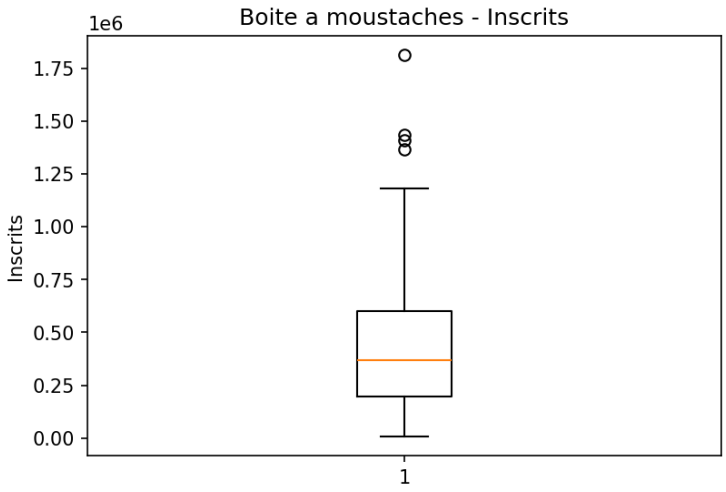
Question 7 et 9 :

colonne	Distance IQ	Distance ID
Inscrits	401050	793988,8
Abstentions	106489	193676,2
Votants	301770,5	602687,2
Blancs	4852,5	8845,8
Nuls	1917	3240,6
Exprimés	296870,5	590169,2
Voix	1517,5	3015,6
Voix.1	6264,5	13104,2
Voix.2	101317	177340,2
Voix.3	7999,5	13813
Voix.4	63342	130094,6
Voix.5	20638,5	43668,8
Voix.6	60743,5	159421,2
Voix.7	4779	10712,2
Voix.8	14833,5	38190,8
Voix.9	13265,5	27686,8
Voix.10	2466	4266,6
Voix.11	6146,5	12311

Ce tableau présente deux mesures de distance (IQ et ID) pour chaque catégorie électorale, ce qui permet d'analyser la dispersion ou l'écart entre les circonscriptions.
Les Inscrits, Votants, et Exprimés affichent des distances très élevées (notamment pour ID), ce qui confirme que certaines circonscriptions, probablement très peuplées ou urbaines, s'éloignent fortement des autres en termes de taille ou de participation. Les Abstentions, bien que moins marquées, montrent aussi des écarts significatifs, suggérant des disparités locales dans l'engagement électoral.
Côté voix, les candidats comme Voix.2, Voix.4, et Voix.6 se distinguent par des distances IQ et ID très élevées, indiquant une répartition très inégale de leurs scores selon les territoires.

Certains candidats obtiennent des résultats concentrés dans quelques circonscriptions, tandis que d'autres, comme Voix.10 ou Voix.11, présentent des distances plus faibles, signe d'une distribution plus homogène mais moins impactante.

Question 8 : La boite à moustache



Cette boîte à moustaches illustre une forte disparité dans le nombre d'inscrits par bureau de vote. La majorité des bureaux se situent entre 100 000 et 600 000 inscrits, avec une médiane aux alentours de 300 000, ce qui montre que la moitié des bureaux ont moins de ce nombre d'électeurs. Cependant, la présence de valeurs extrêmes dépassant 1,5 million d'inscrits révèle que certains bureaux, probablement situés dans des grandes villes ou des zones très peuplées, concentrent un nombre exceptionnellement élevé d'électeurs.
La distribution est asymétrique, avec une queue étirée vers le haut, indiquant que quelques bureaux très peuplés tirent la moyenne globale vers le haut. En résumé, cette représentation met en lumière une géographie électorale inégale, où une minorité de bureaux urbains ou

densément peuplés contrastent fortement avec une majorité de bureaux aux effectifs plus modestes.

Question 9 :

Surface (km²)	count
]0,10]	78423
]10,25]	2327
]25,50]	1164
]50,100]	788
]100,2500]	1346
]2500,5000]	60
]5000,10000]	40
]10000,+inf[	71

Cette répartition montre une forte concentration de petites surfaces parmi les bureaux ou circonscriptions analysés : plus de 78 000 d'entre eux ont une superficie inférieure à 10 km², ce qui suggère une majorité de zones urbaines ou très localisées. Les surfaces intermédiaires (entre 10 et 2500 km²) sont bien moins fréquentes, avec une baisse progressive des effectifs à mesure que la taille augmente. Enfin, les très grandes surfaces (au-delà de 5000 km²) sont rares, avec seulement 171 cas, indiquant des territoires vastes et probablement peu denses, comme des zones rurales étendues ou des circonscriptions spéciales. Cette distribution reflète un déséquilibre marqué entre des petites unités très nombreuses et quelques grandes étendues exceptionnelles.

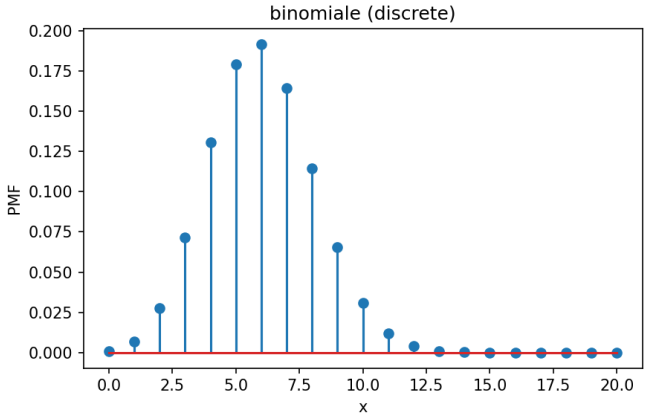
Question Bonus :

Surface (km²)	count
]0,10]	78423
]10,25]	2327
]25,50]	1164
]50,100]	788
]100,2500]	1346
]2500,5000]	60
]5000,10000]	40
]10000,+inf[	71
TOTAL	84219

Séance 4 :

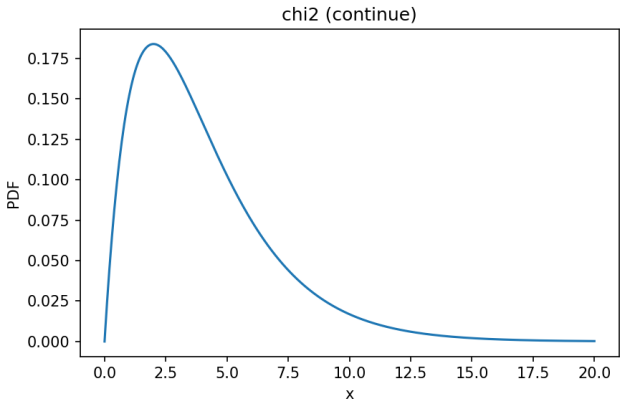
Question 1 :

a. Loi binomiale (discrète)



