# RAPPORT D'ACTIVITÉ

Analyse de données - parcours débutant

par LUCCIARDI Sébastien

Master 1 GAED, parcours GéoSuds – sociétés, territoires, développement.

**Table des matières**

## Séance 1

Objectif du cours :

- Comprendre le format et les objectifs du cours.
- Se familiariser avec les outils : Python, GitHub, Docker.
- Installer les outils et préparer l'espace de travail pour le semestre.

## Séance 2

Objectifs :

- Manipuler un fichier C.S.V.
- Faire des sorties graphiques
- Utiliser les bibliothèques Pandas (données) et Matplotlib (graphiques)
- Calculer des effectifs
- Calculer des fréquences
- Faire des graphiques (diagrammes en bâton et circulaires, et histogrammes)

Question de cours :

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

La géographie utilise les statistiques comme un outil d'analyse. Les statistiques permettent notamment de quantifier, de comparer et d'interpréter les phénomènes spatiaux tels que la population, le climat ou encore l'économie. La géographie donne le sens spatial et territorial aux données chiffrées

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

Effectivement, le hasard existe, mais de manière limitée. Le hasard pur est rare : les phénomènes géographiques ont souvent des causes spatiales, sociales ou naturelles. Cependant, une part d'aléatoire peut exister, par exemple dans les catastrophes naturelles ou les mouvements humains imprévisibles.

**3. Quels sont les types d'information géographique ?**

Il y a trois différents types d'information géographique. Les données quantitatives, c'est-à-dire les nombres avec par exemple la population ou la température. Puis les données qualitatives, qui représentent plutôt des catégories avec le types de sols ou alors les régions. Enfin, les données spatiales (localisation, coordonnées, formes des territoires).

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

La géographie a besoin d'outils statistiques pour décrire, comparer et expliquer les phénomènes dans l'espace. Cela passe notamment par la collecte**,** le traitement et la représentation de données spatiales fiables.

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

D'une part, la statistique descriptive sert à résumer et présenter les données avec des moyennes, des cartes ou des graphiques. D'une autre part, la statistique explicative cherche

à comprendre les relations entre variables et à expliquer pourquoi un phénomène se produit.

**6. Quelles sont les types de visualisation de données en géographie ? Comment choisir celles-ci ?**

En géographie, les principaux types de visualisation de données sont les cartes, les graphiques et les tableaux statistiques. Les cartes servent à représenter la répartition spatiale d'un phénomène et à montrer où il se produit. Les graphiques comme les histogrammes, diagrammes ou courbes, permettent de comparer des valeurs ou de montrer une évolution dans le temps. Enfin, les tableaux statistiques, eux, organisent les données de façon précise et détaillée, souvent avant la création d'une carte ou d'un graphique. Le choix du type de visualisation dépend du type de données c'est-à-dire quantitatives ou qualitatives et de l'objectif de l'analyse. Par exemple, on utilise une carte pour localiser, un graphique pour comparer ou illustrer une tendance, et un tableau pour présenter les chiffres exacts.

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

Les méthodes d'analyse de données se divisent en trois grandes catégories : les méthodes descriptives, qui servent à résumer et visualiser les données ; les méthodes explicatives, qui permettent d'étudier les relations entre une variable à expliquer et des variables explicatives ; et les méthodes de prévision, qui visent à analyser et anticiper l'évolution des phénomènes dans le temps.

**8. Définitions :**

- La population statistique est l'ensemble des individus ou éléments sur lesquels porte l'étude. C'est le groupe que l'on observe pour répondre à une question ou vérifier une hypothèse par exemple.
- Un individu statistique est un élément de la population. C'est l'unité sur laquelle on recueille des informations.
- Un caractère statistique est la propriété étudiée chez les individus. C'est ce qu'on observe, compte ou mesure. Par exemple : l'âge, le sexe, le niveau d'étude.
- Les modalités statistiques sont les différentes valeurs possibles que peut prendre un caractère. Par exemple, pour le caractère sexe, les modalités sont Homme, Femme, Autre

Et on distingue plusieurs types de caractères :
- caractères qualitatifs : Nominal pas d'ordre entre les modalités par exemple la couleur des yeux. Ou alors ordinal il existe un ordre par exemple le niveau d'étude ou le degré de satisfaction

- Caractères quantitatifs : Discret Donc valeurs entières par exemple le nombre d'enfants. Ou alors continu Donc valeurs avec décimales possibles par exemple la température

**9. Comment mesurer une amplitude et une densité ?**

Amplitude d'une classe : pour une classe d'intervalle $]a, b]$, l'amplitude est la longueur $b - a$.

Densité d'une classe : rapport entre l'effectif $ni$ de la classe et son amplitude ; formellement $d = ni\ b–a$ . La densité permet de comparer des classes de largeur

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

Ces formules sont des règles empiriques destinées à guider le choix du nombre de classes $k$ lors de la discrétisation d'une variable continue (préparation d'un histogramme). Elles visent à éviter un découpage trop fin (trop de classes, perte de lisibilité) ou trop grossier (trop peu de classes, perte d'information). Deux formules usuelles : la formule de Sturges $k \approx 1 +$ 3,2222 × log 10 $n$et la formule de Yule $k \approx 2{,}5\sqrt{n}\ 4$ , qui donnent des valeurs approximatives en fonction de la taille de l'échantillon $n$.

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

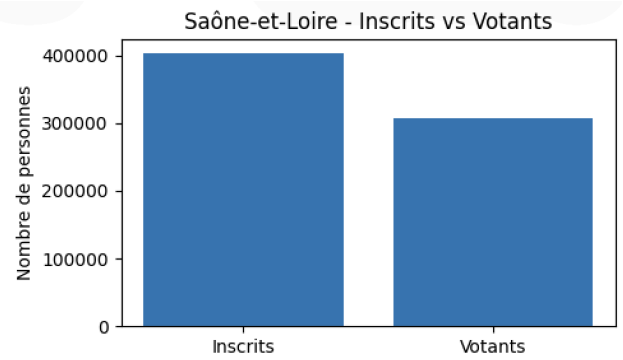
- Effectif $ni$ : nombre d'observations correspondant à une modalité ou à une classe.
- Fréquence relative $fi$ : proportion d'observations pour la modalité $i$, calculée par $fi = ni\ n$ où $n$ est l'effectif total.
- Fréquence cumulée jusqu'à la $k$-ième modalité : somme des fréquences des modalités inférieures ou égales à $k$, soit $Fk = \sum k\ i{=}1\ fi = 1\ n \sum k\ i{=}1\ ni$ . Sur l'ensemble des modalités, la somme des fréquences vaut 1.

