Portier Léa

Rapport d'activité Léa Portier

Python – Niveau débutants

Séance 2 : Principes généraux de la statistique

1. Questions de cours

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

La géographie entretient historiquement une relation complexe avec la statistique. Discipline située à l'interface entre les sciences humaines et sociales et les sciences de la nature, elle a longtemps manifesté une certaine méfiance à l'égard de la formalisation mathématique, considérée comme extérieure à son champ disciplinaire. Pourtant, la géographie produit et mobilise une quantité massive de données, notamment dans le cadre de la géographie humaine (données démographiques, économiques, sociales) et de la géographie physique (données climatiques, hydrologiques, géomorphologiques). Les statistiques apparaissent ainsi comme un outil indispensable permettant d'organiser, de résumer et d'interpréter cette information géographique. Elles dégagent des tendances globales, comparent des territoires et proposent des modèles spatiaux. L'analyse spatiale, en particulier, illustre cette volonté de faire de la géographie une science capable de produire des connaissances généralisables. En revanche, la statistique ne peut remplacer le raisonnement géographique : elle le complète et le structure.

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

Le hasard en géographie relève d'abord d'une question philosophique. Deux positions s'opposent traditionnellement : le déterminisme, pour lequel tout phénomène a une cause explicable, et une vision contingente admettant que certains événements peuvent se produire ou non sans qu'il soit possible de les prévoir précisément. Dans le cadre de la statistique, le hasard n'interdit pas l'analyse scientifique. Il est impossible de prévoir le comportement individuel de chaque acteur ou la réalisation précise de chaque événement, mais il est possible de dégager une certitude globale, c'est-à-dire une tendance statistique. On distingue ainsi un hasard bénin, compatible avec des lois de probabilité stables (notamment la loi normale), et un hasard sauvage, plus irrégulier. La géographie mobilise cette conception du hasard dans un raisonnement multiscalaire : à l'échelle globale, des régularités apparaissent ; à des échelles plus fines, la diversité des situations locales persiste. Le hasard n'empêche donc pas la géographie d'être une science, dès lors qu'elle s'appuie sur des modèles statistiques adaptés.

3) Quels sont les types d'information géographique ?

L'information géographique se décompose en deux grandes catégories statistiques. La première concerne les informations attributaires, qui décrivent les caractéristiques des territoires. Elles relèvent aussi bien de la géographie humaine (population, niveau de vie, emploi, structures sociales) que de la géographie physique (température, précipitations, débits fluviaux). Dans un système d'information géographique (SIG), ces données constituent la base attributaire. La seconde

catégorie concerne les informations géométriques, qui décrivent la forme, la taille, la structure et la morphologie des objets géographiques (surfaces, réseaux, limites). Ces informations relèvent davantage de l'analyse spatiale et géométrique.

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 transformer des données brutes en information interprétable. Cela suppose une production ou une collecte rigoureuse des données, appuyée sur des nomenclatures clairement définies et des méta-données qui permettent un examen critique des sources. L'analyse de données décrit ensuite la structure interne des jeux de données, détecte des contrastes spatiaux et des relations entre variables. Elle contribue également à mesurer la fiabilité des résultats et à confronter les observations aux connaissances théoriques existantes. Enfin, l'analyse statistique est indispensable pour construire des modèles spatiaux et socio-spatiaux et pour comparer des territoires à différentes échelles.

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 dans une population ou un échantillon. Elle vise à produire une image simplifiée de la réalité à l'aide de paramètres numériques (moyenne, médiane, écart type, fréquences) et de représentations graphiques. Elle résume les distributions et prépare les comparaisons et les analyses ultérieures. La statistique explicative, quant à elle, cherche à mettre en relation une variable à expliquer Y avec une ou plusieurs variables explicatives $X_1,\ldots,X_k$. La statistique descriptive est donc une étape préalable indispensable à la statistique explicative.

6) Quels sont les types de visualisation de données en géographie ? Comment les choisir ?

Les visualisations de données dépendent du type de variables étudiées et de l'objectif de l'analyse. Pour les variables qualitatives, on utilise principalement des diagrammes en secteurs ou en barres, qui représentent les fréquences des modalités. Pour les variables quantitatives discrètes, le diagramme en bâtons est le plus adapté. Pour les variables quantitatives continues, on privilégie l'histogramme, le polygone de fréquences, la courbe cumulative ou les boîtes à moustaches. Lorsque plusieurs variables sont analysées simultanément, des méthodes graphiques multidimensionnelles sont utilisées, telles que l'analyse en composantes principales (ACP) pour les variables quantitatives ou l'analyse factorielle des correspondances (AFC, ACM) pour les variables qualitatives. Le choix d'une visualisation repose donc sur la nature des données et sur la question scientifique posée.

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

Les méthodes d'analyse de données se répartissent en trois grandes catégories. Les méthodes descriptives visent à résumer et visualiser les données (ACP, AFC, ACM, classifications). Les méthodes explicatives cherchent à établir des relations entre variables (régression, analyse de la variance, régression logistique, analyse discriminante). Enfin, les méthodes de prévision portent sur l'analyse des séries chronologiques et reposent sur des modèles reliant le présent au passé.

8) Définitions fondamentales

La population statistique est l'ensemble des éléments étudiés. L'individu statistique est un élément de cette population ; en géographie, il correspond souvent à une unité spatiale localisable. Un caractère statistique est une propriété observée sur chaque individu. Les modalités statistiques sont les valeurs prises par ce caractère, elles sont exhaustives et incompatibles. On distingue quatre

types de caractères statistiques : qualitatifs nominaux, qualitatifs ordinaux, quantitatifs discrets et quantitatifs continus. Il existe une hiérarchie implicite entre eux, les caractères quantitatifs offrant des possibilités d'analyse statistique plus étendues.

9) Comment mesurer une amplitude et une densité ?

L'amplitude de classe correspond à la longueur d'un intervalle statistique et se calcule par :

$$A = b - a$$

où a est la borne inférieure et b la borne supérieure de la classe.

La densité de classe permet de tenir compte de l'amplitude de la classe et se calcule par :

$$d = n_i \,/\, b - a$$

où n_i est l'effectif de la classe considérée.

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

Les formules de Sturges et de Yule servent à déterminer un nombre de classes statistiquement pertinent lors de la discrétisation d'une variable quantitative continue. Elles permettent d'éviter un découpage trop fin ou trop grossier, tous deux sources de perte d'information.

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

L'effectif n_i : nombre correspond au nombre d'individus appartenant à une modalité ou à une classe donnée. La fréquence associée est le rapport entre cet effectif et l'effectif total n. La fréquence cumulée est obtenue en additionnant les fréquences des modalités inférieures ou égales à une valeur donnée. Une distribution statistique correspond à la répartition des effectifs ou des fréquences selon les modalités d'un caractère. Elle constitue la base empirique permettant d'identifier une loi de probabilité et d'interpréter statistiquement les données.

2. Code

6) Réponse :
- Nombre de lignes : 107.
- Nombre de colonnes : 56.

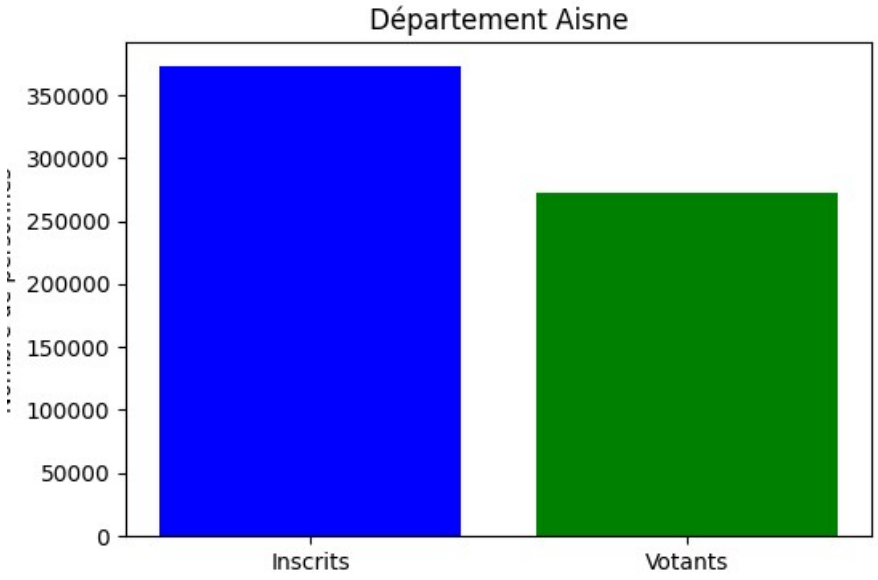
9) 438 109 personnes sont inscrites.

