ALLAOUA SEGURON TESS Master 1 GEOINT

RAPPORT D'ACTIVITÉ ANALYSE DE DONNÉES

QUESTION DE COURS :

Séance 2. Les principes généraux de la statistique

1. Quel est le positionnement de la géographie par rapport aux statistiques ?

La géographie entretient une relation particulière avec la statistique, souvent marquée par des tensions. La statistique est une discipline scientifique issue des mathématiques, qui fournit des outils méthodologiques rigoureux. La géographie, quant à elle, est une discipline en constante construction, dont l'un des rôles majeurs est la production de données liées à l'étude des territoires.

Dans ce cadre, l'outil statistique est indispensable pour analyser ces volumes importants de données. Pourtant, les géographes ont parfois sous-estimé l'apport des méthodes statistiques, ce qui a pu limiter la portée de certaines analyses. Une maîtrise solide des statistiques apparaît donc essentielle pour assurer la validité scientifique des travaux géographiques.

2. Le hasard existe-t-il en géographie ?

La question de l'existence du hasard relève avant tout d'un débat philosophique. Selon le déterminisme, notamment défendu par Laplace, le hasard n'existe pas réellement, car chaque phénomène possède une cause, même si celle-ci n'est pas toujours connue.

À l'inverse, l'approche statistique admet que certains phénomènes sont aléatoires dans leurs réalisations individuelles, tout en restant prévisibles globalement. En géographie, et particulièrement en géographie humaine, il est impossible de prévoir le comportement de chaque acteur, mais il est néanmoins possible de dégager des tendances générales à l'aide des outils statistiques.

3. Quels sont les types d'information géographique ?

L'information géographique se divise en deux grandes catégories.
D'une part, les caractères attribuaires regroupent les informations thématiques associées à un territoire ou à une unité spatiale, comme les données démographiques, sociales ou

économiques en géographie humaine, ou les données climatiques en géographie physique. D'autre part, les données géométriques concernent directement l'espace et décrivent la forme, la localisation et la géométrie des objets géographiques.

4. Quels sont les besoins de la géographie au niveau de l'analyse de données ?

La géographie a besoin de l'analyse de données pour comprendre la structure interne de jeux de données souvent complexes. Cette analyse permet d'organiser et de synthétiser l'information disponible.

Les résultats statistiques doivent ensuite être confrontés aux conditions de production des données ainsi qu'aux connaissances existantes sur le phénomène étudié. L'objectif est de transformer des observations brutes en informations exploitables, notamment dans une perspective opérationnelle, tout en maintenant un lien étroit avec le raisonnement géographique.

5. Quelles sont les différences entre la statistique descriptive et la statistique explicative ?

La statistique descriptive a pour objectif de décrire et de résumer les données observées, qu'elles proviennent d'une population ou d'un échantillon. Elle permet d'obtenir une image simplifiée de la réalité en mettant de l'ordre dans les données et constitue une étape indispensable de toute analyse statistique.

La statistique explicative, aussi appelée statistique inférentielle, s'appuie sur la statistique descriptive et sur les lois de probabilité théoriques afin de modéliser les phénomènes réels. Elle vise à généraliser les résultats obtenus et à produire des prédictions ou des inférences sur la population étudiée.

6. Quelles sont les types de visualisation de données en géographie ? Comment les choisir ?

La visualisation graphique est un élément central de la description statistique en géographie. Le choix de la représentation dépend de la nature du caractère étudié.

Les variables quantitatives continues sont généralement représentées par des histogrammes, des boîtes à moustaches ou des courbes de lissage. Les variables qualitatives sont souvent illustrées par des diagrammes sectoriels, tandis que les variables qualitatives ordinales peuvent être représentées par des histogrammes disjoints.

Le choix de la visualisation doit permettre de représenter fidèlement la distribution des données et de faciliter leur compréhension.

7. Quelles sont les méthodes d'analyse de données possibles ?

Les méthodes d'analyse statistique peuvent être regroupées en trois grandes catégories.
Les méthodes descriptives ont pour objectif de visualiser et de classer les données, notamment dans le cadre d'analyses multidimensionnelles.
Les méthodes explicatives cherchent à établir des relations entre une variable à expliquer et une ou plusieurs variables explicatives.
Enfin, les méthodes de prévision se concentrent sur l'analyse des séries chronologiques afin d'anticiper l'évolution future des phénomènes.

8. Comment définir la population statistique, l'individu statistique, les caractères statistiques et les modalités statistiques ? Quels sont les types de caractères et existe-t-il une hiérarchie ?

La population statistique correspond à l'ensemble des unités sur lesquelles porte l'étude.
L'individu statistique désigne une unité élémentaire appartenant à cette population.
Les caractères statistiques sont les caractéristiques observées sur les individus et deviennent des variables statistiques lorsque leurs modalités sont connues pour chaque individu.
Les modalités représentent les différentes valeurs possibles d'un caractère et doivent être exclusives et exhaustives.

Les caractères statistiques se divisent en variables qualitatives (nominales et ordinales) et quantitatives (discrètes et continues). Les variables quantitatives occupent le niveau de mesure le plus élevé, car elles permettent le calcul de nombreux paramètres statistiques.

9. Comment mesurer une amplitude et une densité ?

L'amplitude correspond à la longueur d'une classe lors de la discrétisation d'une variable quantitative. Elle est calculée comme la différence entre les bornes supérieure et inférieure de la classe.

La densité de classe est le rapport entre l'effectif de la classe et son amplitude. Elle permet de comparer des classes de tailles différentes et d'analyser la concentration des observations.

10. À quoi servent les formules de Sturges et de Yule ?

Les formules de Sturges et de Yule sont utilisées pour déterminer le nombre optimal de classes lors de la discrétisation des variables quantitatives. Elles permettent d'obtenir une représentation équilibrée des données, en évitant à la fois une perte d'information et une fragmentation excessive.

11. Comment définir un effectif, une fréquence, une fréquence cumulée et une distribution statistique ?

L'effectif correspond au nombre d'individus appartenant à une modalité ou à une classe donnée.
La fréquence est la proportion de cet effectif par rapport à l'effectif total de la population.
La fréquence cumulée est obtenue en additionnant progressivement les fréquences jusqu'à une valeur donnée.

La distribution statistique décrit la manière dont les valeurs d'un caractère sont réparties au sein de la population, généralement à l'aide d'un tableau ou d'un graphique.

12. Quel caractère est le plus général : le caractère quantitatif ou le caractère qualitatif ?

Le caractère quantitatif est considéré comme le plus général, car il permet non seulement de classer et d'ordonner les données, mais aussi d'effectuer des opérations arithmétiques et de calculer de nombreux paramètres statistiques. Les caractères qualitatifs, en particulier nominaux, se limitent principalement au calcul des fréquences.