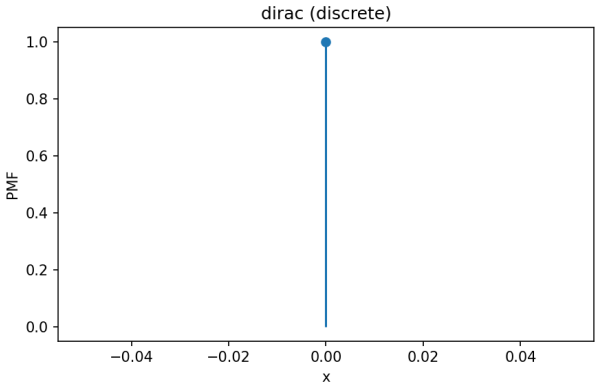
Interprétation :
Ce graphique montre la fonction de masse de probabilité (PMF) d'une loi binomiale discrète. Les pics de probabilité sont concentrés entre 5 et 10, ce qui suggère que les valeurs les plus probables pour cette distribution se situent dans cet intervalle. La probabilité diminue rapidement en dehors de cette plage, indiquant que les valeurs extrêmes (proches de 0 ou 20) sont peu probables. Cela reflète une situation où un événement a une certaine probabilité de se produire plusieurs fois dans un nombre fixe d'essais.

b. Chi2



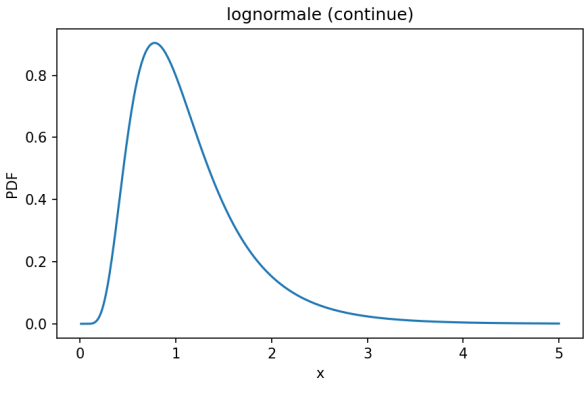
Interprétation :
Ce graphique représente la densité de probabilité (PDF) d'une loi du chi². La courbe montre un pic autour de 3-4, puis décroît progressivement vers la droite. Cela signifie que les valeurs proches de 3-4 sont les plus probables, tandis que les valeurs plus élevées deviennent de moins en moins probables. Cette distribution est souvent utilisée pour tester l'indépendance entre deux variables catégorielles ou pour évaluer l'adéquation d'un modèle à des données observées.

c. Dirac (discrète)



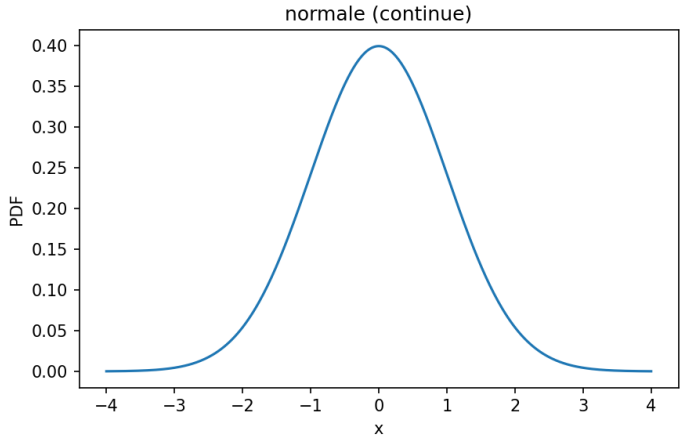
Interprétation :
Ce graphique montre une distribution de Dirac, qui est une distribution discrète concentrée en un seul point (ici, 0). La probabilité est de 1 à ce point et 0 ailleurs. Cela signifie que l'événement ne prend qu'une seule valeur avec certitude. C'est une distribution théorique souvent utilisée pour modéliser des phénomènes déterministes dans un cadre probabiliste.

d. Loi log-normale (continue)



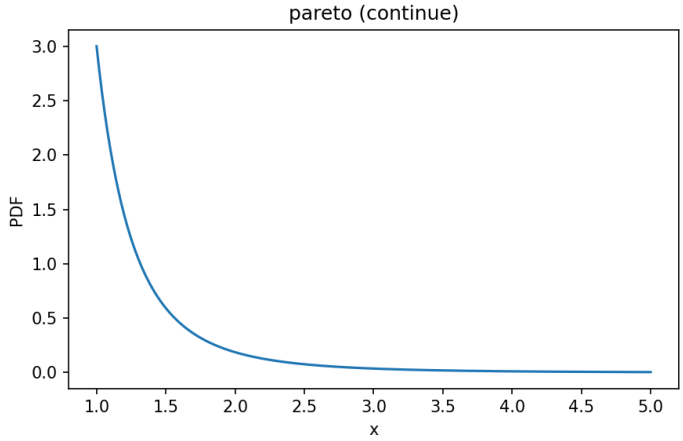
Interprétation :
Ce graphique représente la densité de probabilité (PDF) d'une loi lognormale. La courbe est fortement asymétrique vers la droite, avec un pic autour de 1 et une longue queue vers les valeurs élevées. Cela signifie que les petites valeurs sont plus probables, mais qu'il existe une probabilité non négligeable pour des valeurs beaucoup plus grandes. Cette distribution est souvent utilisée pour modéliser des phénomènes multiplicatifs, comme les revenus ou les tailles de particules.

e. Loi normale (continue)



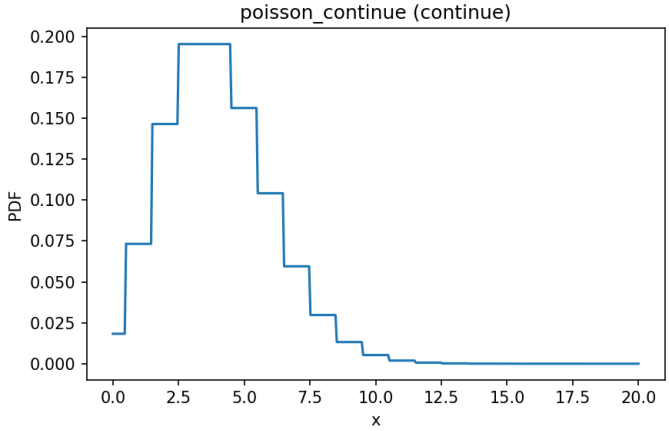
Interprétation :
Ce graphique montre la densité de probabilité (PDF) d'une loi normale (ou gaussienne). La courbe est symétrique et centrée autour de 0, avec un pic à cet endroit. La probabilité diminue de manière égale des deux côtés de la moyenne. Cette distribution est très courante et est utilisée pour modéliser de nombreux phénomènes naturels, sociaux ou physiques, où les valeurs se regroupent autour d'une moyenne.