10) Voici les résultats.

```
[1 rows x 56 columns]
[('Inscrits', 48747876), ('Abstentions', 12824169.0), ('Votants', 35923707.0), ('Blancs', 543609.0), ('Nuls', 247151.0), ('Exprimés', 35132947.0), ('Voix', 197094.0), ('Voix.1', 802422.0), ('Voix.2', 9783058.0), ('Voix.3', 1101387.0), ('Voix.4', 8133828.0), ('Voix.5', 2485226.0), ('Voix.6', 7712520.0), ('Voix.7', 616478.0), ('Voix.8', 1627853.0), ('Voix.9', 1679001.0), ('Voix.10', 268904.0), ('Voix.11', 725176.0)]
```

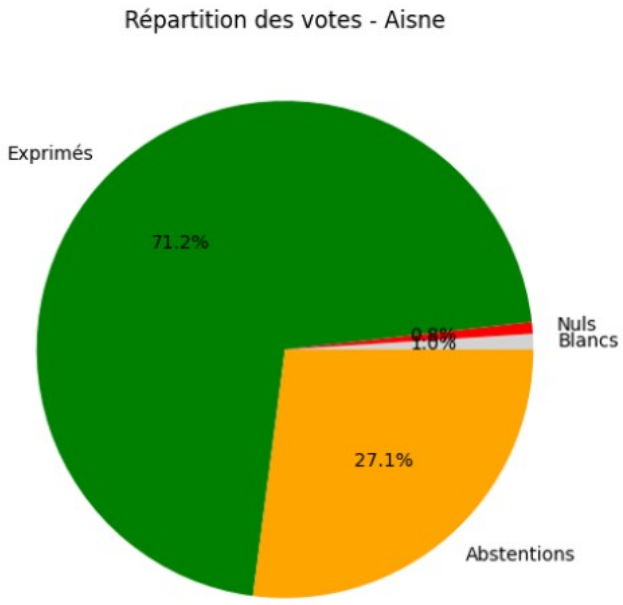
11)

Diagramme en barre obtenu



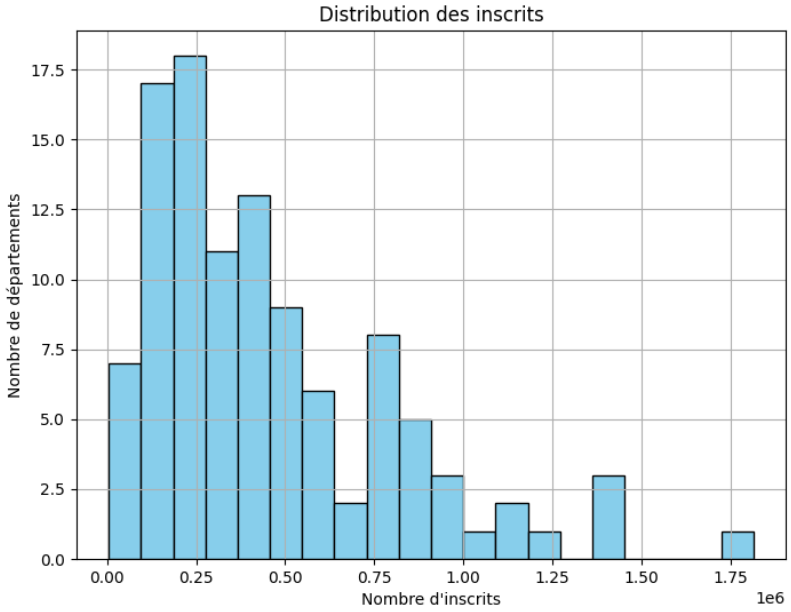
12)

Diagramme circulaire obtenu



13)

Histogramme de distribution des inscrits



3. Commentaire de code

Ce code constitue un exemple d'initiation à l'analyse de données électorales en géographie, en combinant exploration statistique et visualisation graphique. Il montre d'abord l'importance d'organiser le travail de manière rigoureuse, en créant automatiquement des dossiers pour stocker les résultats, ce qui favorise la lisibilité et la reproductibilité de l'analyse. La lecture des données sous forme de DataFrame permet ensuite d'adopter une approche structurée, indispensable pour traiter un jeu de données complexe et multidimensionnel comme des résultats électoraux. Les premières opérations réalisées (affichage, dimensions, types de variables) traduisent une démarche méthodologique pertinente, car elles permettent de comprendre la nature des données avant toute interprétation. Le calcul de sommes globales donne une vision agrégée à l'échelle nationale et sert de point de contrôle de la cohérence des informations. Les visualisations produites ne se limitent pas à une illustration graphique : elles participent réellement à l'analyse en mettant en évidence les écarts entre inscrits et votants, la structure de la participation électorale et le poids de l'abstention selon les départements. Enfin, l'histogramme sur le nombre d'inscrits apporte une lecture synthétique des inégalités démographiques territoriales, montrant que la statistique et la visualisation sont ici utilisées comme de véritables outils d'interprétation géographique et non comme de simples sorties techniques.

Séance 3 : Paramètres statistiques élémentaires

1. Questions de cours

1) Quel caractère est le plus général : le caractère quantitatif ou le caractère qualitatif ? Pourquoi ?

Le caractère qualitatif est le plus général, car toute information statistique peut toujours être décrite sous une forme qualitative, alors que l'inverse n'est pas toujours possible. Un caractère qualitatif permet de classer les individus selon des catégories, même lorsque les données numériques sont absentes ou difficiles à obtenir. En revanche, le caractère quantitatif suppose une mesure numérique précise, ce qui nécessite des conditions d'observation plus strictes. De plus, un caractère quantitatif peut toujours être transformé en caractère qualitatif par regroupement en classes, alors qu'un caractère qualitatif ne peut pas toujours être quantifié sans perte d'information ou hypothèse supplémentaire. C'est pourquoi le caractère qualitatif constitue la forme la plus générale de description statistique.

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

Les caractères quantitatifs discrets prennent des valeurs numériques isolées et dénombrables, généralement issues d'un comptage, comme le nombre d'enfants dans un ménage ou le nombre de communes dans un département. À l'inverse, les caractères quantitatifs continus peuvent prendre une infinité de valeurs dans un intervalle donné, car ils résultent d'une mesure, par exemple la température, la durée ou un revenu. Il est nécessaire de les distinguer car les outils statistiques et graphiques utilisés diffèrent. Les variables discrètes se prêtent à des diagrammes en bâtons et à des calculs exacts, tandis que les variables continues nécessitent souvent une discrétisation en classes, des histogrammes et des méthodes spécifiques de calcul des paramètres statistiques.

3) Paramètres de position

a) Pourquoi existe-t-il plusieurs types de moyenne ?

Il existe plusieurs types de moyenne car aucune ne résume parfaitement toutes les distributions. La moyenne arithmétique est sensible aux valeurs extrêmes, ce qui la rend peu représentative dans des distributions très dissymétriques. D'autres moyennes, comme la moyenne pondérée ou géométrique, prennent mieux en compte certaines contraintes ou structures des données. La diversité des moyennes répond donc à la diversité des situations statistiques rencontrées.

b) Pourquoi calculer une médiane ?

La médiane permet de partager la population en deux parties égales et offre une mesure de position robuste face aux valeurs extrêmes. Elle est particulièrement utile lorsque la distribution est asymétrique ou comporte des valeurs aberrantes. Dans ce cas, la médiane représente souvent mieux la situation « centrale » que la moyenne.

c) Quand est-il possible de calculer un mode ?

Le mode peut être calculé pour tous les types de caractères, qualitatifs comme quantitatifs, à condition que certaines modalités ou valeurs se répètent. Il correspond à la modalité la plus fréquente et est particulièrement pertinent pour les caractères qualitatifs ou les distributions présentant des concentrations marquées.

