DELLIERE
Anaé
M1 GéoSuds

Rapport d'activité Analyse de Données
Niveau Débutant
Semestre 1 - 2025 / 2026
M. Forriez

SOMMAIRE

Séance 2. Les principes généraux de la statistique

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

La relation entre la géographie et les statistiques est paradoxale. Elle se méfie des définitions mathématiques de la statistique, puisqu'elle considère que cela ne fait pas partie de son champ disciplinaire. Pourtant, la géographie produit des séries de données massives qui, sans une étude statistique, ne peuvent être étudiées efficacement. En somme, la géographie sous-estime l'apport statistique. Cette relation paradoxale est aussi issue du fait que les sciences humaines s'opposent aux sciences dites dures, les mathématiques notamment, et donc les statistiques, dans l'imaginaire collectif. Malgré une réticence aux statistiques de la plupart des géographes, surtout de la part de la géographie qualitative, les statistiques apparaissent pourtant indispensables à la discipline. C'est aussi cette apport statistique qui en fait une discipline souvent à la croisée des sciences humaines et des sciences dures.

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

Le hasard est d'abord un concept philosophique. Dans la tradition déterministe, le hasard n'existe pas : tout a une cause. Mais en parallèle, la théorie du chaos considère que certains phénomènes sont aléatoires, ou alors s'expliquent par des causes qu'on ne connaît pas encore. Donc, le hasard est d'abord une limite de la connaissance humaine. Dans les modélisations géographiques, on distingue deux types de hasard. Le hasard bénin est distribué selon une loi normale, c'est à dire qu'il n'empêche pas d'établir une causalité. Le hasard sauvage est quant à lui lié à des distributions atypiques mais il est plus rare. On utilise ces lois probabilistes en géographie pour étudier des phénomènes physiques ou humains. La géographie oscille entre la nécessité et la contingence. Certains phénomènes géographiques obéissent à des lois quasi immuables comme la gravité ou le climat par exemple tandis que d'autres évènements sont contingents telles qu'une avalanche ou une migration entre autres. La contingence entend s'affranchir d'une conception déterministe de la géographie. Pour ce qui concerne la géographie humaine, on considère que le hasard existe puisqu'on ne peut pas prévoir le comportement individuel de chaque acteur sur un territoire. Toutefois, on peut dégager des tendances globales à l'aide des statistiques ou de la sociologie par exemple. Le hasard en géographie n'est donc pas une absence de causalité mais une manifestation de l'incertitude. Plus l'échelle est globale, moins le hasard existe et inversement, plus l'échelle est locale, plus les contingences persistent.

3. Quels sont les types d'information géographique ?

L'information géographique se décompose en deux grandes catégories statistiques. Les informations attributaires ou thématiques sont la première catégorie. Elles concernent les caractéristiques des territoires étudiés et relèvent de la géographie humaine ou de la géographie physique. Dans un Système d'Information Géographique (SIG), elles constituent la base attributaire, c'est-à-dire les données descriptives associées aux objets géographiques tels que le nombre d'habitants d'une commune, le taux de chômage ou encore la pluviométrie annuelle. Les informations géométriques ou morphologiques sont la deuxième catégorie et elles décrivent la forme et la structure spatiale des ensembles géographiques. Il s'agit de la

géométrie des objets, c'est-à-dire des contours, des surfaces, des distances ou des réseaux par exemple. Dans un SIG, elles correspondent aux données géométriques (les points, les lignes, les polygones) qui permettent de représenter les territoires sur une carte tels que le tracé d'un fleuve, les limites administratives ou le réseau routier.

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

D'une part, le traitement des données massives produites par la géographie ne peut se faire qu'avec une analyse des données. La quantité de données produites doit être organisée et analysée. L'organisation de l'analyse des données, avec les métadonnées et la nomenclature, permet de travailler sur des données bien documentées, des sources fiables et critiques. L'analyse de données permet de dégager des tendances globales et favorise l'approche multiscalaire puisque l'on peut étudier les phénomènes à différentes échelles. L'analyse de données permet enfin de modéliser les données en modèles spatiaux ou socio-spatiaux pour comprendre et prévoir les dynamiques territoriales.

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

La statistique descriptive décrit et résume les données observées à l'aide de calculs comme la moyenne, les écarts-types, les histogrammes. Elle permet de simplifier la réalité et de mettre de l'ordre dans les données alors que la statistique explicative permet de relier des variables entre elles qui permettent une explication d'un phénomène. La statistique explicative permet d'établir un modèle et de comprendre les relations causales ou probabilistes entre les variables. On peut utiliser pour cela la régression linéaire, l'analyse de variance ou la segmentation. Autrement dit, la statistique descriptive permet simplement de décrire les données sans chercher la causalité, cette dernière étant recherchée par la statistique explicative.

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

celles-ci ?

Il existe de nombreux types de visualisation de données en géographie. On peut d'abord penser aux histogrammes qui sont surtout utiles pour les variables continues, comme la répartition des températures par exemple. Il existe aussi des diagrammes sectoriels qui sont utilisés pour visualiser des variables qualitatives, comme une répartition par secteur d'activité. Les cartes thématiques produites par les SIG permettent de représenter spatialement des attributs. Les graphiques multidimensionnels sont utilisés pour les représentations des statistiques multivariées. Le choix de la visualisation dépend du type de variable, c'est à dire qu'elle soit qualitative ou quantitative, du nombre de variables, autrement dit que la statistique soit univariée, bivariée et multivariée, et enfin de l'objectif de l'analyse, on peut chercher soit à décrire, comparer ou expliquer.

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

Les méthodes d'analyse de données possibles dépendent de l'objectif de l'analyse. Les méthodes descriptives regroupent l'ACP, l'AFC, l'ACM, la classification hiérarchique ou encore les nuées dynamiques mais aussi la moyenne, la médiane, les écarts-types et les quartiles. Les méthodes explicatives consistent en une régression simple ou multiple, une analyse de variance, des modèles linéaires, une régression logistique et une analyse

discriminante. Les méthodes de prévision sont surtout des séries chronologiques, une extrapolation des tendances et des modèles ARIMA. Enfin, les méthodes mixtes comme l'analyse factorielle de données mixtes et l'analyse factorielle multiple constituent aussi des méthodes d'analyse de données.

8. Comment définiriez-vous : (a) population statistique ? (b) individu statistique ? (c) caractères statistiques ? (d) modalités statistiques ? Quels sont les types de caractères ?

Existe-t-il une hiérarchie entre eux ?

La population statistique est un ensemble d'unités, des individus ou des objets, sur lequel porte l'étude statistique, par exemple une population statistique peut être l'ensemble des habitants d'un territoire. C'est le niveau le plus large puisque c'est l'ensemble global étudié.

L'individu statistique est quant à lui un élément de la population étudiée qu'on appelle aussi unité statistique, comme une catégorie d'habitants d'un territoire selon l'âge par exemple. On parle dans ce cas d'unités spatiales puisqu'elles sont localisables et cartographiables. Elle est donc souvent inférieure à la population puisqu'elle ne comprend pas un ensemble.

Les caractères statistiques sont les caractéristiques observées sur les individus qui peuvent être qualitatives ou quantitatives. On peut analyser une commune selon sa population, sa superficie par exemple. Les caractères peuvent être quantitatifs (ils se mesurent numériquement) ou qualitatifs (ils décrivent une qualité et ne sont pas mesurables numériquement). Les caractères quantitatifs peuvent être continus ou discrets alors que les caractères qualitatifs peuvent être soit nominal ou soit ordinal.

Les modalités statistiques sont les valeurs concrètes prises par un caractère. Ce sont les déclinaisons des caractères.

9. Comment mesurer une amplitude et une densité ?

Une amplitude est l'écart entre deux bornes d'une classe statistique. Elle se calcule de la manière suivante : on soustrait la borne inférieure à la borne supérieure.

La densité correspond au rapport entre l'effectif de la classe et son amplitude. La formule du calcul de la densité est la suivante : Effectif de la classe divisé par l'amplitude de la classe.

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

Ces formules servent à déterminer le nombre optimal de classes lors de la construction d'un histogramme ou d'une distribution statistique. Elles permettent d'éviter un trop grand nombre de classes, ce qui serait illisible, ou un trop petit nombre de classes, ce qui conduirait à une perte d'information.

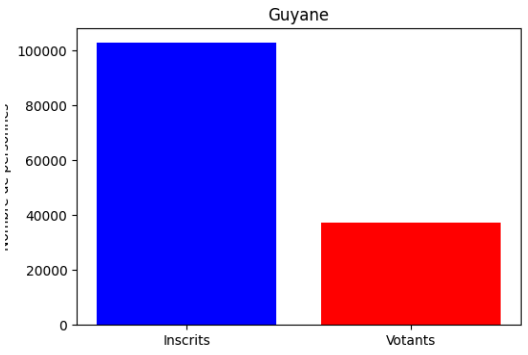
11. Comment définir un effectif ? Comment calculer une fréquence et une fréquence cumulée ? Qu'est-ce qu'une distribution statistique ?

