Compte rendu d'Analyse de données

Maëlys CALURI

Semestre 1 - année 2025/2026

Lien vers le portfolio associé github.com/MaelysCaluri/Caluri-2025-2026-Analyse-de-donnees

Table des matières

1 Introduction

Ce compte rendu s'inscrit dans le cadre d'un travail d'analyse de données mobilisant les outils statistiques fondamentaux et leur mise en œuvre à l'aide du langage de programmation Python. Il a pour objectif de présenter, de manière synthétique et rigoureuse, les notions théoriques abordées au cours des différentes séances, tout en mettant en évidence leur application pratique à travers le codage et l'exploitation de jeux de données.

L'apprentissage est organisé de façon progressive, depuis les principes généraux de la statistique jusqu'à l'analyse plus spécifique des variables qualitatives. Chaque séance contribue à la construction d'une démarche analytique cohérente, fondée à la fois sur la compréhension des concepts statistiques et sur leur traduction opérationnelle en Python.

Le présent compte rendu couvre ainsi les cinq séances suivantes :

Séance 2 : Les principes généraux de la statistique, consacrée aux bases conceptuelles nécessaires à toute analyse de données ;

Séance 3 : Les paramètres statistiques élémentaires, portant sur les indicateurs de tendance centrale et de dispersion ;

Séance 4 : Les distributions statistiques, qui introduit l'étude de la répartition des données et des lois statistiques courantes ;

Séance 5 : Les statistiques inférentielles, dédiée aux méthodes permettant de tirer des conclusions générales à partir d'un échantillon ;

Séance 6 : La statistique d'ordre des variables qualitatives, axée sur l'analyse et le classement des données non quantitatives.

L'ensemble de ces séances constitue le socle méthodologique de l'analyse présentée dans ce document, articulant étroitement théorie statistique et implémentation informatique.

2 Séance 2 - Principes généraux

2.1 Questions de cours

Depuis ses débuts, la géographie ne cesse de chercher sa place vis-a-vis de la statistique.

En effet, ces deux disciplines entretiennent une relation tendue et complexe. S'il est vrai que les géographes ont longtemps sous-estimé l'intérêt des outils de la statistique - parfois en méprisant les définitions mathématiques élémentaires de cette dernière - il n'en demeure pas moins que seule la statistique permet d'étudier les données massives que produit la géographie. Ainsi les statistiques (à différencier de *la* statistique, qui en est la science) sont aujourd'hui essentielles à la compréhension de l'information géographique.

De plus, comme la statistique est capable d'analyser les phénomènes aléatoires, on pressent que la notion de hasard n'est pas absente de l'information géographique.

Lorsqu'un phénomène aléatoire se manifeste, il est par définition impossible d'en prévoir chaque réalisation. En revanche, bien qu'aléatoire, ce phénomène peut suivre des lois, voire une tendance, que la statistique nous permet de mettre en lumière. Il en va de même pour la géographie humaine : il est impossible de prédire chaque action et chaque conséquence sur un territoire donné, mais on peut tout de même les analyser dans leur ensemble afin d'en extraire une connaissance, une certitude à une échelle plus grande. Le hasard fait donc bien partie de l'information géographique, mais il n'empêche pas l'analyse de cette dernière.

L'information géographique se compose de différents types de données. D'une part, les informations attributaires décrivent les caractéristiques des territoires, qu'elles relèvent de la géographie humaine (population, revenus, structures sociales) ou de la géographie physique (température, précipitations, altitude). D'autre part, les informations géométriques concernent la forme, la localisation et l'organisation spatiale des objets géographiques, telles que les surfaces, les distances ou les réseaux. Dans un système d'information géographique, ces deux dimensions sont indissociables, les attributs qualifiant des objets localisés dans l'espace. L'analyse de ces informations nécessite un important travail préalable de production ou de collecte des données, reposant sur des nomenclatures précises et sur des métadonnées détaillées, afin de garantir la fiabilité et la comparabilité des résultats.

Les besoins de la géographie en matière d'analyse de données sont nombreux. Il s'agit avant tout de structurer et de synthétiser des masses de données hétérogènes, de comparer des territoires entre eux et de mettre en évidence des structures spatiales. L'analyse de données permet également de tester des hypothèses, d'explorer les relations entre variables et de préparer des démarches explicatives ou prospectives. Dans ce cadre, la statistique descriptive joue un rôle fondamental en résumant les données à l'aide d'indicateurs numériques et de représentations graphiques. Elle permet de décrire les distributions, d'identifier des valeurs extrêmes et de repérer d'éventuelles anomalies. La statistique explicative, quant à elle, vise à établir des relations entre une variable à expliquer et une ou plusieurs variables explicatives, à travers des modèles tels que la régression ou l'analyse de la variance. Ces deux approches sont complémentaires : la statistique descriptive constitue une étape préalable indispensable à toute analyse explicative.

La visualisation des données occupe une place essentielle en géographie, car elle permet de rendre intelligibles des phénomènes complexes. Le choix des représentations graphiques dépend de la nature des variables étudiées. Les variables qualitatives sont généralement représentées par des diagrammes en barres ou en secteurs, tandis que les variables quantitatives continues sont visualisées à l'aide d'histogrammes ou de boîtes à moustaches. En géographie, ces visualisations statistiques sont souvent complétées par des cartes thématiques, qui intègrent explicitement la dimension spatiale. Le choix des modes de représentation doit être guidé par l'objectif de l'analyse, l'échelle d'étude et la lisibilité du message scientifique que l'on souhaite transmettre.

