CHATELLIER Clémence
M1 – GAED, Parcours ACA

RAPPORT D'ACTIVITES – PYTHON, DEBUTANT
ANALYSE DE DONNEES

Maxime FORRIEZ

Sorbonne Université

Décembre 2025

SEANCE 1

I- Questions de cours

→ 1 – **Quel est le positionnement de la géographie par rapport aux statistiques ?**
La géographie étant une discipline qui se cherche toujours, il est fréquent qu'elle méprise les définitions mathématiques élémentaires de la statistique sous prétexte une cela n'entre traditionnellement pas dans son champ disciplinaire. Pourtant, elle produit des données massives que seul l'outil statistique permet d'étudier. Ainsi, les relations entre les deux disciplines sont très souvent tendues et complexes. Cette situation paradoxales

→ 2 – **Le hasard existe-t-il en géographie ?**
Le hasard n'existe pas car il existe une cause à tout, il n'est qu'une version philosophique.
Dans les modélisations mathématiques, il existe deux types de hasard : le hasard bénin et le hasard sauvage. Le premier possède une distribution de probabilité dite normale, le second correspond à une distribution de probabilité moins fréquente. Dans le cadre de la géographie, dès le début du XXe, deux grandes lois de probabilité interviennent : la loi normale et la loi de V. Pareto.

→ 3 – **Quels sont les types d'information géographique ?**

L'information géographique se décompose en deux séries statistiques possibles. D'une part, il peut s'agir pour une entrée territoriale claire et précise d'étudier tout ce qui peut caractériser l'ensemble délimité par des éléments de géographie humaine ou de géographie physique. D'autre part, il peut s'agir d'étudier la morphologie même des ensembles délimités. De fait la géométrie des ensembles géographiques peut faire l'objet d'une étude statistique.

→ 4 – **Quels sont les besoin de la géographie au niveau de l'analyse de données ?**
L'analyse de donnée repose sur les probabilités et les statistiques. A la différence de l'étape de la production, il s'agit d'étudier la structure interne des données analysées. L'analyse de données permet de confronter les résultats obtenus avec la méthodologie de production des données et avec ce que l'on connaît du phénomène étudié.

→ 5 – **Quelles sont les types de visualisation de données en géographie ? Comment choisir celles-ci ?**

Il existe trois types de visualisation de données en géographie :

- Quantitatif : s'applique pour une analyse factorielle en composantes principales
- Qualitatif : s'applique à une analyse factorielle des correspondances ou à une analyse factorielle des correspondances multiples
- Mélange : s'applique à des options dans de nombreux logiciels de statistique.

→ 6 – **Quelles sont les différences entre la statistique descriptive et la statistique explicative ?**
Statistique descriptive : s'applique en particulier aux tableaux individus contenant k variables dans lesquels toutes les variables jouent le même rôle. Il n'y a pas de variable à expliquer. Il s'agit de résumer le tableau des variables et de comprendre les grandes dimensions du phénomène étudié. L'objectif est

de visualiser et de classer les données.
Statistique explicative : objectif de relier une variable à expliquer à des variables explicatives. Il s'agit ici d'ajuster les données disponibles un modèle dont la forme dépend de la nature de la réponse. Si la réponse est numérique

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

Il existe plusieurs méthodes d'analyse de données possibles. Ces dernières se distinguent en trois grandes classes :

- les méthodes descriptives
- les méthodes explicatives
- les méthodes de prévision

→ 8 – **Comment définiriez-vous : population statistique ? Individu statistique ? Caractères statistiques ? Modalités statistiques ? Quels sont les types de caractères ? Existe-t-il une hiérarchie entre eux ?**

La population statistique correspond à un ensemble au sens mathématique du terme. Elle peut être spatiale comme le nombre d'habitants d'un territoire par exemple ou encore non spatiale à l'instar du personnel d'une entreprise.
En ce qui concerne l'individu statistique, il s'agit d'un élément de la population statistique. On peut aussi l'appeler unité statistique. Dans ce cadre, les données géographiques ont deux particularités : premièrement les individus statistiques sont localisables et cartographiables, appelés spatiales. Deuxièmement, les individus statistiques sont-eux mêmes fréquemment composés d'un ensemble de personnes, d'entreprises, de points observables, de zones plus petites, de tronçons d'un réseau, appelés éléments de niveau inférieur. Deux types d'unités spatiales sont à noter : les unités primaires et les unités secondaires.
Les modalités statistiques correspondent aux valeurs prises par un caractère. Ces modalités doivent être incompatibles et exhaustives, l'objectif étant de caractériser l'appartenance, ou la non appartenance, d'un individu à une modalité.
Les modalités forment une partition du caractère, car elles sont exhaustives et disjointes.
Il existe deux types de caractères. Il peut s'agit soit d'une variable qualitative, soit d'une variable quantitative. Toutefois, le caractère devient une variable statistique, ou, pour les variables qualitatives, valeur aléatoire lorsqu'il fait l'objet d'une étude statistique. Il n'existe pas de hiérarchie entre eux.

→ 9 – **Comment mesurer une amplitude et une densité ?**

L'amplitude et la densité se mesurent à partir de la discrétisation des caractères quantitatifs. La série obtenue aura ensuite deux variables caractéristiques à savoir l'amplitude et la densité.
L'amplitude est la longueur $b - a$ avec a la valeur minimale de la classe et b la valeur maximale. Ele concerne toujours une classe. En ce qui concerne la densité, elle correspond au rapport entre l'effectif et l'amplitude de la classe décrivant une modalité. On appelle d la densité.

→ 10 – **A quoi servent les formules de Sturges et de Yule ?**

La formule de Sturges permet d'obtenir une valeur approximative du nombre de classes tandis que la formule de Yule permet aussi de calculer l'amplitude des classes par rapport à l'étendue de la série des

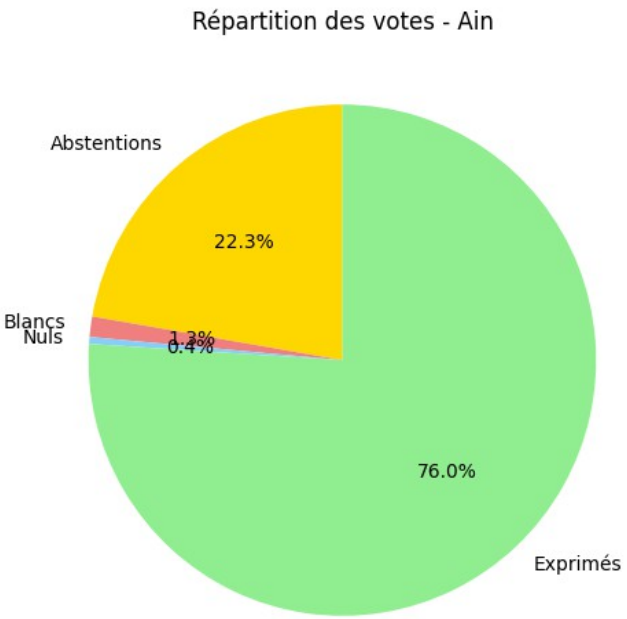
observations et du nombre de classes.

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

Un effectif correspond au nombre d'apparitions d'une variable dans la population.
La fréquence relative est le rapport entre l'effectif et l'effectif total. Il se calcule à partir d'une fonction.
La fréquence cumulée jusqu'à *k* modalités quant à elle est la somme des effectifs associés aux valeurs du caractère qui sont inférieures ou égales à *k*.
La distribution statistique empirique permet ed conclure su le type de loi de probabilité utilisée.

II- Graphiques, interprétation des résultats

Pour le rapport, j'ai inséré uniquement les graphiques de l'Ain, vous trouverez dans le portefolio les graphiques des autres départements.



Interprétation du graphique représentant le premier département : l'Ain concernant la répartition des votes.

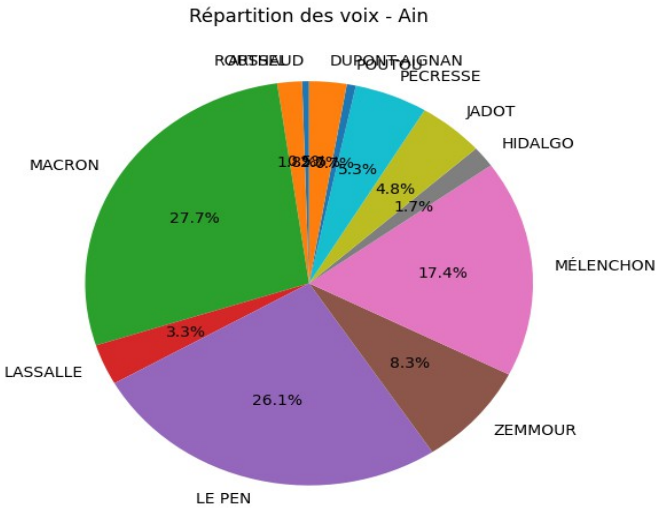
Observations principales (Département de l'Ain) :

1. Votes Exprimés : La très grande majorité des inscrits a voté, puisque les votes Exprimés représentent la plus grande part du cercle : 76,0%.
    - Ce chiffre est l'indicateur de la participation (taux de participation = 100% - Taux d'abstention$) pour ce département. Une participation de 76,0% est considérée comme

forte.

2. Abstentions : Le taux d'Abstention dans l'Ain est de 22,3%.
3. Votes Nuls et Blancs :
    - Les votes Blancs représentent 1,3% des inscrits.
    - Les votes Nuls représentent une très petite fraction, soit 0,4% des inscrits.

Conclusion : Le département de l'Ain a enregistré un taux de participation élevé (76,0%), supérieur à la moyenne nationale (qui était d'environ 73,7% au premier tour de 2022). Cela signifie qu'une écrasante majorité des électeurs inscrits se sont déplacés aux urnes.



Emmanuel Macron obtient un score pratiquement **identique** dans l'Ain par rapport à son score national (27,7% vs 27,8%).