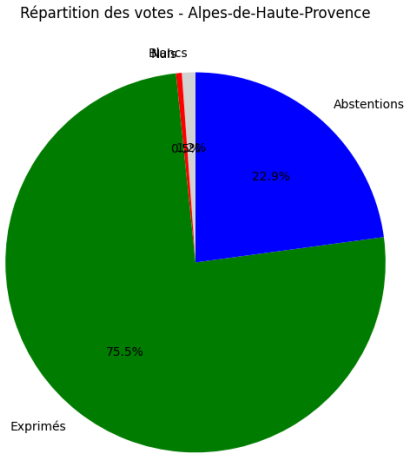
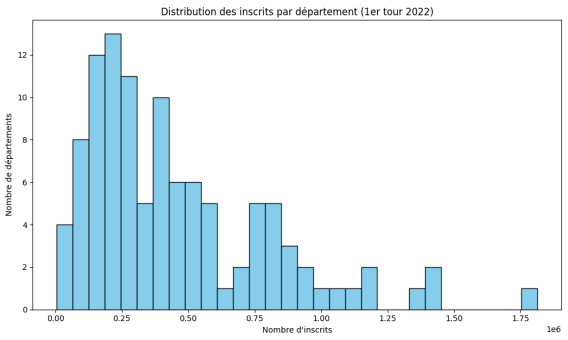
Un effectif est un nombre d'individus appartenant à une modalité ou une classe donnée. Par exemple, si 25 communes ont entre 1 000 et 2 000 habitants, l'effectif de cette classe est 25.

La fréquence est le rapport entre l'effectif d'une modalité et l'effectif total. On divise l'effectif de la modalité par l'effectif total pour obtenir la fréquence. La fréquence cumulée est quant à elle la somme des fréquences des classes successives jusqu'à une borne donnée. Elle permet de savoir quelle proportion d'individus se situe en dessous d'une valeur. On a donc seulement à additionner les fréquences successives pour calculer la fréquence cumulée.

Une distribution statistique est la répartition des effectifs ou des fréquences selon les différentes modalités d'un caractère. Elle peut être représentée par un tableau, un histogramme ou une courbe.

Code de la Séance 2

Question 6 : Dimensions du fichier	- Nombre de lignes = 107 - Nombre de colonnes = 56
Question 7 : Types de colonnes	- int (effectifs : inscrits, votants, blancs, nuls, exprimés, abstentions, voix candidats) - str (noms des départements, libellés) - float (si pourcentages présents)
Question 8 : Aperçu des données	Colonnes principales : Libellé du département, Inscrits, Votants, Blancs, Nuls, Exprimés, Abstentions, Prénom, Nom etc
Question 9 et Question 10	Nombre d'Inscrits = 48747876 Abstentions : 12824169.0 Votants : 35923707.0 Blancs : 543609.0 Nuls : 247151.0 Exprimés : 35132947.0
Question 11 (Ici, un diagramme à titre d'exemple mais le code a généré un diagramme par département)	

Question 12 (même remarque)	
Question 13	

Commentaire des résultats :

Cette séance est celle qui m'a pris le plus de temps mais elle m'a permis d'avancer plus vite par la suite car j'avais compris les principaux mécanismes du codage. J'ai passé d'abord du temps à télécharger les bibliothèques *pandas* et *matplotlib* qui permettent respectivement la gestion des tableaux statistiques et la visualisation graphique. J'ai passé beaucoup de temps ensuite pour que Python arrive à localiser et ouvrir mon fichier CSV. Le fichier CSV correspond aux statistiques électorales par département. Chaque ligne correspond à un département, un individu statistique, et chaque colonne correspond aux caractères statistiques donc les Inscrits, les Votants etc. J'ai ensuite affiché le nombre de colonnes et de lignes grâce à la fonction *len* puis *print* qui m'a permis de connaître le nombre de variables et la taille de la population statistique. J'ai ensuite créé une boucle avec *for* pour le tri des colonnes et la définition de leur nature statistique, du type de variable. *Int* et *Float* correspondent aux variables quantitatives alors que *str* correspond aux variables qualitatives. Pour la question 11, j'ai créé une boucle afin de créer des diagrammes en barre par département. J'ai choisi les couleurs bleu et rouge et la taille 6/4 pour les diagrammes (ci-dessus).On observe de manière générale que les votants sont assez proportionnels par rapport aux nombres d'inscrits dans l'ensemble des départements sauf dans les DROM où le nombre de votants par rapport au nombre d'inscrits est inférieur à la moitié. J'ai ensuite créé des diagrammes circulaires en montrant la répartition relative des votes. On peut faire le même constat que pour les

diagrammes en barres précédemment. Pour l'ensemble des départements métropolitains, il y a un taux d'environ 20 à 25% d'abstention alors que dans les DROM, l'abstention est d'environ 60 à 70%. Enfin, l'histogramme permet d'observer la forme de la distribution statistique, c'est -à -dire de voir la distribution du nombre d'inscrits par département. On remarque que la répartition des inscrits par département est fortement inégale avec une asymétrie à droite et une concentration des départements autour de valeurs moyennes et par quelques départements très peuplés. Il y a donc une grande hétérogénéité territoriale de la population électorale.

Séance 3. Les paramètres statistiques élémentaires

1. Quel caractère est le plus général : le caractère quantitatif ou le caractère qualitatif ? Justifier pourquoi.

Le caractère qualitatif est plus général parce qu'il englobe toutes sortes d'attributs qui ne sont pas nécessairement mesurables, alors que le caractère quantitatif est une sous-catégorie spécifique de caractère, restreinte aux propriétés mesurables, donc le qualitatif englobe le quantitatif.

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

Les caractères quantitatifs discrets sont comptables, ce sont des valeurs dénombrables et entières. Les caractères quantitatifs continus sont des caractères qui peuvent prendre toutes les valeurs possibles dans un intervalle donné, y compris des valeurs décimales ou fractionnaires. Ils sont infiniment divisibles. Les méthodes statistiques diffèrent selon que les caractères soient quantitatifs ou qualitatifs et les calculs ne s'interprètent pas de la même manière.

3. Paramètres de position

— Pourquoi existe-t-il plusieurs types de moyenne ?

Il existe plusieurs types de moyennes car la variable n'est pas toujours de même nature. La moyenne varie selon la nature des données.

— Pourquoi calculer une médiane ?

La médiane est la valeur qui partage une série ordonnée en deux groupes de même effectif. Elle est utile quand les données sont très dispersées ou quand certaines valeurs extrêmes faussent la moyenne.

— Quand est-il possible de calculer un mode ?

Le mode est la valeur la plus fréquente dans une série. Il est possible de le calculer uniquement si les données présentent des répétitions.

4. Paramètres de concentration

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

La médiale est la valeur qui partage la population en deux parties égales en termes de volume cumulé alors que l'Indice de Gini permet de mesurer l'inégalité ou la concentration. 0 signifie qu'il y a une égalité parfaite et 1 signifie qu'il y a une inégalité totale. Ces deux outils permettent d'évaluer la répartition d'une ressource, comme les revenus ou le patrimoine par exemple.

5. Paramètres de dispersion

— Pourquoi calculer une variance à la place de l'écart à la moyenne ?

L'écart à la moyenne montre à quel point une valeur s'éloigne du centre, mais la somme de tous ces écarts est toujours nulle car la moyenne équilibre la série. Pour éviter ce problème, on peut élever les écarts au carré avant de les additionner : c'est ce qui donne la variance. Elle mesure la dispersion globale autour de la moyenne.

- **Pourquoi la remplacer par l'écart type ?**

La variance s'exprime en unités au carré, ce qui est peu intuitif. Pour revenir à l'unité d'origine, on prend sa racine carrée, c'est l'écart type.

— Pourquoi calculer l'étendue ?

Elle indique la dispersion totale de la série, c'est-à-dire l'amplitude des données. C'est la différence entre la plus grande et la plus petite des valeurs. Elle permet de repérer la variabilité d'un ensemble de données et détecter les valeurs extrêmes.

— À quoi sert-il de créer un quantile ? Quel(s) est (sont) le(s) quantile(s) le(s) plus utilisé(s) ?

Un quantile partage une série ordonnée en groupes de même effectif. Il sert à décrire la répartition des valeurs sans être influencé par les extrêmes. Les types de quantiles les plus utilisés sont la médiane (division en deux groupes) et le quartile (division en 4 groupes).

— Pourquoi construire une boîte de dispersion ? Comment l'interpréter ?

Une boîte de dispersion est une représentation graphique qui résume visuellement la médiane, le premier quartile (Q1), le troisième quartile (Q3), les valeurs extrêmes (min, max), et parfois les valeurs aberrantes. Pour ce qui est de l'interprétation, une boîte centrée et symétrique signifie que la répartition est équilibrée. Une boîte déplacée vers le bas ou le haut signifie que les données sont asymétriques. Une moustache longue ou une valeur isolée signifie que la dispersion est forte.

6. Paramètres de forme

— Quelle différence faites-vous entre les moments centrés et les moments absolus ? Pourquoi les utiliser ?

Un moment centré mesure la dispersion ou la forme d'une série par rapport à sa moyenne. On élève les écarts à la moyenne à une puissance donnée. Les moments absolus utilisent la valeur absolue des écarts à la moyenne. Les moments centrés donnent des informations sur la dispersion et la symétrie tandis que les moments absolus sont plus robustes aux valeurs extrêmes.

— Pourquoi vérifier la symétrie d'une distribution et comment faire ?