4) Paramètres de concentration

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

La médiale permet d'identifier le point à partir duquel une moitié de la somme totale est atteinte, ce qui apporte une information différente de celle de la médiane, centrée sur les individus. L'indice de concentration de Gini mesure le degré d'inégalité dans la répartition d'une variable, comme les revenus ou les populations. Il est particulièrement utile pour comparer des distributions et évaluer leur niveau de concentration ou d'hétérogénéité.

5) Paramètres de dispersion

a) Pourquoi calculer une variance plutôt que l'écart à la moyenne ? Pourquoi utiliser ensuite l'écart type ?

La variance est calculée car la somme des écarts simples à la moyenne est toujours nulle, ce qui empêche toute mesure de dispersion. En élevant ces écarts au carré, on obtient une mesure globale de la dispersion. L'écart type, racine carrée de la variance, est ensuite utilisé car il est exprimé dans la même unité que la variable étudiée, ce qui le rend plus facilement interprétable.

b) Pourquoi calculer l'étendue ?

L'étendue mesure l'écart entre la valeur maximale et la valeur minimale d'une distribution. Elle donne une première indication simple de la dispersion, mais elle est très sensible aux valeurs extrêmes. Elle est donc souvent utilisée comme indicateur complémentaire plutôt que comme mesure principale.

c) À quoi sert un quantile ? Quels sont les plus utilisés ?

Un quantile permet de découper une distribution en parts égales en nombre d'individus, ce qui facilite les comparaisons et l'analyse de la structure des données. Les quantiles les plus utilisés sont les quartiles, en particulier le premier et le troisième quartile, qui encadrent la moitié centrale de la population.

d) Pourquoi construire une boîte de dispersion ? Comment l'interpréter ?

La boîte de dispersion, ou boîte à moustaches, synthétise graphiquement la position centrale, la dispersion et les valeurs extrêmes d'une distribution. Elle permet de comparer rapidement plusieurs distributions et de détecter d'éventuelles asymétries ou valeurs aberrantes. Sa lecture repose sur la position de la médiane, la taille de la boîte (dispersion centrale) et la longueur des moustaches.

6) Paramètres de forme

a) Quelle différence entre moments centrés et moments absolus ? Pourquoi les utiliser ?

Les moments absolus sont calculés par rapport à l'origine et décrivent la distribution dans son ensemble, tandis que les moments centrés sont calculés par rapport à la moyenne et mettent en évidence la structure interne de la distribution. Les moments centrés sont particulièrement utiles pour analyser la dispersion, l'asymétrie et l'aplatissement d'une distribution.

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

Vérifier la symétrie d'une distribution permet de mieux comprendre sa forme et de choisir les outils statistiques adaptés. Une distribution symétrique justifie souvent l'usage de la moyenne et de l'écart type, tandis qu'une distribution dissymétrique invite à privilégier la médiane et les quantiles. La symétrie peut être évaluée graphiquement (histogramme, boîte à moustaches) ou à l'aide d'indicateurs de dissymétrie fondés sur les moments.

2. Code

5) Voici les résultats :

```
Type                42271.21
Trait de côte (km)      12.82
Surface (km²)          117.72
Latitude                9.58
Longitude              21.17
dtype: float64
Type                42279.00
Trait de côte (km)      2.23
Surface (km²)           0.18
Latitude               10.47
Longitude              25.11

Calcul des modes:
Type                    1.00
Trait de côte (km)      0.12
Surface (km²)           0.00
Latitude              -73.00
Longitude              22.00

Calcul des écarts-types:
Type                24414.57
Trait de côte (km)    224.70
Surface (km²)        8997.06
Latitude               97.56
Longitude              36.12

Calcul des écarts absolus moyens:
Type                21148.57
Trait de côte (km)     17.57
Surface (km²)         226.71
Latitude               86.90
Longitude              30.30

Calcul des étendues:
Type                85062.00
Trait de côte (km)    39577.14
Surface (km²)       2117507.76
Latitude              359.97
Longitude             162.80
```

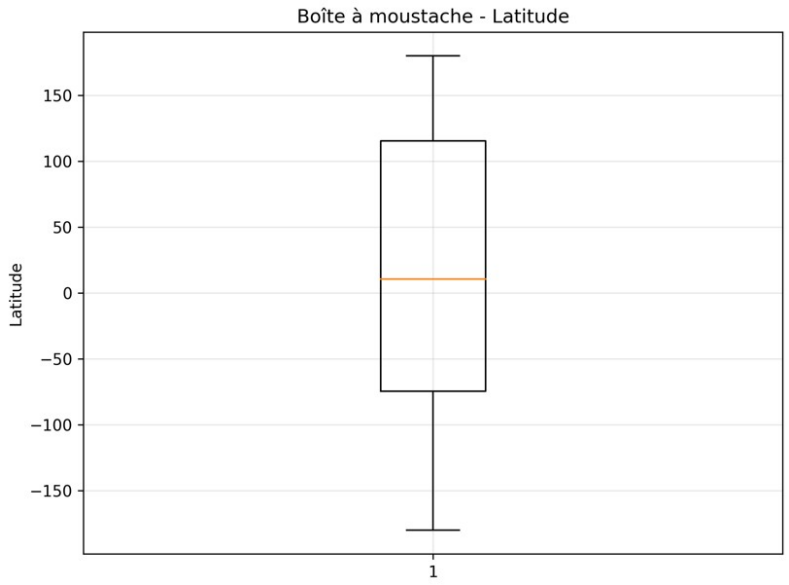
7) Voici les résultats :

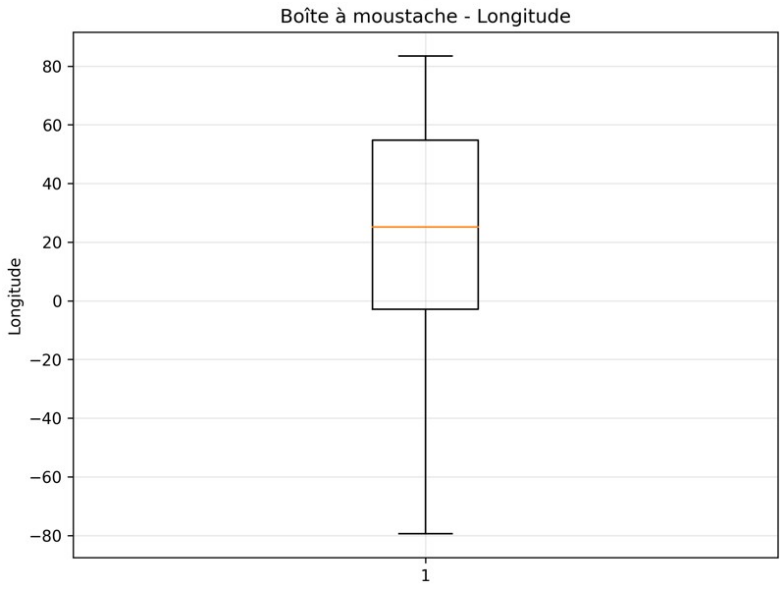
Calcul de la distance interquartile (IQR) :
Type 42303.00
Trait de côte (km) 3.88
Surface (km²) 0.78
Latitude 189.96
Longitude 57.63

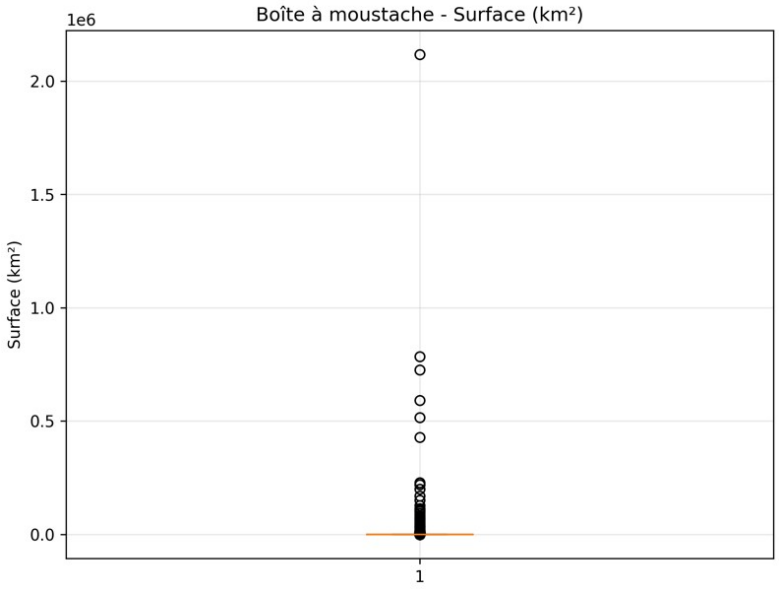
Calcul de la distance interdécile :
Type 67637.40
Trait de côte (km) 13.20
Surface (km²) 4.87
Latitude 240.24
Longitude 98.56