13. Que sont les caractères quantitatifs discrets et continus ? Pourquoi les distinguer ?

Les caractères quantitatifs discrets prennent des valeurs isolées et dénombrables, tandis que les caractères quantitatifs continus peuvent prendre une infinité de valeurs dans un intervalle donné. Cette distinction est essentielle pour la discrétisation des données et pour le choix des représentations graphiques adaptées.

Séance 3. Les paramètres statistiques élémentaires

14. Pourquoi existe-t-il plusieurs types de moyenne ?

Il existe plusieurs types de moyennes afin de s'adapter à la nature des phénomènes étudiés. La moyenne arithmétique ne permet pas toujours de rendre compte correctement de la

régularité des observations, ce qui justifie l'utilisation d'autres moyennes, comme la moyenne géométrique ou harmonique, selon les objectifs de l'analyse.

15. Pourquoi calculer une médiane ?

La médiane permet d'identifier la valeur qui partage la distribution en deux parts égales. Contrairement à la moyenne, elle est peu sensible aux valeurs extrêmes, ce qui en fait une mesure pertinente lorsque la distribution est asymétrique ou comporte des valeurs aberrantes.

16. Quand est-il possible de calculer un mode ?

Le mode correspond à la valeur ou à la classe la plus fréquente d'une distribution. Il peut être calculé pour tous les types de variables, y compris les variables qualitatives nominales, contrairement à la moyenne ou à la médiane.

17. Quel est l'intérêt de la médiale et de l'indice de Gini ?

La médiale permet de diviser la masse totale d'une variable en deux parts égales, ce qui est utile pour analyser les inégalités. L'indice de Gini mesure le degré d'inégalité de la répartition d'une variable au sein d'une population, avec des valeurs comprises entre égalité parfaite et inégalité maximale.

18. Pourquoi calculer une variance plutôt que l'écart à la moyenne, et pourquoi utiliser l'écart type ?

La somme des écarts à la moyenne étant toujours nulle, elle ne permet pas de mesurer la dispersion. La variance corrige ce problème en élevant les écarts au carré. L'écart type, qui est la racine carrée de la variance, permet de retrouver l'unité de mesure initiale et facilite l'interprétation.

19. Pourquoi calculer l'étendue ?

L'étendue correspond à la différence entre la valeur maximale et la valeur minimale d'une série. Elle fournit une indication simple et immédiate de l'amplitude totale des valeurs observées.

20. À quoi sert un quantile et quels sont les plus utilisés ?

Les quantiles permettent de diviser une distribution en parts égales en termes de fréquence. Ils sont utilisés pour analyser la dispersion et identifier des populations atypiques. Les quantiles les plus couramment utilisés sont les quartiles.

21. Pourquoi construire une boîte de dispersion et comment l'interpréter ?

La boîte de dispersion synthétise la position, la dispersion et l'asymétrie d'une variable quantitative. Elle repose sur les quartiles et permet d'identifier la médiane, l'intervalle interquartile, l'étendue des valeurs non aberrantes et la présence éventuelle de valeurs extrêmes.

22. Quelle est la différence entre les moments centrés et les moments absolus ?

Les moments centrés reposent sur les écarts à la moyenne et servent à analyser la forme de la distribution, notamment l'asymétrie et l'aplatissement. Les moments absolus utilisent les écarts absolus et sont moins sensibles aux valeurs extrêmes.

23. Pourquoi vérifier la symétrie d'une distribution et comment le faire ?

La vérification de la symétrie est essentielle car de nombreux tests statistiques supposent une distribution normale. Elle peut être évaluée à l'aide du coefficient d'asymétrie, issu du moment centré d'ordre trois.

Séance 4. Les distributions statistiques

1. Quels critères mettriez-vous en avant pour choisir entre une distribution statistique avec des variables discrètes et une distribution statistique avec des variables continues?

Selon moi, le critère principal à mettre en avant est le type de donnée, qui déterminera la méthode de l'enquête. Si ce sont des valeurs quantitatives discrètes, ce sont des valeurs numériques limitées et comptables. Par exemple, le nombre de séances de cours d'analyse de données dans un semestre.

En revanche, les données quantitatives continues sont des mesures qui peuvent prendre une infinité de valeurs dans un intervalle donné, ce qui permet une représentation très précise des valeurs. Par exemple, le nombre de lignes de code Python réalisées au cours de l'histoire du numérique.

2. Expliquez selon vous quelles sont les lois les plus utilisées en géographie:

Selon moi, les lois les plus utilisées en géographie sont la loi normale, la loi de Poisson, la loi binomiale, la loi de Zipf et la loi log-normale, car elles permettent de modéliser les phénomènes spatiaux. Par exemple, la loi normale pour les altitudes ou températures, la loi de Poisson pour les séismes, et la loi de Zipf pour la répartition des villes.

Séance 5. Les statistiques inférentielles

Comment définir l'échantillonnage et pourquoi ne pas utiliser la population entière ?

L'échantillonnage consiste à sélectionner une partie de la population statistique afin d'en tirer des conclusions sur l'ensemble de cette population. Il s'agit d'une démarche centrale en statistique inférentielle, qui permet d'étudier un phénomène sans avoir à observer tous les individus.

L'étude de la population entière est souvent impossible ou peu réaliste, en raison de contraintes matérielles, financières et temporelles. De plus, certaines populations sont trop vastes ou évoluent trop rapidement pour être observées de manière exhaustive. L'échantillonnage permet donc de réduire les coûts et les délais tout en produisant des résultats exploitables, à condition que l'échantillon soit représentatif.

Quelles sont les méthodes d'échantillonnage et comment les choisir ?

Les méthodes d'échantillonnage peuvent être regroupées en deux grandes catégories.
Les méthodes probabilistes reposent sur un tirage aléatoire et garantissent que chaque individu a une probabilité connue d'être sélectionné. Elles incluent notamment l'échantillonnage aléatoire simple, l'échantillonnage stratifié et l'échantillonnage en grappes.

Les méthodes non probabilistes reposent sur des choix empiriques, comme l'échantillonnage par quotas ou par convenance. Le choix de la méthode dépend des objectifs de l'étude, des contraintes disponibles, de la structure de la population et du niveau de précision attendu.

Comment définir un estimateur et une estimation ?

Un estimateur est une règle de calcul ou une fonction mathématique appliquée aux données issues d'un échantillon afin d'approcher un paramètre inconnu de la population. Il s'agit donc d'un outil théorique défini avant l'observation des données.