La symétrie d'une distribution donne une idée de sa forme : Une distribution symétrique a une moyenne, une médiane et un mode identiques. Une distribution asymétrique indique un déséquilibre : soit une asymétrie à droite c'est à dire que la moyenne est supérieure à la médiane, soit une asymétrie à gauche, c'est à dire que la moyenne est inférieure à la médiane. La symétrie influence le choix des indicateurs de position (moyenne ou médiane), le choix des modèles statistiques et l'interprétation des écarts et des valeurs extrêmes.

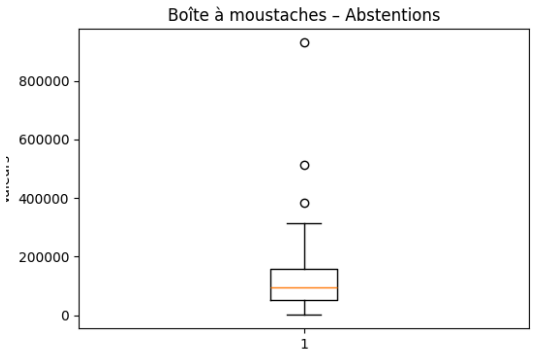
Il y a plusieurs façons de mesurer la symétrie. On peut utiliser le coefficient d'asymétrie d'abord. Dans ce cas, si = 0, la distribution est symétrique, si > 0, il y a une asymétrie à droite et si < 0, il y a une asymétrie à gauche. On peut aussi comparer la moyenne, la médiane et le mode. Si la moyenne est égale à la médiane, la distribution est symétrique. Si la moyenne est supérieure à la médiane, il y a une queue à droite et si la moyenne est inférieure à la médiane, il y a une queue à gauche. On peut aussi utiliser une représentation graphique telle que la boîte à moustaches.

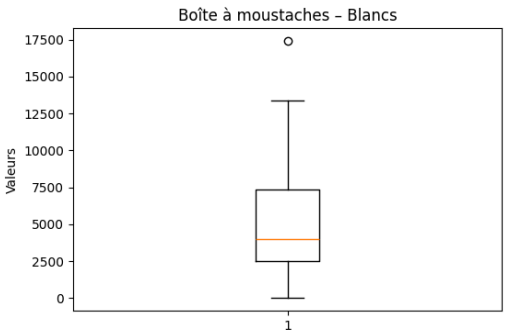
Code de la Séance 3

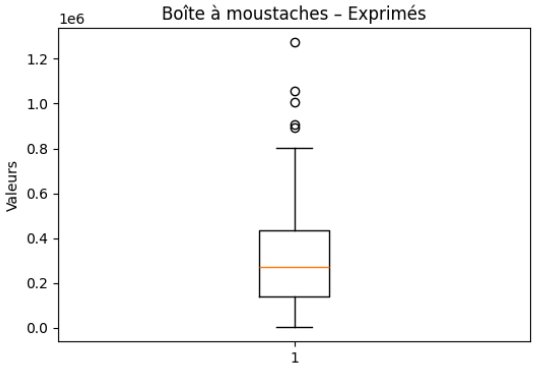
Question 5	**- Sélection des colonnes quantitatives :** ['Inscrits', 'Abstentions', 'Votants', 'Blancs', 'Nuls', 'Exprimés', 'Voix', 'Voix.1', 'Voix.2', 'Voix.3', 'Voix.4', 'Voix.5', 'Voix.6', 'Voix.7', 'Voix.8', 'Voix.9', 'Voix.10', 'Voix.11'] - **Moyennes** : [455587.63, 119852.05, 335735.58, 5080.46, 2309.82, 328345.3, 1842.0, 7499.27, 91430.45, 10293.34, 76017.08, 23226.41, 72079.63, 5761.48, 15213.58, 15691.6, 2513.12, 6777.35] - **Médianes** : [366859.0, 95369.0, 274372.0, 4001.0, 2039.0, 268568.0, 1627.0, 5968.0, 67831.0, 8944.0, 64543.0, 16885.0, 51556.0, 4881.0, 9561.0, 11918.0, 2118.0, 6152.0] - **Modes** : [5045.0, 2272.0, 2773.0, 4577.0, 17.0, 2701.0, 1203.0, 19.0, 534.0, 17010.0, 459.0, 9657.0, 501.0, 75.0, 72.0, 51.0, 3663.0, 7271.0] - **Ecart-type** : [351003.78, 117017.8, 258393.81, 3492.52, 1501.38, 253758.58, 1268.37, 6501.29, 77226.14, 7464.32, 60278.1, 20760.6, 66210.68, 4581.79, 14807.62, 13027.13, 1781.41, 4636.02] - **Ecart absolu à la moyenne** : [272240.72, 74959.07, 201517.17, 2817.95, 1131.99, 197762.2, 977.36, 4474.96, 59929.14, 5140.37, 42514.72, 15278.36, 49157.01, 3333.34, 11136.57, 9432.01, 1404.5, 3689.5] - **Étendues** : [1808861.0, 929183.0, 1297100.0, 17389.0, 8236.0,

	1272080.0, 7651.0, 45883.0, 372286.0, 48168.0, 372668.0, 108537.0, 316871.0, 22826.0, 80196.0, 69513.0, 8686.0, 20535.0]
Question 7	- **Distance interquartile** (IQR) : [401050.0, 106489.0, 301770.5, 4852.5, 1917.0, 296870.5, 1517.5, 6264.5, 101317.0, 7999.5, 63342.0, 20638.5, 60743.5, 4779.0, 14833.5, 13265.5, 2466.0, 6146.5] - **Distance interdécile** (IDR) : [793988.8, 193676.2, 602687.2, 8845.8, 3240.6, 590169.2, 3015.6, 13104.2, 177340.2, 13813.0, 130094.6, 43668.8, 159421.2, 10712.2, 38190.8, 27686.8, 4266.6, 12311.0]
Question 10	Catégorie :]0;10] 78423]10;25] 2327]25;50] 1164]50;100] 788]100;2500] 1346]2500;5000] 60]5000;10000] 40]10000;+∞[71

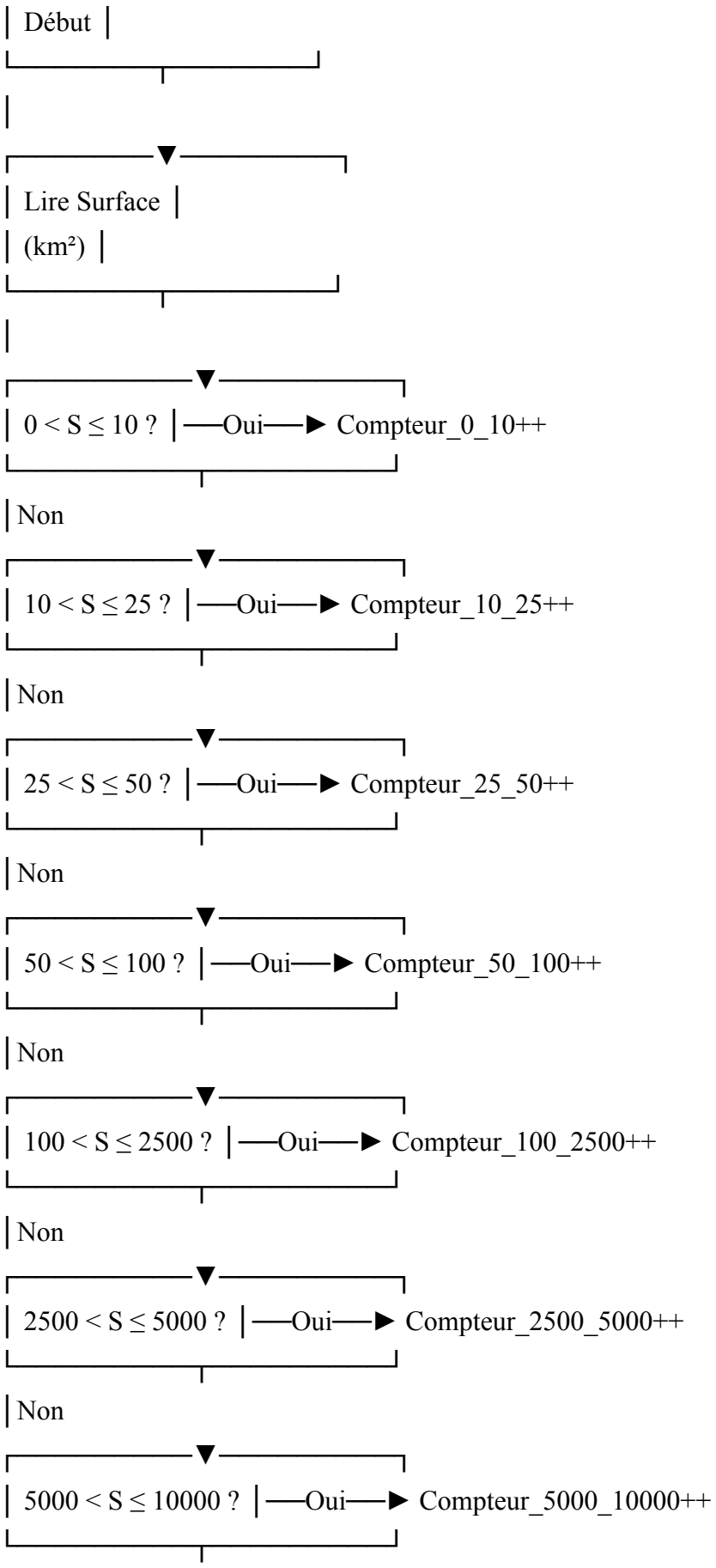
Question 8

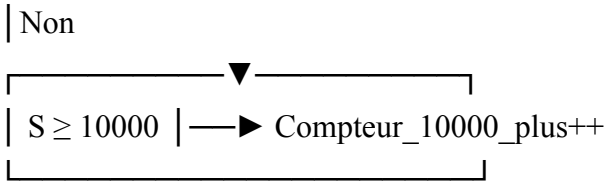






Algorithme de la question 11 :