- Distribution statistique : représentation empirique de la répartition des effectifs ou fréquences sur les modalités ou classes. Elle constitue le lien observable entre les données et les lois de probabilité théoriques, et sert de base à la description et au choix des modèles statistiques.

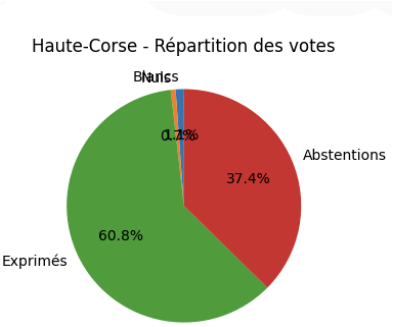
<u>Mise en oeuvre python :</u>

Lors de cette séance, j'ai appris à utiliser Python pour analyser un jeu de données réel à l'aide des bibliothèques Pandas et Matplotlib. J'ai commencé par ouvrir le fichier CSV contenant les résultats du premier tour de l'élection présidentielle de 2022. Cela m'a permis de découvrir la structure du tableau, le nombre de lignes et de colonnes, ainsi que la nature des variables étudiées. J'ai notamment identifié des variables quantitatives comme le nombre d'inscrits, de votants, d'abstentions ou encore de bulletins blancs et nuls.

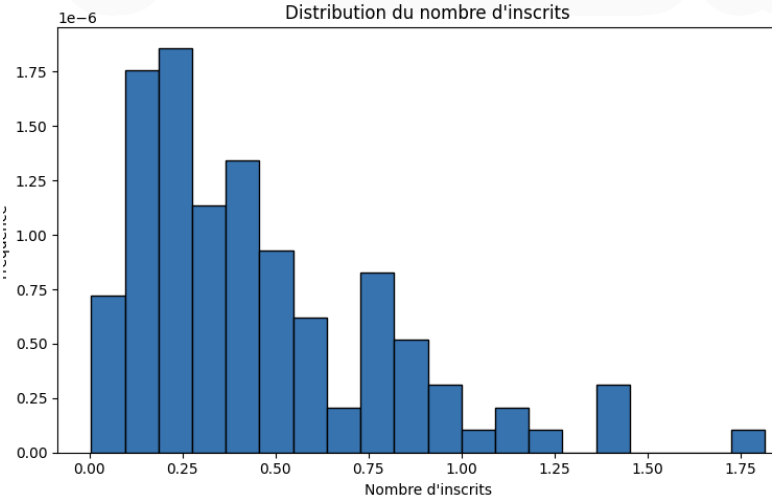
Ensuite, j'ai réalisé des diagrammes en barres comparant le nombre d'inscrits et le nombre de votants pour chaque département. Ces graphiques montrent clairement que le nombre de votants est toujours inférieur au nombre d'inscrits. L'écart entre les deux barres correspond aux abstentions. On observe que cet écart varie selon les départements, ce qui traduit des différences de participation électorale sur le territoire.




J'ai également créé des diagrammes circulaires représentant la répartition des votes entre les suffrages exprimés, les votes blancs, les votes nuls et les abstentions. Ces graphiques montrent que les suffrages exprimés sont majoritaires, mais que les abstentions représentent une part importante des inscrits. Les votes blancs et nuls restent minoritaires, mais ils ne sont pas négligeables et traduisent une forme de contestation ou de désengagement politique.




Enfin, j'ai réalisé un histogramme du nombre d'inscrits par département. Cet histogramme permet d'observer la distribution du nombre d'inscrits et met en évidence une asymétrie. La majorité des départements ont un nombre d'inscrits relativement faible, tandis que quelques départements très peuplés concentrent un nombre élevé d'inscrits. Cet outil permet donc de visualiser la dispersion des données et les inégalités démographiques entre les départements.



## Séance 3

Objectifs :

- Découvrir les méthodes de Pandas permettant de calculer les paramètres d'une série statistique
- Tracer une boîte de dispersion

Questions de cours :

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

Pour moi, le caractère qualitatif est le concept le plus général en statistique. Un caractère représente simplement une propriété observée sur un individu, comme la couleur des yeux, l'âge ou la taille.

- Caractères qualitatifs : ils décrivent des catégories ou attributs non mesurables numériquement (ex. : "Bleu", "Femme", "Ville").

- Caractères quantitatifs : ils sont un cas particulier, décrivant des propriétés mesurables par un nombre (ex. : 1,75 m, 20 ans).

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

Les caractères quantitatifs peuvent être discrets ou continus selon l'ensemble des valeurs qu'ils peuvent prendre :

- Discrets : ne prennent que des valeurs précises et dénombrables, souvent entières. On peut utiliser directement les valeurs observées et les représenter avec des diagrammes en barres.
- Continus : peuvent prendre toutes les valeurs possibles dans un intervalle donné. Ils nécessitent un regroupement en classes pour être représentés, par exemple via un histogramme.

Il est important de les distinguer car les méthodes de traitement et de représentation sont différentes.

**3. Paramètres de position**

Ces paramètres permettent d'identifier le centre d'une distribution de données.

- Multiplicité des moyennes : il existe plusieurs types de moyennes (arithmétique, quadratique, harmonique, géométrique). La moyenne arithmétique est la plus courante mais sensible aux valeurs extrêmes. D'autres moyennes, comme la

géométrique, sont utilisées selon le contexte (ex. taux de croissance).
Rôle de la médiane : c'est la valeur qui partage les données en deux parties égales après classement. Elle est robuste aux valeurs extrêmes et utile pour les distributions asymétriques.