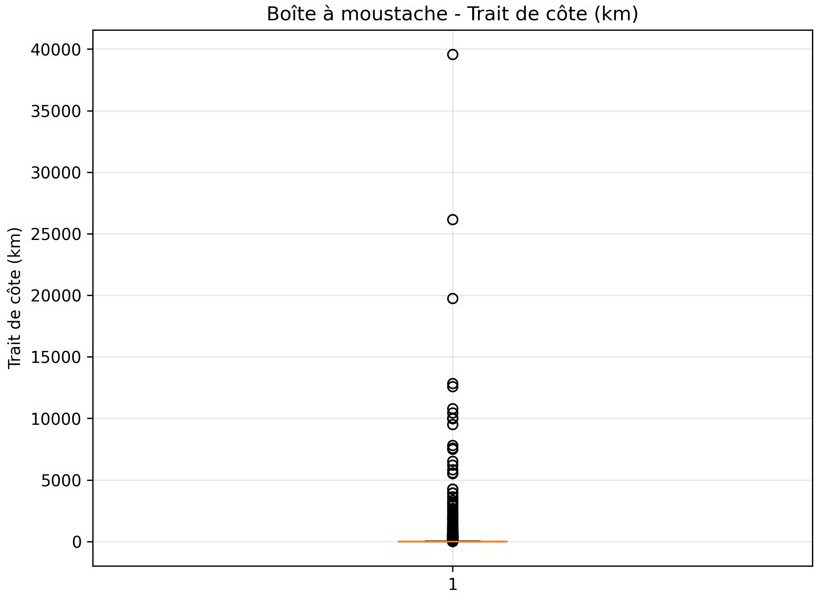
8)

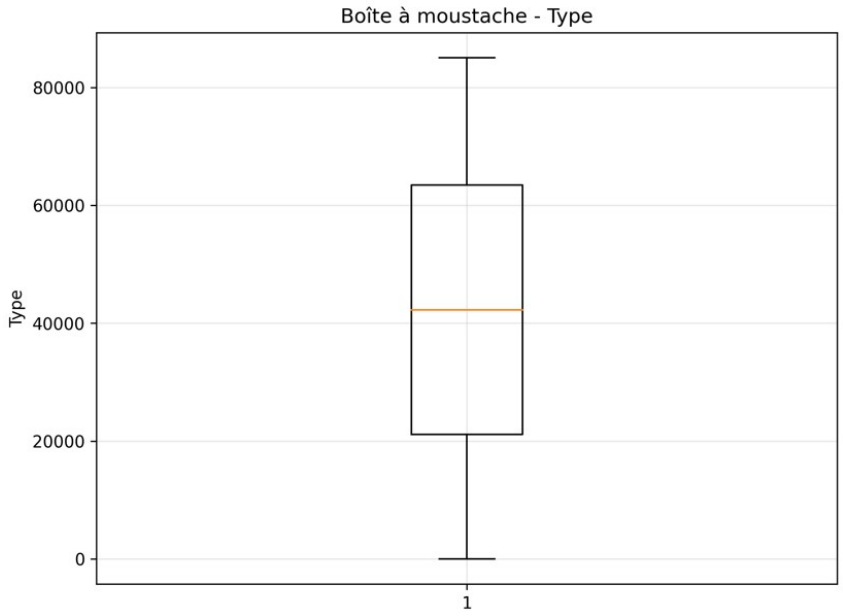
<u>Boîtes à moustaches obtenues</u>





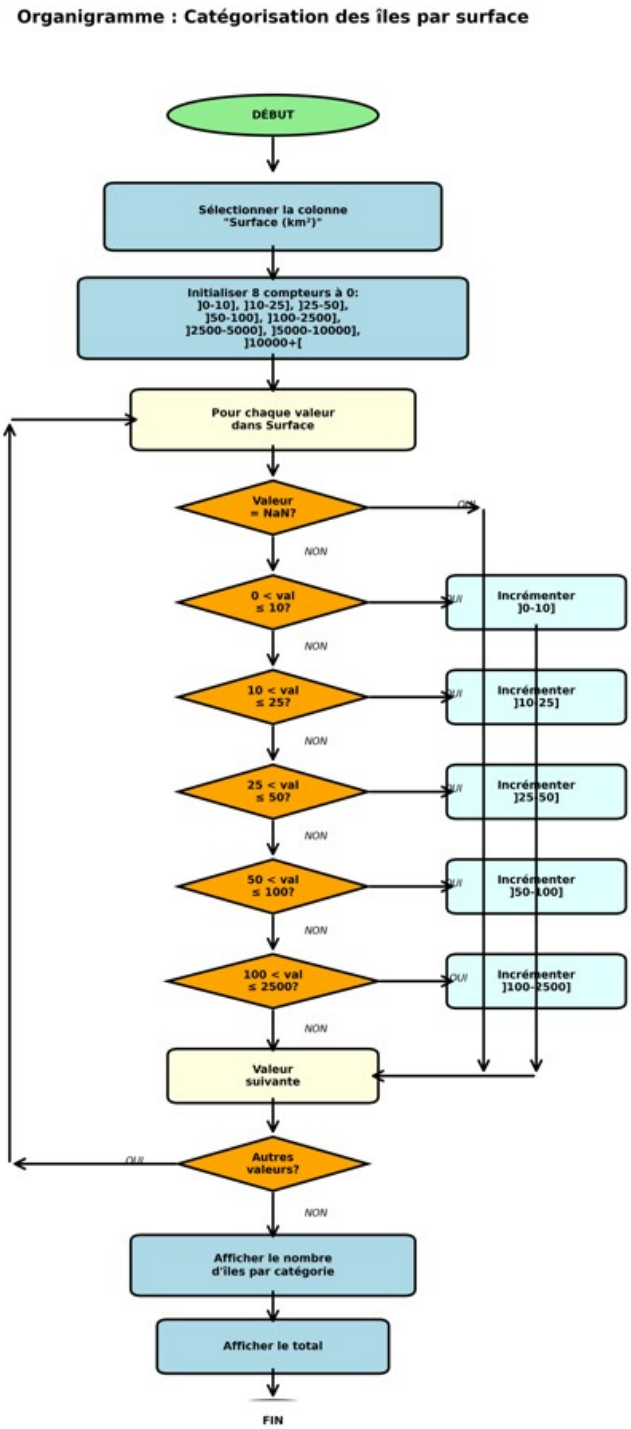

Boîte à moustache - Surface (km²)
1e6
2.0
1.5
1.0
0.5
0.0
Surface (km²)
1


Boîte à moustache - Trait de côte (km)
40000
35000
30000
25000
20000
15000
10000
5000
0
Trait de côte (km)
1


Boîte à moustache - Type
80000
60000
40000
20000
0
Type
1

10)

Organigramme obtenu



3. Commentaire de code

Ce script met en œuvre une démarche complète d'analyse exploratoire appliquée à un jeu de données géographiques sur les îles du monde. Il ne se limite pas à produire des indicateurs statistiques isolés, mais cherche à caractériser la structure globale des données, notamment leur forte dispersion et leur asymétrie. Les résultats numériques montrent clairement une concentration

massive des îles de très petite surface, tandis que quelques îles de grande taille influencent fortement les moyennes et les écarts-types. La comparaison entre moyenne, médiane et mode met ainsi en évidence l'inadéquation de la moyenne comme indicateur central pour ce type de distribution très déséquilibrée. La catégorisation par classes de surface renforce cette lecture en fournissant une vision synthétique et interprétable de la répartition des îles, plus pertinente qu'une simple liste de valeurs. L'organigramme final ne sert pas seulement à illustrer le code, mais à formaliser l'algorithme et à rendre explicite la logique de décision, ce qui est essentiel dans un contexte pédagogique ou méthodologique. D'un point de vue général, ce script est utile pour comprendre et illustrer les limites des statistiques classiques lorsqu'elles sont appliquées à des données hétérogènes et extrêmes. Il constitue un bon outil d'aide à l'analyse et à la prise de recul critique sur les indicateurs utilisés. En revanche, il ne permet pas de tirer des conclusions causales ni d'expliquer les phénomènes observés, car il ne dépasse pas le cadre descriptif. Son intérêt est donc surtout analytique et pédagogique, plutôt qu'opérationnel ou prédictif.