Début
Lire Surface (km²)
0 < S ≤ 10 ?
Oui
Compteur_0_10++
Non
10 < S ≤ 25 ?
Oui
Compteur_10_25++
Non
25 < S ≤ 50 ?
Oui
Compteur_25_50++
Non
50 < S ≤ 100 ?
Oui
Compteur_50_100++
Non
100 < S ≤ 2500 ?
Oui
Compteur_100_2500++
Non
2500 < S ≤ 5000 ?
Oui
Compteur_2500_5000++
Non
5000 < S ≤ 10000 ?
Oui
Compteur_5000_10000++



Commentaire des résultats :

Cette séance m'a paru plus facile que la précédente. J'ai commencé par télécharger les bibliothèques notamment *numpy* qui permet des calculs numériques. La question 5 permet d'identifier les variables quantitatives et exclut les variables qualitatives donc. J'ai ensuite calculé les indicateurs de position, à savoir la moyenne, la médiane et le mode qui permettent de résumer la distribution. J'ai ensuite calculé les indicateurs de dispersion, à savoir l'écart-type, l'écart absolu et l'étendue. La question permet le calcul des distances interquartile et interdécile. La question 8 m'a permis à l'aide de la bibliothèque *os* et d'une boucle de créer une boîte à moustaches pour chaque variable quantitative. La question 11 permet de transformer les valeurs continues en classes statistiques à l'aide d'intervalles de classe, de calculer les effectifs par classe et d'analyser la structure de la distribution.

<u>Séance 4:</u> 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 ?

Il existe plusieurs critères essentiels permettant de choisir entre une loi discrète et une loi continue. Le tout premier critère concerne la nature du phénomène étudié, ce qui est un élément fondamental du choix. Ainsi, lorsque les valeurs prises par la variable sont dénombrables, séparées et finies, le nombre de boules rouges tirées, le nombre d'habitants classés par rang, ou encore le nombre d'essais avant un succès par exemple, la variable est discrète. Alors, la fonction de répartition prend la forme d'une fonction en escalier qui progresse par sauts successifs. Une variable continue, à l'inverse, admet une infinité de valeurs possibles et sa fonction de répartition est dérivable presque partout. Pour une variable continue, la probabilité d'observer exactement une valeur donnée est nulle, ce qui justifie l'utilisation d'une densité de probabilité f(x) telle que $Pr(x \leq X \leq x + dx) = f(x)\,dx$.

Il y a un deuxième critère qui est la forme empirique de la distribution. Cette étape est cruciale : si la distribution observée prend des valeurs isolées, la modélisation continue n'aurait pas de sens. Si, à l'inverse, l'histogramme empirique révèle une forme lisse ou une continuité, une loi continue devient pertinente.

Le troisième critère concerne les caractéristiques mathématiques du phénomène, telles que l'espérance, la variance, les coefficients d'asymétrie et d'aplatissement. Par exemple, des phénomènes asymétriques à longue queue peuvent être mieux modélisés par des lois

continues comme la loi log-normale ou la loi de Pareto, alors que des phénomènes binaires ou comptables s'accordent mieux avec des lois discrètes comme Bernoulli ou Poisson.

Un quatrième critère essentiel est le nombre de paramètres des lois utilisables. Par exemple, les lois continues comme la loi bêta ou la loi gamma possèdent plusieurs paramètres permettant d'ajuster finement la courbe à une grande variété de phénomènes. Les lois discrètes, quant à elles, disposent parfois de paramètres limités mais bien adaptés aux phénomènes comptables simples.

Enfin, un critère pratique repose sur le contexte d'utilisation ou le domaine d'application. Il y a différents domaines d'usage typiques des lois discrètes comme les jeux de hasard, les sondages d'opinion, les phénomènes biologiques, les files d'attente ou encore les états de matériels. Ces exemples montrent que les lois discrètes décrivent des phénomènes où les valeurs sont naturellement dénombrées. À l'inverse, les lois continues sont typiquement utilisées pour modéliser des phénomènes comme des mesures physiques (distance, temps, poids), des intensités, des grandeurs naturelles ou économiques. Toutes les lois continues présentées — normale, log-normale, Pareto, Gumbel, Weibull — sont destinées à des phénomènes mesurables sur un continuum.

Donc le choix entre variable discrète ou continue dépend d'abord de la structure des données (dénombrables ou mesurables), puis de la forme empirique, des paramètres observables et enfin du contexte du phénomène étudié.

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

Il existe plusieurs lois de probabilité qui ont une grande importance en géographie, notamment à travers l'étude des grandeurs spatiales, des structures urbaines, ou des phénomènes naturels.

Une première loi très utilisée en géographie est la loi de Zipf. "*Cette loi se rencontre dans les lois rang-taille confrontant, au sein d'un territoire, le nombre d'habitants d'une ville avec son rang".* Cela signifie que la distribution des tailles de villes suit souvent une relation en puissance où la ville la plus grande est environ deux fois plus grande que la deuxième, trois fois plus grande que la troisième, etc. Cette loi est donc centrale en géographie urbaine et en analyse des systèmes de villes.

La loi log-normale occupe également une place importante. Elle est pertinente pour les phénomènes de croissance multiplicative, tels que l'évolution de la taille des villes, le développement économique des régions ou encore les répartitions de longueurs naturelles. C'est une loi dérivant d'une transformation exponentielle d'une variable normale, ce qui correspond bien au caractère cumulatif, non linéaire et multiplicatif de nombreux phénomènes géographiques comme la croissance démographique ou l'évolution de surfaces naturelles ou urbaines.

De même, les lois de type Pareto sont importantes en géographie car elles modélisent les distributions scalantes, souvent utilisées pour décrire des phénomènes spatiaux comme la taille des centres urbains, la longueur des cours d'eau, la distribution des montagnes entre autres. La loi de Pareto est ainsi particulièrement adaptée aux distributions à « longue queue », fréquentes dans la nature, où un petit nombre d'entités constitue une large part de l'ensemble (par exemple, quelques grandes villes concentrent une majorité de la population).

Les lois de Fréchet et de Gumbel, liées aux valeurs extrêmes sont très utilisées en géographie physique pour l'étude d'événements rares mais majeurs : crues exceptionnelles, valeurs maximales de précipitations, températures extrêmes, amplitudes de tremblements de terre par exemple.

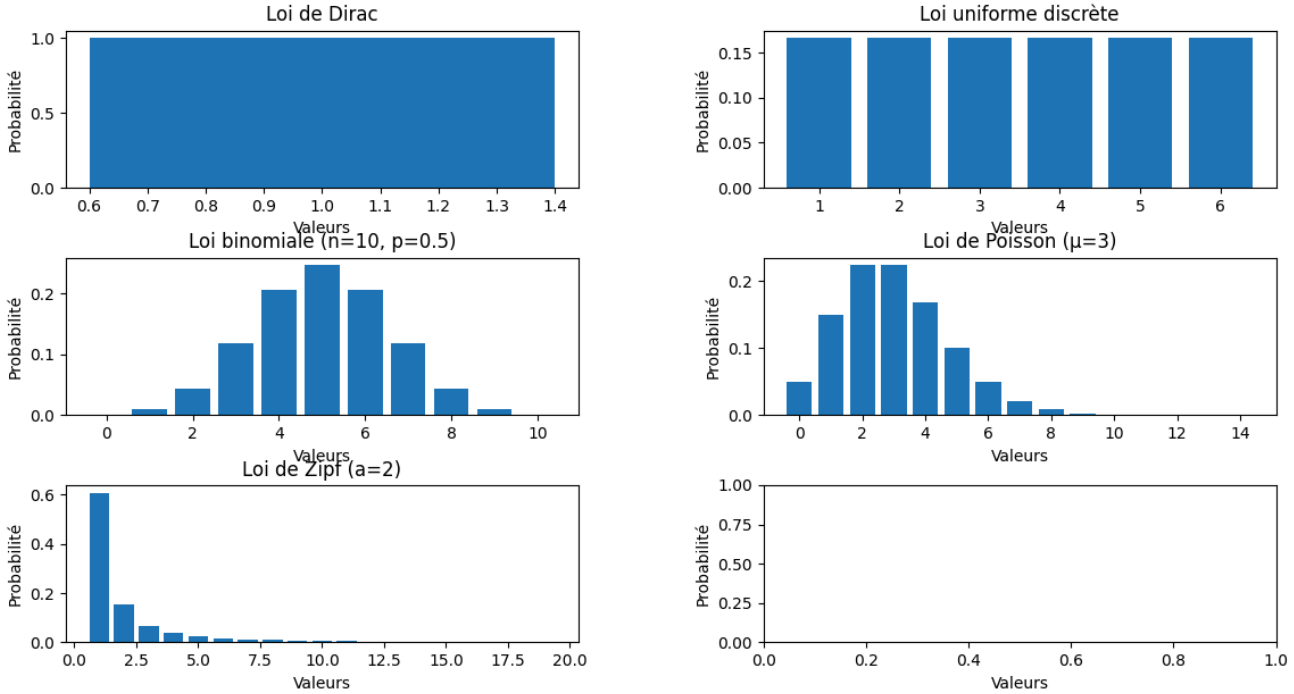
La loi de Benford, quant à elle, trouve aussi des applications géographiques. Cette loi, relative au premier chiffre significatif, permet par exemple de vérifier l'intégrité de données géographiques ou socio-économiques et d'étudier des phénomènes qui s'étendent sur plusieurs ordres de grandeur.