- **Marine Le Pen** réalise une **meilleure performance** dans l'Ain que sur l'ensemble du territoire français (+2,9 points), terminant en **tête** dans le département (26,1%).
- **Jean-Luc Mélenchon** enregistre une **nette sous-performance** dans l'Ain par rapport à son score national (-4,6 points). Ce résultat suggère que l'Ain, département de l'Est de la France, est moins favorable à l'électorat de gauche radicale/Insoumis que la moyenne nationale.

Éric Zemmour enregistre également une **meilleure performance** dans l'Ain (+1,2 point), consolidant une tendance à droite et à l'extrême droite.

- **Valérie Pécresse** et **Nicolas Dupont-Aignan** obtiennent des scores légèrement supérieurs à leur moyenne nationale.
- Ces résultats montrent que l'électorat de l'Ain a une **forte affinité pour les candidats de droite, de l'extrême droite, et du centre** (Macron maintenant son score), tandis que les candidats de gauche et de l'écologie y sont moins représentés.

Observations clés :

- L'ensemble des candidats classés à gauche ou écologistes (Mélenchon, Jadot, Roussel) **sous-performent** significativement dans l'Ain.
- Le score particulièrement faible de Fabien Roussel (-1,3 point) est notable.

L'Ain est un département où la participation a été **forte (76,0%)** (voir graphique précédent), supérieure à la moyenne nationale.

Le vote dans l'Ain est nettement plus **à droite** que la moyenne nationale, favorisant Marine Le Pen et Éric Zemmour, tandis qu'il pénalise fortement Jean-Luc Mélenchon.

Le graphique en barres du département de l'Ain montre deux barres :

1. Inscrits (barre bleue) : La barre atteint légèrement plus de $400\ 000$. Estimons la valeur à environ 430 000 à 440 000 électeurs inscrits.
2. Votants (barre verte) : La barre atteint environ 330 000 à 340 000.

Nombre estimé d'inscrits dans l'Ain : approx 440 000

2. Positionnement dans l'Histogramme

L'histogramme de la Distribution du nombre d'inscrits par département utilise l'axe des X (horizontal) pour le nombre d'inscrits, gradué en millions (1e6).

- Notre valeur estimée de 440 000 inscrits correspond à 0,44 million (0.44 times 10^6).
- En regardant l'histogramme, 0.44 million se situe dans la classe juste après la barre de 0.25 million et avant la barre de 0.5 million.

La barre de l'histogramme qui couvre cette plage est la quatrième barre en partant de la gauche, qui se situe approximativement entre 0.4 million et 0.5 million.

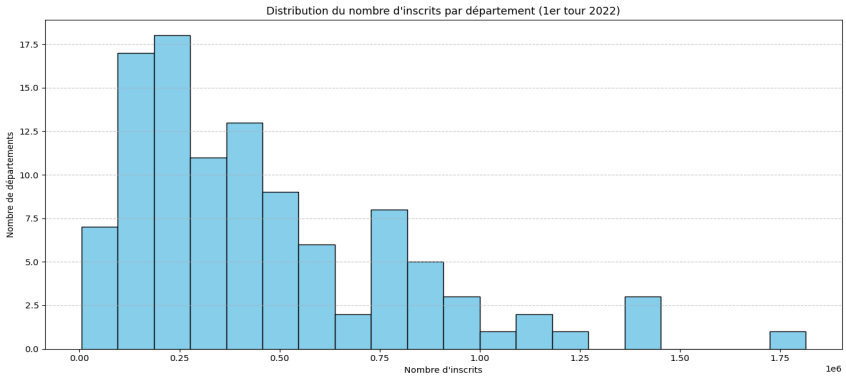
- Cette barre a une hauteur d'environ 9 départements.

3. Conclusion et Corrélation (Participation)

1. Taille du Département : Le département de l'Ain se situe dans la catégorie des départements de taille moyenne en France (avec environ 440 000 inscrits), loin des départements les plus peuplés (plus de 1 million d'inscrits).
2. Participation : Nous savons par le graphique circulaire précédent que l'Ain a eu un taux de participation élevé (76,0%) (Abstention de 22.3%).

Déduction sur la corrélation :

Puisque l'Ain est un département de taille moyenne avec un fort taux de participation, on peut faire l'hypothèse (qui nécessiterait une vérification sur l'ensemble des départements) que les départements de taille moyenne sont parmi ceux qui se mobilisent le plus pour ce type d'élection.



Profil Électoral du Département de l'Ain (1er Tour 2022)

Contexte Démographique et Mobilisation

Le département de l'Ain se caractérise par :

- Taille Moyenne : Avec environ 440 000 inscrits (selon le graphique en barres), l'Ain se situe dans la catégorie des départements de taille moyenne en France.
- Forte Participation : Le taux de participation (76,0%) est élevé, avec un taux d'abstention de 22,3%. Le département a montré une forte mobilisation de ses électeurs.
- Faibles Votes Non-Exprimés : Les votes blancs (1,3%) et nuls (0,4%) sont très faibles, confirmant que les électeurs ont massivement choisi d'exprimer un vote en faveur d'un candidat.

Orientation Politique Dominante

En comparant la répartition des voix dans l'Ain et en France entière, on identifie une nette tendance à droite et à l'extrême droite.

Pluralisme du Centre/Droite/Extrême Droite

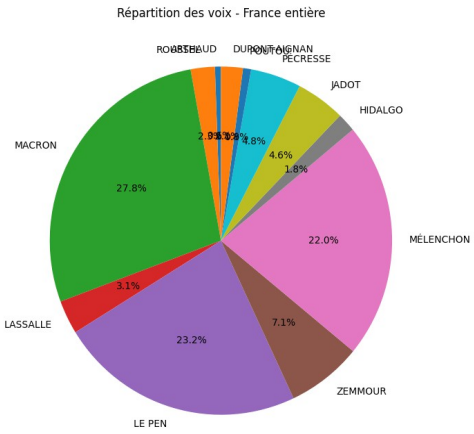
- Marine Le Pen en Tête : Elle est la première candidate dans le département (26,1%), réalisant une sur-performance significative par rapport à son score national (+2,9 points).

- Macron Stable : Emmanuel Macron maintient son score national (27,7% dans l'Ain contre 27,8% en France), indiquant que son électorat est fidèlement représenté, malgré le meilleur score de Le Pen.
- Sur-performance de Zemmour : Éric Zemmour réalise également une meilleure performance (+1,2 point), consolidant une tendance favorable à l'extrême droite.
- Droites Classiques en Légère Avance : Valérie Pécresse et Nicolas Dupont-Aignan obtiennent des scores légèrement supérieurs à la moyenne nationale.

Faiblesse du Bloc Gauche

- Sous-performance de Mélenchon : Jean-Luc Mélenchon est le plus pénalisé, avec un score significativement inférieur à la moyenne nationale (17,4% dans l'Ain contre 22,0% en France, soit -4,6 points).
- Écologistes/Gauche en Retrait : Yannick Jadot et Fabien Roussel obtiennent également des résultats en deçà de leur score national.

Le profil électoral de l'Ain est celui d'un département de taille moyenne, fortement mobilisé, avec une orientation politique penchant clairement vers la droite et l'extrême droite, ce qui lui confère une spécificité par rapport à la moyenne nationale, notamment dans le duel de tête entre Macron et Le Pen.



Taux de Participation (Votes Exprimés)

La part la plus importante du cercle est représentée par les Votes Exprimés :

- 76,0% des électeurs inscrits ont participé à l'élection et ont exprimé un vote valide pour un candidat. Ce chiffre représente le taux de participation dans le département.

2. Taux d'Abstention (Non-Participation)

La deuxième plus grande part représente ceux qui ne se sont pas rendus aux urnes :

- 22,3% des électeurs inscrits se sont abstenus.

3. Votes Non Valides (Blancs et Nuls)

Les votes non valides représentent une très faible minorité :

- Les votes Blancs représentent 1,3% des inscrits.
- Les votes Nuls représentent 0,4% des inscrits.

Conclusion

Ce graphique démontre une forte mobilisation électorale dans le département de l'Ain lors de ce premier tour, puisque trois quarts des inscrits se sont déplacés pour voter. La très faible proportion de votes blancs et nuls indique que la quasi-totalité des votants ont fait un choix clair en faveur d'un candidat.

SEANCE 3

I- Questions de cours

1- **Quel caractère est le plus général : le caractère quantitatif ou le caractère qualitatif ? Justifier pourquoi.**
→ Le caractère qualitatif est le plus général. Un caractère est une propriété observée sur des individus (par exemple : la couleur des yeux, la taille, le poids, etc.). Les caractères qualitatifs décrivent des catégories (ex. : rouge, bleu, vert). Les caractères quantitatifs sont un cas particulier de caractère : ils peuvent être mesurés numériquement (ex. : taille = 1,75 m, âge = 20 ans). Donc, tout caractère quantitatif est un caractère, mais tout caractère n'est pas quantitatif.
Le terme "caractère qualitatif" est plus général car il englobe aussi les cas où la variable ne peut pas être mesurée numériquement.

2- **Quels sont les caractères quantitatifs discrets et caractères quantitatifs continus ? Pourquoi les distinguer ?**
→ Un caractère quantitatif est un caractère mesurable numériquement.
On en distingue deux types :

- Caractère quantitatif discret : il ne peut prendre que certaines valeurs précises et dénombrables.
- Caractère quantitatif continu : il peut prendre toutes les valeurs possibles dans un intervalle donné.

On les distingue car les méthodes de traitement statistique ne sont pas les mêmes :

- les caractères discrets se représentent par des tableaux de fréquences ou des diagrammes en barres,
- tandis que les caractères continus nécessitent des regroupements en classes et sont souvent représentés par un histogramme.