f. Pareto (continue)



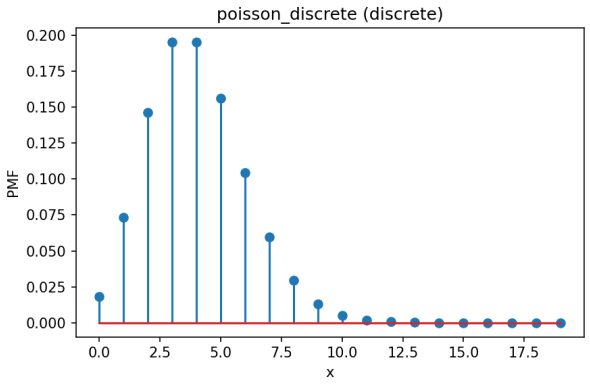
Interprétation :
Ce graphique montre la densité de probabilité (PDF) d'une loi de Pareto. La courbe décroît rapidement et suit une loi de puissance, indiquant que les petites valeurs sont très probables, tandis que les grandes valeurs deviennent de plus en plus rares, mais restent possibles. Cette distribution est souvent utilisée pour modéliser des phénomènes comme la répartition des richesses ou la taille des villes, où une minorité de valeurs extrêmes a un impact significatif.

g. Loi de Poisson (continue)



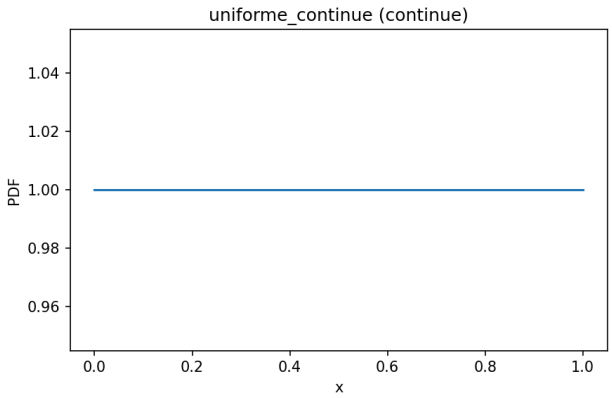
Interprétation :
Ce graphique représente une densité de probabilité (PDF) d'une loi de Poisson approximée en continu. La courbe montre une concentration des probabilités autour de x = 5, avec une décroissance progressive vers les valeurs plus élevées. Cela signifie que les événements modélisés par cette loi (comme le nombre d'occurrences d'un événement rare dans un intervalle de temps) sont les plus probables autour de cette valeur centrale. La loi de Poisson est souvent utilisée pour décrire des phénomènes comme le nombre d'appels reçus dans un centre d'appels ou le nombre de défauts dans un processus de fabrication.

h. Loi de Poisson (discrète)



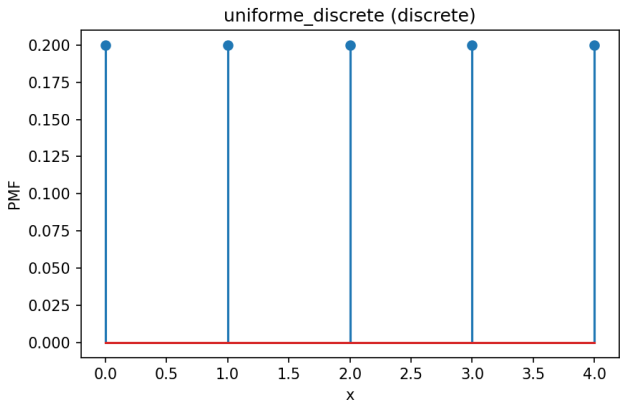
Interprétation :
Ce graphique montre la fonction de masse de probabilité (PMF) d'une loi de Poisson discrète. Les pics de probabilité sont concentrés autour de x = 5, avec des valeurs maximales proches de 0.2. Cela signifie que les événements modélisés (comme le nombre de fois qu'un événement se produit dans un intervalle donné) sont les plus probables autour de cette valeur. Les probabilités diminuent rapidement pour des valeurs plus élevées ou plus basses, ce qui est typique d'une loi de Poisson.

i. Uniforme (continue)



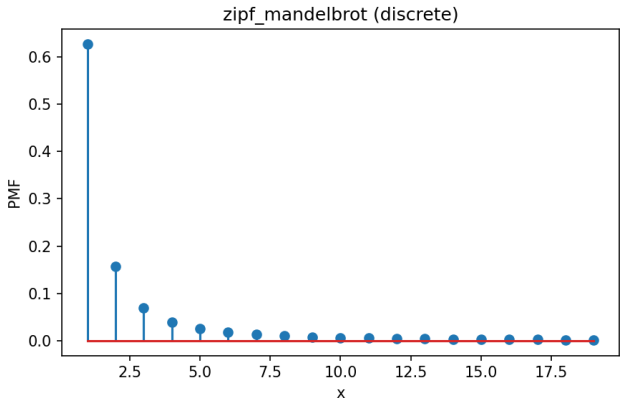
Interprétation :
Ce graphique représente la densité de probabilité (PDF) d'une loi uniforme continue. La courbe est une ligne droite horizontale à PDF = 1, indiquant que toutes les valeurs entre 0 et 1 ont la même probabilité. Cela signifie que chaque résultat dans cet intervalle est également probable. La loi uniforme est souvent utilisée pour modéliser des phénomènes où chaque résultat a la même chance de se produire, comme le tirage aléatoire d'un nombre entre 0 et 1.

j. Uniforme (discrète)



Interprétation :
Ce graphique montre la fonction de masse de probabilité (PMF) d'une loi uniforme discrète. Les probabilités sont égales pour chaque valeur discrète de x (0, 1, 2, 3, 4) avec une valeur de PMF = 0.2. Cela signifie que chaque valeur a la même chance de se produire. Cette distribution est souvent utilisée pour modéliser des événements où chaque résultat discret est également probable, comme le lancer d'un dé équilibré.

k. Zipf-Mandelbrot (discrète)



Interprétation :
Ce graphique représente la fonction de masse de probabilité (PMF) d'une loi de Zipf-Mandelbrot. La probabilité est très élevée pour x = 1 et décroît rapidement pour les valeurs plus grandes. Cela signifie que quelques événements sont très probables (comme les mots fréquents dans un texte), tandis que la majorité des événements sont rares. Cette loi est souvent utilisée pour modéliser des phénomènes comme la fréquence des mots dans un texte ou la taille des villes.