Les géographes utilisent surtout des lois asymétriques ou en puissance (Pareto, log-normale, Zipf) plutôt que des lois symétriques comme la normale.

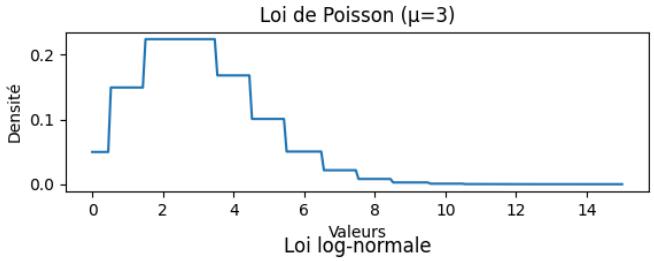
Ainsi, les lois les plus courantes en géographie sont notamment la loi de Zipf pour les systèmes urbains et les relations rang-taille, la loi log-normale pour les grandeurs issues de processus multiplicatifs, les lois de Pareto pour les distributions scalantes (surfaces, longueurs, tailles), la loi de Benford pour les grandeurs à large échelle (fleuves, pays, élections) et les lois extrêmes (Gumbel, Fréchet) pour les phénomènes naturels rares.

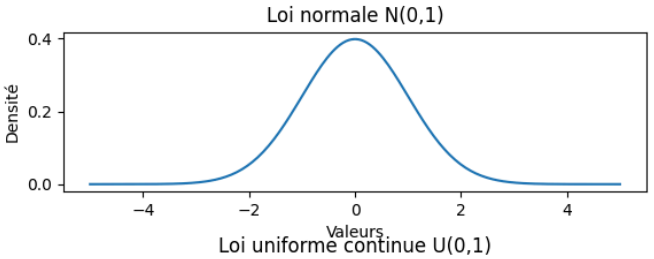
Code de la Séance 4 :

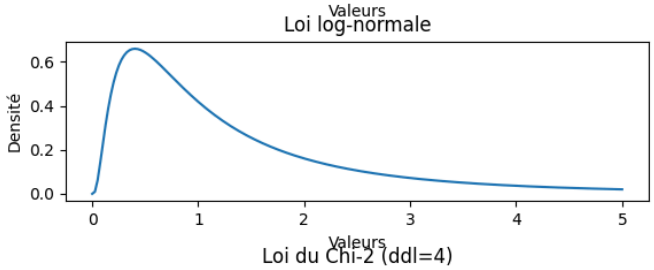
Le code permet de visualiser et simuler les différentes lois de probabilité, discrètes et continues. J'ai commencé par importer les bibliothèques notamment scipy.stats qui permet d'implanter les lois de probabilités. La première partie du code concerne la loi de Dirac qui correspond à un phénomène non aléatoire dans lequel la variable prend toujours la même valeur. La deuxième partie du code correspond à la loi uniforme c'est à dire que tous les évènements sont équiprobables. La troisième partie concerne la loi binomiale et la quatrième partie la loi de Poisson discrète et permet de modéliser le nombre d'événements rares sur un intervalle donné. La loi de Zipf ensuite est utilisée pour les lois rang-taille. J'ai ensuite créé un document général qui regroupe l'ensemble de ces lois. Le graphique de la loi Dirac représente une seule valeur, 1, avec une probabilité 1. Il modélise une certitude absolue, il n'y a pas de variabilité. Le graphique de la loi uniforme discrète affiche la probabilité égale pour chaque valeur. Le graphique de la loi binomiale met en avant une distribution symétrique centrée autour de 5. Le graphique de la loi de Poisson montre un pic autour de 3 et une décroissance ensuite. Enfin, la loi de Zipf modélise la très forte probabilité pour les petites valeurs avec une décroissance rapide.

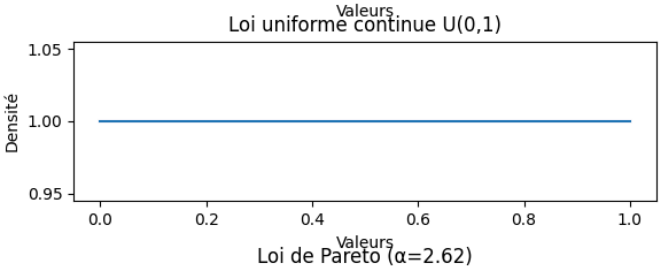


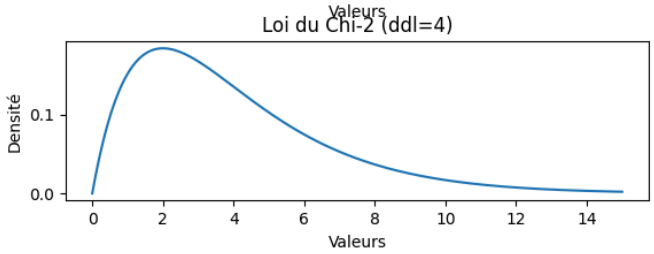
Les seconds graphiques correspondent aux lois continues dont la loi de Poisson, la loi normale, la loi log-normale, la loi uniforme continue, la loi du Chi-deux et la loi de Pareto. La loi de Poisson représente une distribution discrète en densité lissée. On remarque un pic autour de la valeur 3 puis une décroissance rapide ensuite. La loi normale est représentée par une courbe en cloche parfaitement symétrique dont la moyenne est égale à 0 et l'écart-type à 1. Le graphique de la loi log-normale est asymétrique avec une longue "traîne" à droite et les valeurs sont toujours positives. La loi uniforme continue représente une densité constante entre 0 et 1 et chaque valeur dans l'intervalle a la même probabilité. Le graphique de la loi du Chi-2 représente une distribution asymétrique avec un pic vers les faibles valeurs. Enfin, le graphique de la loi de Pareto montre une forte concentration de probabilité sur les petites valeurs.

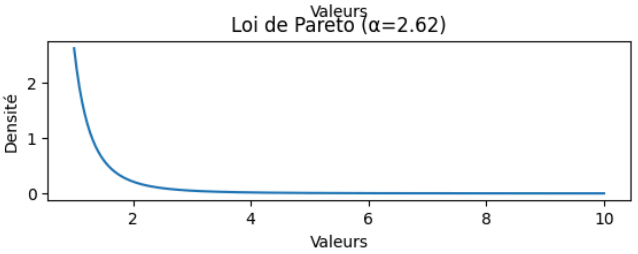












Séance 5. Les statistiques inférentielles

1. Comment définir l'échantillonnage. Pourquoi ne pas utiliser la population en entier ?

Quelles sont les méthodes d'échantillonnage ? Comment les choisir ?

L'échantillonnage est le prélèvement d'un sous-ensemble d'individus dans la population mère afin de tirer des conclusions sur l'ensemble. On ne peut pas travailler sur la population entière car la taille est souvent trop grande, cela prendrait trop de temps ou coûterait trop cher par exemple. C'est pour cela que l'échantillon doit être représentatif pour que l'on puisse dégager des tendances générales.

Il existe deux types de méthodes d'échantillonnage. La première est aléatoire, c'est à dire que l'on effectue un tirage au sort simple qui garantit une équiprobabilité. La seconde méthode est non aléatoire, c'est -à -dire qu'on définit des quotas ou un échantillonnage systématique par exemple. Le choix de l'une ou l'autre des méthodes dépend de la disponibilité ou non d'une base de sondage, du coût et du besoin de précision de l'enquête.

2. Comment définir un estimateur et une estimation ?

Un estimateur est une fonction des données, construite pour approcher la valeur d'un paramètre inconnu de la population telle que la moyenne, la variance ou la proportion. C'est une variable aléatoire qui dépend des observations. Une estimation est la valeur numérique obtenue lorsque l'on applique l'estimateur aux données observées. L'estimateur est donc théorique, l'estimation est concrète et calculée.

3. Comment distingueriez-vous l'intervalle de fluctuation et l'intervalle de confiance ?

L'intervalle de fluctuation est lié à un échantillonnage, il suppose que la proportion théorique est connue. Il décrit la variabilité attendue des fréquences observées dans un échantillon alors que l'intervalle de confiance est lié à une estimation, il est construit pour encadrer un paramètre inconnu (par exemple la moyenne ou la proportion) avec une probabilité donnée.

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

Dans la théorie de l'estimation, un estimateur est, comme nous l'avons vu, une fonction des données destinée à approcher un paramètre inconnu de la population. Un biais est la différence systématique entre la valeur attendue d'un estimateur, ce que l'on espérait, et la vraie valeur du paramètre. Un estimateur biaisé ne converge pas vers la valeur réelle, même avec un grand nombre d'observations. Le biais peut conduire à des conclusions fausses ou trompeuses sur la population mère. L'objectif est donc de choisir des estimateurs non biaisés ou de corriger le biais. La théorie de l'estimation cherche à minimiser ou supprimer ce biais pour garantir que les résultats obtenus à partir d'un échantillon soient fiables et généralisables à la population entière.