L'estimation correspond à la valeur numérique obtenue une fois que l'estimateur est appliqué aux données observées. Ainsi, l'estimateur est un concept abstrait, tandis que l'estimation est un résultat concret, dépendant de l'échantillon étudié.

Comment distinguer l'intervalle de fluctuation et l'intervalle de confiance ?

L'intervalle de fluctuation est utilisé lorsque le paramètre de la population est connu. Il permet de déterminer l'intervalle dans lequel une statistique observée est susceptible de varier du fait de l'aléa d'échantillonnage. Il est principalement mobilisé dans le cadre des tests statistiques.

À l'inverse, l'intervalle de confiance est utilisé lorsque le paramètre de la population est inconnu. Il fournit une fourchette de valeurs dans laquelle ce paramètre a une forte probabilité de se situer. L'intervalle de confiance permet donc d'exprimer l'incertitude associée à une estimation.

Qu'est-ce qu'un biais dans la théorie de l'estimation ?

Un biais correspond à une erreur systématique introduite dans l'estimation d'un paramètre. Il se manifeste lorsque l'espérance mathématique d'un estimateur est différente de la valeur réelle du paramètre que l'on cherche à estimer.

Un biais peut provenir de plusieurs sources, notamment d'un échantillonnage non représentatif, d'erreurs de mesure ou d'un modèle statistique inadapté. Un estimateur biaisé conduit à des conclusions incorrectes, même lorsque la taille de l'échantillon est importante.

Comment appelle-t-on une statistique portant sur la population totale et quel est le lien avec les données massives ?

Une statistique portant sur l'ensemble de la population est qualifiée de statistique exhaustive. Dans ce cas, il n'y a pas d'incertitude liée à l'échantillonnage, puisque tous les individus sont observés.

Les données massives, ou « big data », s'inscrivent partiellement dans cette logique d'exhaustivité, car elles reposent sur la collecte d'un volume très important de données.

Toutefois, malgré cette abondance, des traitements statistiques restent nécessaires pour structurer, analyser et interpréter l'information.

Quels sont les enjeux liés au choix d'un estimateur ?

Le choix d'un estimateur est crucial car il conditionne la qualité des résultats obtenus. Un bon estimateur doit être peu biaisé, précis et convergent, c'est-à-dire qu'il doit se rapprocher de la valeur réelle du paramètre lorsque la taille de l'échantillon augmente.

Il doit également être robuste face aux valeurs extrêmes et adapté à la nature des données étudiées. Un mauvais choix d'estimateur peut conduire à des interprétations erronées et remettre en cause la validité scientifique de l'analyse.

Quelles sont les méthodes d'estimation d'un paramètre et comment en sélectionner une ?

Les principales méthodes d'estimation incluent l'estimation ponctuelle, qui fournit une valeur unique du paramètre, et l'estimation par intervalle, qui exprime l'incertitude autour de cette valeur. D'autres méthodes, comme le maximum de vraisemblance ou l'estimation bayésienne, reposent sur des hypothèses probabilistes spécifiques.

Le choix d'une méthode dépend du cadre théorique, de la taille de l'échantillon, des hypothèses sur la distribution des données et des objectifs de l'étude. Il doit également prendre en compte les contraintes pratiques et la facilité d'interprétation des résultats.

Quels sont les tests statistiques existants, à quoi servent-ils et comment en construire un ?

Les tests statistiques peuvent être paramétriques ou non paramétriques. Les tests paramétriques reposent sur des hypothèses fortes concernant la distribution des données, tandis que les tests non paramétriques sont plus souples mais souvent moins puissants.

Les tests statistiques servent à prendre une décision à partir des données observées, en comparant une hypothèse nulle à une hypothèse alternative. La construction d'un test repose sur plusieurs étapes : formulation des hypothèses, choix d'une statistique de test, fixation d'un seuil de risque, calcul de la statistique et prise de décision.

Que penser des critiques de la statistique inférentielle ?

Les critiques adressées à la statistique inférentielle portent principalement sur le non-respect de ses hypothèses, l'interprétation abusive des résultats et la dépendance aux modèles théoriques. Ces critiques sont légitimes lorsque les méthodes sont appliquées de manière mécanique ou sans réflexion.

Cependant, lorsqu'elle est utilisée de façon rigoureuse et contextualisée, la statistique inférentielle constitue un outil essentiel pour analyser des phénomènes complexes et incertains. Elle permet de quantifier l'incertitude et d'éclairer la prise de décision, à condition de rester consciente de ses limites.

Séance 6. La statistique d'ordre des variables qualitatives

1. Qu'est-ce qu'une statistique ordinale ? À quelle autre statistique catégorielle s'oppose-t-elle ? Quel type de variables utilise-t-elle ? En quoi permet-elle de matérialiser une hiérarchie spatiale ?

La statistique ordinale, aussi appelée statistique d'ordre, consiste à classer une série d'observations selon un ordre donné, le plus souvent croissant. Elle repose sur l'ordination des valeurs observées afin de mettre en évidence la position relative de chaque entité au sein d'un ensemble. L'objectif principal de cette approche est d'évaluer l'évolution d'une entité dans un classement, par exemple pour déterminer si elle progresse, stagne ou recule par rapport aux autres.

La statistique ordinale s'oppose à la statistique nominale, qui concerne des variables qualitatives sans ordre naturel entre les modalités. Dans le cas des variables nominales, il n'est pas pertinent d'établir un classement, contrairement aux variables ordinales pour lesquelles un ordre hiérarchique a du sens.

Elle mobilise principalement des variables ordinales, mais elle peut également être appliquée à des variables quantitatives, continues ou discrètes, lorsque celles-ci sont transformées en rangs. Dans ce cas, l'analyse ne porte plus sur les valeurs elles-mêmes, mais sur leur position relative.

En géographie, la statistique ordinale permet de matérialiser des hiérarchies spatiales en mettant en évidence des classements d'objets géographiques. Elle est particulièrement centrale en géographie humaine, notamment pour analyser des phénomènes comme la hiérarchie urbaine à travers le classement des villes selon leur population ou leur poids économique.

2. Quel ordre est à privilégier dans les classifications ?

Dans les classifications statistiques, l'ordre à privilégier est l'ordre croissant qui est en realité l'ordre naturel. Cet ordre facilite la lecture des distributions et permet d'identifier plus aisément les valeurs extrêmes, qu'elles soient exceptionnellement faibles ou élevées.