- Calcul du mode : c'est la valeur la plus fréquente. Elle peut être unique ou multiple (distribution bimodale ou multimodale). Le mode est le seul paramètre applicable aux données qualitatives nominales.

**4. Paramètres de concentration**

Ils mesurent comment la masse totale d'un caractère est répartie parmi les individus.

- Médiale : proche de la médiane, mais elle divise la masse totale de la variable en deux parties égales. Par exemple, pour les revenus, elle sépare les individus selon la moitié des revenus totaux. La comparaison avec la médiane permet d'évaluer la concentration.
- Indice et courbe de Gini : ils décrivent la concentration d'une population (souvent pour les revenus). L'indice de Gini varie entre 0 (égalité parfaite) et 1 (concentration maximale).

**5. Paramètres de dispersion**

Ils mesurent l'étalement des données autour d'un centre.

- Variance et écart-type : la variance calcule la moyenne des carrés des écarts à la moyenne, donnant plus de poids aux valeurs éloignées. L'écart-type (racine carrée de la variance) permet de revenir à l'unité originale.
  Petit écart-type : données regroupées autour de la moyenne.
  Grand écart-type : données dispersées.

- Étendue : différence entre la valeur maximale et minimale. Simple mais dépend uniquement des extrêmes.

- Quantiles et écart interquartile : divisent la série ordonnée en parties égales. L'écart interquartile ($Q3 - Q1$) contient 50% des observations centrales.

- Boîte à moustaches : représente graphiquement la valeur minimale, Q1, la médiane, Q3 et la valeur maximale. Elle permet de visualiser la dispersion, l'asymétrie et de comparer plusieurs séries.

**6. Paramètres de forme**

Les moments caractérisent la forme globale d'une distribution (symétrie, aplatissement).

- Moments centrés : moyenne des puissances des écarts à la moyenne, utilisés pour mesurer l'asymétrie (moment d'ordre 3).

- Moments absolus : moyenne des valeurs absolues des écarts à la moyenne, moins sensibles aux valeurs extrêmes.

- Vérification de la symétrie : importante pour choisir les méthodes statistiques.
  Comparer moyenne, médiane et mode : différences importantes → forte asymétrie.
  Observation graphique : histogrammes ou boîtes à moustaches.
  Coefficients d'asymétrie basés sur les moments centrés pour une mesure quantitative.

Mise en oeuvre python :

Lors de cette séance, j'ai travaillé sur les résultats du premier tour de l'élection présidentielle de 2022 à partir d'un fichier CSV issu du site data.gouv.fr. J'ai commencé par importer les bibliothèques nécessaires, notamment NumPy, Pandas et Matplotlib, afin de pouvoir manipuler les données, réaliser des calculs statistiques et produire des graphiques.

J'ai ensuite chargé le fichier CSV contenant les résultats électoraux à l'aide de Pandas. Une fois le tableau importé, j'ai identifié automatiquement les variables quantitatives du jeu de données. Pour cela, j'ai analysé le type de chaque colonne et sélectionné uniquement celles correspondant à des valeurs numériques (entiers et nombres réels).

Sur ces variables quantitatives, j'ai calculé plusieurs indicateurs statistiques que j'ai regroupés sous forme de tableaux : la moyenne, la médiane, le mode, l'écart-type, l'écart absolu moyen, l'étendue, l'écart interquartile, l'écart interdécile.

Ces résultats permettent de résumer l'information contenue dans les données et de mieux comprendre la dispersion et la variabilité des effectifs électoraux selon les départements.

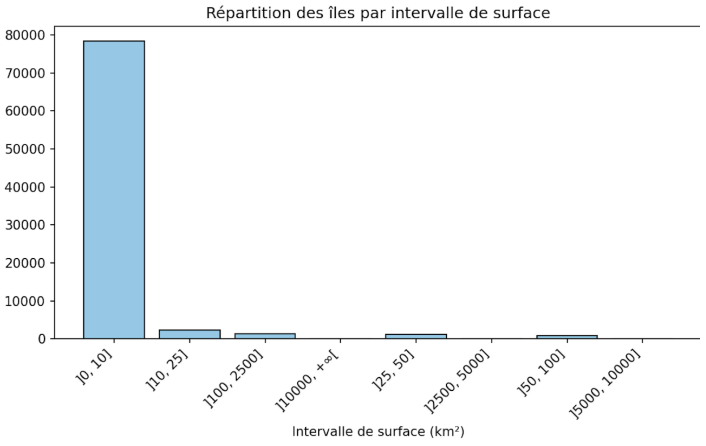
Les moyennes et médianes mettent en évidence des écarts importants entre les départements, ce qui s'explique par des différences démographiques. Les écarts-types et les étendues sont relativement élevés pour certaines variables, ce qui montre une forte dispersion des données. Les écarts interquartiles et interdéciles confirment que la distribution n'est pas homogène et que certains départements se distinguent nettement par des effectifs très supérieurs à la majorité des autres.

Dans une seconde partie, j'ai travaillé sur un autre jeu de données portant sur la surface des îles, issu du fichier island-index.csv. J'ai importé ce fichier avec Pandas puis j'ai classé la

variable Surface (km²) en intervalles afin de transformer une variable quantitative continue en variable qualitative ordonnée.

J'ai défini plusieurs classes de surface (par exemple ]0,10], ]10,25], ]25,50], etc.) et attribué chaque île à une catégorie en fonction de sa superficie. Cette nouvelle variable a ensuite été ajoutée au tableau sous le nom Classe_Surface.

J'ai ensuite compté le nombre d'îles appartenant à chaque intervalle et représenté ces résultats sous la forme d'un diagramme en barres, que j'ai enregistré sous forme d'image.




Le diagramme en barres montre que la majorité des îles possède une surface relativement faible. Les classes correspondant aux petites superficies regroupent le plus grand nombre d'îles, tandis que les îles de très grande surface sont beaucoup plus rares. Cette représentation permet de visualiser clairement la répartition des îles selon leur taille et met en évidence une forte asymétrie de la distribution. Ce type de graphique est particulièrement adapté pour comparer des effectifs par catégorie et synthétiser l'information de manière lisible.