5. Comment appelle-t-on une statistique travaillant sur la population totale ? Faites le lien avec la notion de données massives ?

Lorsque l'on travaille non pas sur un échantillon mais sur tous les individus de la population mère, on parle de recensement. Le recensement est une enquête exhaustive : il consiste à observer ou interroger l'ensemble des unités de la population. Contrairement au sondage ou à l'échantillonnage, il ne repose pas sur une sélection partielle mais sur une couverture complète. C'est donc la forme la plus directe de statistique descriptive, car elle ne nécessite pas d'inférence : les paramètres calculés (moyenne, variance, proportions) sont les valeurs exactes de la population.

Dans la pratique, on considère presque toujours la population étudiée comme un échantillon, car il est très difficile, voire impossible, de travailler sur la totalité des individus. Or, avec l'essor des données massives (big data), on se rapproche de situations où l'on dispose d'informations exhaustives ou quasi exhaustives sur une population. Par exemple, des bases de données administratives, des capteurs numériques ou des plateformes en ligne peuvent collecter des millions de données couvrant presque toute la population.

6. Quels sont les enjeux autour du choix d'un estimateur ?

Le choix de l'estimateur est une étape cruciale de la statistique car il conditionne la fiabilité des conclusions que l'on tire sur la population mère. Un des principaux enjeux est la convergence. Un bon estimateur doit être convergent, sa valeur doit se rapprocher du vrai paramètre ce qui signifie que plus l'on collecte de données, plus l'estimation devient précise. L'absence de biais est un autre enjeu de taille. L'estimateur doit être non biaisé, c'est-à-dire

que son espérance mathématique doit être égale à la vraie valeur. Si l'estimateur est biaisé, il introduit une erreur systématique qui restera même avec un très grand nombre d'observations. L'efficacité est aussi un des enjeux de l'estimateur. Il est nécessaire de privilégier celui qui a la variance la plus faible, soit celui qui produit les estimations les plus stables et les moins dispersées autour du paramètre. La simplicité et la faisabilité sont aussi très importantes car le calcul de l'estimateur doit être réalisable avec les données disponibles et les méthodes doivent rester simples. Enfin, la robustesse de l'estimateur est primordiale. L'estimateur doit résister aux fluctuations d'échantillonnage et ne peut pas être trop sensible car sinon il risquerait de donner des résultats trompeurs.

7. Quelles sont les méthodes d'estimation d'un paramètre ? Comment en sélectionner une ?

L'estimation ponctuelle est une des deux approches possibles. Elle consiste à donner une valeur unique qui serait une approximation du paramètre. Elle a l'avantage d'être simple et rapide mais elle ne donne aucune information sur la précision ou l'incertitude de l'estimation. La seconde méthode consiste à estimer par intervalle, ce qui revient à encadrer le paramètre par un intervalle de confiance construit à partir des données de l'échantillon. Contrairement à la première méthode, elle fournit une marge d'erreur et une probabilité de confiance mais elle est plus complexe à calculer car elle nécessite des hypothèses sur la loi de la variable. Il existe en parallèle des méthodes plus spécifiques telles que le principe de vraisemblance, l'inférence bayésienne ou encore la méthode de Monte Carlo.

Le choix de la méthode dépend de plusieurs critères. Il dépend de la nature du paramètre (moyenne, variance, proportion entre autres) qui ont chacun leurs estimateurs privilégiés. La taille de l'échantillon est aussi un critère puisque pour un petit échantillon par exemple, les intervalles de confiance sont plus adaptés car ils définissent une marge d'erreur. Les hypothèses sur la loi entrent en compte aussi dans le choix de la méthode. Si la variable suit une loi normale, on utilise des formules spécifiques, sinon on peut recourir à des méthodes plus générales. On choisit aussi la méthode en fonction de la précision recherchée. Enfin, on privilégie les estimateurs qui sont non biaisés et dont la variance est minimale pour obtenir des résultats fiables et stables.

8. Quels sont les tests statistiques existants ? À quoi servent-ils ? Comment créer un test ?

Un test statistique est une procédure permettant de vérifier une hypothèse concernant un paramètre d'une population (par exemple la moyenne, la variance ou une proportion). Il repose sur la comparaison entre une valeur observée dans un échantillon et une valeur théorique attendue selon une loi de probabilité. Les tests basés sur l'intervalle de fluctuation sont un des types de tests statistique largement utilisés. Le seuil de fluctuation est fixé à 95% dans la plupart des cas. Le test de Student est un autre type de test qui permet de comparer une moyenne théorique à une moyenne réelle observée. Les tests liés aux lois de probabilité

permettent de déterminer le type de loi étudiée et de savoir si les données étudiées correspondent à une des ces lois.

Les tests permettent de valider ou rejeter une hypothèse (appelée hypothèse nulle). Ils sont aussi utilisés pour mesurer la signification statistique des résultats, c'est-à-dire savoir si une différence observée est due au hasard ou à une véritable caractéristique de la population. Ils servent dans des domaines variés tels que les sondages électoraux, les études médicales, ou encore les analyses économiques.

Il y a plusieurs étapes de création d'un test. Tout d'abord, il faut formuler une hypothèse nulle appelée H0. Ensuite, on choisit une statistique de test adaptée au paramètre étudié, comme la moyenne, la proportion ou la variance. On détermine dès lors la loi de probabilité qui y est associée, telle que la loi normale, la loi de Student ou du chi² notamment. Puis, on fixe un seuil de risque, souvent de 5% et, enfin on compare la valeur observée à la valeur théorique pour pouvoir valider ou non l'hypothèse nulle de départ.

9. Que pensez-vous des critiques de la statistique inférentielle ?

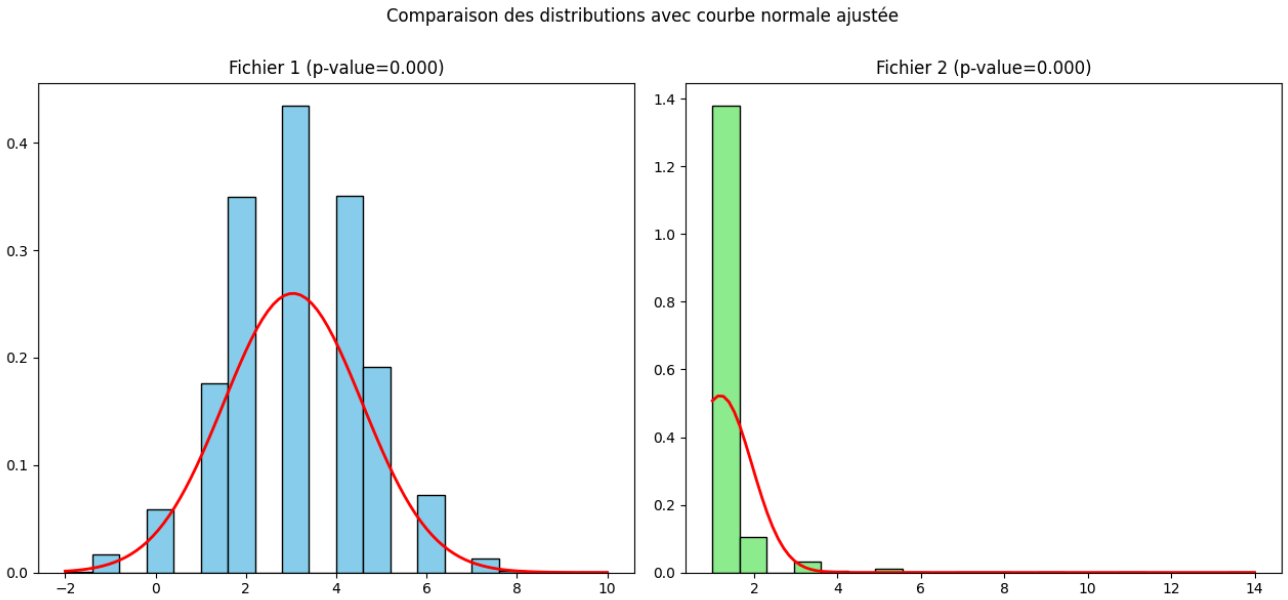
Les critiques sont principalement liées au fait qu'il existe des fluctuations et des incertitudes dans l'inférence statistique. En effet, elles dépendent des échantillons et donc la fiabilité dépend de l'échantillon choisi alors que l'on sait que même un échantillon représentatif comporte inévitablement une marge d'erreur. Une autre critique repose sur le fait que la statistique inférentielle se fonde beaucoup sur des hypothèses ce qui est peu fiable car si cette hypothèse est fausse, les résultats peuvent être biaisés. L'absence de certitude absolue rejoint aussi cette idée que les données obtenues reposent sur des probabilités et non des certitudes établies. De plus, l'interprétation peut être délicate car une mauvaise compréhension des biais ou des intervalles de confiance par exemple conduit à des conclusions erronées. Certes, un recensement complet serait le plus sûr et permettrait de pallier aux défauts de la statistique inférentielle mais comme nous l'avons vu, il est difficile à mettre en place et pas toujours nécessaire quand on cherche à dégager des tendances générales et non vérités scientifiques certaines. Si l'on interprète les conclusions de la statistique inférentielle en connaissance de cause et avec prudence, les risques de mauvaise manipulation des données ensuite seront limitées.