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 valeurs de moyenne et d'écart-type pour différentes distributions statistiques, offrant ainsi une comparaison de leur dispersion et de leur tendance centrale.
La distribution uniforme discrète a une moyenne de 3,5, ce qui est attendu pour une distribution symétrique sur un intervalle discret, avec un écart-type de 1,7078, reflétant une dispersion modérée autour de cette moyenne. La distribution binomiale montre une moyenne de 5, ce qui correspond probablement à un nombre d'essais et une probabilité de succès typiques, avec un écart-type de 1,5811, indiquant une dispersion relativement faible.
La distribution de Poisson (discrète) a une moyenne de 3, ce qui est cohérent avec sa propriété où la moyenne et la variance sont égales, et un écart-type de 1,7320, confirmant cette caractéristique. La version continue de la distribution de Poisson a une moyenne et un écart-type très proches de ceux de la version discrète, ce qui est attendu pour une approximation continue.
La distribution de Zipf se distingue par une moyenne de 2,2540 et un écart-type élevé de 2,7294, indiquant une forte dispersion. Cela est typique des distributions de Zipf, où quelques valeurs ont une probabilité très élevée, tandis que la majorité des valeurs ont une probabilité très faible.
La distribution normale a une moyenne de 0, ce qui est typique pour une distribution centrée, et un écart-type proche de 1 (0,9994), ce qui est attendu pour une distribution normale standard.
La distribution log-normale a une moyenne de 1,1294 et un écart-type de 0,5920, indiquant une asymétrie positive, typique de cette distribution où les valeurs sont concentrées vers la gauche mais avec une queue étendue vers la droite.
La distribution uniforme continue a une moyenne de 3 et un écart-type de 1,1547, ce qui est cohérent avec une distribution uniforme sur un intervalle continu.
La distribution du Chi² montre une moyenne de 4,9722 et un écart-type de 3,1016, ce qui est typique pour une distribution du Chi² avec plusieurs degrés de liberté, où la moyenne est proche du nombre de degrés de liberté et l'écart-type est relativement élevé.
Enfin, la distribution de Pareto a une moyenne de 1,5176 et un écart-type de 0,6478, indiquant une forte concentration des valeurs autour de la moyenne, mais avec une queue épaisse vers les valeurs élevées, typique de cette distribution.

Questions de cours :

Choix et utilisation des lois statistiques en géographie : une analyse complète
En analyse de données, le choix entre une loi discrète et une loi continue repose avant tout sur la nature du phénomène étudié. Ce choix est guidé par plusieurs critères : la structure des valeurs observées, la forme de la distribution empirique (visuellement ou statistiquement testable), ainsi que les caractéristiques de la série (espérance, médiane, variance, asymétrie). Certaines lois s'adaptent mieux que d'autres en fonction de leur flexibilité et du nombre de paramètres qu'elles intègrent.

Lois discrètes : comptages et événements dénombrables
Les lois discrètes sont utilisées lorsque les valeurs possibles sont dénombrables, souvent limitées à des entiers. Elles sont particulièrement adaptées aux comptages : nombre d'événements, de succès ou d'échecs, ou encore d'individus. Parmi les lois discrètes les plus courantes, on retrouve la loi binomiale, qui modélise le nombre de succès dans une série d'essais indépendants, la loi de Bernoulli, qui décrit un essai unique à deux issues, et la loi de Poisson, essentielle pour les événements rares. Cette dernière est particulièrement utile en

géographie pour modéliser des occurrences ponctuelles dans l'espace ou le temps, comme le nombre d'accidents, de séismes, ou d'autres phénomènes localisés.

Lois continues : mesures et phénomènes sur un intervalle
À l'inverse, les lois continues s'appliquent lorsque la variable peut prendre toutes les valeurs d'un intervalle, non dénombrables. Elles sont utilisées pour des mesures continues telles que le temps, la distance, l'altitude, ou la température. Parmi les lois continues les plus répandues, on trouve la loi normale (ou loi de Gauss), décrite comme « la plus fréquente ». Elle constitue souvent la distribution limite de nombreux phénomènes naturels et sociaux, comme la répartition des hauteurs, des températures, des revenus, ou des rendements agricoles. La loi log-normale, quant à elle, est essentielle pour modéliser des variables multiplicatives et asymétriques, comme la taille des villes, la surface des parcelles, les revenus, ou l'intensité de certains flux. Enfin, la loi exponentielle est adaptée aux processus liés au temps d'attente ou aux risques, comme les phénomènes de fiabilité et de survie. Elle est utile pour modéliser les durées d'événements naturels ou techniques, ou encore la probabilité de défaillance entre deux occurrences.

Lois spécifiques à la géographie
En géographie, certaines lois statistiques sont particulièrement importantes. La loi de Poisson, par exemple, est décrite comme « indispensable pour les événements rares ». Elle permet de modéliser des occurrences ponctuelles dans l'espace ou le temps, comme le nombre de séismes ou d'accidents dans une région donnée. La loi normale, en tant que distribution limite, est omniprésente et permet d'approximer de nombreuses variables naturelles et sociales, comme les hauteurs, les températures, ou les revenus.

La loi log-normale est cruciale pour analyser des variables asymétriques et multiplicatives, comme la taille des villes ou les revenus. Elle est souvent utilisée pour étudier la loi de Galton-Gibrat, qui décrit la croissance proportionnelle des entités (ex. : les villes). Les lois rang-taille, comme celles de Zipf ou Zipf-Mandelbrot, sont quant à elles utilisées pour analyser les distributions rang-taille, notamment pour les tailles de villes. Elles permettent de modéliser la hiérarchie urbaine et d'analyser la structure polarisée d'un territoire.
Enfin, la loi exponentielle est employée pour étudier les processus liés au temps d'attente ou aux risques. Elle est particulièrement adaptée aux phénomènes de fiabilité et de survie, comme les durées d'événements naturels ou techniques, ou encore la modélisation de la probabilité de défaillance entre deux occurrences.

Critères de choix et applications pratiques
Le critère majeur pour choisir entre une loi discrète et une loi continue reste la nature du phénomène et la structure des valeurs observées. Ce choix est également appuyé par la forme empirique de la distribution et les paramètres statistiques (espérance, variance, asymétrie). En géographie, ces lois permettent non seulement de décrire les phénomènes, mais aussi de les modéliser, les comprendre, et même les prédire.
Par exemple, pour étudier la répartition des séismes dans une région, la loi de Poisson sera privilégiée en raison de la nature discrète et rare des événements. À l'inverse, pour analyser la distribution des revenus ou des tailles de villes, les lois continues comme la loi normale ou log-normale seront plus adaptées, car elles permettent de capturer la variabilité et l'asymétrie des données.