Grâce à ce travail, j'ai appris à manipuler des jeux de données réels avec Python, à calculer des indicateurs statistiques descriptifs et à représenter graphiquement des résultats. L'utilisation conjointe de Pandas et Matplotlib m'a permis de produire des tableaux synthétiques et des graphiques clairs, facilitant l'analyse et l'interprétation des données d'un point de vue statistique.

## Séance 4

Objectifs :

- Savoir afficher une distribution statistique. Ce savoir est utilisé pour comparer une distribution observée avec une distribution théorique.

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 choix entre une distribution statistique discrète ou continue dépend de la nature du caractère étudié et de l'objectif de l'analyse. Les critères de décision sont :

La nature du phénomène étudié :
**-** Pour modéliser des événements comptables (ex. : nombre de naissances, nombre d'accidents), une distribution discrète est adaptée.
- Pour modéliser un phénomène mesurable en continu (ex. : température, altitude, distance), une distribution continue est préférable.

La forme de la distribution empirique :
L'analyse de l'histogramme ou du nuage de points des données permet de suggérer la loi théorique la plus appropriée (symétrie, asymétrie, concentration autour d'une valeur).

La connaissance et interprétation des caractéristiques :
Le choix dépend également de la loi qui offre la meilleure adéquation entre ses paramètres (moyenne, variance…) et les caractéristiques observées des données.

Le nombre de paramètres des lois :
Certaines lois de probabilité comportent plus ou moins de paramètres. Une loi avec davantage de paramètres peut parfois s'adapter plus facilement à la distribution observée.

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

En géographie, qui étudie des phénomènes naturels, sociaux et économiques, un large éventail de lois de probabilité est utilisé pour modéliser la répartition d'un phénomène.

Lois continues (pour les mesures) :

- Loi normale (ou de Gauss) : Modélise les phénomènes centrés autour d'une moyenne avec des valeurs extrêmes moins fréquentes. Exemples : températures moyennes, altitudes, revenus dans des populations homogènes.

- Loi log-normale : Adaptée aux phénomènes strictement positifs et fortement dispersés. Exemples : taille des villes, débits de cours d'eau, surfaces.
- Loi de Pareto : Modélise les phénomènes concentrés spatialement ou socio-économiquement, où une petite partie de la population détient une grande part du phénomène. Exemples : répartition de la richesse, taille des villes dominantes.
- Loi exponentielle : Utilisée pour modéliser la distance ou le temps entre deux événements aléatoires successifs, avec probabilité constante dans le temps ou l'espace. Exemples : temps entre séismes, distance entre commerces.
- Loi uniforme : Référence théorique pour des situations où toutes les valeurs d'un intervalle ont la même probabilité. Elle modélise une répartition homogène.
- Loi gamma : Appliquée aux phénomènes naturels liés au climat ou à l'hydrologie. Exemples : intensité des précipitations, durée des sécheresses, débits de rivières.

**Lois discrètes pour les dénombrements :**

- Loi de Poisson **:** Modélise le nombre d'événements rares sur une unité d'espace ou de temps donnée. Exemples : nombre de séismes dans une région, incendies, accidents sur un tronçon de route.

D'autres lois, comme Zipf ou Zipf-Mandelbrot, sont utilisées en géographie urbaine et démographie pour décrire hiérarchies et classements (ex. : relation entre le rang d'une ville et sa taille), mais elles restent moins fréquemment employées que les lois fondamentales présentées ci-dessus.

<u>Mise en oeuvre python :</u>

Lors de cette séance, j'ai exploré différentes lois statistiques discrètes et continues à l'aide de Python et des bibliothèques NumPy, Matplotlib et SciPy. L'objectif était de visualiser graphiquement ces distributions et de calculer leurs paramètres principaux (moyenne et écart-type).

J'ai donc représenté les lois discrètes : Dirac, uniforme discrète, binomiale, Poisson, Zipf-Mandelbrot et les lois continues : normale, log-normale, uniforme continue, chi², Pareto

Chaque loi a été tracée dans un graphique adapté : stem pour les lois discrètes et plot pour les lois continues.

**Les lois discrètes**

- Loi de Dirac :

  La distribution est concentrée sur une seule valeur (x=0). Cela représente un événement certain. La moyenne est exactement 0 et l'écart-type est nul, ce qui

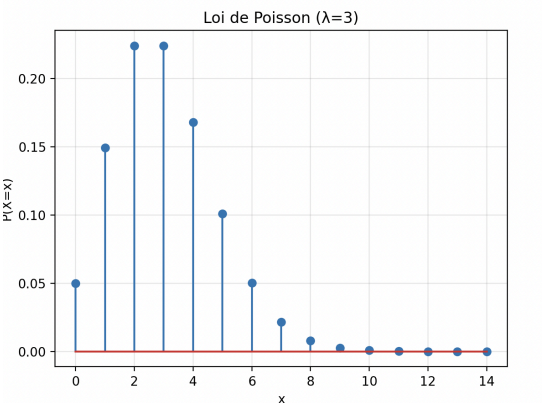
correspond parfaitement à la définition théorique de cette loi.

- Loi uniforme discrète :

  Tous les résultats possibles (de 1 à 6) ont la même probabilité. Le graphique montre que les barres sont de hauteur identique, ce qui illustre que chaque issue est également probable. La moyenne est au milieu de l'intervalle et l'écart-type reflète la dispersion uniforme autour de cette moyenne.

- Loi binomiale :

  Cette distribution modélise le nombre de succès dans un nombre fixe d'essais indépendants. Le graphique montre que la probabilité est centrée autour de 5, ce qui correspond à n*p. La distribution est symétrique pour p=0.5 et l'écart-type illustre l'étalement autour de la moyenne.




- Loi de Poisson :

  La probabilité augmente pour les petites valeurs de k et atteint un maximum autour de k=3, puis diminue. Cela montre que le paramètre λ=3 définit à la fois la moyenne et la variance. La distribution est asymétrique à droite, comme attendu.

- Loi de Zipf-Mandelbrot :

  La probabilité diminue fortement avec le rang. La première valeur a la probabilité la plus élevée, et les valeurs suivantes décroissent rapidement. Cette distribution illustre une forte inégalité entre les valeurs.