3. Quelle est la différence entre une corrélation des rangs et une concordance de classements ?

La corrélation des rangs vise à mesurer le degré de dépendance entre deux classements ou deux variables ordinales. Elle permet de déterminer si deux classements évoluent de manière similaire ou non. Les coefficients de Spearman et de Kendall sont utilisés dans ce cadre pour tester l'indépendance ou la liaison entre deux séries ordonnées.

La concordance de classements, quant à elle, s'applique à un nombre supérieur ou égal à deux classements. Elle cherche à évaluer le niveau d'accord global entre plusieurs classements portant sur les mêmes objets. Le coefficient de concordance de Kendall (W) constitue une généralisation du coefficient de Kendall et permet d'apprécier la cohérence globale de plusieurs critères de classement.

4. Quelle est la différence entre les tests de Spearman et de Kendall ?

Les tests de Spearman et de Kendall ont un objectif commun : comparer deux classements afin de déterminer s'ils sont similaires ou non. Il s'agit dans les deux cas de tests non paramétriques, ce qui signifie qu'ils ne reposent pas sur des hypothèses fortes concernant la distribution des données.

Le test de Spearman est fondé sur le coefficient de corrélation classique appliqué aux rangs. Il repose sur le calcul des différences entre les rangs associés aux deux variables, puis sur la somme des carrés de ces différences. Ce test suppose l'absence de rangs ex æquo, ou nécessite une correction lorsque ceux-ci sont présents.

Le test de Kendall repose sur une logique différente. Il s'appuie sur le comptage des paires de rangs concordantes et discordantes afin de calculer un score global. Le coefficient τ est obtenu en rapportant ce score au nombre total de paires possibles. L'un des principaux avantages du test de Kendall réside dans sa capacité à être généralisé à plusieurs classements grâce au coefficient de concordance W.

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

Les coefficients de Goodman-Kruskal et de Yule sont utilisés pour mesurer l'association entre des variables qualitatives, en particulier lorsqu'elles sont ordinales ou dichotomiques.

Le coefficient de Goodman-Kruskal (Γ) mesure l'excès de paires concordantes par rapport aux paires discordantes. Il prend des valeurs comprises entre -1 et +1 et s'interprète de manière similaire à un coefficient de corrélation : une valeur proche de +1 indique une forte association positive, tandis qu'une valeur proche de -1 traduit une association négative. Ce coefficient est particulièrement adapté à l'analyse du lien entre deux variables ordinales.

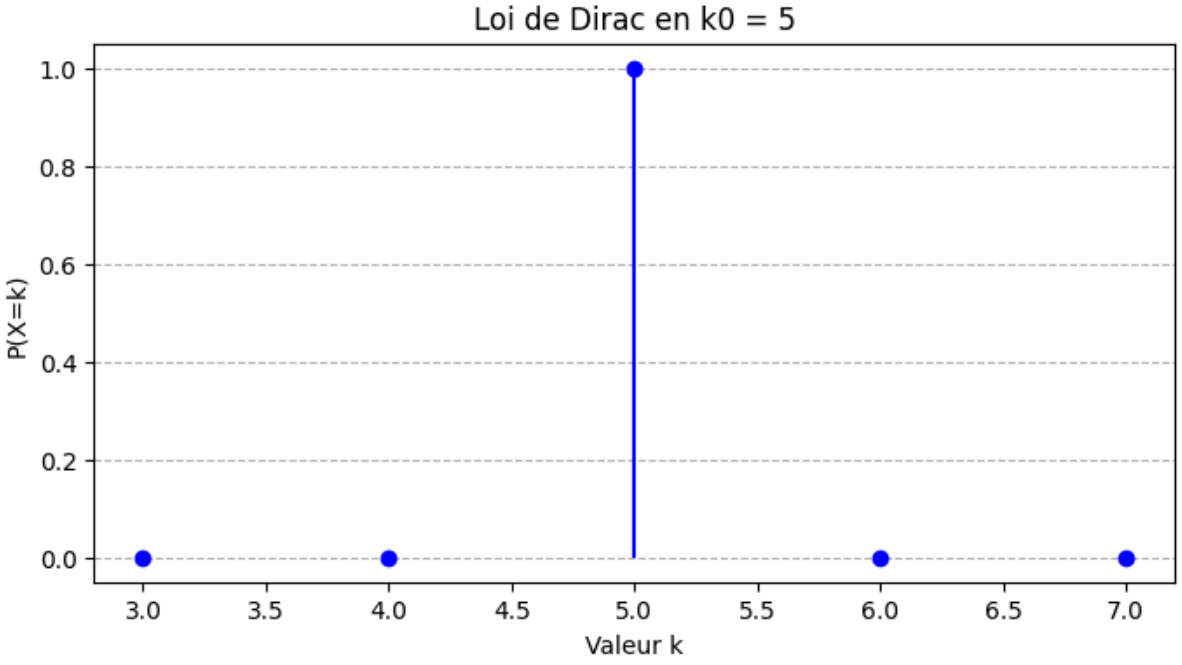
Le coefficient Q de Yule constitue un cas particulier du coefficient de Goodman-Kruskal. Il est spécifiquement conçu pour les tableaux de contingence de dimension 2×2 et permet de mesurer l'association entre deux variables dichotomiques. Comme le coefficient Γ, il varie entre -1 et +1 et peut également être interprété à partir du rapport de cotes, ce qui en fait un outil fréquent dans l'analyse des relations entre variables binaires.

PARTIE II : Description des figures obtenus tout au long des séance :