Conclusion
En résumé, le choix entre une loi discrète et une loi continue dépend de la nature des données et du phénomène étudié. Les lois discrètes sont idéales pour les comptages et les événements rares, tandis que les lois continues conviennent mieux aux mesures et aux phénomènes distribués sur un intervalle. En géographie, ces outils statistiques sont indispensables pour modéliser des dynamiques spatiales complexes, qu'il s'agisse d'analyser des événements ponctuels (séismes, accidents) ou des phénomènes continus (revenus, tailles de villes). Leur maîtrise permet une analyse rigoureuse et une prise de décision éclairée, essentielle pour les politiques publiques et l'aménagement du territoire.

Séance 5 :

Questions de cours :

L'échantillonnage consiste à sélectionner un sous-ensemble d'individus, appelé échantillon, à partir d'une population mère, selon une procédure aléatoire ou systématique. L'objectif est d'inférer les caractéristiques de la population à partir des informations fournies par cet échantillon.
L'étude de la population entière est souvent irréalisable en pratique. Les populations peuvent être trop vastes ou difficiles d'accès, et les contraintes de temps, de coût ou de logistique rendent le recensement exhaustif peu réaliste. L'échantillonnage permet alors d'obtenir des résultats fiables à condition que l'échantillon soit représentatif de la population.

On distingue deux grandes familles de méthodes d'échantillonnage.

Les méthodes aléatoires reposent sur un mécanisme probabiliste garantissant à chaque individu une probabilité connue de sélection. Le sondage aléatoire simple assure l'équiprobabilité, tandis que le tirage peut s'effectuer avec ou sans remise. L'échantillonnage systématique consiste à sélectionner des individus selon un pas fixe dans une liste ordonnée. La méthode des quotas vise à reproduire dans l'échantillon les proportions de certains sous-groupes de la population.

Les méthodes non aléatoires ne reposent pas sur un tirage probabiliste. L'échantillonnage par convenance privilégie les individus facilement accessibles, au détriment de la représentativité. La méthode de Monte Carlo utilise des simulations aléatoires pour estimer des paramètres, principalement dans des contextes numériques ou théoriques.

Le choix d'une méthode dépend de l'objectif de l'étude, de la disponibilité d'une base de sondage exhaustive, des contraintes pratiques et de la taille de l'échantillon. En général, un échantillon aléatoire et représentatif est préférable à un échantillon plus large mais biaisé.

Comment définir un estimateur et une estimation ?
Un estimateur est une variable aléatoire, c'est-à-dire une fonction des données de l'échantillon, destinée à approcher un paramètre inconnu de la population, comme une moyenne ou une variance. Par exemple, la moyenne empirique $\bar{X}$ est un estimateur de la moyenne μ de la population.

Une estimation correspond à la valeur numérique obtenue lorsque l'estimateur est calculé à partir d'un échantillon donné. Ainsi, si la moyenne empirique vaut $\bar{X}=5$, alors 5 est une estimation de la moyenne μ.

Comment distinguer un intervalle de fluctuation et un intervalle de confiance ?
Un intervalle de fluctuation est un intervalle théorique qui encadre les valeurs possibles de la fréquence observée dans un échantillon, en supposant que la proportion p de la population est connue. Il permet de vérifier si une fréquence observée est compatible avec l'hypothèse portant sur p.
Un intervalle de confiance est un intervalle calculé à partir des données de l'échantillon, destiné à encadrer un paramètre inconnu de la population avec un certain niveau de confiance. Il fournit une estimation de la plage de valeurs plausibles du paramètre.
La différence essentielle réside dans le statut du paramètre : l'intervalle de fluctuation suppose le paramètre connu et teste la compatibilité des données, tandis que l'intervalle de confiance suppose le paramètre inconnu et vise à l'estimer.

Qu'est-ce qu'un biais dans la théorie de l'estimation ?
Un biais correspond à l'écart entre l'espérance mathématique d'un estimateur et la vraie valeur du paramètre qu'il cherche à estimer. Un estimateur est dit sans biais lorsque son espérance est égale au paramètre réel ; il est biaisé dans le cas contraire.
Un biais systématique altère la qualité de l'estimation, car il conduit à des valeurs systématiquement trop élevées ou trop faibles. Même avec un grand échantillon, un estimateur biaisé peut produire des résultats peu fiables.

Quelle statistique utilise la population totale ? Quel lien avec les données massives ?
Une statistique exhaustive, ou recensement, consiste à observer l'ensemble des individus d'une population. Elle fournit des résultats exacts, sans recourir à l'inférence statistique. Un exemple typique est le recensement national de la population.
Les données massives (Big Data) permettent parfois d'analyser des populations entières à partir de traces numériques ou de capteurs, réduisant voire supprimant le recours à l'échantillonnage. Elles offrent une grande précision et éliminent le biais d'échantillonnage.
Cependant, ces approches soulèvent des défis importants en termes de collecte, de stockage, de traitement des données et de respect de la vie privée. Ainsi, l'échantillonnage reste une méthode privilégiée lorsqu'une analyse exhaustive est trop complexe, tandis que les données massives ouvrent de nouvelles perspectives lorsqu'une couverture quasi complète est possible.

Quels sont les enjeux autour du choix d'un estimateur ?
Le choix d'un estimateur est un enjeu central de la statistique inférentielle. Il s'agit d'approcher les paramètres inconnus d'une population mère à partir d'un échantillon nécessairement imparfait. Toute estimation est donc entachée d'incertitude.
Un premier enjeu consiste à limiter l'erreur d'estimation. L'échantillon ne fournissant qu'une information partielle, les estimations sont affectées par les fluctuations d'échantillonnage. L'objectif est alors de minimiser l'écart entre la valeur estimée et la valeur réelle du paramètre.
Un second enjeu concerne le compromis biais–variance. Un estimateur peut être biaisé lorsque son espérance diffère du paramètre réel, ou présenter une variance élevée lorsque les

estimations sont très dispersées. Le critère fondamental est l'erreur quadratique moyenne (ERQM), qui combine biais et variance : ERQM(θ^)=V(θ^)+(biais)2.

L'enjeu est donc de privilégier un estimateur non biaisé ou faiblement biaisé, et de variance minimale.
Enfin, un bon estimateur doit vérifier des propriétés de convergence et de consistance. Lorsque la taille de l'échantillon augmente, il doit converger vers le vrai paramètre, avec un biais et une variance tendant vers zéro. Cela garantit la fiabilité de l'inférence à long terme.
L'enjeu global est ainsi d'obtenir une estimation fiable, précise, stable et interprétable, malgré l'incertitude inhérente à l'échantillonnage.