**Les lois continues**

- Loi normale :

  Le graphique montre la classique courbe en cloche centrée sur 0. La moyenne est 0 et l'écart-type 1. La distribution est symétrique et les probabilités diminuent rapidement à mesure que l'on s'éloigne de la moyenne.

- Loi log-normale :

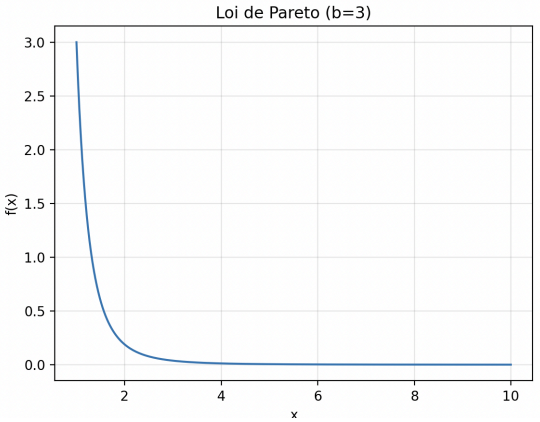
  La distribution est fortement asymétrique à droite. La plupart des valeurs sont proches de 1, mais il existe une longue queue vers les grandes valeurs, ce qui indique que des valeurs élevées sont rares mais possibles.

- Loi uniforme continue :

  Toutes les valeurs dans l'intervalle [0,1] ont la même probabilité. Le graphique est plat, ce qui illustre la constance de la densité de probabilité. La moyenne est au

centre de l'intervalle et l'écart-type reflète l'étalement uniforme.

- Loi du $\chi^2$ :
  La distribution est asymétrique à droite avec un maximum proche de 3. Elle est utilisée notamment pour les tests statistiques. L'écart-type est supérieur à la moyenne, ce qui confirme l'asymétrie.




- Loi de Pareto :
  La distribution montre qu'il existe de nombreuses petites valeurs et quelques valeurs très grandes (longue queue à droite). Cela correspond à des phénomènes où la majorité des observations sont faibles mais où des valeurs extrêmes sont possibles.

| Loi | Moyenne | Écart-type |
|---|---|---|
| Dirac | 0 | 0 |
| Uniforme discrète | 4 | 1.71 |
| Binomiale B(10,0.5) | 5 | 1.58 |
| Poisson λ=3 | 3 | 1.73 |
| Zipf a=2 | 1.64 | 0.83 |
| Normale N(0,1) | 0 | 1 |
| Log-normale s=0.954 | 1.31 | 1.37 |
| Uniforme continue [0,1] | 0.5 | 0.29 |
| Chi² df=5 | 5 | 3.16 |
| Pareto b=3 | 1.5 | 0.90 |

Les valeurs sont calculées automatiquement avec la fonction calcul_moyenne_et_ecart().

J'ai pu visualiser et comparer différentes lois statistiques, comprendre leur forme et leur dispersion. Les lois discrètes montrent comment la probabilité est concentrée sur des valeurs entières, tandis que les lois continues présentent des distributions lisses sur un intervalle. Les moyennes et écarts-types obtenus confirment les propriétés théoriques de chaque loi. Les graphiques m'ont permis de mettre en évidence la symétrie, l'asymétrie, les queues longues ou courtes et d'illustrer la différence entre des distributions uniformes,

concentrées ou fortement inégalitaires. Cette séance m’a permis de lier théorie statistique et pratique Python de manière concrète.

## Séance 5

Objectifs :
- Manipuler harmonieusement les fonctions natives avec les méthodes Pandas
- Comprendre les trois théories permettant de valider un résultat en analyse de données

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 consiste à prélever un sous-ensemble représentatif d'une population afin d'en déduire des conclusions sur l'ensemble. Il est souvent impossible ou trop coûteux d'étudier la population entière. L'objectif est de généraliser les résultats tout en maîtrisant l'erreur d'échantillonnage.

- Méthodes aléatoires : tirage au sort simple, avec ou sans remise. Elles assurent la représentativité et permettent l'application rigoureuse des lois statistiques.
- Méthodes non aléatoires : échantillonnage systématique, par quotas, Monte Carlo, adaptées quand la base de sondage est incomplète.

Le choix de la méthode dépend de l'objectif de l'étude, du coût, de la taille de la population et du niveau de précision souhaité.

**2. Comment définir un estimateur et une estimation ?**

Un estimateur est une fonction des données observées permettant d'évaluer un paramètre inconnu de la population (moyenne, variance, proportion…)
L'estimation est la valeur numérique obtenue à partir de l'échantillon. Par exemple, la moyenne de l'échantillon X¯\bar{X}X¯ est un estimateur de la moyenne de la population μ\muμ, et la valeur calculée x¯x̄x¯ constitue une estimation.

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

- Intervalle de fluctuation : zone où la fréquence observée est susceptible de se situer quand la proportion théorique ppp est connue. C'est un outil d'échantillonnage pour mesurer les variations dues au hasard.
- Intervalle de confiance : estimation du paramètre inconnu avec un certain niveau de certitude. Il sert à estimer un paramètre inconnu.

Le premier se calcule à partir d'un paramètre connu, le second permet d'estimer un paramètre inconnu.

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

Le biais d'un estimateur est la différence entre son espérance mathématique et la vraie valeur du paramètre :

Biais=E(θ^)−θBiais = E(\hat{\theta}) - \thetaBiais=E(θ^)−θ

Un estimateur est sans biais si cette différence est nulle. Le biais traduit une erreur systématique. Un bon estimateur doit être sans biais, convergent et de variance minimale.

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

Une statistique calculée sur l'ensemble de la population est une statistique exhaustive. Elle ne repose plus sur l'inférence mais sur une mesure complète des paramètres. Avec le big data, cette approche devient plus fréquente, car les bases massives permettent une observation quasi exhaustive, limitant le recours aux inférences.