Séance 4 : Distributions statistiques

1. Questions de cours

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 ?

Le premier critère fondamental pour choisir entre une distribution discrète et une distribution continue est la nature même de la variable étudiée. Une variable discrète correspond à un phénomène qui ne peut prendre que des valeurs isolées et dénombrables, généralement issues d'un comptage. À l'inverse, une variable continue représente une grandeur mesurable pouvant théoriquement prendre une infinité de valeurs dans un intervalle donné. Un second critère important est le mécanisme de production des données. Les lois discrètes sont adaptées aux situations où l'on répète des expériences élémentaires (succès/échec, présence/absence, événement rare) où l'on observe un nombre d'occurrences. Par exemple, le nombre d'îles dans une classe de surface, le nombre de séismes dans une région donnée ou le nombre de communes par département relèvent naturellement de lois discrètes comme la binomiale ou la loi de Poisson. Un troisième critère est le niveau de précision de la mesure. Même lorsqu'un phénomène est continu, la mesure peut le rendre discret (arrondis, classes, seuils administratifs). Cependant, il faut distinguer la nature réelle du phénomène de la façon dont on l'observe. Par exemple, la population d'une ville est une variable discrète, mais la densité de population est continue car elle résulte d'un rapport de grandeurs mesurables. Enfin, l'objectif de l'analyse joue un rôle décisif. Les distributions discrètes sont particulièrement adaptées à l'étude d'événements rares, de répartitions par catégories ou de phénomènes de comptage. Les distributions continues, quant à elles, sont plus pertinentes pour décrire des phénomènes spatiaux, des variations de taille, de distance ou d'intensité, comme les surfaces, les altitudes ou les précipitations. Le choix entre discret et continu conditionne donc directement l'interprétation des résultats et la validité des conclusions statistiques.

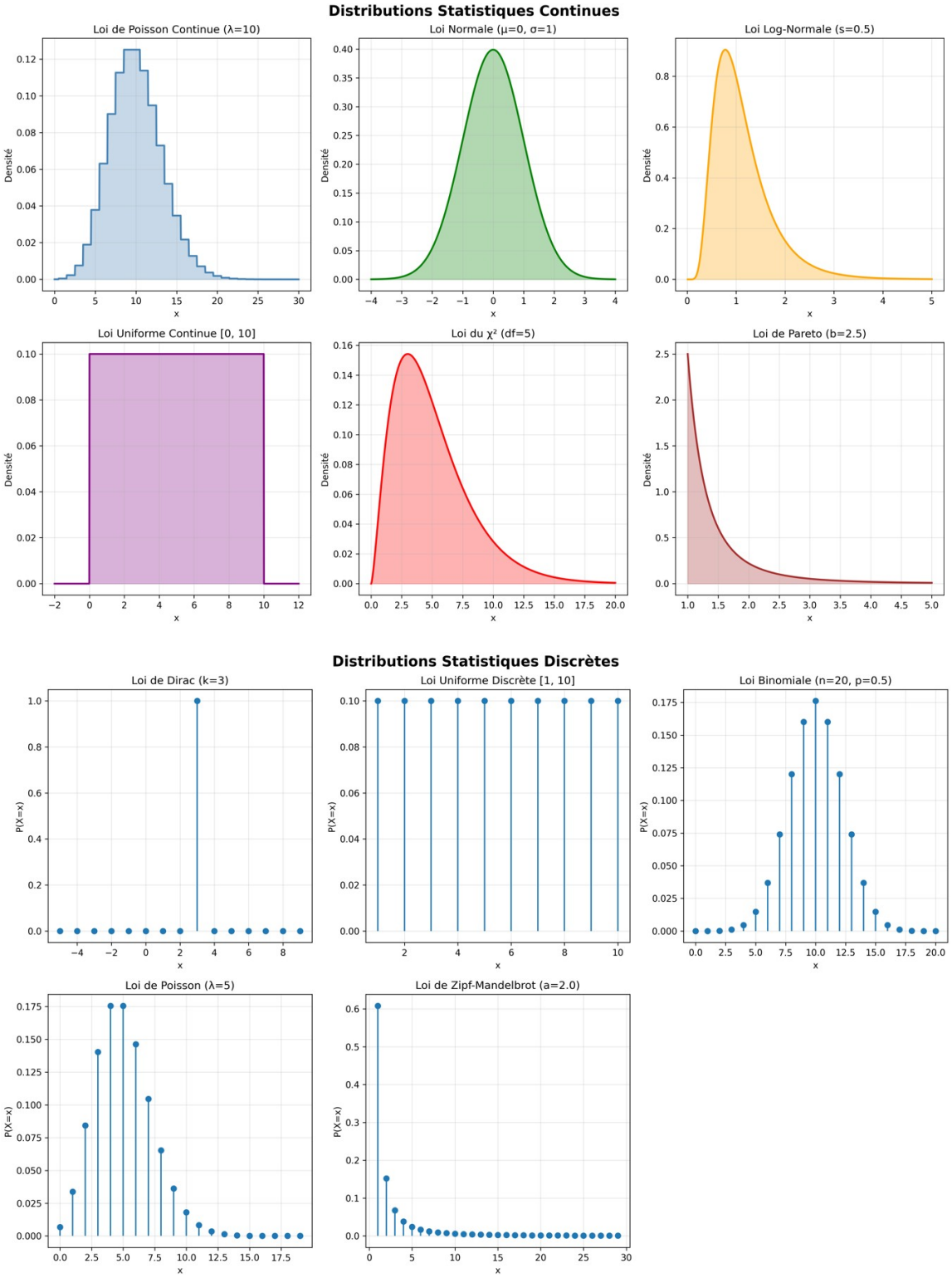
2) Expliquez selon vous quelles sont les lois les plus utilisées en géographie ?

En géographie, les lois statistiques les plus utilisées sont celles qui permettent de décrire des phénomènes spatiaux hétérogènes, asymétriques et souvent dominés par des valeurs extrêmes, ce que le cours met fortement en évidence. La loi normale reste largement utilisée, notamment pour des phénomènes résultant de la combinaison de nombreux facteurs indépendants. Elle est fréquemment mobilisée pour modéliser des erreurs de mesure, certaines variables climatiques (comme les températures moyennes) ou des phénomènes relativement homogènes à grande échelle. Les lois asymétriques à droite, comme la loi log-normale, occupent une place centrale en géographie. Elles sont particulièrement adaptées aux variables positives résultant de processus multiplicatifs, tels que les surfaces des îles, la taille des bassins versants, les revenus régionaux ou les volumes urbains. Ces lois expliquent pourquoi quelques valeurs très grandes coexistent avec une multitude de petites valeurs, situation extrêmement fréquente dans les systèmes géographiques.

Les lois de puissance, notamment les lois de Pareto, Zipf et Zipf-Mandelbrot, sont également essentielles. Elles sont utilisées pour analyser la hiérarchie urbaine, la distribution des tailles de villes, la fréquence des toponymes ou la répartition des équipements. Ces lois traduisent des mécanismes d'auto-organisation et d'inégalités structurelles, où les extrêmes jouent un rôle déterminant. Enfin, certaines lois discrètes comme la loi de Poisson sont couramment employées pour modéliser des événements spatiaux rares et localisés, par exemple les séismes, les crues exceptionnelles ou les accidents industriels. Ces lois permettent d'étudier la fréquence des événements sans supposer une régularité spatiale parfaite. En résumé, la géographie mobilise principalement des lois capables de rendre compte de la diversité, de l'asymétrie et des extrêmes, ce

qui explique l'importance des distributions non gaussiennes dans l'analyse des phénomènes spatiaux.