Quelles sont les méthodes d'estimation d'un paramètre ? Comment en sélectionner une ?
L'estimation ponctuelle consiste à associer à un paramètre inconnu une statistique calculée à partir de l'échantillon. Par exemple, la moyenne empirique estime l'espérance, la variance empirique corrigée estime la variance, et la fréquence empirique estime une proportion. Cette approche fournit une valeur unique, sans indication directe sur l'incertitude.
L'estimation par intervalle, ou intervalle de confiance, permet d'encadrer le paramètre avec un niveau de confiance donné. Par exemple : [μ^ ±tα/2 n σ^]

Cette méthode est essentielle pour quantifier l'incertitude liée à l'échantillonnage.
Les méthodes fondées sur la vraisemblance reposent sur un principe fondamental de l'inférence statistique : choisir les paramètres qui rendent les données observées les plus probables. Elles conduisent à des estimateurs efficaces et font le lien avec les statistiques exhaustives et l'information de Fisher.

Les estimateurs robustes visent à limiter l'influence des valeurs aberrantes. Ils incluent notamment la médiane, les quartiles, les moyennes tronquées et les M-estimateurs (Huber, bicarré).
Le choix d'une méthode d'estimation dépend de plusieurs critères : la nature du paramètre étudié, les propriétés statistiques de l'estimateur (absence de biais, variance minimale, convergence), la taille de l'échantillon, la sensibilité aux valeurs extrêmes et la quantité d'information conservée. On privilégie ainsi un estimateur sans biais, convergent, efficace et robuste, adapté au contexte empirique.

Quels sont les tests statistiques ? À quoi servent-ils ? Comment en construire un ?
Les tests statistiques permettent de prendre une décision sous incertitude à partir d'un échantillon, concernant un paramètre de la population ou une hypothèse théorique (égalité, différence, effet). Ils constituent une application centrale de l'inférence statistique.
On distingue notamment les tests portant sur une moyenne (basés sur la loi normale ou la loi de Student), les tests sur une proportion, les tests fondés sur des intervalles de fluctuation, ainsi que les tests reposant sur des résultats asymptotiques, en particulier le théorème central limite.
La construction d'un test statistique suit une procédure rigoureuse. On commence par formuler les hypothèses : l'hypothèse nulle H0 et l'hypothèse alternative H1. On choisit ensuite une statistique de test adaptée au paramètre étudié, puis on détermine sa loi sous H0 (normale, Student ou asymptotique). Un niveau de risque α est alors fixé, généralement à 5 %.

Enfin, on définit une région critique et on compare la statistique observée au seuil critique afin de décider du rejet ou non de H0. Un test statistique est donc une procédure décisionnelle formalisée, fondée sur les propriétés probabilistes des estimateurs.

Que penser des critiques de la statistique inférentielle ?
La statistique inférentielle fait l'objet de plusieurs critiques. Tout d'abord, ses résultats sont probabilistes et jamais certains. Elle dépend fortement des hypothèses de modèle (normalité, indépendance, homogénéité), et peut être affectée par des échantillons biaisés. Elle est également sensible aux valeurs aberrantes lorsque les outils utilisés ne sont pas adaptés.
Cependant, la statistique inférentielle intègre ses propres limites. Elle quantifie l'erreur d'estimation, contrôle le risque des tests, mobilise l'information de Fisher et la borne de Cramér–Rao, et propose des estimateurs robustes ainsi que des intervalles de confiance.
La statistique inférentielle n'est pas une science de la certitude, mais une science de la décision raisonnée sous incertitude. Les critiques sont pertinentes lorsque ses hypothèses sont ignorées, mais injustes lorsqu'on néglige la rigueur méthodologique qu'elle impose. Elle demeure indispensable dès lors qu'un recensement exhaustif est impossible.

Séance 6 :

Questions de cours :

Qu'est-ce qu'une statistique ordinale ? À quoi s'oppose-t-elle ?
La statistique ordinale (ou statistique d'ordre) est essentielle en géographie humaine. Elle consiste à classer des entités géographiques — villes, régions, pays — selon un ordre précis : qui monte, qui descend, ou qui stagne dans un classement. Contrairement à la statistique nominale, qui se contente de catégoriser sans hiérarchie (comme des couleurs ou des noms), la statistique ordinale introduit une notion de rang ou de niveau.

Quels types de variables utilise-t-elle ?
Elle s'appuie sur des variables qualitatives ordinales, c'est-à-dire des données qui expriment un ordre ou une intensité. Par exemple :
Un classement des villes selon leur attractivité économique.
Une échelle de développement (très développé, développé, peu développé).
Comment cela crée-t-il une hiérarchie spatiale ?
La statistique ordinale permet de structurer l'espace en établissant des relations de domination ou de subordination entre les territoires. Par exemple :

Centre vs. périphérie : Une métropole peut être classée comme "centre dominant", tandis qu'une zone rurale est considérée comme "périphérique".
Urbanisation : Du cœur dense d'une ville aux zones rurales les moins peuplées.
Risques naturels : En géographie physique, elle permet de classer l'intensité des crues d'un fleuve ou la force des séismes dans une région.

À quoi ça ressemble sur une carte ?
Cartographiquement, cela se traduit par des cartes en classes ordonnées (du plus fort au plus faible). Ces représentations visuelles mettent en évidence des relations de pouvoir (centralité, domination) ou des niveaux territoriaux (rangs, statuts).

En résumé, la statistique ordinale révèle les hiérarchies spontanées qui émergent dans les sociétés et les espaces, qu'elles soient économiques, sociales ou environnementales.

Quel ordre privilégier dans les classifications ?
En règle générale, on utilise l'ordre croissant (ou ordre naturel), car il correspond à une progression logique (du plus petit au plus grand). Cependant, il existe des exceptions, comme la loi rang-taille en géographie, qui étudie la répartition des villes selon leur taille.

À quoi sert l'ordination ?
Elle permet de repérer les valeurs aberrantes – celles qui sont trop élevées ou trop basses par rapport à la tendance générale. Par exemple, une ville dont la croissance démographique est anormalement rapide ou lente dans un classement.

Corrélation des rangs vs. concordance des classements : quelle différence ?

Corrélation des rangs :
Elle mesure l'intensité et la direction de la relation entre deux séries de classements. Par exemple, si on compare le classement des villes selon leur PIB et selon leur population, on cherche à savoir si les villes les plus riches sont aussi les plus peuplées. On utilise souvent le coefficient de Spearman pour cette analyse.

Concordance des classements :
Elle évalue le degré d'accord entre plusieurs classements (plus de deux). Par exemple, si trois études classent les mêmes villes selon leur attractivité, leur qualité de vie et leur dynamisme économique, la concordance mesure si ces classements se ressemblent globalement. Contrairement à la corrélation, elle ne se limite pas à deux variables.