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

Le choix d'un estimateur nécessite un compromis entre :

- Fidélité : absence de biais
- Précision : faible variance
- Robustesse : résistance aux valeurs aberrantes
- Efficacité : minimisation de l'erreur quadratique moyenne

  Selon les théorèmes de Rao-Blackwell et Lehmann-Scheffé, un estimateur sans biais et de variance minimale est optimal

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

- Estimation ponctuelle : une seule valeur, ex. moyenne de l'échantillon.
- Estimation par intervalle : plage de valeurs avec forte probabilité d'inclure le paramètre (intervalle de confiance).

Les méthodes peuvent être par vraisemblance, par moments ou Bayésiennes (intégrant une connaissance a priori). Le choix dépend du modèle probabiliste, de la nature du paramètre et de la disponibilité des données.

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

Les tests statistiques vérifient la validité d'une hypothèse sur un paramètre de la population et permettent de décider avec un risque d'erreur contrôlé si l'observation confirme ou infirme l'hypothèse.

- Tests paramétriques : supposent une loi continue (Student, Khi-deux, Fisher…).
- Tests non paramétriques : utilisés lorsque la loi de distribution n'est pas connue.

**Création d'un test** :

1. Formulation des hypothèses H0H_0H0 et H1H_1H1
2. Choix d'une statistique test
3. Détermination du seuil de signification α\alphaα
4. Décision selon la valeur observée

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

Les critiques concernent :

- La dépendance aux hypothèses théoriques (normalité, indépendance, homogénéité des variances).
- La sensibilité aux biais d'échantillonnage et aux valeurs aberrantes.
- L'interprétation abusive des seuils de signification
- La pertinence réduite à l'ère du big data, où les populations peuvent être observées quasi totalement

Cependant, l'inférence reste essentielle pour tirer des conclusions généralisables à partir de données partielles. Ces critiques invitent surtout à une lecture critique et réfléchie des résultats plutôt qu'à abandonner la méthode.

<u>Mise en oeuvre python :</u>

Lors de cette séance, j'ai travaillé sur un fichier contenant 100 échantillons aléatoires issus d'une population mère de 2185 individus. Les réponses se répartissent en trois catégories : « Pour », « Contre » et « Sans opinion ». Mon objectif était de comprendre comment les échantillons reflètent la population, de calculer des intervalles de fluctuation et de confiance, puis de vérifier la normalité de certaines distributions.

En commençant par les moyennes des 100 échantillons, j'ai pu observer que les valeurs arrondies sont très proches des proportions réelles de la population mère. En effet, la moyenne des individus « Pour » est de 391, celle des « Contre » est de 416 et celle des

« Sans opinion » est de 193. En comparant ces chiffres aux proportions effectives dans la population mère, on remarque que les fréquences observées dans les échantillons (0,39 pour « Pour », 0,42 pour « Contre » et 0,19 pour « Sans opinion ») sont quasiment identiques aux fréquences théoriques. Cela montre que, même avec un nombre limité de données aléatoires, l'échantillonnage permet d'obtenir une estimation fiable et représentative de la population réelle.

J'ai ensuite calculé les intervalles de fluctuation à 95 % pour chaque catégorie. Ces intervalles indiquent la plage dans laquelle on peut attendre que les fréquences observées d'un échantillon aléatoire se situent. Les fréquences obtenues pour le premier échantillon se trouvent toutes à l'intérieur de ces intervalles, ce qui confirme que la répartition des individus dans les échantillons est cohérente avec celle de la population mère. On peut donc conclure que l'échantillonnage aléatoire, malgré ses variations inhérentes, fournit des résultats statistiquement fiables.

Dans un deuxième temps, j'ai analysé le premier échantillon individuellement. En calculant les fréquences des différentes catégories et les intervalles de confiance à 95 %, j'ai pu vérifier que les proportions observées sont compatibles avec celles de la population. Les intervalles de confiance englobent les fréquences effectives, ce qui signifie que même un échantillon unique peut offrir une bonne estimation des paramètres de la population. Cela illustre la puissance de la théorie de l'échantillonnage et son rôle central dans les études statistiques.

Pour compléter cette analyse, j'ai appliqué le test de normalité de Shapiro-Wilk sur deux fichiers de données. Pour le fichier « Loi-normale-Test-1.csv », la p-value obtenue est très faible, ce qui entraîne le rejet de l'hypothèse de normalité. Les données ne suivent donc pas une distribution normale et présentent probablement une asymétrie notable. En revanche, le fichier « Loi-normale-Test-2.csv » a une p-value supérieure à 0,05, ce qui signifie que l'hypothèse de normalité ne peut pas être rejetée. Les valeurs de ce second fichier sont donc compatibles avec une loi normale. Cette distinction est cruciale pour déterminer quelles méthodes statistiques sont appropriées pour analyser les données.

Enfin, j'ai réalisé une analyse descriptive des distributions non normales. Pour le fichier « Loi-normale-Test-1.csv », la moyenne est nettement différente de la médiane, ce qui indique une forte asymétrie et suggère que les données pourraient suivre une loi exponentielle ou une autre loi asymétrique. Pour « Loi-normale-Test-2.csv », la moyenne et la médiane sont proches et l'écart-type est cohérent avec l'étendue des données, ce qui est caractéristique d'une distribution normale.

En conclusion, cette séance m'a permis de mettre en pratique les concepts de la théorie de l'échantillonnage, de l'estimation et de la prise de décision statistique. Les échantillons analysés fournissent des estimations fiables de la population mère, les intervalles de fluctuation et de confiance valident la représentativité des données, et le test de

Shapiro-Wilk permet de déterminer la normalité des distributions étudiées. L'ensemble de ces étapes montre l'importance d'une approche structurée pour analyser des données réelles et tirer des conclusions robustes.

## Séance 6 :

Objectifs :
- Manipuler des fonctions locales et comprendre la nécessité de factoriser son code en une liste de fonctions ou de procédures exécutant une tâche unique
- Créer des fonctions locales spécifiques au traitement d'un problème
- Comprendre l'analyse de variables qualitatives ordinale

Questions de cours :

**1. Qu'est-ce qu'une statistique ordinale ? À quelle autre statistique catégorielle s'oppose-t-elle ?**

Une statistique ordinale correspond aux méthodes basées sur le classement des objets ou individus, c'est-à-dire sur l'ordre des observations plutôt que sur leurs valeurs absolues. Les observations sont rangées selon leurs rangs X(1)≤⋯≤X(n)X_{(1)} \le \cdots \le X_{(n)}X(1)≤⋯≤X(n). Elle s'oppose aux statistiques nominales, qui se contentent de répartir les individus en catégories sans hiérarchie ni ordre.