2. Code



Résultats des calculs pour les distributions statistiques de variables discrètes continues

--- DISTRIBUTIONS DISCRÈTES ---

Loi de Dirac (k=3):
  Moyenne: 3.0000
  Écart-type: 0.0000

Loi Uniforme Discrète (a=1, b=10):
  Moyenne: 5.5000
  Écart-type: 2.8723
  [scipy.stats: 5.5000, 2.8723]

Loi Binomiale (n=20, p=0.5):
  Moyenne: 10.0000
  Écart-type: 2.2361
  [scipy.stats: 10.0000, 2.2361]

Loi de Poisson (λ=5):
  Moyenne: 5.0000
  Écart-type: 2.2361
  [scipy.stats: 5.0000, 2.2361]

Loi de Zipf (a=2.0):
  Moyenne: inf
  Écart-type: inf
  [scipy.stats: inf, inf]

Résultats des calculs pour les distributions statistiques de variables continues

--- DISTRIBUTIONS CONTINUES ---

Loi Normale (μ=0, σ=1):
  Moyenne: 0.0000
  Écart-type: 1.0000
  [scipy.stats: 0.0000, 1.0000]

Loi Log-Normale (s=0.5):
  Moyenne: 1.1331
  Écart-type: 0.6039
  [scipy.stats: 1.1331, 0.6039]

Loi Uniforme Continue (a=0, b=10):
  Moyenne: 5.0000
  Écart-type: 2.8868
  [scipy.stats: 5.0000, 2.8868]

Loi du Chi-deux (df=5):
  Moyenne: 5.0000
  Écart-type: 3.1623
  [scipy.stats: 5.0000, 3.1623]

Loi de Pareto (b=2.5):
  Moyenne: 1.6667
  Écart-type: 1.4907
  [scipy.stats: 1.6667, 1.4907]

3. Commentaire de code

Ce script s'inscrit dans une démarche de compréhension globale des distributions statistiques plutôt que dans une simple démonstration technique. Il met en évidence l'opposition fondamentale entre variables discrètes et variables continues, non seulement sur le plan théorique, mais aussi à travers leurs comportements statistiques et graphiques. En mobilisant un large éventail de lois, il montre que le choix d'une distribution n'est jamais neutre : chaque loi correspond à une hypothèse précise sur la nature du phénomène étudié, qu'il s'agisse de comptages, de mesures continues ou de phénomènes hiérarchisés. L'intérêt principal du code réside dans sa capacité à confronter visuellement et numériquement ces modèles. Les graphiques révèlent immédiatement des différences de structure, notamment l'asymétrie, la concentration des valeurs ou l'importance des extrêmes, éléments centraux dans l'analyse des phénomènes réels. Les calculs de moyenne et d'écart-type permettent ensuite de souligner que certains indicateurs classiques perdent leur sens pour certaines lois, comme celles à queues lourdes. D'un point de vue général, ce script est pertinent car il encourage une lecture critique des données : il rappelle que l'on ne peut pas appliquer mécaniquement une loi normale à tout phénomène. Il constitue ainsi un outil utile pour l'analyse exploratoire et pour la réflexion méthodologique, notamment dans des disciplines comme la géographie, où les distributions observées sont souvent très éloignées de l'homogénéité théorique.

Séance 5 : Statistiques inférentielles

1. Questions de cours

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 une méthode statistique qui consiste à étudier un sous-ensemble d'individus extrait d'une population mère afin d'en inférer les caractéristiques globales. Il n'est généralement pas possible d'étudier la population entière, soit parce que sa taille est trop grande, soit parce que l'enquête serait trop coûteuse ou trop longue à réaliser. On distingue les méthodes d'échantillonnage aléatoires, fondées sur un tirage au sort, et les méthodes non aléatoires, comme l'échantillonnage systématique ou la méthode des quotas. Le choix de la méthode dépend notamment de la disponibilité d'une base de sondage, des contraintes matérielles et du niveau de précision attendu.

2) Comment définir un estimateur et une estimation ?

Un estimateur est une variable aléatoire construite à partir des observations issues d'un échantillon et destinée à approcher un paramètre inconnu de la population, comme la moyenne, la variance ou une proportion. Une estimation correspond à la valeur numérique prise par cet estimateur une fois l'échantillon observé. Ainsi, l'estimateur est un objet théorique, tandis que l'estimation est le résultat concret obtenu à partir des données.

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

L'intervalle de fluctuation est utilisé lorsque la valeur théorique d'un paramètre, notamment une proportion, est connue, et permet de déterminer si une fréquence observée dans un échantillon est compatible avec cette valeur. Il s'inscrit dans une logique de contrôle et de décision. À l'inverse, l'intervalle de confiance est employé lorsque le paramètre est inconnu et vise à encadrer une valeur plausible de ce paramètre à partir d'un estimateur, avec un risque d'erreur fixé à l'avance.

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

Dans la théorie de l'estimation, un biais correspond à la différence entre l'espérance mathématique d'un estimateur et la valeur réelle du paramètre à estimer. Lorsqu'un estimateur est biaisé, il produit des estimations qui s'écartent systématiquement de la valeur vraie. Un estimateur sans biais, au contraire, donne en moyenne la bonne valeur du paramètre, même si chaque estimation individuelle peut fluctuer.

5) Comment appelle-t-on une statistique travaillant sur la population totale ? Faites le lien avec la notion de données massives 1 ?

Une statistique travaillant sur la population totale est issue d'un recensement, c'est-à-dire d'une enquête exhaustive portant sur tous les individus. Les données massives s'inscrivent dans une logique proche du recensement, car elles visent à collecter un très grand volume de données. Toutefois, même dans ce contexte, des méthodes statistiques restent nécessaires pour traiter les biais, interpréter les résultats et modéliser l'incertitude.

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

Le choix d'un estimateur est un enjeu central en statistique inférentielle, car il conditionne la qualité des conclusions tirées à partir des données. Un bon estimateur doit être sans biais ou asymptotiquement sans biais, présenter une variance faible et converger vers la vraie valeur du paramètre lorsque la taille de l'échantillon augmente. Un estimateur mal choisi peut conduire à des estimations imprécises ou trompeuses, même avec un échantillon de grande taille.

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

Les méthodes d'estimation d'un paramètre incluent l'estimation ponctuelle, qui fournit une valeur unique, l'estimation par intervalle de confiance, qui tient compte de l'incertitude, et les méthodes de simulation comme Monte-Carlo ou le bootstrap. Le choix d'une méthode dépend de la nature du paramètre à estimer, de la taille de l'échantillon, des hypothèses sur la loi de la variable étudiée et du compromis entre biais et variance que l'on souhaite privilégier.

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

Les tests statistiques sont des outils permettant de prendre une décision à partir des données, tout en contrôlant le risque d'erreur. Ils servent notamment à vérifier la compatibilité entre des observations et un modèle théorique ou une hypothèse donnée. La construction d'un test repose sur la définition d'une hypothèse nulle, le choix d'une statistique de test adaptée, la détermination de sa loi de probabilité et la fixation d'un seuil de signification.

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

Les critiques de la statistique inférentielle soulignent souvent la dépendance aux hypothèses de modèle, les biais possibles dans les échantillons et les risques de mauvaise interprétation des résultats. Cependant, le cours montre que ces limites sont bien identifiées et que la statistique inférentielle demeure indispensable pour étudier des populations inaccessibles dans leur totalité. Elle permet surtout de quantifier l'incertitude et d'encadrer rigoureusement les décisions statistiques.

2. Code

A) Théorie de l'échantillonnage

Moyennes obtenues

Moyenne arrondie par colonne :
Pour 391
Contre 416
Sans opinion 193

Fréquences obtenues

	Moyenne_arrondie	Frequence_echantillon	Frequence_population
Pour	391.0	0.39	0.39
Contre	416.0	0.42	0.42
Sans opinion	193.0	0.19	0.19

	Freq_population	Freq_echantillon	Inf_95%	Sup_95%
Pour	0.39	0.39	0.39	0.39

Contre	0.42	0.42	0.42	0.42
Sans opinion	0.19	0.19	0.19	0.19

L'intervalle de fluctuation à 95 % sert à vérifier si les fréquences observées dans les échantillons sont compatibles avec les fréquences réelles de la population mère. Il encadre les valeurs que l'on peut attribuer aux variations dues au hasard de l'échantillonnage. Dans notre cas, les fréquences issues des échantillons sont exactement égales aux fréquences de la population mère et se situent à l'intérieur des intervalles de fluctuation, qui sont ici réduits à une valeur unique. Cela montre que les écarts observés ne sont pas dus au hasard, mais que les échantillons reproduisent parfaitement la structure de la population. On peut donc conclure que les échantillons utilisés sont totalement représentatifs de la population mère. Toutefois, cette absence de fluctuation limite l'intérêt de l'intervalle de fluctuation, qui est normalement destiné à mesurer l'incertitude liée à l'échantillonnage.