Test de Spearman vs. test de Kendall : quelles différences ?

Test de Spearman :
Mesure une corrélation monotone entre deux séries de rangs.
Compare les rangs transformés des variables et calcule une corrélation sur ces rangs.
Sensible aux valeurs extrêmes et aux ex æquo (égalités dans les rangs).

Test de Kendall :
Mesure l'accord entre les paires d'observations.
Compare chaque couple de données pour voir s'ils sont concordants (même ordre) ou discordants (ordre inverse).
Moins sensible aux ex æquo et aux valeurs aberrantes que Spearman.

Spearman est plus simple et rapide, mais moins robuste face aux données atypiques.
Kendall est plus précis pour analyser les accords entre paires, mais plus complexe à calculer.

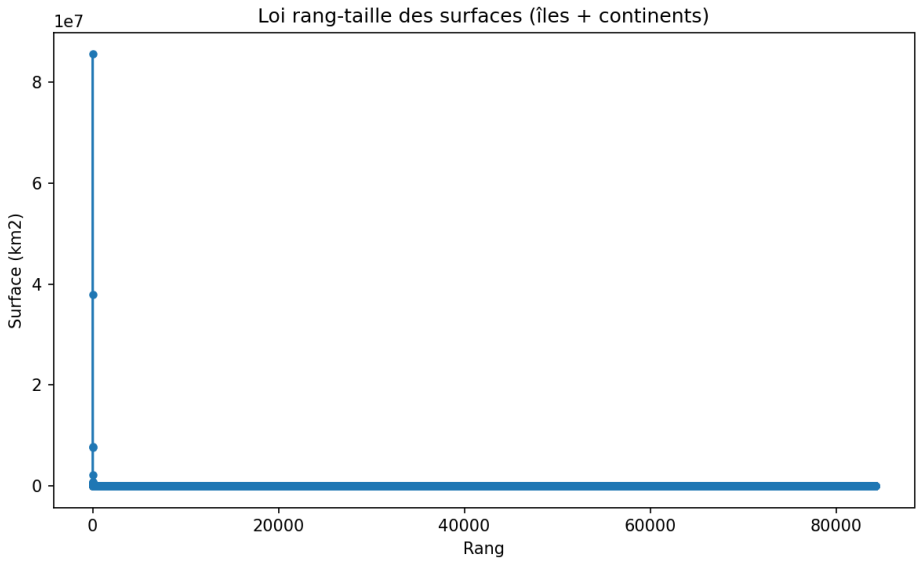
À quoi servent les coefficients de Goodman-Kruskal et de Yule ?

Coefficient de Goodman-Kruskal :
Mesure l'excès de paires concordantes par rapport aux paires discordantes.

Exprime une proportion d'accord entre deux classements, en tenant compte de toutes les paires possibles.

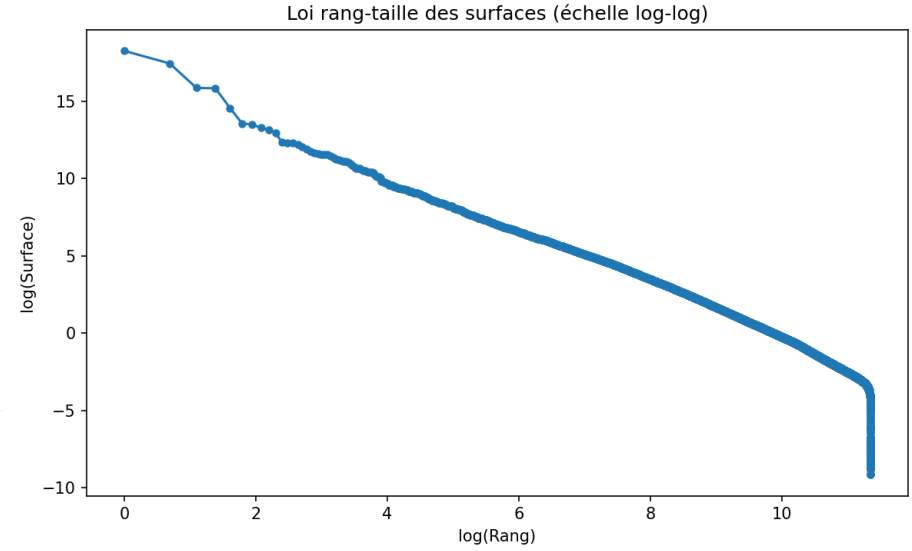
Coefficient de Yule :
Cas particulier de Goodman-Kruskal, appliqué aux tableaux 2x2 (matrices de contingence).
Évalue la fréquence des événements en comparant les cases d'un tableau croisé (ex. : présence/absence de deux phénomènes).
Utile pour étudier des relations binaires (oui/non, présent/absent).
En géographie, ces coefficients permettent de vérifier si deux hiérarchies spatiales (comme le classement des villes selon leur richesse et leur niveau d'éducation) sont cohérentes ou indépendantes.

Question 5 : La loi rang-taille



Interprétation :
Ce graphique illustre la relation entre le rang et la taille des surfaces (îles et continents) en échelle linéaire. On observe une décroissance brutale des surfaces dès les premiers rangs, reflétant la présence de quelques très grandes entités – probablement les continents – suivies d'une multitude de surfaces beaucoup plus petites. La courbe montre que la majorité des îles et archipels ont des superficies négligeables comparées aux continents, formant une "longue traîne" proche de zéro. Cela met en lumière une distribution très inégale des surfaces terrestres, où une poignée de grandes masses domine largement.

Question 6 : La loi rang-taille log



caractéristique d'une loi de
uit un modèle fractal, où la
rts mineurs à cette droite,
imites dans la distribution,
firme que la répartition des
, typique des phénomènes naturels.

Question 14 :

Coefficient de Spearman : 0,9862
Tau de Kendall : 0,9043

Coefficient de Spearman : 0,9673
Tau de Kendall : 0,8588

Les résultats des tests de corrélation de rangs révèlent une forte cohérence entre les deux classifications étudiées.
Pour le premier couple de variables, le coefficient de Spearman atteint 0,986 et le tau de Kendall 0,904, avec des p-valeurs quasi nulles (respectivement 1,07e-134 et 2,09e-69). Ces valeurs indiquent une relation statistiquement très significative, confirmant une concordance presque parfaite entre les deux classements.
Le second couple de variables affiche également une corrélation marquée, avec un coefficient de Spearman de 0,967 et un tau de Kendall de 0,859, accompagnés de p-valeurs tout aussi significatives (3,69e-103 et 8,63e-63). Ces résultats renforcent l'idée que les départements les plus peuplés tendent à être les plus denses. La légère différence entre les coefficients de Spearman et de Kendall est normale, car ce dernier est généralement plus conservateur, car il prend en compte les paires concordantes de manière distincte.
Les p-valeurs quasi nulles permettent de rejeter sans ambiguïté l'hypothèse d'indépendance entre les rangs. Cela valide statistiquement l'existence d'une structure commune dans l'organisation démographique des départements français.