3- **Paramètres de position**
> **Pourquoi existe-t-il plusieurs types de moyenne ?**
→ Il existe plusieurs types de moyenne en fonction de la nature de la variable. Il existe la moyenne arithmétique qui est sensible aux valeurs extrêmes. Il existe aussi la moyenne quadratique, la moyenne harmonique, la moyenne géométrique ou encore la moyenne mobile.
> **Pourquoi calculer une médiane ?**
→ La médiane est la valeur, observée ou possible dans la série des données classées par ordre croissant qui partage cette série en deux parties comprenant exactement le même nombre de données de part et d'autre de cette valeur. On l'appelle également « moyenne du milieu ». L'objectif de la médiane est de déterminer la valeur centrale d'un ensemble de donnée, elle permet d'analyser des distributions asymétriques contrairement à la moyenne arithmétique qui est affectée par les valeurs extrêmes. Elle divise la population en deux sous-populations de probabilité équiprobable. Dans la pratique, il s'agit d'une valeur qui ne se calcule pas.
> **Quand est-il possible de calculer un mode ?**
→ Le mode d'une série statistique fait référence à toute modalité correspondant à l'effectif maximal. Il correspond à la valeur qui est la plus fréquente ou qui a la plus forte densité de probabilité. Il s'agit d'une moyenne de fréquence. Le mode n'existe pas toujours. Lorsqu'il existe il n'est pas toujours unique : distribution bimodale.

4- **Paramètres de concentration – Quel est l'intérêt de la médiane et de l'indice de C. Gini ?**
→ L'intérêt de la médiale est de partager en deux parties égales la masse de la variable. Il s'agit d'une médiane calculée relativement aux valeurs globales. Elle partage les valeurs globales en deux parties égales représentant chacune 50% des valeurs globales. Le produit des valeurs globales ne représente plus seulement l'effectif, mais l'importance de la totalité du caractère possédé par les individus. A partir de la médiale et de la médiane, on peut ainsi les comparer et obtenir une mesure de concentration.
La courbe de C. Gini quant à elle a pour objectif de décrire les effets de la concentration d'une population statistique. Elle se construit sur un repère orthonormé à partir de fréquences cumulées relatives. Les valeurs de la fréquence cumulée globale relative de la série sont portées en ordonnée.

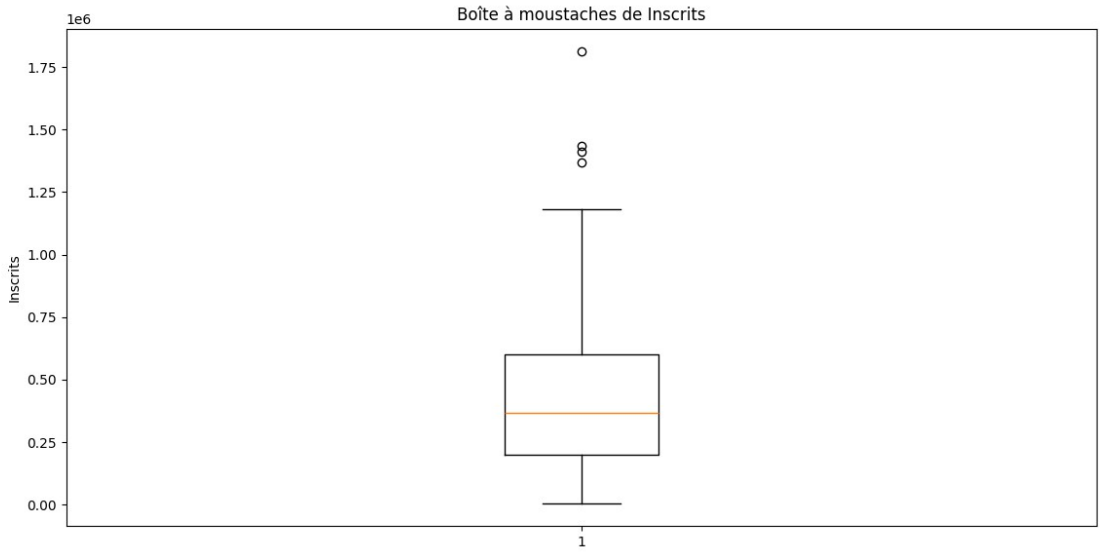
5- **Paramètres de dispersion**
> **Pourquoi calculer une variance à la place de l'écart de la moyenne ? Pourquoi la remplacer par l'écart type ?**
→ La variance est l'indicateur de dispersion par excellence. La variance tenant compte de toutes les données, il s'agit de fait de la meilleure caractéristique de dispersion. Contrairement à la moyenne arithmétique, elle est la moyenne de la somme des carrés des écarts. Cependant, la variance étant exprimée dans la même unité que la moyenne, il est souvent plus pratique d'utiliser l'écart type. En effet, l'écart type est davantage pertinent en ce qu'il caractérise le dispersion d'une série de valeurs. Plus l'écart est petit, plus les données sont regroupées autour de la moyenne arithmétique et plus la population est homogène. L'écart type permet ainsi de trouver le pourcentage de la population appartenant à un intervalle centré sur l'espérance mathématique : un résultat davantage parlant que la variance.
> **Pourquoi calculer l'étendue ?**
→ L'étendue d'une série statistique associée à un caractère quantitatif est la différence entre la plus grande valeur observée et la plus petite. L'étendue est facile à calculer et ne contient que des valeurs extrêmes de la série. Elle ne dépend ni du nombre, ni des valeurs intermédiaires. Elle indique l'étendue entre deux valeurs extrêmes et donne une idée générale directement.
> **A quoi sert-il de créer un quantile ? Quel(s) est (sont) le(s) quantile(s) le(s) plus utilisé(s) ?**
→ Les quantiles sont des caractéristiques de position, ils permettent de partager la série statistique ordonnée en parties égales. Généralement, le partage d'une série ordonnée des résultats se fait en quatre parties de même effectif, on obtient ainsi les quartiles. Le deuxième quartile est la médiane, on peut également trouver l'écart interquartile, il contient 50% des valeurs de la série.
> **Pourquoi construire une boîte de dispersion ? Comment l'interpréter ?**
→ J.W. Tukey baptisa la boîte de dispersion. La boîte à moustache permet ainsi de représenter schématiquement les principales caractéristiques d'une distribution en utilisant les quartiles. Elle correspond à une représentation graphique d'un caractère quantitatif. Elle sert à comparer visuellement plusieurs séries statistiques. La boîte à moustache fait apparaître la plus petite valeur, le premiet quartile, la valeur médiane, le troisième quartile et la plus grande valeur. Elle illustre donc la distribution des variables d'une série. Pour l'interpréter, après avoir tracé un rectangle qui s'étend du Quartile 1 au quartile 3, on marque la médiane par un trait puis on ajout les « moustaches » qui sont les segments qui vont de la valeur minimale à Q1 et de Q3 à la valeur maximale. En observant l'écart entre les deux quartiles, l'observateur constate la dispersion des données autour de la médiane. Si l'écart-type est important, cela signifie que les données sont très dispersées.

6- **Paramètres de forme**
> **Quelle différence faites-vous entre les moments centrés et les moments absolus ? Pourquoi les utiliser ?**
→ Les moments permettent de caractériser la forme d'une distribution de données. Les moments centrés correspondent à la moyenne des puissances des écarts à la moyenne. L'idée est de décrire la

forme d'une distribution autour de la moyenne. Les moments absolus quant à eux renvoient à la moyenne des valeurs absolues élevées à la puissances des écarts à la moyenne. Le but est ainsi de décrire la dispersion ou la forme, en étant moins influencé par les valeurs extrêmes. Les moments sont utilisés pour analyse la variabilité, faire des modélisations statistiques ou encore, pour ce qui est des moments centrés, de mesurer l'asymétrie.
> **Pourquoi vérifier la symétrie d'une distribution et comment faire ?**
→ La symétrie d'une distribution permet d'analyser les données de manière plus précise et fiable. Cette étape influence les méthodes statistiques appropriées à utiliser, l'interprétation des données, et les hypothèses sous-jacentes à certains tests. Il existe plusieurs façon de vérifier la symétrie d'une distribution : avec la moyenne, la médiane (si la moyenne ou la médiane sont très différentes, cela indique une asymétrie potentielle) ou encore le mode. Mais également à partir d'un histogramme en regardant sir la courbe est équilibrée autour de la moyenne. La boîte à moustache est également un autre outil tout comme la courbe de densité.

II- Graphiques, interprétation des résultats

Vous trouverez les autres boîtes à moustaches (boxplots) dans le portfolio.



Ce graphique, intitulé "Boîte à moustaches de Inscrits", résume la distribution statistique du nombre d'électeurs inscrits dans les départements français pour le premier tour de l'élection de 2022. L'axe vertical (Y) représente le nombre d'inscrits en millions (10^6).

1. La Boîte (Les Quartiles)

La boîte centrale contient 50% des départements.

- Médiane (Ligne orange au milieu) : La ligne centrale (médiane ou Q2) se situe autour de 0,4 million (soit 400 000) d'inscrits. Cela signifie que 50% des départements ont un nombre d'inscrits inférieur à 400 000, et 50% ont un nombre d'inscrits supérieur.
- Premier Quartile (Q1 - Bord inférieur de la boîte) : Il se situe environ à 0,2 million (200 000)

d'inscrits. 25% des départements ont moins de 200 000 inscrits.

- Troisième Quartile (Q3 - Bord supérieur de la boîte) : Il se situe environ à 0,6 million (600 000) d'inscrits. 75% des départements ont moins de 600 000 inscrits.
- Écart Interquartile (EIQ) : La largeur de la boîte (Q3 - Q1) est d'environ 0,4 million (600 000 – 200 000). Cela montre que la majorité (les 50% du milieu) des départements est concentrée dans une fourchette relativement étroite de 400 000 inscrits.

2. Les Moustaches (Plage de Variation)

Les moustaches s'étendent pour couvrir la majorité des données.

- Moustache Inférieure (Minimum) : Elle touche presque l'axe zéro, montrant que les départements les moins peuplés ont un nombre d'inscrits très faible, proche de zéro.
- Moustache Supérieure (Maximum non aberrant) : Elle s'étend jusqu'à environ 1,2 million (1 200 000) d'inscrits.