Code de la Séance 5

Théorie de l'échantillonnage Expliquez le lien entre l'intervalle de fluctuation et les valeurs réelles de la population mère. Que pouvez-vous en conclure par rapport aux échantillons	- Moyennes par colonne (arrondies) : Pour -> Moyenne : 391 Contre -> Moyenne : 416 Sans opinion -> Moyenne : 193 - Somme des moyennes : 1000

utilisés pour le calcul ? L'intervalle de fluctuation sert à vérifier si un échantillon est représentatif d'une population mère. On part donc des proportions réelles de la population mère (Pour 39%, Contre 42%, Sans opinion 19%). Le code compare les fréquences de l'échantillon aux fréquences de la population mère. En somme, l'intervalle de fluctuation est outil de contrôle de la qualité des échantillons. Les résultats montrent qu'il y a une parfaite adéquation entre les données et la théorie statistique. Il y a une égalité stricte entre les fréquences de l'échantillon et celles de la population mère. Pour chaque catégorie, la fréquence observée se situe exactement au centre de l'intervalle.	- Fréquences de l'échantillon : Pour -> 0.39 Contre -> 0.42 Sans opinion -> 0.19 - Comparaison des fréquences : Pour -> Échantillon : 0.39 | Population : 0.39 Contre -> Échantillon : 0.42 | Population : 0.42 Sans opinion -> Échantillon : 0.19 | Population : 0.19 - Intervalles de fluctuation (95%) Pour -> Fréquence échantillon : 0.39, Intervalle : [0.294, 0.486] Contre -> Fréquence échantillon : 0.42, Intervalle : [0.323, 0.517] Sans opinion -> Fréquence échantillon : 0.19, Intervalle : [0.113, 0.267]
Théorie de l'estimation On remarque avec cette théorie que la précision est accrue, les intervalles sont beaucoup plus resserrés qu'avec la théorie de l'échantillonnage. Les fréquences obtenues montrent que la qualité de l'échantillon est plutôt bonne car elles sont très proches des paramètres de la population mère. La marge d'erreur est plus étroite qu'avec la théorie précédente. Si on prend la ligne 2 ou 3 du fichier, les fréquences varient légèrement (par exemple, 0.38 ou 0.41 pour "Pour"). Toutefois, bien que les fréquences oscillent d'une ligne à l'autre, l'intervalle de confiance calculé sur chaque ligne devrait, dans 95% des cas, contenir la valeur réelle de la population mère (0.39, 0.42, 0.19). Les résultats sont donc systématiquement proches des valeurs de la	- Fréquences de l'échantillon : {'Pour': np.float64(0.395), 'Contre': np.float64(0.396), 'Sans opinion': np.float64(0.209)} - Intervalles de confiance (95%) : Pour -> Fréquence : 0.40, Intervalle : [0.365, 0.425] Contre -> Fréquence : 0.40, Intervalle : [0.366, 0.426] Sans opinion -> Fréquence : 0.21, Intervalle : [0.184, 0.234]

population mère, ce qui confirme que les échantillons ne sont pas biaisés.	
Théorie de la décision Selon le test, aucune de ces distributions n'est normale puisque dans les deux cas l'hypothèse nulle de départ est rejetée. Cependant, le premier graphique présente une forme en cloche qui se rapproche fortement d'une distribution normale alors que la courbe du second graphique ne correspond pas du tout à une distribution normale puisqu'elle est asymétrique et fortement biaisé à droite.	Test de Shapiro-Wilk : Fichier 1 -> W = 0.964, p-value = 0.000 Fichier 2 -> W = 0.261, p-value = 0.000 Le fichier 1 ne suit pas une loi normale (on rejette H0). Le fichier 2 ne suit pas une loi normale (on rejette H0)



Commentaire des résultats :

Mon code se base sur la statistique inférentielle, c'est -à -dire le passage des données observées à des conclusions sur la population mère grâce aux trois théories vues ci-dessus. J'ai commencé par charger le fichier dans lequel chaque colonne représente une modalité (Pour, Contre, Sans opinion) et chaque ligne correspond à un échantillon observé. J'ai ensuite calculé les moyennes grâce à la formule correspondante qui permettait d'estimer l'effectif moyen observé pour chaque colonne. Ensuite, à l'aide de la formule "fréquences échantillon" j'ai cherché à estimer les fréquences, ce qui permet de transformer les effectifs en

proportions. On compare ensuite les fréquences empiriques (l'échantillon) avec les fréquences théoriques (la population mère). Avec la formule de l'intervalle de fluctuation, on cherche à vérifier si la fréquence observée est compatible avec la théorie. Ici, z=1,96 correspond à un niveau de confiance de 95%. On calcule ensuite l'intervalle de confiance grâce à la formule ou z=1,96 équivaut toujours à un niveau de confiance de 95%. On réalise le test de Shapiro-Wilk pour les deux fichiers pour savoir s'ils suivent une loi normale ou non. J'ai ensuite décidé de réaliser un graphique pour permettre une validation visuelle du test statistique précédent.

Séance 6 : La statistique d'ordre des variables qualitatives

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

La statistique ordinale, que l'on appelle aussi statistique d'ordre est une méthode qui permet d'ordonner une série d'observations selon un rang croissant ou naturel. Cette statistique est au cœur de la géographie humaine puisque tous les classements possibles sont effectués pour comparer des objets géographiques selon un critère donné. La statistique ordinale utilise donc des variables ordinales, c'est à dire des variables qualitatives que l'on peut classer selon un ordre particulier. Elle s'oppose donc à la statistique catégorielle non ordinale, c'est à dire aux statistiques basées sur des variables qualitatives nominales, donc sans ordre. Le caractère ordinal permet de mettre en avant des rangs donc des positions relatives entre les unités spatiales. La statistique ordinale permet d'identifier les valeurs extrêmes et les relations hiérarchiques entre les territoires et les espaces, comme le classement des villes par exemple ou l'intensité d'un phénomène.

2. Quel ordre est à privilégier dans les classifications ?

L'ordre à privilégier est l'ordre croissant que l'on appelle aussi ordre naturel car il permet une lecture cohérente entre les rangs et facilite la détection des valeurs dites aberrantes.

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

La corrélation des rangs, c'est à dire la corrélation de Spearman et Kendall, cherche à répondre à la question suivante : Les classements opérés sont-ils identiques ? Elle mesure dès lors la force et la direction de la relation entre deux séries ordonnées. La corrélation de rangs évalue en quoi deux classements attribuent les mêmes ordres aux objets dans l'ensemble.

La concordance de classements consiste à comparer plusieurs classements en même temps, au-delà de deux contrairement à la corrélation de rangs.

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

Le test de Spearman permet de comparer deux classements en regardant leur écart. On prend les rangs du premier classement et on prend les rangs du second classement puis on calcule la différence entre les deux. Grâce à cela, on mesure l'ampleur des écarts entre les objets.

Le test de Kendall permet de comparer les rangs en comptant les paires concordantes et discordantes. On mesure la cohérence paire par paire. Si les deux classements sont dans le même ordre, on dit qu'il y a concordance, sinon on parle de discordance. Donc Spearman mesure à quel point les rangs diffèrent tandis que Kendall mesure à quelle fréquence les deux classements ordonnent les objets dans le même sens.

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

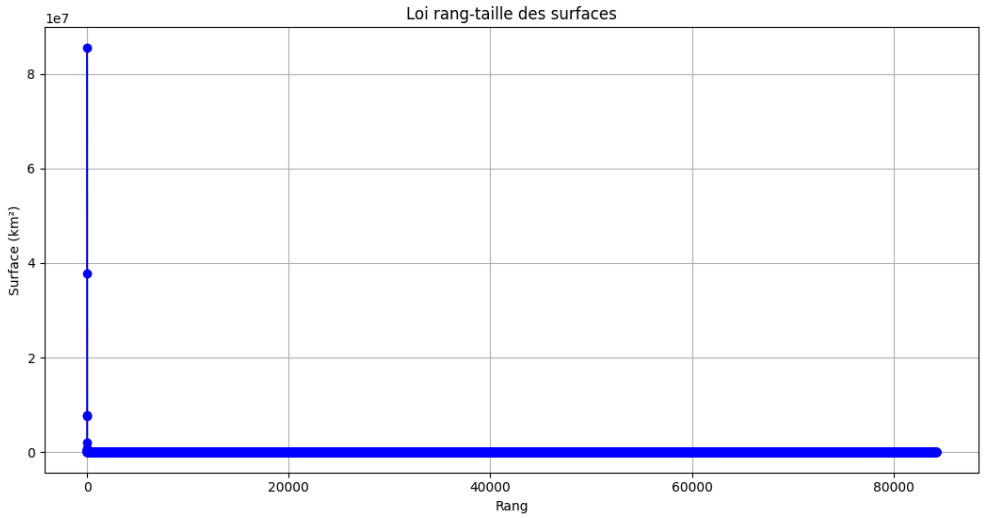
Les deux coefficients servent à mesurer une association entre deux variables mais pas dans les mêmes situations.

Le coefficient de Goodman-Kruskal sert à mesurer l'association entre deux classements ou deux variables ordinales. Il se base sur le nombre de paires concordantes et le nombre de paires discordantes.