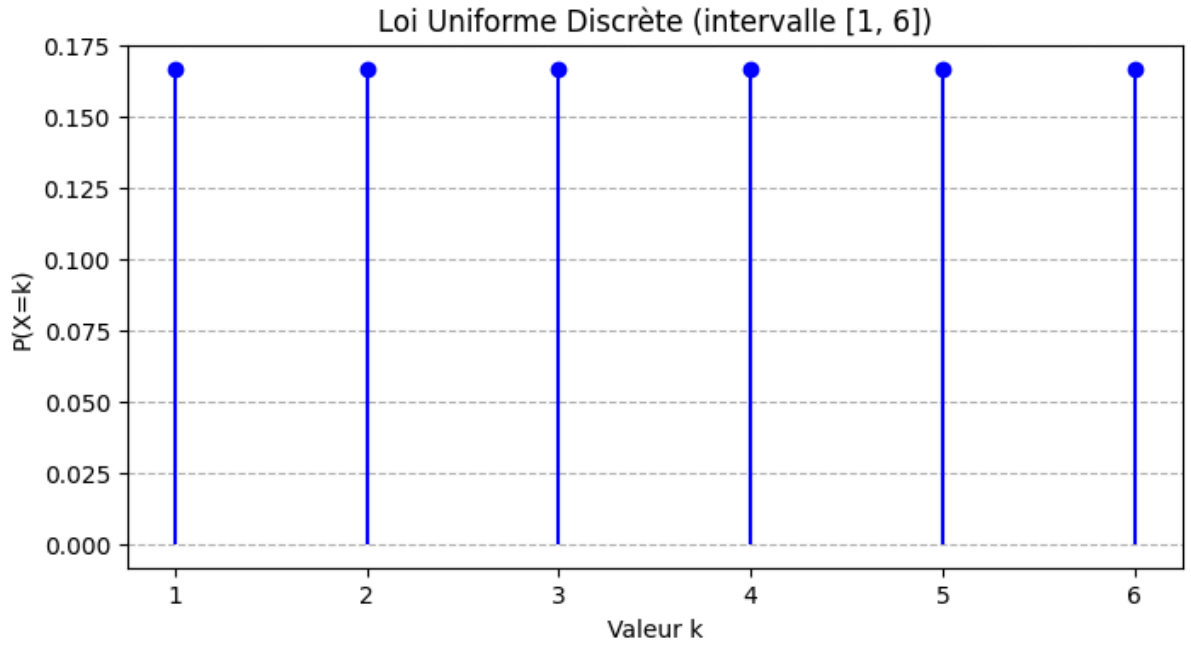
Séance 4. Les distributions statistiques

NB: Les lois discrètes sont utilisées pour modéliser :

1. les résultats des jeux de hasard;
2. les sondages d'opinion;
3. les phénomènes biologiques;
4. les processus aléatoires tels que les files d'attente, l'évolution de l'état de matériels



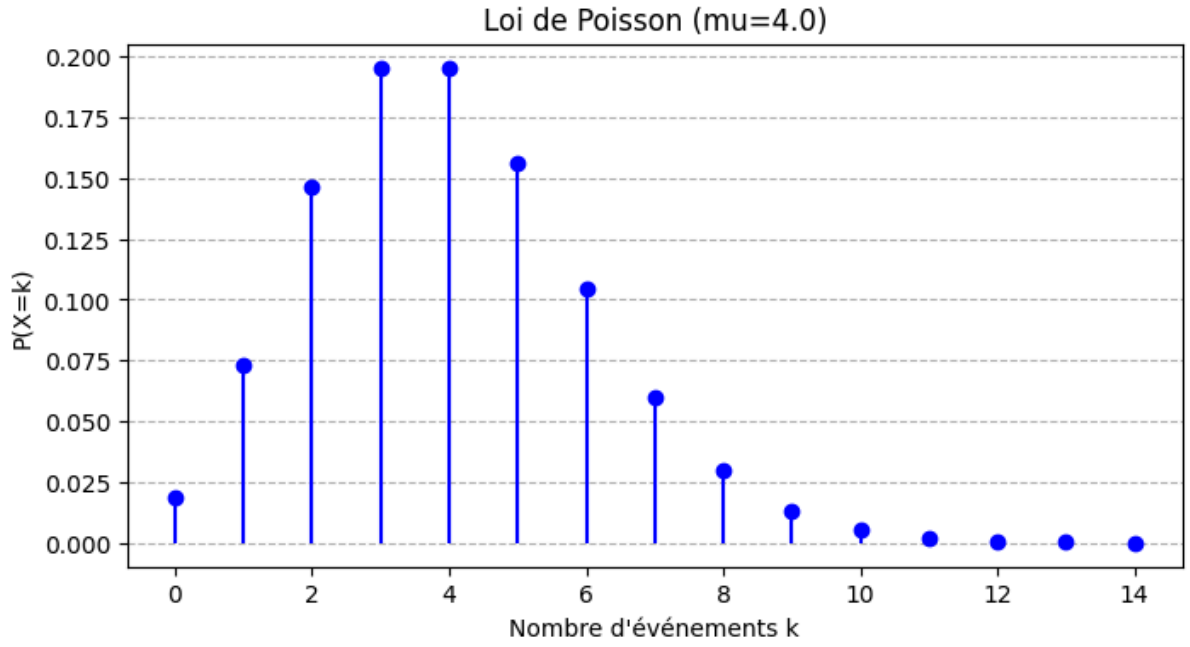
Le graphique représente une loi de Dirac centrée en k0=5=. La variable aléatoire X prend une seule valeur possible, égale à 5, avec une probabilité de 1. Toutes les autres valeurs ont une probabilité nulle. Cette loi modélise une situation déterministe, dans laquelle le résultat est connu à l'avance et ne comporte aucune incertitude.



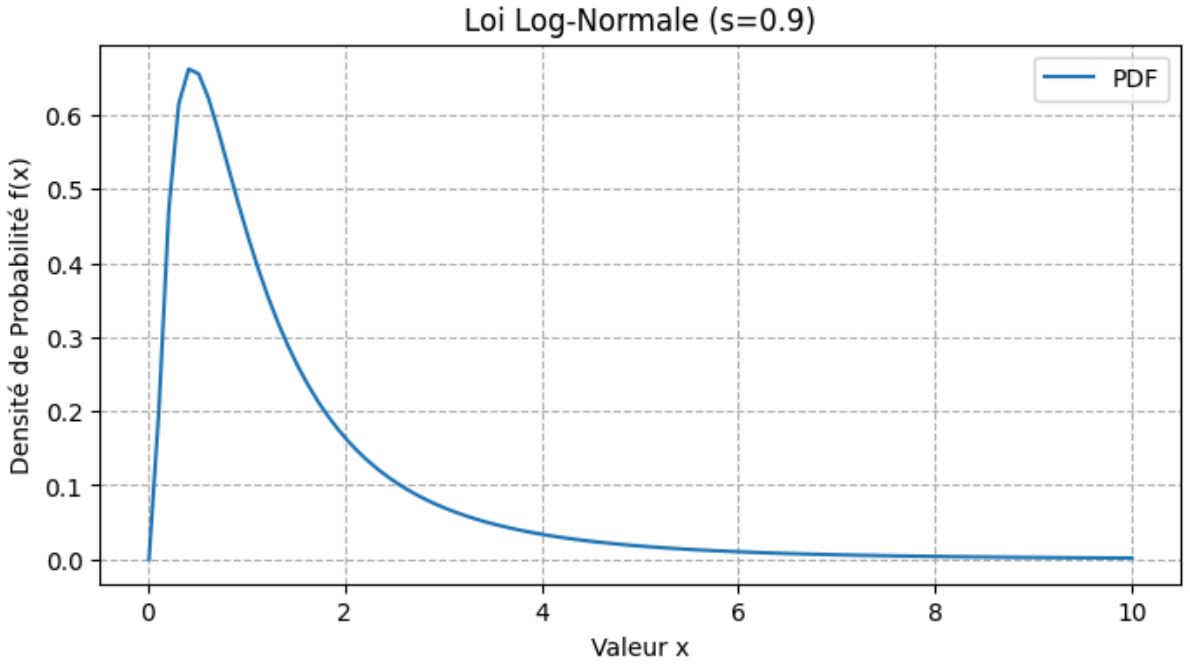
Le graphique représente une loi uniforme discrète définie sur l'intervalle [1,6][1,6][1,6]. La variable aléatoire X peut prendre les valeurs 1, 2, 3, 4, 5 ou 6. Chacune de ces valeurs possède la même probabilité d'occurrence, égale à

après calculs, résultat final ≈0,167.