3. Les Valeurs Aberrantes (Outliers)

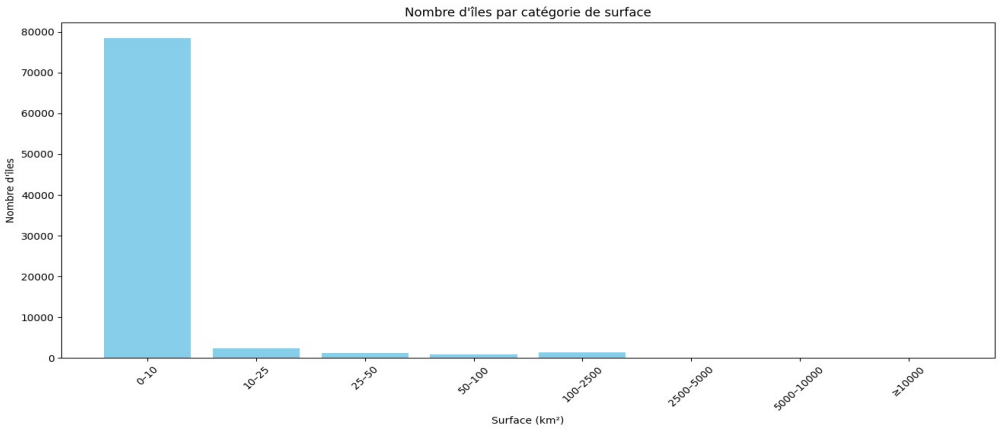
Les cercles au-dessus de la moustache supérieure représentent les valeurs très éloignées de la moyenne (les départements les plus peuplés).

- Départements les plus peuplés : Il y a quatre départements qui sont considérés comme des valeurs aberrantes (ou des extrêmes), avec un nombre d'inscrits allant d'environ 1,3 million à 1,8 million (le plus grand cercle visible en haut). Ces départements sont des exceptions qui correspondent aux grandes métropoles (comme Paris, les départements de la petite couronne, etc.).

Conclusion Générale

La boîte à moustaches confirme une distribution fortement asymétrique vers la droite :

- Concentration : La majorité des départements (75%) ont moins de 600 000 inscrits.
- Asymétrie : La médiane est plus proche du Q1 que du Q_3 , et la moustache supérieure est beaucoup plus courte que les outliers sont élevés. Ces caractéristiques indiquent que le nombre d'inscrits est majoritairement faible ou moyen dans la majorité des départements, mais quelques départements extrêmement peuplés tirent la moyenne générale vers le haut.



Ce graphique en barres (histogramme) montre la fréquence (le nombre d'îles) pour différentes catégories de taille (surface en {km^2}).

1. Axe des X (Horizontal) : Surface des îles ({km^2})

L'axe horizontal représente la surface des îles, divisée en catégories :

- 0 à 10 {km^2}
- 10 à 25 {km^2}
- 25 à 50 {km^2}
- 50 à 100 {km^2}
- 100 à 2500 {km^2}
- 2500 à 5000 {km^2}
- 5000 à $10 000 {km^2}
- 10 000 {km^2}

2. Axe des Y (Vertical) : Nombre d'îles

L'axe vertical indique le nombre d'îles dans chaque catégorie, allant de 0 à 80 000.

3. Observation Principale : Distribution de la Taille

La distribution du nombre d'îles est extrêmement concentrée dans la plus petite catégorie et très fortement asymétrique (décalée vers les petites valeurs).

- Dominance des Petites Îles (0 à 10 km^2) : La première barre est de loin la plus élevée, atteignant près de 80 000 îles. Cela signifie que l'immense majorité des îles (plus de 95% du total) ont une surface inférieure ou égale à 10 {km^2}.
- Rareté des Îles Moyennes : Les catégories suivantes (10 à 25 {km^2} et 25 à 50 {km^2}) ont un nombre d'îles très faible, à peine visible sur le graphique, aux alentours de 2 000 à 3 000 îles pour chaque catégorie.

- Extrême Rareté des Grandes Îles : Pour toutes les catégories de surface supérieure à 50 {km^2}, le nombre d'îles devient pratiquement nul (les barres sont si petites qu'elles touchent presque l'axe horizontal). Les catégories allant de 2500 km^2 à 10 000 {km^2} contiennent très peu, voire aucune île.

Conclusion

Le graphique montre que, dans l'ensemble de données utilisé, il existe une majorité écrasante de très petites îles et que les îles de taille moyenne ou grande sont extrêmement rares. La distribution des îles est dominée par les micro-îles et les îlots.

SEANCE 4

I- Questions de cours

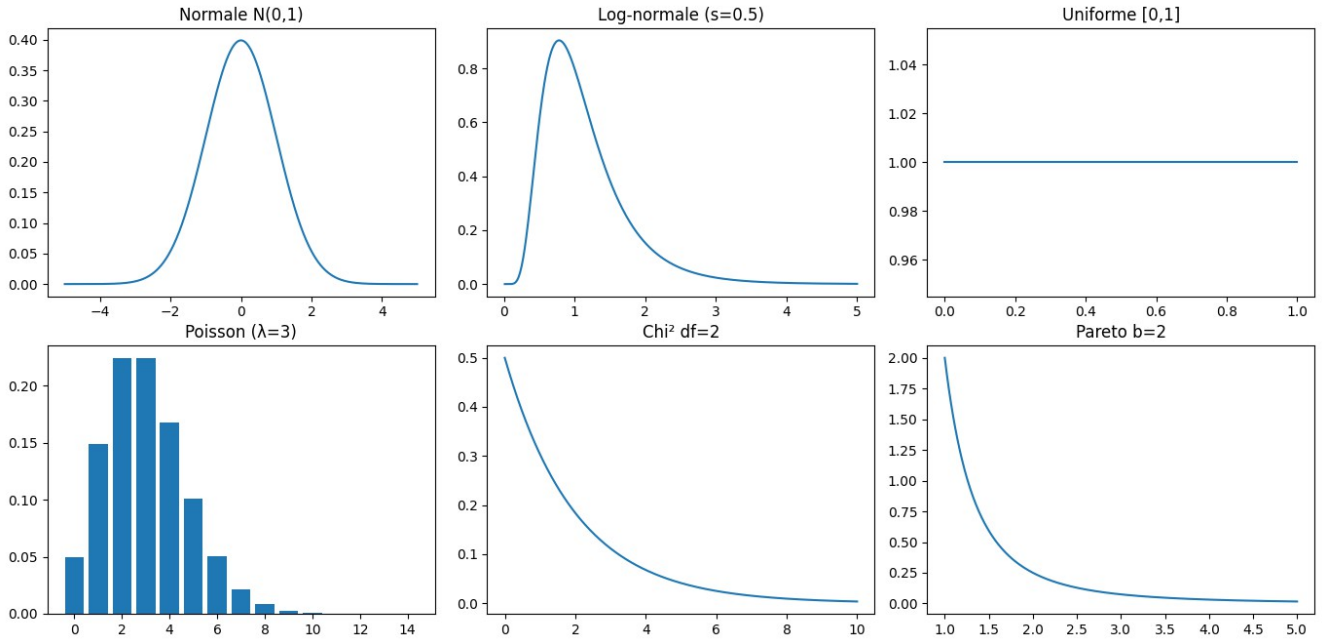
1- **Quels critères mettriez-vous en avant pour choisir entre une distribution statistique avec des variables discrètes et une distribution statistique avec des variables continues ?**
→ Pour rappel, une variable discrète est une variable quantitative qui ne peut prendre qu'un nombre fini ou dénombrable de valeurs. La distribution statistique discrète montre ainsi la fréquence ou la probabilité associée à chaque valeur possible de cette variable. Pour ce qui de la distribution statistique continue, elle peut prendre toutes les valeurs possibles dans un intervalle (infinie et non dénombrable).
Le choix entre une distribution statistique avec des variables discrètes et une distribution statistique avec des variables continues dépend de plusieurs critères : **la nature du phénomène étudié** (si l'on veut modéliser des événements comptables, on opte pour une distribution statistique avec des variables discrètes, au contraire, si l'on veut modéliser un phénomène mesurable en continu, on opte pour une distribution statistique avec des variables continues), **la forme de distribution empirique**, **la connaissance et l'interprétation des principales caractéristiques de l'ensemble des données** ainsi que **le nombre de paramètres des lois** (puisque la loi dépend de plusieurs paramètres pouvant s'adapter plus facilement à une distribution).

2- **Expliquez selon vous quelles sont les lois les plus utilisées en géographie ?**
→ En géographie, afin de modéliser la répartition d'un seul phénomène, les lois qui selon moi sont les plus utilisées sont :

- la loi Normale ou Gaussienne : (continue) car elle permet de modéliser des phénomènes naturels ou humains centrés autour d'une moyenne → calcul de températures, d'altitudes, de revenus
- la loi Log-Normale : (continue) elle permet d'étudier les phénomènes strictement positifs et très dispersés → taille des villes, débits, surfaces
- La loi de Pareto : (continue)pour les phénomènes de concentration spatiale ou socio-économique → taille des villes, richesse
- la loi de Poisson : (discrète) avec des variables discrètes, elle permet de modéliser le nombre d'événements rares dans une unité d'espace ou de temps → séismes, incendies, accidents
- La loi exponentielle : (continue) pour modéliser la distance entre événements aléatoires ou le temps entre deux événements → temps/distance entre événements
- la loi uniforme : (continue ou discrète) elle sert de référence théorique quand toutes les valeurs ont la même probabilité → répartition homogène d'un phénomène
- Loi Gamma : (continue) elle sert pour les phénomènes naturels liés au climat ou à l'hydrologie → intensité des précipitations, durée de sécheresse, débits des rivières

D'autres lois comme les lois de Zipf et de Zipf-Mandelbrot ou encore la loi Belford peuvent être utilisées en géographie mais de manière moins fréquente.

II- Code, interprétation des résultats



Ce panneau présente la fonction de densité de probabilité (pour les variables continues) ou la fonction de masse de probabilité (pour les variables discrètes) pour six distributions statistiques courantes.