B) Théorie de l'estimation

Somme de la ligne et fréquences obtenues

Somme de l'échantillon : 1000
Fréquences : [0.4, 0.4, 0.21]

Intervalles obtenus

Opinion 1 : IC 95 % = (0.36, 0.43)
Opinion 2 : IC 95 % = (0.37, 0.43)
Opinion 3 : IC 95 % = (0.18, 0.23)

Dans la majorité des travaux en sciences humaines et sociales, la population mère n'est pas entièrement connue et l'analyse repose uniquement sur un échantillon. Pour pouvoir accorder une confiance aux résultats obtenus, il est nécessaire d'évaluer l'incertitude liée à l'échantillonnage. La construction d'intervalles de confiance permet précisément de répondre à cette problématique. Ces intervalles encadrent les valeurs plausibles du paramètre réel de la population à partir des données observées dans l'échantillon. Si l'échantillon est suffisamment grand et correctement construit, l'intervalle de confiance fournit une estimation fiable de la population mère, tout en rendant explicite la marge d'erreur associée. Ainsi, cette méthode permet de transformer un résultat issu d'un échantillon en une information statistiquement exploitable et interprétable.

C) Théorie de la décision

L'intervalle de confiance calculé à partir d'un seul échantillon dépend uniquement de la taille totale de celui-ci. Les intervalles obtenus pour chaque opinion sont plus larges que ceux calculés précédemment à partir de la moyenne de l'ensemble des échantillons, ce qui traduit une incertitude plus importante. Cette différence s'explique par le fait qu'un échantillon isolé est davantage soumis aux fluctuations aléatoires qu'un ensemble d'échantillons moyennés. En comparant les résultats obtenus sur plusieurs lignes de l'échantillon, on observe que les fréquences et les intervalles varient d'un échantillon à l'autre, bien que les valeurs restent globalement cohérentes. Cette variabilité confirme que chaque échantillon fournit une estimation différente de la population mère, mais que ces estimations deviennent plus stables lorsque l'on augmente le nombre d'échantillons ou la taille de ceux-ci. On peut ainsi conclure que l'utilisation d'un seul échantillon donne une information exploitable mais moins fiable que l'approche précédente, qui reposait sur la moyenne de plusieurs échantillons et conduisait à des estimations plus robustes.

Quelle est la distribution normale ?

Dans les deux cas, la p-value est strictement inférieure aux seuils usuels de significativité (5 % ou 1 %). On rejette donc l'hypothèse nulle de normalité pour les deux séries. Autrement dit, aucune des deux distributions ne peut être considérée comme normale au sens du test de Shapiro-Wilk. Il est important d'expliquer ce résultat. Le test de Shapiro-Wilk est très sensible, en particulier lorsque la taille de l'échantillon est importante. Même de légers écarts à la normalité théorique peuvent conduire à une p-value extrêmement faible, affichée ici comme 0.0 après arrondi. Ainsi, une distribution qui est visuellement proche d'une loi normale peut néanmoins être rejetée par le test, car elle n'est pas strictement normale. En conclusion, d'un point de vue statistique et au regard du test de Shapiro-Wilk, aucune des deux séries ne suit une loi normale, puisque l'hypothèse de normalité est rejetée dans les deux cas. Ce résultat rappelle qu'un test de normalité doit toujours être interprété avec prudence et, idéalement, complété par une analyse graphique (histogramme, Q-Q plot) pour apprécier la nature des écarts à la normalité.

3. Commentaire de code

Le code proposé s'inscrit dans une démarche statistique cohérente qui articule simulation, estimation et décision, tout en exploitant efficacement les outils offerts par Pandas et SciPy. Le recours à des données simulées permet de maîtriser les paramètres de la population mère et de confronter directement les résultats empiriques aux valeurs théoriques, ce qui constitue un cadre pédagogique solide pour tester la validité des méthodes statistiques. L'utilisation de moyennes calculées sur un grand nombre d'échantillons montre une volonté explicite de réduire l'effet des fluctuations aléatoires, conformément aux principes de la théorie de l'échantillonnage. Le code met ainsi en évidence que la stabilité des résultats ne provient pas d'un échantillon isolé, mais de la répétition des tirages. Le choix de travailler en fréquences plutôt qu'en effectifs facilite la comparaison avec la population mère et permet une interprétation immédiatement probabiliste. La construction d'intervalles de fluctuation et d'intervalles de confiance illustre clairement la différence entre contrôle d'un modèle connu et estimation d'un paramètre inconnu. Le code souligne que l'incertitude statistique ne disparaît jamais totalement et qu'elle dépend fortement de la taille de l'échantillon, ce qui est un point central en sciences humaines et sociales. Enfin, l'intégration du test de Shapiro-Wilk traduit une approche décisionnelle rigoureuse, fondée sur des hypothèses statistiques explicites. Le rejet systématique de la normalité rappelle que les tests sont sensibles aux écarts au modèle théorique et qu'un résultat numérique ne peut être interprété sans réflexion critique. Globalement, ce code ne se limite pas à produire des résultats : il met en évidence les limites, les conditions de validité et les enjeux interprétatifs des outils statistiques mobilisés.

Séance 6 : Statistiques d'ordre des variables qualitatives

1. Questions de cours

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

Une statistique ordinale est une statistique qui repose sur le classement des individus ou des objets selon un ordre donné, sans nécessairement tenir compte des écarts numériques entre eux. Elle s'oppose à la statistique nominale, qui se limite à des catégories sans ordre intrinsèque. La statistique ordinale utilise donc des variables qualitatives ordinales, c'est-à-dire des variables dont les modalités sont hiérarchisées. En géographie, ce type de statistique permet de matérialiser une hiérarchie spatiale en classant, par exemple, des villes, des régions ou des territoires selon leur importance relative, leur taille ou leur attractivité.

2) Quel ordre est à privilégier dans les classifications ?

Dans les classifications, l'ordre à privilégier est l'ordre croissant, également appelé ordre naturel, car il facilite la lecture, la comparaison et l'interprétation des données. Cet ordre permet notamment d'identifier les valeurs extrêmes, trop faibles ou trop élevées, et d'étudier les phénomènes liés aux maxima et minima d'une série d'observations. Certaines exceptions existent en géographie, comme la loi rang-taille, mais l'ordre croissant reste la règle générale.

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

La corrélation des rangs mesure le degré de dépendance statistique entre deux variables ordinales en comparant leurs rangs respectifs. La concordance de classements, quant à elle, évalue la similarité globale entre plusieurs classements en analysant le nombre de paires concordantes et discordantes. Ainsi, la corrélation des rangs s'intéresse à la relation entre deux variables ordinales, tandis que la concordance porte davantage sur l'accord entre des classements, éventuellement multiples, d'un même ensemble d'objets.

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

La différence principale entre les tests de Spearman et de Kendal réside dans leur mode de calcul et leur champ d'application. Le test de Spearman repose sur le coefficient de corrélation des rangs et s'appuie sur les différences entre les rangs attribués à chaque individu, ce qui le rapproche d'une corrélation classique. Le test de Kendal, en revanche, est fondé sur le comptage des paires concordantes et discordantes, ce qui le rend plus robuste et plus facilement généralisable à plusieurs classements. Le coefficient de Kendal présente également l'avantage de pouvoir être étendu au cas de plus de deux classements.