Cela signifie qu'aucun résultat n'est favorisé par rapport aux autres. Les domaines d'utilisation sont : — des jeux de pile ou face; — des jeux de dé non pipé; — des jeux de cartes; — des loteries; — des sondages

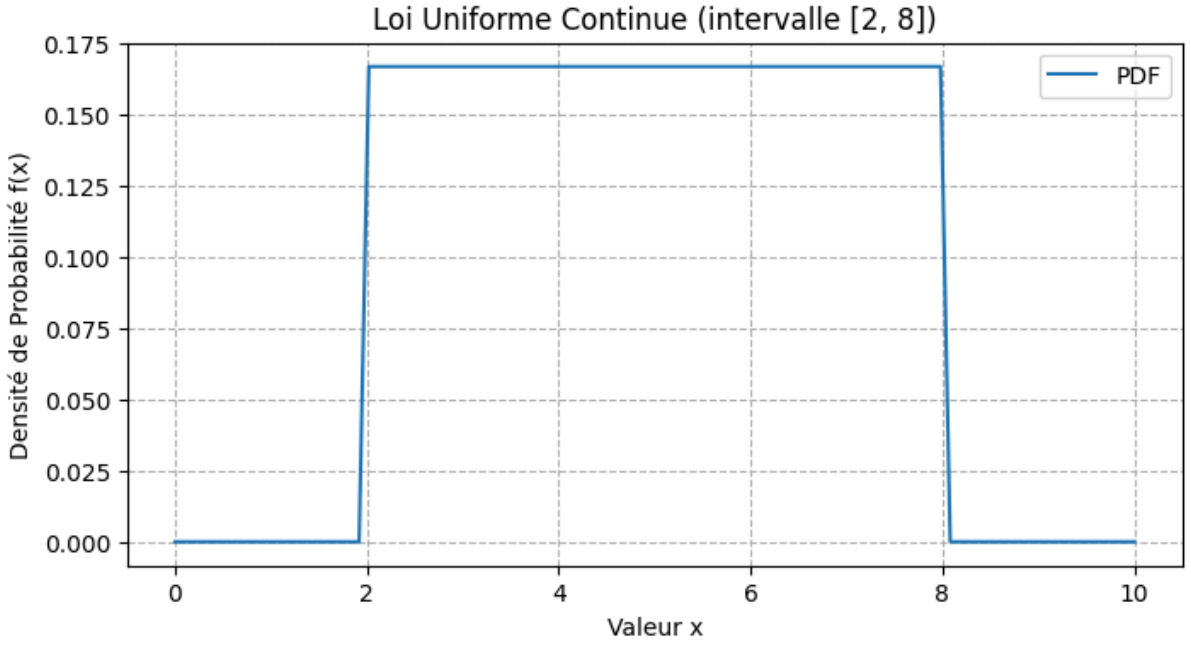


Le graphique représente une loi de Poisson de paramètre μ=4. Cette loi discrète modélise le nombre d'événements pouvant se produire dans un intervalle donné. La probabilité maximale est atteinte pour des valeurs proches de 4, ce qui correspond à la valeur moyenne. Les probabilités diminuent lorsque le nombre d'événements s'éloigne de cette valeur. Cette loi est souvent utilisée La loi de S. D. Poisson correspond à la loi des événements rares (ou la loi des petites probabilités). Le domaine d'utilisation sert pour determiner l'approximation de la loi de distribution binomiale (k petit, pourvu que = np, et que p soit petit). autrement dit elle permet d'estimer la réalisation d'événements peu probables dans une succession d'épreuves très nombreuses (au moins 50).

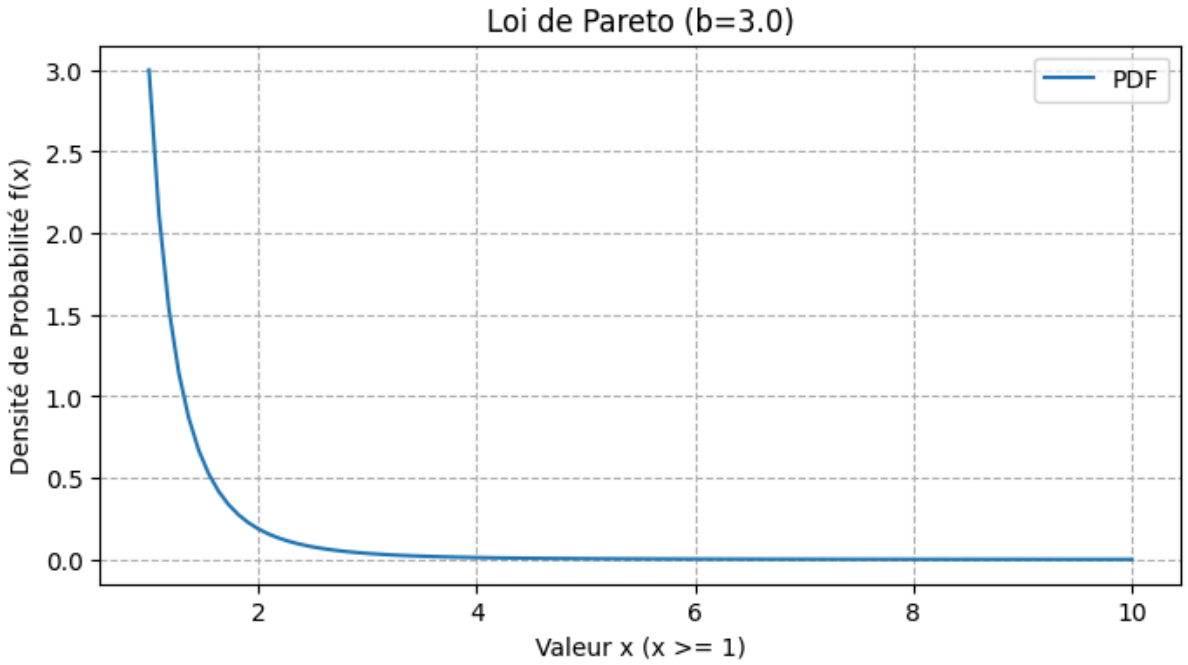


La figure ci-dessus représente une loi log-normale de paramètre s=0,9s = 0{,}9s=0,9. Cette loi continue est définie uniquement pour des valeurs positives. On observe une distribution asymétrique, avec un maximum pour une valeur faible de x, puis une décroissance progressive vers la droite. La majorité des valeurs est concentrée autour des petites valeurs, tandis que des valeurs élevées restent possibles mais peu probables.