La statistique ordinale utilise des variables ordinales, c'est-à-dire des variables qualitatives pour lesquelles un ordre naturel peut être établi, croissant ou décroissant. L'ordre croissant est généralement privilégié, sauf pour certains cas particuliers comme l'analyse selon la loi rang-taille.

Dans le domaine spatial, ce type de statistique permet de matérialiser une hiérarchie spatiale, car de nombreux phénomènes géographiques se prêtent naturellement au classement : taille des villes, intensité de phénomènes naturels (crues, séismes), dynamisme socio-économique, etc. Le classement permet d'identifier les entités dominantes, intermédiaires ou marginales, révélant ainsi l'organisation hiérarchique d'un territoire.

**2. Quel ordre est à privilégier dans les classifications ?**

L'ordre à privilégier est l'ordre croissant, également appelé ordre naturel. Il facilite l'analyse des rangs, la détection des valeurs extrêmes et l'étude de certaines distributions, comme la valeur maximale d'une série.

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

- La corrélation des rangs mesure la similarité globale entre deux séries ordonnées, en comparant les rangs attribués à chaque individu. Les principaux indicateurs sont le coefficient de Spearman et le coefficient $\tau$ de Kendall, qui permettent de savoir si les

rangs sont proches, inversés ou indépendants.

- La concordance de classements s'intéresse au nombre de paires concordantes ou discordantes entre deux classements. Une concordance est dite « complète » si toutes les paires respectent le même ordre, et « nulle » si le nombre de concordances est égal au nombre de discordances.

Ainsi, la corrélation mesure une proximité globale entre deux ordres, tandis que la concordance examine la cohérence paire par paire.

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

- Le test de Spearman utilise directement les rangs et calcule une corrélation à partir des différences (ui−vi)2(u_i - v_i)^2(ui−vi)2 entre deux séries. Il est sensible aux ex-aequo et sa distribution peut être approximée par une loi normale pour n>30n > 30n>30.

- Le test de Kendall se base sur le nombre de paires concordantes et discordantes, et compare l'ordre de chaque paire d'individus. Conceptuellement plus simple, il peut se généraliser facilement à plusieurs classements.

En résumé, Spearman mesure la proximité quantitative des rangs, tandis que Kendall mesure la cohérence qualitative de l'ordre. Les deux tests sont complémentaires pour analyser les hiérarchies spatiales.

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

Le coefficient de Goodman-Kruskal (Γ) évalue la force de l'association d'ordre entre deux variables ordinales, en comparant le nombre de paires concordantes (Na) et discordantes (Nd). Il varie de -1 à +1 :

- Γ = +1 : concordance parfaite
- Γ = -1 : inversion totale
- Γ = 0 : absence d'association
  Il est conceptuellement proche du coefficient τ de Kendall.

Le coefficient de Yule est un cas particulier du Γ, réservé aux tableaux de contingence 2 x 2, pour des variables dichotomiques (oui/non, présent/absent). Il indique également l'association entre -1 et +1.

En résumé, Goodman-Kruskal fournit une mesure générale de l'association d'ordre sur des couples classés, tandis que Yule est un outil ciblé pour des variables binaires. Ces coefficients

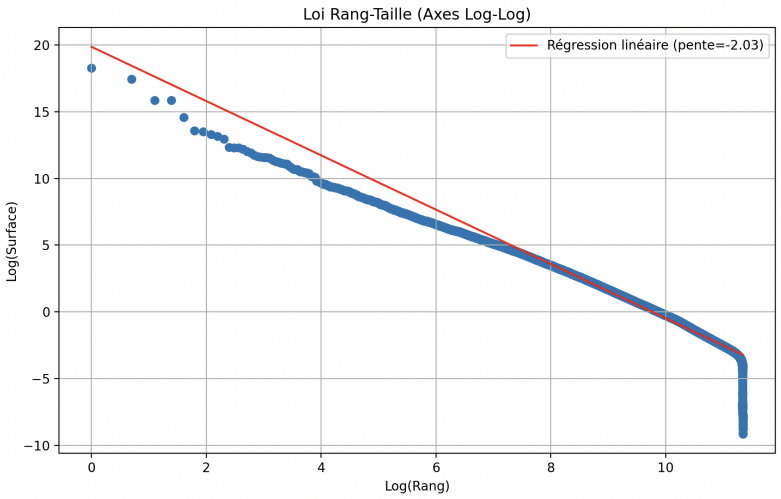
permettent de quantifier la force des relations entre variables catégorielles et d'interpréter les hiérarchies ou dépendances dans des données géographiques.

Mise en oeuvre Python :

Lors de cette séance, j'ai travaillé sur deux ensembles de données principaux : un fichier relatif aux îles du monde et un fichier sur les populations mondiales entre 2007 et 2025. L'objectif était d'analyser les distributions de surface, d'étudier la loi de rang-taille, et d'évaluer la concordance des rangs des populations et des densités sur plusieurs années.

En commençant par les îles, j'ai extrait la colonne « Surface (km²) » et ajouté les superficies des masses continentales pour compléter l'analyse. La liste des surfaces a été triée par ordre décroissant et visualisée dans un premier graphique sur des axes normaux. On observe que les plus grandes superficies dominent largement, tandis que les îles plus petites sont beaucoup plus nombreuses et concentrées sur des valeurs faibles. Cette distribution suit une loi de type rang-taille, où les territoires les plus importants sont peu nombreux, et la majorité des entités sont de petites tailles.

Pour faciliter l'interprétation des données, j'ai ensuite transformé les axes en échelle logarithmique. Sur le graphique log-log, le nuage de points présente une tendance décroissante claire, confirmée par la régression linéaire qui donne une pente négative indicative d'une loi proche de Zipf. Cette représentation met en évidence la proportionnalité approximative entre rang et taille en échelle logarithmique et justifie l'utilisation de tests basés sur les rangs pour analyser la relation entre surface et autres caractéristiques des îles.