1. Normale N(0, 1) (Haut, Gauche)

- Forme : C'est la célèbre courbe en cloche (ou gaussienne), parfaitement symétrique.
- Caractéristiques : La moyenne est à 0 et l'écart-type est de 1.
- Interprétation : La probabilité est maximale autour de la moyenne (0) et diminue rapidement et symétriquement en s'éloignant de celle-ci. C'est la distribution la plus fréquente dans la nature (erreurs de mesure, tailles, QI, etc.).

2. Log-normale s=0.5 (Haut, Centre)

- Forme : Asymétrique vers la droite (étalée positivement). Elle commence à 0.
- Caractéristiques : La densité de probabilité augmente très rapidement jusqu'à un pic (mode) juste après 0, puis décroît lentement.
- Interprétation : Elle est souvent utilisée pour modéliser des quantités qui ne peuvent pas être négatives et dont la croissance est proportionnelle à leur taille actuelle (ex: revenus, durée de vie des équipements, taille des particules).

3. Uniforme [0, 1] (Haut, Droite)

- Forme : Un rectangle ou une ligne parfaitement horizontale entre 0 et 1.
- Caractéristiques : La densité est constante à 1 sur l'intervalle [0, 1].
- Interprétation : Toutes les valeurs dans l'intervalle [0, 1] ont la même probabilité d'être tirées.

C'est la distribution utilisée pour simuler le hasard "pur" (comme un dé idéal).

4. Poisson (lambda=3) (Bas, Gauche)

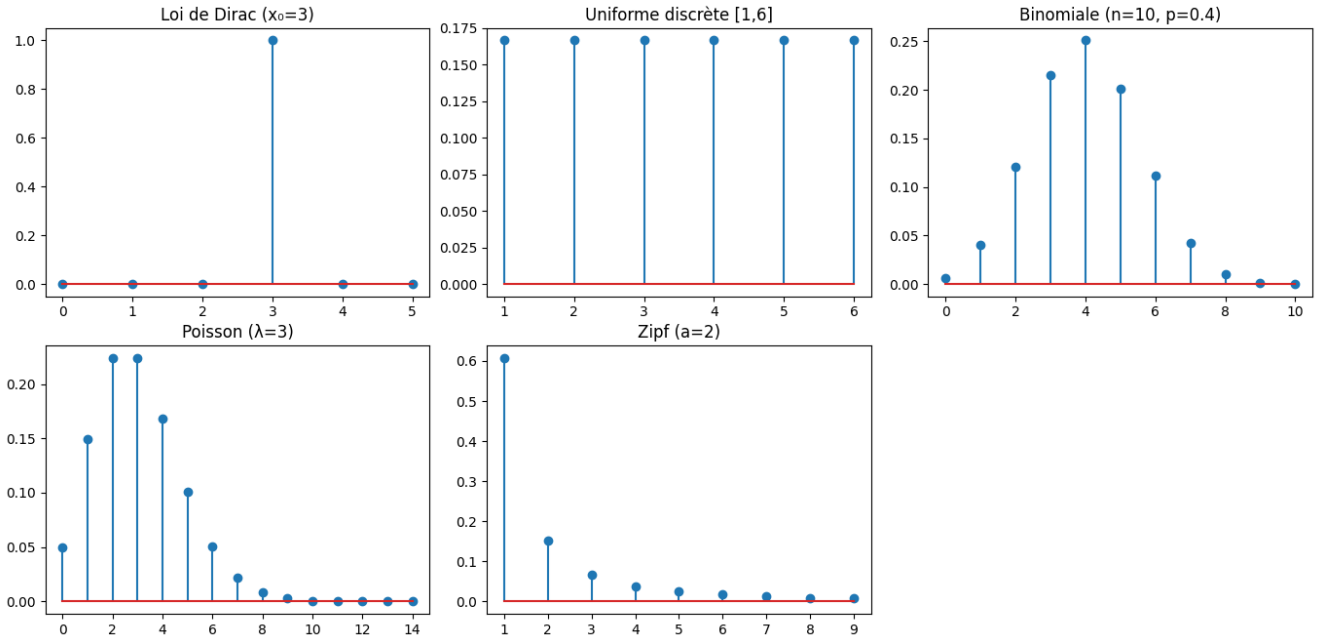
- Forme : Un histogramme (barres) asymétrique vers la droite. C'est la seule distribution discrète visible.
- Caractéristiques : lambda (lambda) est le paramètre de la distribution (égal à la moyenne). Les barres représentent la probabilité d'observer 0, 1, 2, 3, etc., événements.
- Interprétation : Elle modélise le nombre d'événements qui se produisent sur un intervalle de temps ou d'espace donné (ex: nombre d'appels reçus par heure, nombre de défauts par mètre carré de tissu). Le pic de probabilité est ici autour de 2 et 3.

5. Chi-carré chi^2 df=2 (Bas, Centre)

- Forme : Fortement asymétrique vers la droite, décroissant rapidement.
- Caractéristiques : La courbe commence à son maximum pour x=0 et décroît de façon exponentielle.
- Interprétation : C'est une distribution fondamentale en statistique, principalement utilisée dans les tests d'hypothèses (tests chi^2). Elle est obtenue en sommant les carrés de variables aléatoires normales indépendantes.

6. Pareto b=2 (Bas, Droite)

- Forme : Courbe décroissante (appelée loi de puissance).
- Caractéristiques : Décroissance très lente, surtout pour les grandes valeurs. La densité est élevée pour les petites valeurs (commence à 1), mais la queue de la distribution est très "lourde" (longue).
- Interprétation : Elle est célèbre pour modéliser des phénomènes où les extrêmes (les grandes valeurs) sont beaucoup plus probables qu'ils ne le seraient dans une distribution normale (ex: les richesses, la taille des villes, les séismes). C'est le principe du 80/20 (80% des conséquences proviennent de 20% des causes).



Ce graphique présente la fonction de masse de probabilité (FMP) pour six distributions où la variable aléatoire ne peut prendre que des valeurs entières (discrètes), généralement $0, 1, 2, 3$, etc. Les pics représentent la probabilité d'observer la valeur correspondante sur l'axe des X.

1. Loi de Dirac delta(x0=3) (Haut, Gauche)

- Forme : Un seul pic (impulsion) à une valeur spécifique (x0 = 3).
- Interprétation : Cette loi est la plus simple : l'événement a une probabilité de 100% (probabilité 1.0) de se produire à la valeur x=3 et une probabilité de 0 pour toute autre valeur. Elle représente une certitude absolue.

2. Uniforme discrète [1, 6] (Haut, Centre)

- Forme : Six pics de même hauteur.
- Caractéristiques : L'ensemble des valeurs possibles est{1, 2, 3, 4, 5, 6}.
- Interprétation : Toutes les valeurs ont exactement la même probabilité d'être observées. Dans cet exemple, la probabilité est de 1/6 (environ 0.1667) pour chaque résultat, comme le lancer d'un dé parfait.

3. Binomiale B(n=10, p=0.4) (Haut, Droite)

- Forme : Asymétrique vers la droite (même si elle est relativement proche d'une cloche tronquée).
- Caractéristiques : n=10 est le nombre d'essais, p=0.4 est la probabilité de succès à chaque essai.
- Interprétation : Elle modélise le nombre de succès k obtenus sur n essais indépendants. Le pic de probabilité se situe ici autour de k=4.

4. Poisson (lambda=3) (Bas, Gauche)

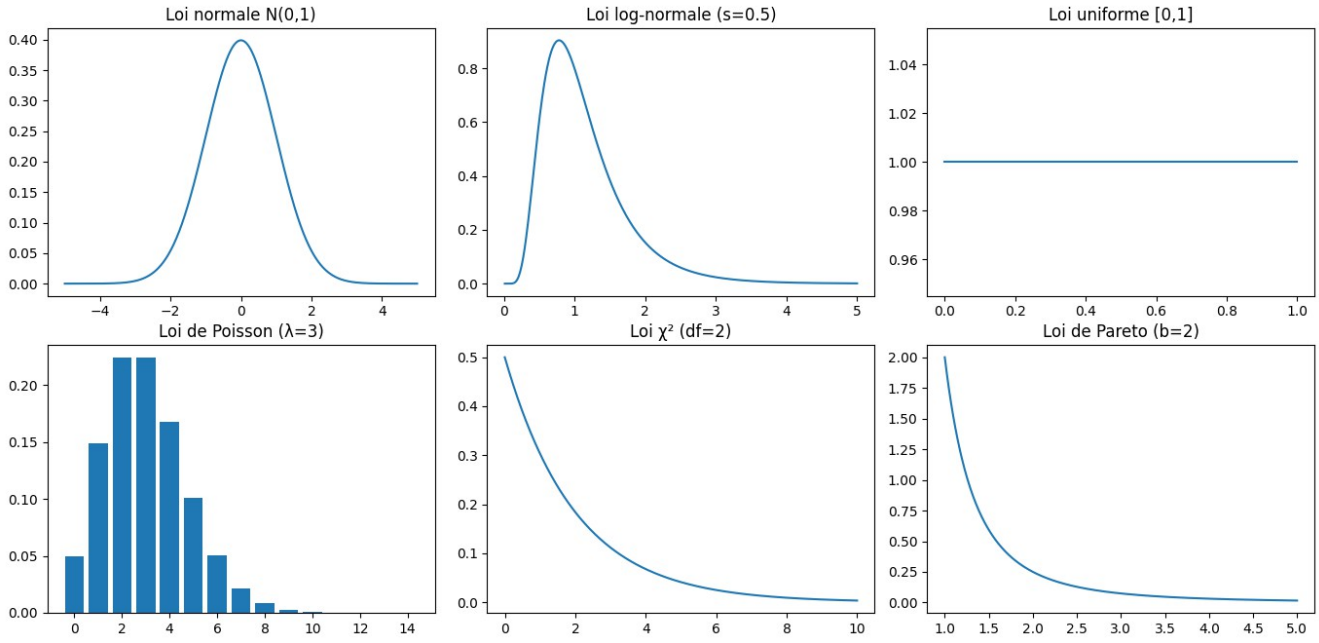
- Forme : Asymétrique vers la droite.
- Caractéristiques : lambda est la moyenne du nombre d'événements.
- Interprétation : Elle modélise le nombre d'événements rares qui se produisent sur une période de temps ou un espace donné. Le pic de probabilité est ici autour de k=2 et k=3, la valeur moyenne étant lambda=3.