La figure ci-dessus illustre une loi uniforme continue définie sur l'intervalle [2,8][2, 8][2,8]. La densité de probabilité est constante sur cet intervalle et nulle en dehors, ce qui signifie que toutes les valeurs comprises entre 2 et 8 ont la même probabilité d'être observées. Cette loi modélise une situation d'incertitude totale sur un intervalle donné, sans valeur privilégiée. Elle est souvent utilisée lorsque l'on sait uniquement que la variable aléatoire appartient à un intervalle précis, sans information supplémentaire sur sa distribution.

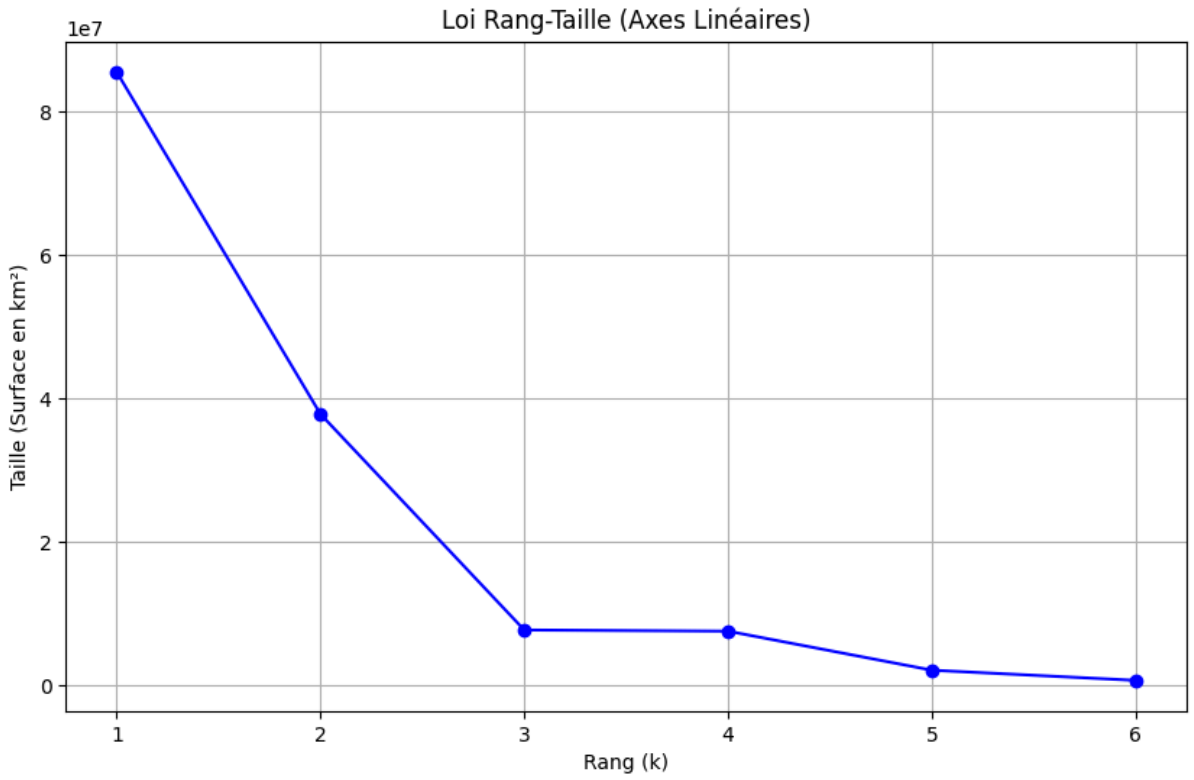


La figure ci-dessus illustre une loi uniforme continue définie sur l'intervalle [2,8][2, 8][2,8]. La densité de probabilité est constante sur cet intervalle et nulle en dehors, ce qui signifie que toutes les valeurs comprises entre 2 et 8 ont la même probabilité d'être observées. Cette loi modélise une situation d'incertitude totale sur un intervalle donné, sans valeur privilégiée. Elle est souvent utilisée lorsque l'on sait uniquement que la variable aléatoire appartient à un intervalle précis, sans information supplémentaire sur sa distribution.



Le graphique ci-dessus illustre une loi de Pareto avec un paramètre de forme b=3.0b = 3.0b=3.0. La distribution est définie pour des valeurs supérieures ou égales à 1, avec une densité maximale au seuil minimal (x=1x = 1x=1) puis une décroissance rapide. Cette forme en "loi de puissance" montre que la majorité des observations se concentre sur des valeurs faibles, tandis qu'une minorité d'événements peut atteindre des valeurs plus élevées.