Dans la seconde partie, j'ai étudié les populations mondiales. J'ai comparé les rangs des États en termes de population et de densité pour les années 2007 et 2025. L'analyse des corrélations de rang à l'aide de Spearman et Kendall révèle une forte concordance des classements. Le coefficient de Spearman est très proche de 1, ce qui indique que les pays les mieux classés en population le sont également en densité, et le coefficient de Kendall confirme cette tendance avec une forte concordance des rangs. Les p-values associées sont extrêmement faibles, montrant que ces corrélations sont statistiquement significatives.

Cette observation met en évidence la robustesse des hiérarchies entre États et la stabilité des classements sur la période étudiée.

J'ai ensuite appliqué ces analyses sur toutes les années de 2007 à 2025. Les coefficients de corrélation Spearman et Kendall sont restés très élevés, illustrant que les classements des populations sont globalement stables sur le long terme. Le graphique de l'évolution annuelle des coefficients montre une tendance quasi-constante, confirmant que les rangs relatifs des États ne subissent que de légères variations sur cette période.

Enfin, une analyse plus spécifique sur les îles, comparant leur surface et la longueur du trait de côte, a permis de calculer la corrélation des rangs. Le coefficient de Spearman proche de 0,94 et le coefficient de Kendall autour de 0,85 indiquent une forte relation monotone entre surface et trait de côte. Les îles les plus vastes possèdent naturellement un trait de côte plus long, mais la distribution n'est pas parfaitement linéaire, ce qui justifie l'usage des mesures de corrélation des rangs plutôt que sur les valeurs absolues.

En conclusion, cette étude illustre l'importance des analyses basées sur les rangs pour comprendre les distributions inégalement réparties, comme les surfaces d'îles ou les populations d'États. Les représentations graphiques (axes normaux et log-log) permettent de visualiser la loi rang-taille et de détecter les tendances globales, tandis que les coefficients de Spearman et Kendall offrent un outil robuste pour comparer la stabilité et la concordance des classements. L'ensemble de ces résultats démontre que, malgré la variabilité des valeurs absolues, les relations structurelles entre entités sont solides et statistiquement significatives.

## Réflexion sur les humanités numériques

Ayant suivi une double licence Histoire-Géographie, je n'ai jamais eu l'occasion de suivre des cours d'outils informatiques, ces compétences n'étant pas intégrées dans le cursus. À l'époque, je trouvais cela frustrant, car il me semblait que la maîtrise de logiciels ou de langages de programmation aurait pu constituer un atout précieux, notamment pour l'analyse de données ou la cartographie. Aujourd'hui, confronté aux sciences des données et aux humanités numériques, je comprends mieux la complexité de ce domaine et je réalise qu'il ne s'agit pas seulement d'apprendre des commandes ou des formules, mais bien de développer une manière de penser et d'exploiter les données de façon rigoureuse et efficace.

Ce qui m'a le plus surpris, c'est la puissance de ces outils. Avec Python, il est possible de traiter d'immenses volumes de données, d'automatiser des analyses, de produire rapidement des graphiques et des visualisations, et de mettre en lumière des tendances qui seraient quasiment impossibles à détecter manuellement. Pour un géographe, ces compétences ouvrent des perspectives inédites : elles permettent d'analyser finement les dynamiques territoriales, de croiser différents types d'informations et de produire des analyses à la fois quantitatives et visuelles.

Même si l'apprentissage est exigeant et parfois intimidant, je mesure à quel point ce domaine peut enrichir notre approche des sciences sociales et de la géographie. Ce qui me paraissait auparavant comme une lacune dans mon parcours devient aujourd'hui une opportunité passionnante : comprendre et manipuler les données permet d'aller au-delà de la simple description des phénomènes pour en saisir les mécanismes, les interactions et les tendances à grande échelle.

## Retour général sur le cours

Le cours d'analyse de données et d'humanités numériques m'a paru assez complexe, notamment à cause de la pédagogie inversée. Pour des étudiants comme moi, qui n'avaient jamais utilisé Python ou d'autres outils informatiques, se retrouver confronté directement aux exercices et devoir progresser en autonomie a été un véritable défi. La matière demandait non seulement de comprendre de nouveaux concepts, mais aussi de les mettre en pratique sans toujours bénéficier d'un encadrement immédiat, ce qui a rendu l'apprentissage plus difficile. Heureusement, le travail en équipe pendant les cours nous a permis de surmonter certaines difficultés et de progresser plus efficacement.

Je comprends néanmoins que ce type de cours reste difficile à enseigner. Les problématiques propres à l'université, comme des ordinateurs qui ne fonctionnent pas ou des configurations différentes selon les étudiants, compliquent davantage l'enseignement, surtout pour une matière pratique qui nécessite des manipulations informatiques. Ce n'est

pas comparable à un cours où l'on doit lire des textes ou analyser des documents : ici, la technique et la mise en pratique sont centrales, ce qui multiplie les sources de difficultés pour les étudiants comme pour les enseignants.

Par ailleurs, la charge de travail en dehors des cours m'a semblé particulièrement lourde, surtout au regard de l'importance de la matière dans le parcours GAED, qui reste secondaire pour la majorité d'entre nous. Bien que le contenu soit extrêmement intéressant et que les compétences acquises soient puissantes et utiles, le temps nécessaire pour assimiler les notions et produire des analyses dépassait parfois ce que l'on pourrait attendre d'une matière de ce type dans le cursus.

Pour autant, malgré ces difficultés, j'ai pu apprécier le potentiel offert par Python et les sciences des données. La capacité à produire rapidement de nombreux graphiques, à analyser de vastes ensembles de données et à en tirer des conclusions pertinentes montre à quel point ces outils peuvent être précieux pour un géographe. La difficulté du cours est donc contrebalancée par la richesse et la puissance des méthodes enseignées.

Enfin, je tiens à vous remercier pour le temps que vous nous avez accordé et pour toutes les notions que vous avez pu nous transmettre, qui nous permettront de progresser dans l'utilisation de ces outils à l'avenir.