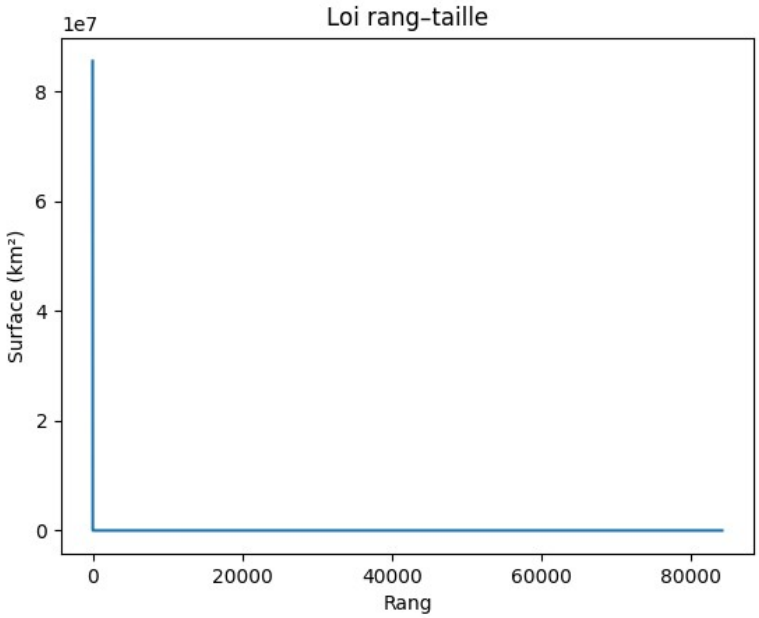
5) À quoi servent les coefficients de Goodman-Kursdal et de Yule ?

Les coefficients de Goodman-Kruskal et de Yule servent à mesurer l'association entre des variables ordinales ou qualitatives, en se fondant sur la comparaison entre paires concordantes et discordantes. Le coefficient de Goodman-Kruskal évalue le surplus de paires concordantes par rapport aux paires discordantes et s'interprète comme un coefficient de corrélation. Le coefficient de Yule est un cas particulier appliqué aux tableaux de contingence de dimension 2×2 et permet de mesurer la force et le sens de l'association entre deux variables qualitatives dichotomiques.

2. Code

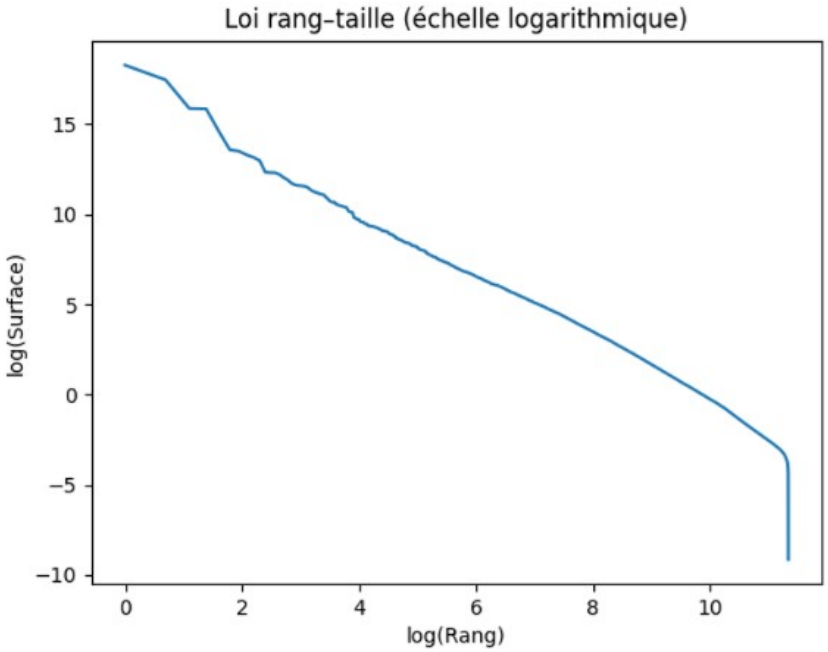
5)

Loi rang-taille obtenue



6)

Loi rang-taille obtenue grâce à la fonction locale conversion log()



14)

Résultats

Population :
  Spearman = 0.9861941451289199
  Kendall = 0.9042567659458726

Densité :
  Spearman = 0.9672541347992915
  Kendall = 0.8588331293349655

Commentaire des résultats

Les coefficients de corrélation de Spearman et de Kendal obtenus pour la population entre 2007 et 2025 sont très élevés ($\rho \approx 0{,}99$; $\tau \approx 0{,}90$), ce qui traduit une très forte stabilité des classements des États en fonction du nombre d'habitants sur la période étudiée. Les pays les plus peuplés en 2007 restent globalement en tête du classement en 2025, malgré une croissance démographique inégale. Pour la densité de population, les coefficients restent également élevés ($\rho \approx 0{,}97$; $\tau \approx 0{,}86$), indiquant une forte concordance des rangs, mais légèrement inférieure à celle observée pour la population totale. Cela suggère que la densité est un indicateur plus sensible aux évolutions démographiques et territoriales, avec des réajustements de rangs plus marqués entre 2007 et 2025.Globalement, ces résultats montrent que la hiérarchie mondiale des États est relativement stable sur la période étudiée, tout en laissant apparaître des dynamiques plus contrastées pour la densité que pour la population totale.

3. Commentaire du code

Le code mis en œuvre vise à analyser la structure et l'évolution des hiérarchies territoriales à partir de données démographiques, en comparant les classements des États selon la population totale et la densité de population entre 2007 et 2025. Il s'appuie sur une chaîne de traitements cohérente allant de l'importation et du nettoyage des données jusqu'à l'analyse statistique des rangs. La méthodologie repose sur la construction de classements décroissants à l'aide de fonctions locales dédiées, permettant d'associer chaque valeur quantitative à un État et de transformer ces valeurs en rangs ordonnés. Ce choix méthodologique permet de se concentrer sur les positions relatives des pays plutôt que sur les valeurs absolues, ce qui est particulièrement pertinent pour comparer des systèmes hétérogènes à grande échelle.

Les classements obtenus pour différentes dates et indicateurs sont ensuite alignés pays par pays afin de garantir la comparabilité des rangs. Cette étape est indispensable pour éviter les biais liés aux valeurs manquantes ou aux différences de périmètre. L'utilisation conjointe des coefficients de Spearman et de Kendal permet d'évaluer à la fois la corrélation globale des hiérarchies et la concordance locale des rangs, offrant ainsi une lecture complémentaire de la stabilité ou des recompositions observées. Ce type d'approche est généralisable à de nombreux domaines où l'on cherche à analyser des classements dans le temps, tels que l'économie, la géographie quantitative ou les sciences sociales. Il constitue un outil robuste pour identifier des dynamiques structurelles tout en limitant l'influence des valeurs extrêmes et des distributions asymétriques.

Réflexion personnelle – Bilan fin de semestre

Ce semestre, mon initiation à l'analyse de données en Python m'a permis de mieux comprendre le lien étroit entre les sciences des données et les humanités numériques. En tant que débutante en programmation, j'ai d'abord perçu Python comme un outil essentiellement technique, parfois difficile à maîtriser, mais j'ai progressivement compris qu'il constitue un véritable moyen d'exploration des phénomènes humains à partir de données numériques. L'un des principaux enjeux rencontrés a été la gestion et l'organisation des données. Dans les humanités numériques, les données sont souvent nombreuses, hétérogènes et parfois désordonnées, qu'il s'agisse de textes, d'archives ou de traces numériques. Apprendre à les nettoyer, les structurer et les analyser m'a fait prendre conscience que la qualité de l'analyse dépend directement du traitement préalable des données. La programmation devient alors indispensable pour rendre ces corpus exploitables et comparables.

Un autre défi important a été l'interprétation des résultats. Les données ne sont jamais neutres et nécessitent une lecture critique, surtout lorsqu'elles renvoient à des réalités humaines, culturelles ou sociales. Ce semestre m'a montré que l'analyse de données en humanités numériques ne se limite pas à produire des chiffres ou des graphiques, mais qu'elle doit s'inscrire dans une réflexion théorique et contextuelle afin d'éviter les interprétations simplistes. La visualisation des données a également constitué un enjeu central. Il ne s'agit pas seulement de représenter des résultats, mais de les rendre lisibles et pertinents pour répondre à une problématique de recherche. Enfin, j'ai pu développé au cours de ce semestre à la fois des compétences techniques et une rigueur méthodologique, tout en renforçant ma capacité à articuler outils numériques et questionnements issus des sciences humaines. J'ai compris que les sciences des données et les humanités numériques sont profondément complémentaires pour analyser et comprendre les phénomènes humains contemporains.