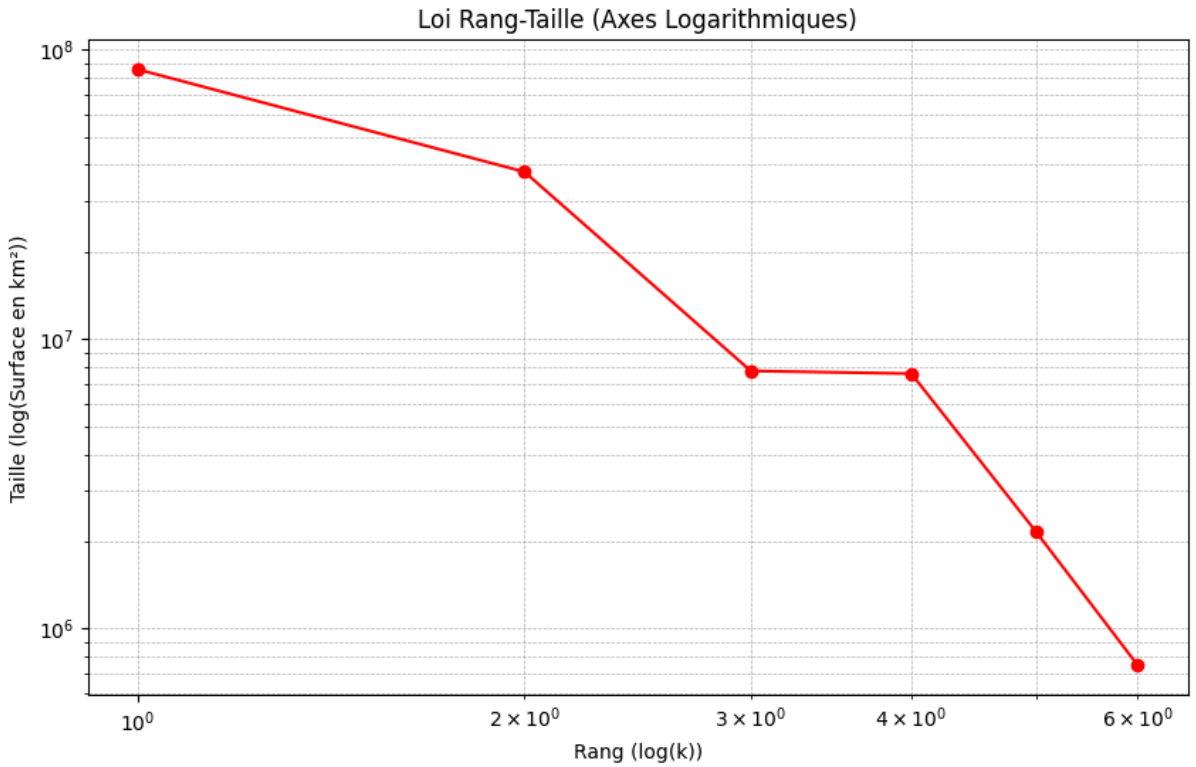
Séance 6. La statistique d'ordre des variables qualitatives



Le graphique illustre la répartition des surfaces des îles contenues dans le fichier *island-index.csv*. L'axe horizontal représente la taille des îles, allant des plus petites aux plus grandes, tandis que l'axe vertical indique le nombre d'îles correspondant à chaque intervalle de surface.

On observe que la majorité des îles possède une surface inférieure à 10 km², ce qui se traduit par un pic marqué à gauche du graphique. Au-delà de cette plage, le nombre d'îles diminue rapidement, révélant la rareté des îles de taille moyenne et grande. La partie droite du graphique, presque horizontale, correspond aux très grandes îles, peu nombreuses mais représentant une proportion significative de la surface totale.

Cette distribution illustre le principe de Pareto : une minorité d'îles de grande taille contribue de manière disproportionnée à la surface totale, tandis qu'une majorité d'îles reste de petite dimension. Le graphique constitue ainsi une représentation claire de l'inégalité naturelle dans la répartition des surfaces des îles, confirmant que celles-ci ne sont pas distribuées de manière uniforme mais selon une loi fortement inégalitaire.



L'analyse des indicateurs socio-économiques du fichier *États du monde* met en évidence une distribution conforme à la loi de Pareto. Une forte asymétrie est observable : la majorité des États se concentre dans la zone de faible densité, correspondant aux niveaux démographiques ou économiques les plus bas, tandis qu'une minorité restreinte de nations constitue la « longue traîne » du graphique. Ce schéma mathématique illustre une concentration structurelle des ressources, où un petit nombre de pays dominants représente la part prépondérante de la valeur totale de l'échantillon.

PARTIE III

Réflexion personnelle sur les sciences des données et les humanités numériques en fonction des exercices de votre parcours:

En tant qu'étudiant en master GEOINT, cette double approche, qui combine techniques de traitement des données et analyse contextuelle, m'a particulièrement interpellé. Les exercices réalisés au cours du semestre m'ont permis de comprendre à quel point la maîtrise des sciences des données est devenue incontournable pour nos futures carrières, notamment dans des domaines où l'analyse de grandes quantités d'informations est essentielle à la prise de décision.

Parallèlement, il a été extrêmement enrichissant de s'intéresser aux humanités numériques, qui offrent des outils et des méthodes pour interpréter et comprendre des phénomènes complexes dans leur contexte social, culturel ou historique. Ces deux dimensions sont indissociables : la puissance analytique des sciences des données ne peut atteindre son plein potentiel sans une réflexion critique et contextuelle, et l'analyse contextuelle bénéficie grandement de l'appui des outils numériques.

Il y a encore quelques années, il aurait pu sembler surprenant d'associer le numérique et les humanités, mais il me paraît aujourd'hui essentiel que toute formation en sciences sociales, et en particulier en master GEOINT, intègre ces compétences. Ne pas maîtriser ces outils et approches reviendrait à ignorer la profonde transformation technologique et numérique qui se déroule sous nos yeux, alors même que ces innovations redéfinissent nos méthodes d'analyse, de communication et de prise de décision.

Cette hybridité entre données et contexte, entre technologie et réflexion critique, représente selon moi un véritable atout pour appréhender les enjeux contemporains et se préparer efficacement à des carrières où la capacité à croiser informations quantitatives et compréhension qualitative est devenue centrale.

Difficultés rencontré durant le semestre

La principale difficulté que j'ai rencontrée au cours de ce semestre, et je pense que vous l'avez bien remarquée, a été le manque de matériels opérationnels. Après avoir essayé sur mon Macintosh puis sur les ordinateurs prêtés par l'université, aucun des logiciels proposés dans le cadre du cours ne fonctionnait correctement. J'ai tenté de coder directement dans le terminal et d'installer Python « dans le dur », mais le temps nécessaire à ces démarches m'a rapidement fait perdre du temps face au volume de cours. C'est ainsi que j'ai dû apprendre à coder sur Google Colab, ce qui m'a permis de suivre les exercices plus efficacement.

Heureusement, j'ai pu bénéficier d'une aide précieuse autour de moi. Céleste Mounier et Jeanne Lopez m'ont permis de rattraper mon retard et de terminer les séances destinées aux débutants. Au-delà de l'obtention d'une note convenable, il était essentiel pour moi de comprendre au moins les grandes lignes du cours afin d'acquérir une initiation à Python que je pourrai valoriser lors de mes futurs entretiens de stage et d'alternance.

Concernant les difficultés d'apprentissage, je pense que ma position était quelque peu biaisée, car ce cours m'a paru particulièrement difficile à appréhender. Le volume horaire limité, bien que encadré par vos soins, ne permet pas toujours de saisir pleinement la matière ni de déconstruire toutes les appréhensions initiales, surtout pour des étudiants issus d'une formation en lettres. Cependant, lors des derniers cours, j'ai constaté que vous étiez particulièrement à l'écoute de nos difficultés et attentif aux problèmes que nous rencontrions, ce qui a grandement facilité ma progression.