Question Bonus :

Spearman : 0,1446

Kendall : 0,0965

L'analyse de corrélation entre le classement des îles par leur surface et celui par la longueur de leur littoral met en évidence une relation statistiquement significative, bien que faible.
Les coefficients de corrélation obtenus sont bas : 0,145 pour le coefficient de Spearman et 0,097 pour le tau de Kendall, avec des p-valeurs inférieures à 0,001. Ces résultats indiquent que, même si la corrélation est statistiquement significative, la taille d'une île ne permet pas de prédire efficacement la longueur de son littoral. Malgré cette faiblesse, la significativité statistique confirme l'existence d'une tendance générale, bien que limitée.

Réflexion personnelle : Les sciences des données et les humanités numériques

Les sciences des données et les humanités numériques représentent deux facettes d'une même révolution : celle de la transformation des savoirs par le numérique. Là où les premières s'attachent à extraire des motifs et des prédictions à partir de masses de données, les secondes interrogent les implications culturelles, historiques et sociales de ces mêmes données. Pour un·e géographe, cette complémentarité est particulièrement féconde. La géographie, discipline à la croisée des sciences naturelles et des sciences humaines, a toujours eu pour ambition de décrire et d'expliquer les interactions entre les sociétés et leurs environnements. Aujourd'hui, avec l'avènement des données

massives (big data), des systèmes d'information géographique (SIG) et des outils d'analyse spatiale, cette ambition prend une nouvelle dimension. Les données ne sont plus seulement des supports d'analyse, mais des objets de recherche à part entière, porteurs de biais, de silences et de récits qu'il s'agit de décrypter.

L'un des apports majeurs des sciences des données en géographie réside dans leur capacité à révéler des dynamiques spatiales invisibles à l'œil nu. Grâce à l'analyse de données satellitaires, de flux de mobilité ou de réseaux sociaux, il devient possible de cartographier des phénomènes aussi variés que l'étalement urbain, les migrations climatiques ou les inégalités d'accès aux ressources. Par exemple, l'utilisation d'algorithmes de machine learning pour traiter des images satellites permet de suivre en temps réel la déforestation en Amazonie ou la montée des eaux dans les zones côtières. Ces outils offrent une précision et une granularité inédites, mais ils soulèvent aussi des questions éthiques et méthodologiques : comment garantir la représentativité des données ? Comment éviter que les modèles ne reproduisent – voire n'amplifient – les biais existants ? C'est ici que les humanités numériques interviennent, en rappelant que les données ne sont jamais neutres. Elles sont produites dans des contextes historiques, politiques et techniques qui influencent leur collecte, leur traitement et leur interprétation.

Les humanités numériques, en insistant sur la dimension critique et réflexive, invitent les géographes à ne pas se contenter d'appliquer des méthodes quantitatives, mais à interroger leur pertinence et leurs limites. Par exemple, la cartographie participative, qui associe les citoyen·ne·s à la production de données géographiques, illustre cette approche hybride. Des projets comme OpenStreetMap montrent comment les outils numériques peuvent être mobilisés pour donner la parole aux populations marginalisées et produire des connaissances plus inclusives. À l'inverse, l'analyse des "silences" des données – ces zones ou ces groupes absents des bases de données officielles – révèle les rapports de pouvoir qui structurent l'information géographique. Ainsi, les humanités numériques rappellent que la géographie ne se réduit pas à une science des localisations, mais qu'elle est aussi une science des représentations et des récits.
Enfin, l'articulation entre sciences des données et humanités numériques ouvre des perspectives passionnantes pour la recherche en géographie, notamment dans le domaine de la visualisation et de la narration. Les cartes interactives, les storytelling cartographiques ou les atlas numériques ne sont pas de simples supports de communication : ils deviennent des outils de médiation entre les données brutes et le grand public. Par exemple, un projet comme "Dear Data", où deux designers échangent des postcards basées sur des données personnelles collectées manuellement, montre comment l'art et la science peuvent dialoguer pour rendre tangibles des phénomènes complexes. Pour les géographes, ces approches offrent une opportunité de repenser la manière dont on raconte les territoires, en combinant rigueur analytique et sensibilité narrative.

En conclusion, les sciences des données et les humanités numériques ne sont pas deux mondes parallèles, mais deux faces d'une même médaille : celle d'une géographie augmentée, à la fois plus précise et plus réflexive. Pour les étudiant·e·s et les chercheur·e·s en géographie, maîtriser ces outils et ces questionnements est devenu indispensable. Non seulement pour analyser les transformations rapides de nos environnements, mais aussi pour contribuer à une production de savoirs plus juste et plus ouverte. Dans un monde où les données sont de plus en plus centralisées entre les mains de quelques acteurs privés ou étatiques, cette double compétence – technique et critique – est un levier pour démocratiser l'accès à l'information géographique et en faire un bien commun.

Options :

Pour mener à bien ce projet, j'ai rapidement réalisé que travailler en groupe serait un atout majeur. En effet, les défis techniques et conceptuels étaient nombreux, et discuter avec mes camarades m'a permis de mieux comprendre les attentes, d'échanger des solutions et de progresser plus efficacement. Cette collaboration a été essentielle pour surmonter les obstacles, d'autant plus que la pédagogie inversée, bien que stimulante, m'a parfois mise en difficulté. N'ayant jamais abordé cette matière auparavant, j'aurais apprécié pouvoir passer plus de temps en cours à en comprendre les bases, avec des démonstrations claires, avant de me lancer dans des travaux pratiques aussi autonomes.
Parmi les difficultés les plus marquantes, l'installation des dépendances, notamment SpaCy, a été un frein important. Malgré plusieurs tentatives, cette étape bloquait

systématiquement sur mon ordinateur, ce qui a ralenti ma progression. J'ai donc dû commenter cette dépendance pour pouvoir continuer l'installation et avancer dans le travail demandé.
Côté environnement de travail, j'ai utilisé un Mac et adapté mon espace en privilégiant Visual Studio Code et son extension Python, plutôt que Docker. Cependant, j'ai conservé les fichiers Docker nécessaires pour les évaluations. Pour sécuriser mes travaux et faciliter les retours en arrière, j'ai créé un environnement virtuel dédié avec venv.
Enfin, pour finaliser le dépôt de l'ensemble de mes dossiers sur GitHub, j'ai utilisé l'IA Claude. Cet outil m'a guidée pas à pas dans les démarches, ce qui m'a permis de conclure cette étape sereinement.