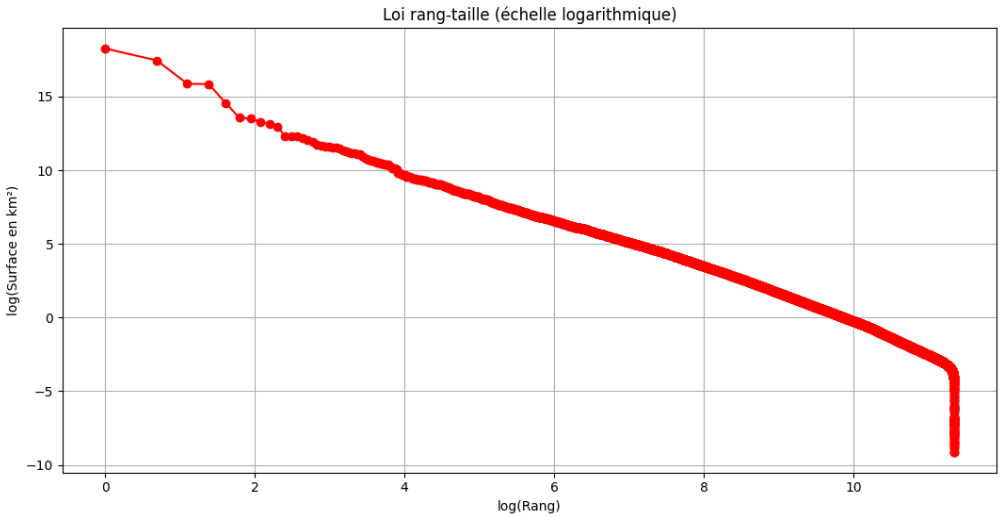
Le coefficient de Yule est un cas particulier du coefficient de Goodman-Kruskal. Il mesure l'association de variables binaires dans un tableau 2x2.

Code de la Séance 6 :

Question 5 :



Question 6 :



Question 7

Est-il possible de faire un test sur les rangs ?

Oui, il est possible de faire un test sur les rangs. Comme les rangs sont des données qualitatives ordinales, on peut utiliser des tests de corrélation non-paramétriques pour mesurer la liaison entre deux classements. Le code utilise d'ailleurs le coefficient de Spearman et le Tau de Kendall pour comparer les rangs de population et de densité.

Question 12	Comparaison (État, rang Pop2007, rang Dens2007) : ('Andorre', 0, 0) ('Antigua-et-Barbuda', 1, 1) ('Chine', 2, 10) ('Brésil', 3, 34) ('Bangladesh', 4, 2) ('Allemagne', 5, 8) ('Corée du Sud', 6, 4) ('Afrique du Sud', 7, 31) ('Colombie', 8, 30) ('Argentine', 9, 39) ('Algérie', 10, 38) ('Canada', 11, 47) ('Afghanistan', 12, 28) ('Arabie Saoudite', 13, 42) ('Corée du Nord', 14, 9) ('Australie', 15, 45) ("Côte d'Ivoire", 16, 26) ('Cameroun', 17, 32) ('Chili', 18, 35) ('Angola', 19, 40) ('Cambodge', 20, 20) ('Burkina Faso', 21, 27) ('Cuba', 22, 15) ('Belgique', 23, 5) ('Biélorussie', 24, 29) ('Bolivie', 25, 44)

	('Bénin', 26, 21) ('Azerbaïdjan', 27, 18) [...] ('Roumanie', 161, 169) ('Sri Lanka', 162, 157) ('Sénégal', 163, 180) ('Tuvalu', 164, 171) ('Vanuatu', 165, 172) ('Viêt-nam', 166, 173) ('Ukraine', 167, 177) ('Venezuela', 168, 185) ('Yémen', 169, 182) ('Syrie', 170, 161) ('Zimbabwe', 171, 184) ('Zambie', 172, 189) ('Tunisie', 173, 170) ('Tchad', 174, 193) ('Serbie', 175, 175) ('Rwanda', 176, 106) ('Suède', 177, 186) ('Somalie', 178, 190) ('Suisse', 179, 160) ('Salvador', 180, 156) ('Tadjikistan', 181, 181) ('Togo', 182, 174) ('Sierra Leone', 183, 176) ('Slovaquie', 184, 166) ('Turkménistan', 185, 191) ('Singapour', 186, 153) ('Uruguay', 187, 187) ('Slovénie', 188, 167) ('Trinité-et-Tobago', 189, 163) ('Eswatini (ex-Swaziland)', 190, 179) ('Timor-Oriental', 191, 178) ('Suriname', 192, 194) ('São-Tomé-et-Principe', 193, 159) ('Seychelles', 194, 164)
Question 14	Corrélation de Spearman : 0.9448105649872954 p-value Spearman : 1.7731507631631262e-95 Ce résultat signifie que nous sommes proches du maximum théorique de 1 ce qui veut dire qu'il y a une quasi-parité entre le rang d'un pays en termes de population et son rang en termes de densité : plus un pays est peuplé par rapport aux autres, plus il tend à être dense. Concordance de Kendall : 0.7928628072957972 p-value Kendall : 7.136382867294459e-61 Ce test analyse la probabilité que les paires de pays soient classées dans le même ordre. La valeur obtenue est relativement élevée ce qui confirme les conclusions du test précédent, les paires ont une très grande probabilité de concordance.

	En somme, il existe une corrélation quasi-systématique entre le poids démographique d'un Etat et sa densité.

Commentaire du code :

Après avoir chargé le fichier, j'ai commencé par extraire la colonnes des surfaces et la convertir directement en *float* pour la transformer en valeur quantitative. J'ai ensuite, avec la même méthode *"float"* ajouté les nouvelles valeurs pour les continents pour pouvoir ensuite étudier une loi rang-taille. Avec la fonction "*ordre décroissant*", j'ai trié les valeurs des surfaces sous forme de liste, ce qui permet ensuite une analyse par rangs.

Ensuite, il était nécessaire de créer une fonction rang "*range*" pour créer des rangs statistiques pour la loi rang-taille. Le plus grand objet a le rang 1 et le plus petit a le rang N. On crée ensuite le graphique permettant de visualiser la loi rang-taille sachant que chaque point correspond à une entité géographique. Mais, ce graphique étant illisible car les points étaient beaucoup trop serrés entre eux, on le convertit en logarithme grâce à la fonction "*conversionLog*". Ensuite, pour le second fichier, on convertit les valeurs numériques en *float* et les noms des Etats en *str* avant de classer les données par ordre décroissant. Après ce tri décroissant, on classe les valeurs par population et densité puis à l'aide de la fonction comparaison, on crée un rang pour la population et un rang pour la densité avant de réaliser les deux tests de corrélation par rangs.

Conclusion personnelle :

Si au début j'avais quelques réticences mentales face au codage et aux statistiques qui me paraissaient hors de ma portée, grâce à l'entraide entre élèves, j'ai progressivement pris confiance en moi. L'aide de Zara Huston, particulièrement, a été très précieuse pour le codage et la mise en place de GitHub. Nous avons organisé des réunions collectives en dehors des cours dans des salles de la bibliothèque afin que l'on puisse tous et toutes s'entraider. J'ai aussi utilisé l'intelligence artificielle quand je rencontrais des difficultés pour coder, lorsqu'une erreur s'affichait et que je n'arrivais pas à la comprendre. Au fur et à mesure des séances, j'ai compris que les statistiques faisaient en réalité partie intégrante de la géographie et qu'il était donc nécessaire d'en maîtriser les compétences de base. J'ai aussi pris conscience que Python pouvait être un excellent outil pour gagner du temps.

L'ensemble des exercices réalisés m'ont dès lors permis de mieux comprendre la portée et les limites des outils statistiques et informatiques appliqués aux données. L'ensemble du parcours m'a montré que l'analyse de données ne se réduit pas à des calculs ou à des représentations graphiques, mais qu'elle constitue une véritable démarche scientifique, où la rigueur méthodologique (surtout dans l'écriture du code par exemple) et l'interprétation critique sont essentielles. Les notions d'échantillonnage, d'estimation, de tests statistiques ou

encore de visualisation m'ont permis de saisir l'importance de la fiabilité et de la représentativité des résultats. En parallèle, les humanités numériques m'ont ouvert une perspective complémentaire : elles montrent que les données ne sont jamais neutres. Elles sont produites, collectées et interprétées dans des contextes sociaux, culturels et historiques. Les exercices de mon parcours m'ont ainsi sensibilisé à la nécessité de croiser les approches quantitatives avec une réflexion qualitative et critique. L'usage des statistiques dans les sciences humaines ne doit pas seulement viser l'efficacité technique, mais aussi la compréhension des enjeux éthiques, politiques et culturels liés aux données. Cette double dimension m'amène à considérer l'analyse de données comme un champ interdisciplinaire. Les sciences des données et les humanités numériques se complètent : les premières apportent des outils de mesure et de modélisation, les secondes rappellent que les données sont des constructions sociales et qu'elles doivent être interrogées dans leur contexte.

Par ailleurs, le format du cours inversé était une bonne idée pour favoriser l'entraide entre élèves mais sans avoir aucune base ni en statistique ni en codage pour beaucoup d'entre nous, il est vrai que la marche était haute et quelques cours au format classique auraient peut-être permis de poser les bases ensemble afin d'être plus à l'aise par la suite.