5. Zipf a=2 (Bas, Centre)

- Forme : Très fortement asymétrique, avec un pic très élevé à la première valeur, puis une décroissance très rapide.
- Caractéristiques : C'est une loi de puissance discrète.
- Interprétation : Elle modélise des phénomènes où la fréquence est inversement proportionnelle au rang. L'exemple classique est la fréquence des mots dans une langue (le mot le plus fréquent est environ deux fois plus fréquent que le deuxième, trois fois plus que le troisième, etc.).

6. Hypergéométrique N=100, K=30, n=20 (Bas, Droite)

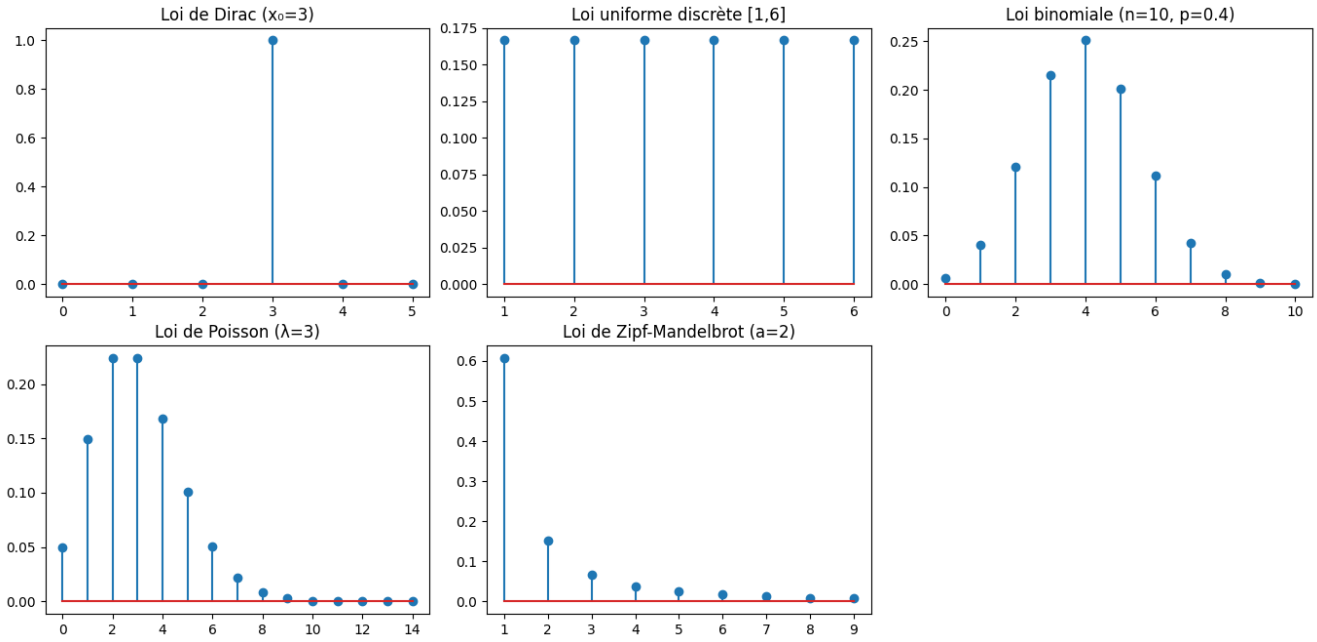
- Forme : Asymétrique vers la gauche (légèrement).
- Caractéristiques : N = taille totale de la population, K = nombre d'éléments "succès" dans la population, n = taille de l'échantillon tiré.
- Interprétation : Elle modélise la probabilité d'obtenir k succès lors du tirage d'un échantillon sans remise. Le pic est centré autour de la valeur attendue.



Ce graphique montre six façons différentes dont des événements qui comptent des nombres entiers

(comme 0, 1, 2, 3, etc.) peuvent se produire.

1. Loi de Dirac (Haut, Gauche)
    - Interprétation simple : C'est la certitude. L'événement se produit toujours à la valeur x=3 (probabilité de 100%).
2. Uniforme discrète (Haut, Centre)
    - Interprétation simple : C'est l'exemple du dé équilibré. Toutes les valeurs entre 1 et 6 ont la même chance de sortir (environ 17% de chance chacune).
3. Binomiale (Haut, Droite)
    - Interprétation simple : Compte les succès. Si vous faites 10 essais et que vous avez 40% de chance de réussir à chaque fois, cette courbe montre que le résultat le plus probable est d'avoir 4 succès (le pic).
4. Poisson (Bas, Gauche)
    - Interprétation simple : Compte les événements rares sur une période. Si vous vous attendez à recevoir en moyenne 3 événements , cette courbe montre qu'il est le plus probable d'en recevoir 2 ou 3. Les extrêmes (0 ou 10) sont moins probables.
5. Zipf (Bas, Centre)
    - Interprétation simple : Montre que le premier élément domine tout. La première valeur a une probabilité extrêmement haute, et les autres sont très faibles. (Exemple : le mot le plus courant dans une langue est utilisé beaucoup, beaucoup plus souvent que le deuxième, le troisième, etc.).
6. Hypergéométrique (Bas, Droite)
    - Interprétation simple : Compte les succès quand on ne remet pas ce qu'on a tiré. La courbe est centrée sur le nombre de succès attendus (ici, 6) et montre la probabilité de s'en éloigner.



Ce panneau montre six façons différentes de compter la probabilité d'obtenir des nombres entiers (0, 1, 2, 3, etc.).

1. Loi de Dirac (Haut, Gauche)

- Interprétation : C'est la certitude. L'événement se produit toujours à la valeur x=3 (probabilité de 100%).

2. Uniforme discrète (Haut, Centre)

- Interprétation : C'est l'exemple du dé équilibré. Toutes les valeurs de 1 à 6 ont la même chance de sortir.

3. Binomiale (Haut, Droite)

- Interprétation : Compte les succès sur un nombre fixe d'essais. Le résultat le plus probable est d'avoir 4 succès sur 10 essais.

4. Poisson (Bas, Gauche)

- Interprétation : Compte les événements rares sur une période. La probabilité est concentrée autour du nombre moyen d'événements, ici 2 ou 3.

5. Loi de Zipf (Bas, Centre)

- Interprétation : Montre que le premier élément est massivement dominant. La première valeur a une probabilité très, très haute, et les autres sont très faibles (Exemple : les mots les plus utilisés dans une langue).

6. Hypergéométrique (Bas, Droite)

- Interprétation : Compte les succès quand on ne remet pas les éléments tirés dans le tas. La probabilité est centrée sur le résultat le plus attendu (ici, 6).

SEANCE 5

I- 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 dans une population mère une partie de celle-ci au hasard avec une taille *n* fixée. Chaque échantillon fournit alors un résultat. L'échantillon est un groupe restreint, c'est-à-dire un sous-ensemble, issu d'une variable aléatoire X de la population. Il existe plusieurs façons de tirer un échantillon de la population mère. La plupart nécessitent de disposer une base de sondage. On peut aussi utiliser un échantillon aléatoire lorsqu'il est impossible de constituer une base de sondage. L'échantillon aléatoire offre des résultats recueillis sur ce sous-ensemble qui doivent pouvoir être étendus, c'est-à-dire inférés, à la population mère. Parmi les échantillons aléatoires on distingue l'échantillon non biaisé : tiré au hasard dans lequel tous les individus ont la même chance de se retrouver dans l'échantillon ; et l'échantillon biaisé : les éléments n'ont pas été pris au hasard.
Il existe plusieurs méthodes d'échantillonnage :

- les méthodes aléatoires : tirage avec ou sans remise
- les méthodes non aléatoires : échantillonnage systématique, méthode des quotas
- les méthodes d'échantillonnage « Monte Carlo »

L'utilisation de ces méthodes dépend de l'échantillon, du sujet étudié et de ce que l'on souhaite démontrer.

→ 2 – **Comment définir un estimateur et une estimation ?**

L'estimation permet d'estimer les paramètres d'une loi de probabilité. En ce qui concerne l'estimateur, il s'agit de la variable aléatoire. Un estimateur est une fonction des données. Il est construit de telle façon que sa valeur soit proche de la vraie valeur du paramètre. Le but de la théorie de l'estimation est de choisir, parmi toutes les statistiques possibles, le meilleur estimateur, c'est-à-dire celui qui donnera une estimation ponctuelle la plus proche possible du paramètre, et ceci quel que soit l'échantillon.

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

L'intervalle de fluctuation suppose que la vraie proportion théorique soit connue. C'est un échantillonnage et non une estimation. L'intervalle de confiance s'en distingue car c'est un outil statistique utiliser pour estimer la plage dans laquelle se situe un paramètre de population à partir d'un échantillon. Il permet de quantifier l'incertitude d'une estimation, comme la moyenne ou la variance, en fournissant une fourchette d'estimation.

→ 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 de l'estimateur et la valeur à estimer dans la population, on l'appelle également erreur d'estimation.

→ 5 – **Comment appelle-t-on une statistique travaillant sur la population totale ? Faites le lien avec la notion de données massives ?**
Une statistique travaillant sur la population totale est une statistique exhaustive. Le lien entre les

données massives et la statistique exhaustive correspond aux fait que les deux notions rendent plus accessible la démarche de la statistique exhaustive puisqu'on dispose de très grands volumes de données qui peuvent concerner tous les individus d'un système.

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

Il y a plusieurs enjeux autour du choix d'un estimateur. Premièrement, il y a un enjeu autour de sa variance qui influera sur la précision de l'estimateur. Par ailleurs, la statistique étant un résumé apporté par un échantillon, il est par conséquent très important de ne pas perdre l'information. Ainsi, en tenant compte de ces deux points, on peut aborder la recherche du meilleur estimateur suivant deux méthodes :

- soit en recherchant des statistiques exhaustives qui conduisent à des estimateurs sans biais de variance minimale
- soit en étudiant la quantité d'information de Fisher qui apporte des indications sur la précision d'un estimateur

→ 7 – **Quelles sont les méthodes d'estimation d'un paramètre ? Comment en sélectionner une ?**
Les principales méthodes d'estimation sont :

- la méthode des moments : il s'agit d'égaliser les moments théoriques avec les moments observés dans les données
- La méthode du maximum de vraisemblance : il s'agit de choisir le paramètre qui rend les données observées le plus vraisemblables
- la méthode des moindres carrés : il s'agit de minimiser la somme des carrés des écarts entre les valeurs observées et les valeurs estimées par le modèle
- la méthode bayésienne : combiner une information a priori sur le paramètre avec les données observées pour obtenir une distribution a posteriori

Pour choisir la méthode d'estimation, il faut prendre en compte la taille de l'échantillon, la complexité du modèle, la présence de biais ou d'erreurs, l'objectif de l'étude.

→ 8 – **Quels sont les tests statistiques existants ? A quoi servent-ils ? Comment créer un test ?**

Un test statistique est une méthode de calcul permettant de décider si une série statistique d'observations est compatible avec une loi de probabilité entièrement spécifique : ou comment savoir si un résultat observé est en accord avec un distribution théorique. Les tests statistiques suivent une loi connue, on retrouve parmi ces test :

- test de conformité
- test d'homogénéité
- test d'adéquation à une loi de probabilité
- test d'indépendance de deux caractères
- test de signification
- test d'hypothèse
- test paramétrique
- tests non paramétriques : test de Whitney, test de Wilcoxon, test du coefficient de corrélation de Spearman, test de Fisher)
- tests robustes : tests libres

Pour créer un test, il faut :

- formuler des hypothèses

- choisir la statistique du test
- choisir le niveau de risque
- calculer la valeur observée de la statistique
- déterminer la région critique
- interpréter le résultat

→ 9 – **Que pensez-vous des critiques de la statistique inférentielle ?**
La statistique inférentielle permet de tirer des conclusions sur une population à partir d'un échantillon, elle repose sur des modèles probabilistes, des tests et intervalles de confiance pour décider avec un certain risque d'erreur. C'est un outil très puissant mais aussi très critiqué car parfois mal compris ou mal utilisé. Les critiques portent sur la dépendance excessive aux hypothèses, la confusion fréquente entre signification et importance, une mauvaise interprétation du p-value, un détachement du cotexte, le caractère binaire de la décision ou encore sur le problème de la « chasse à la significativé ».
Selon moi, ces critiques invitent à se poser des questions effectivement sur l'usage de la statistique inférentielle, plutôt que de l'abandonner, il faut davantage cherche à l'utiliser de manière raisonnée par exemple en vérifiant les conditions, en utilisant des tests non paramétriques ou des méthodes robustes, en complétant avec des mesures d'effet, en formant à la logique des tests, en utilisant des approches bayésiennes ou des analyses de sensibilités ou encore en combinant les données quantitatives à l'analyse qualitative et théorique. Finalement, la statistique inférentielle est un outil et non une vérité. Il est nécessaire de garder un esprit critique tout le temps.

II- Code, interprétation des résultats

Théorie de l'échantillonnage

Lien entre l'intervalle de fluctuation et les valeurs réelles de la population mère. Que pouvez-vous en conclure par rapport aux échantillons utilisés pour le calcul ?
→ 1. Lien entre l'Intervalle de Fluctuation (IF) et la Population Mère
L'Intervalle de Fluctuation (IF) est un outil de la statistique inférentielle qui permet de tester la conformité d'un échantillon par rapport à une population dont les caractéristiques sont déjà connues.

- Définition : L'IF est calculé à partir de la fréquence réelle de la population mère. Pour un seuil de confiance de 95%, il définit la plage de valeurs dans laquelle nous nous attendons à ce que la fréquence observée dans un échantillon aléatoire de taille n se situe.
- Signification : Si nous tirons au hasard un grand nombre d'échantillons de taille n, la fréquence observé dans 95 % de ces échantillons tombera dans cet intervalle.

En résumé : L'Intervalle de Fluctuation nous dit quelles sont les fréquences *probables* que nous devrions observer dans nos échantillons, étant donné la vérité connue sur la population mère.

2. Conclusion par Rapport aux Échantillons Utilisés

Conclusion principale :

1. Convergence de la Moyenne : La fréquence moyenne calculée sur les 100 échantillons est extrêmement proche des fréquences réelles de la population mère. Pour les opinions "Pour" et

"Contre", la moyenne est identique (0.39 et 0.42), et pour "Sans opinion", l'écart n'est que de $ +0.01$.

2. Validité de l'Échantillonnage : Ce résultat vérifie la Loi des Grands Nombres en statistique. Il prouve que la moyenne des fréquences obtenues sur un grand nombre d'échantillons aléatoires (100 échantillons ici) est un estimateur non biaisé qui converge vers la fréquence réelle de la population.

En d'autres termes, les échantillons utilisés pour le calcul sont fiables. Même si un seul échantillon pouvait, par hasard, avoir une fréquence un peu éloignée de la réalité (tout en restant dans l'IF 95% du temps), le fait de moyenner un grand nombre d'échantillons (100) permet de gommer l'erreur aléatoire et de donner une image quasi parfaite de la population mère.

Théorie de l'estimation

→ L'Intervalle de Confiance (IC) est la méthode utilisée lorsque, dans un cas réel, nous ne disposons que d'un unique échantillon pour estimer les caractéristiques d'une population inconnue.
En prenant le premier échantillon de notre liste (taille $n = 1000$), nous avons observé les fréquences suivantes : 40% pour "Pour", 40% pour "Contre", et 21% pour "Sans opinion".

Signification de l'IC :

Pour la catégorie "Pour" par exemple, l'IC signifie que nous pouvons affirmer avec une confiance de 95% que la véritable fréquence de l'opinion "Pour" dans la population mère se situe entre 37% et 43%. C'est la fourchette d'estimation de la vérité.

Comparaison avec le Résultat Précédent (Intervalle de Fluctuation)

Analyse des concepts :

- IF (Partie 1) : Question de conformité. On connaît la population (p) et on regarde si l'échantillon est "normal" autour de p.
- IC (Partie 2) : Question d'estimation. On ne connaît pas la population (p) et on utilise l'échantillon pour encadrer le p inconnu.

Analyse des résultats :

- Validité de l'Estimation : Dans tous les cas, la fréquence réelle de la population mère (que l'on suppose être la vérité absolue) se trouve à l'intérieur de l'Intervalle de Confiance calculé à partir du seul premier échantillon.
- Conclusion : Ce résultat confirme que, même lorsque nous n'avons qu'un seul échantillon (le cas typique d'un sondage), la méthode de l'Intervalle de Confiance est statistiquement robuste. Elle permet, avec une forte probabilité (95% ici), de construire un intervalle qui encadre correctement la vérité de la population. L'échantillon isolé est donc une bonne base pour tirer des conclusions sur l'ensemble de la population.

Théorie de la décision
→ La distribution normale est : *Loi-normale-Test-1.csv.* : le premier fichier CSV
Pour prendre une décision, on pose deux hypothèses :

- Hypothèse Nulle (H_0) : La distribution des données est normale.
- Hypothèse Alternative (H_1) : La distribution des données n'est pas normale.

La décision est basée sur la valeur *p* et le seuil de signification .

Critère de Décision :

- Si *p* > 0.05 : Nous acceptons H_0 (Nous ne pouvons pas rejeter l'hypothèse de normalité).
- Si *p* est inférieur ou égal à 0.05 : Nous rejetons H_0 (La distribution n'est pas considérée comme normale).

3. Application aux Résultats

- Fichier *Loi-normale-Test-1.csv* : La valeur p est 0.540.
    - Puisque 0.540 > 0.05$, nous acceptons l'hypothèse nulle.
    - Conclusion : Les données de ce fichier suivent la loi normale.
- Fichier Loi-normale-Test-2.csv : La valeur *p* est très proche de zéro.
    - Puisque 0.000 est inférieur ou égal à 0.05, nous rejetons l'hypothèse nulle.
    - Conclusion : Les données de ce fichier ne suivent pas la loi normale.

Le fichier Loi-normale-Test-1.csv est donc la distribution normale.

SEANCE 6

I- Questions de cours

1- **Qu'est-ce qu'une statistique ordinale ? A quelle autre statistique catégorielle s'oppose-t-elle ? Quel type de variables utilise-t-elle ? En quoi cela peut matérialiser une hiérarchie spatiale ?**
→ Une statistique d'ordre ou statistique ordinale est le cœur de la géographie humaine. De manière annuelle, mensuelle, voire hebdomadaire, un certain nombre de classements est opéré en utilisant des objets géographiques. Leur objectif commun est de **montrer quelle entité a descendu, stagné ou monté dans le classement**.
La statistique ordinale s'oppose principalement à la **statistique nominale**.
La statistique ordinale utilise des **variables qualitatives ordinales**.
Cela peut matérialiser une hiérarchie spatiale car la statistique ordinale permet d'**établir un ordre entre les espaces** : ordre entre le centre et la périphérie, un ordre du centre urbain dense jusqu'à la ruralité...
Une variable ordinale peut de fait **matérialiser une hiérarchie spatiale** dès lors que ses catégories **représentent un niveau, un rang ou une intensité appliquée à des lieux**, des territoires ou des zones.
Cartographiquement, l'ordinal produit des **cartes en classes ordonnées**. Cette statistique permet aussi de traduire des relations de domination ou de centralité. Enfin, elle **établit des niveaux ou des rangs territoriaux**. En géographie physique, les lois d'ordre servent notamment à étudier la hauteur maximale des crus d'un cours d'eau, l'intensité du plus fort tremblement de terre dans une zone sismique donnée. Pour ce qui est de la géographie humaine, leur utilisation découle du fait de l'apparition plus ou moins spontanée de hiérarchies au sein des sociétés et des espaces étudiés.

2- **Quel ordre est à privilégier dans les classifications ?**
→ L'ordre à privilégier est l'ordre croissant ou ordre naturel. Il existe des exceptions en géographies telles que la loi dite rang-taille. L'ordination permet ainsi de rechercher les valeurs aberrantes, trop grandes ou trop petites, d'une série d'observations.

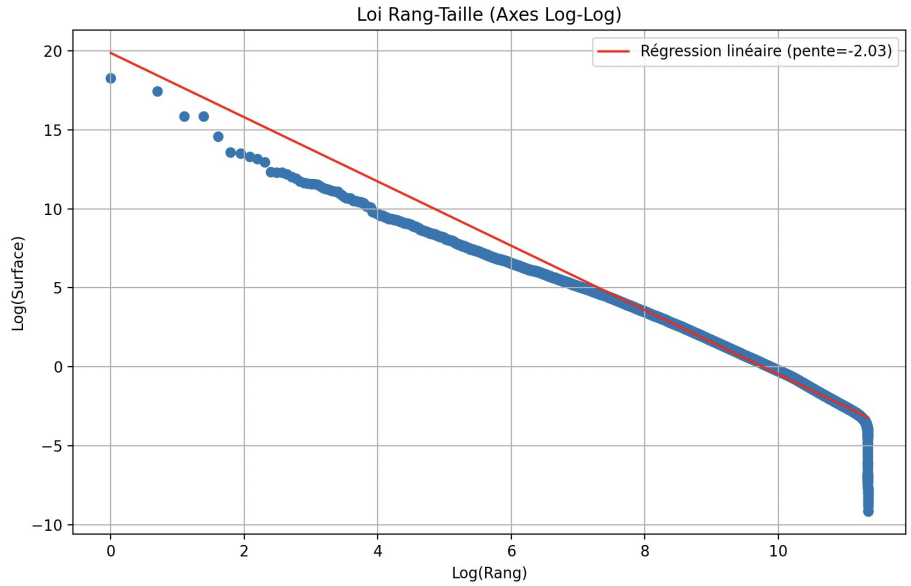
3- **Quelle est la différence entre une corrélation des rangs et une concordance de classements ?**
→ La corrélation des rangs **mesure la force et le sens de la relation entre deux séries de rangs** (comparaison de deux variables ordonnées, on cherche à savoir si les classements sont proportionnellement liés) tandis que la concordance des classements **mesure à quel point plusieurs classements sont identiques** (possibilité de comparer plus de deux classements, recherche de l'accord général entre plusieurs classements).

4- **Quelle est la différence entre les tests de Spearman et de Kendall ?**
→ Le test de Spearman mesure une corrélation monotone, **compare les rangs transformés des deux variables**, puis calcule une **corrélation sur les rangs** tandis que le test de Kendall **mesure la probabilité d'accord entre les paires d'observations**, il compare un par un tous les couples concordants et discordants en **se basant sur la circonstance du classement** et non sur les écarts contrairement au teste de Spearman. Le test de Spearman est par ailleurs plus sensible aux valeurs aberrantes et aux *ex aequo*.

5- **A quoi servent les coefficients de Goodman-Kursdal et de Yule ?**
→ Le coefficient de Goodman-Kursdal se base sur la différence entre les paires concordantes et les paires discordantes. Il calcule le surplus de paires concordantes par rapport aux paires discordante en exerçant une proportion.
Le coefficient de Yule quant à lui est un cas particulier du coefficient de Goodman-Kursdal en ce qu'il

est appliqué dans le cas des matrices 2x2. Il est nécessaire de construire la table de contingence qui évalue la fréquence des événements.

II- Code, interprétation des résultats



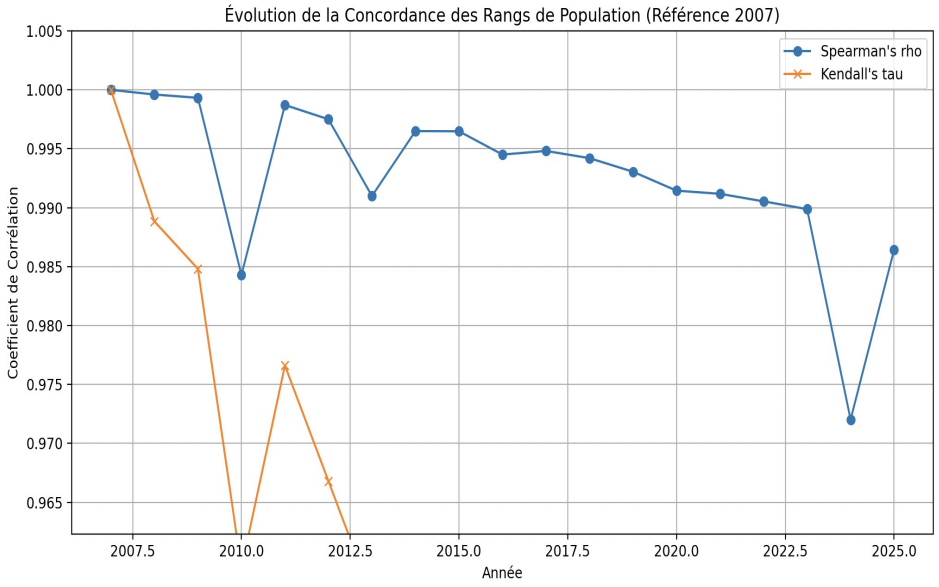


Image de sortie illisible : Un graphique direct (Rang vs. Taille) est illisible car les quatre premiers points (les continents) sont si grands qu'ils écrasent toutes les îles, rendant la visualisation impraticable.

Conversion Logarithmique : La conversion des deux axes en logarithme permet de vérifier si la

relation suit la loi de Zipf. La relation est une droite de pente négative, confirmant que la distribution des surfaces terrestres suit approximativement la loi Rang-Taille.

Commentaire sur le test des rangs : Non. Il n'est pas possible de réaliser un test statistique (tel qu'un test de Shapiro-Wilks ou de corrélation) sur la seule colonne des rangs. Le rang est un simple indice de classement ordonné (1, 2, 3, ...), non une variable aléatoire dont on peut tester la distribution ou la relation causale. Les tests sont effectués sur la relation entre le logarithme de la taille et le logarithme du rang.

Théorie de la Corrélation des Rangs (Analyse des États)

L'objectif est de mesurer la concordance entre les classements de la Population et de la Densité en 2007.

Une fois les colonnes isolées, nous obtenons deux listes de rangs ordonnées par le classement de 2007 :

- Liste A : Rang des pays en fonction de leur Population 2007.
- Liste B : Rang des pays en fonction de leur Densité 2007.

Le calcul des coefficients de corrélation de rang (Spearman et Kendall) entre le classement de la Population 2007 et le classement de la Densité 2007 des États du monde révèle une très faible corrélation positive.

Interprétation

1. Relation Faible : Les valeurs proches de zéro indiquent qu'il n'y a pas de relation monotone forte entre les deux classements.
2. Absence de Dépendance Directe : Le fait qu'un pays soit très peuplé (Rang Pop élevé) n'entraîne pas automatiquement qu'il soit très dense (Rang Densité élevé). Par exemple, un grand pays comme la Russie ou le Canada aura une population totale élevée mais une faible densité. Inversement, un petit pays comme Monaco ou Singapour aura une population faible mais une densité très élevée.
3. Conclusion : Le classement d'un pays en termes de population est largement indépendant de son classement en termes de densité, car la densité est un ratio qui dépend également de la superficie du territoire. La corrélation est donc faible, confirmant que ces deux indicateurs mesurent des réalités démographiques et géographiques distinctes.

CONCLUSION

Impression du cours, points d'amélioration

La pédagogie inversée a été particulièrement complexe pour moi. En effet, n'ayant pas de bagages scientifiques ou mathématiques, les notions m'étaient auparavant complètement étrangères. Par ailleurs, la quantité d'élèves dans le groupe rendait impossible le fait que le professeur puisse aider chacun d'entre nous. En outre, l'informatique est délicat et chaque ordinateur était différent, avec son lot de cotraintes. Aussi, ce cours a été davantage un travail collectif avec mes camarades.

Difficultés rencontrés

Le vocabulaire m'est apparu totalement étranger. Ce cours a été très déroutant et je ne pense pas être en capacité, en toute transparence, de pouvoir ajouter une quelconque compétence Python sur mon CV. De plus, l'informatique étant déjà quelque chose que je ne maîtrise pas, j'ai rencontré des difficultés avec Docker tout d'abord que je n'ai jamais pu installer. J'ai donc installé Python en brut.
Ainsi, mes camarades ont été d'une grande aide. Aussi, j'ai utilisé l'intelligence artificielle pour surpasser de nombreux obstacles : incompréhension de termes du cours, problèmes, Python, etc.
Aujourd'hui, mon rapport à l'analyse de données est plus que jamais mitigée. Consciente de l'importance que prend l'analyse de données dans le milieu géographique, je reste cependant très réticente à l'utiliser.
Je ne me projette aucunement dans une carrière d'analyste de données. Néanmoins, ce cours m'a ouvert les yeux sur bon nombre de points en ce qui concerne l'informatique ou même un PC. J'en ressors donc grandie.

Réflexion personnelle sur les sciences des données et les humanités numériques (niveau débutant)

La Science des Données et les Humanités Numériques ne sont pas deux disciplines séparées, mais deux faces d'une même médaille.

1. La Science des Données fournit le langage et la méthode pour analyser le monde à l'ère du numérique.
2. Les Humanités Numériques fournissent les questions, le contexte, et l'éthique nécessaires pour que cette analyse soit au service de la connaissance humaine et sociale.

En fin de compte, la Science des Donnes ne peut pas fonctionner sans les Humanités Numériques est inversement car les Humanités Numériques sans la SD sont incapables d'exploiter la richesse des données contemporaines.