Les méthodes d'analyse de données mobilisées en géographie sont variées. Les méthodes descriptives multidimension-

nelles, telles que l'analyse en composantes principales ou l'analyse factorielle des correspondances, permettent d'explorer des tableaux de données complexes et de réduire la dimension de l'information. Les méthodes de classification, comme la classification ascendante hiérarchique, servent à regrouper des individus ou des territoires présentant des caractéristiques similaires. Les méthodes explicatives permettent de modéliser les relations entre variables, tandis que les méthodes de prévision, notamment l'analyse des séries chronologiques, visent à anticiper l'évolution future d'un phénomène à partir de ses dynamiques passées.

Toute analyse statistique repose sur un vocabulaire précis. La population statistique désigne l'ensemble des unités étudiées, par exemple un ensemble de communes ou de régions. L'individu statistique correspond à une unité élémentaire de cette population, telle qu'une commune prise individuellement. Les caractères statistiques sont les propriétés observées sur chaque individu, comme la population, la superficie ou le revenu moyen. Les modalités statistiques correspondent aux valeurs prises par ces caractères. Les caractères peuvent être qualitatifs ou quantitatifs, ces derniers étant eux-mêmes discrets ou continus. Cette typologie n'est pas neutre, car elle conditionne les traitements statistiques possibles et les outils mobilisables. Il existe en effet une forme de hiérarchie entre les types de caractères, non pas en termes de valeur scientifique, mais en termes de richesse informationnelle et de possibilités d'analyse. Les variables qualitatives nominales sont les moins structurées, car elles permettent uniquement de distinguer des catégories sans ordre ni mesure, et se limitent au calcul d'effectifs et de fréquences. Les variables qualitatives ordinales introduisent un ordre entre les modalités, autorisant des analyses plus fines comme le calcul de médianes ou l'étude de rangs. Les variables quantitatives discrètes permettent le dénombrement et ouvrent l'accès à des indicateurs numériques de position et de dispersion sous certaines conditions. Enfin, les variables quantitatives continues constituent le niveau le plus élaboré de cette hiérarchie, car elles autorisent l'ensemble des traitements statistiques, l'ajustement à des lois de probabilité et la mise en œuvre de modèles explicatifs et prédictifs. Cette hiérarchie structure donc fortement les choix méthodologiques en analyse de données géographiques.

Dans le traitement des variables quantitatives continues, les notions d'amplitude et de densité sont essentielles. L'amplitude d'une classe se mesure en calculant la différence entre sa borne supérieure et sa borne inférieure ; elle exprime l'étendue de l'intervalle de valeurs couvert par la classe. La densité d'une classe se calcule en rapportant l'effectif de cette classe à son amplitude, ce qui permet de tenir compte de la largeur des intervalles lors de l'analyse des distributions. Cette mesure est indispensable lorsque les classes statistiques n'ont pas la même largeur, notamment dans les histogrammes, car elle permet de comparer correctement les hauteurs des rectangles et d'éviter des interprétations biaisées. La construction de ces classes repose souvent sur des règles empiriques, telles que les formules de Sturges ou de Yule, qui aident à déterminer un nombre de classes équilibré afin de conserver une information pertinente tout en assurant la lisibilité des distributions.

Enfin, l'analyse statistique repose sur des notions fondamentales telles que l'effectif, la fréquence, la fréquence cumulée et la distribution statistique. L'effectif correspond au nombre d'individus statistiques appartenant à une modalité ou à une classe donnée ; il s'agit d'un dénombrement brut qui constitue la base de toute description statistique. À partir de cet effectif, on calcule la fréquence en rapportant l'effectif d'une modalité à l'effectif total de la population étudiée, ce qui permet d'exprimer le poids relatif de cette modalité dans l'ensemble des données. La fréquence cumulée est obtenue en additionnant progressivement les fréquences des modalités ou des classes selon un ordre croissant ou décroissant ; elle permet d'analyser la répartition globale des données et de déterminer, par exemple, la part de la population située en dessous ou au- dessus d'un certain seuil. L'ensemble des modalités ou des classes associées à leurs effectifs, fréquences et fréquences cumulées constitue une distribution statistique. Celle-ci offre une représentation synthétique de la structure des données et constitue un support essentiel pour l'interprétation des phénomènes géographiques, la comparaison entre territoires et la mise en œuvre d'analyses statistiques plus avancées.

2.2 Manipulations avec Python

Question 1 à 3

Dans le dossier *src*, nous avons créé le dossier *data* et y avons introduit le fichier *resultats-elections-presidentielles-2022-1er-tour.csv*.

Ce fichier contient sous forme d'un tableau, l'ensemble des informations relatives au 1er tour des élections présidentielles de 2022. Chaque ligne renseigne les résultats obtenus pour chaque département français. Appuyons nous sur ce fichier pour mener cette première étude.

Dans le dossier *src*, nous avons introduit le fichier *main.py*. Et nous avons choisi l'éditeur de code *VS Code* pour le travailler.

Question 4 et 5 : Extraction d'un fichier CSV

Grâce à l'instruction `with` et la méthode `open()` on ouvre le fichier CSV. La méthode `read_csv()` de la bibliothèque `Pandas` permet ensuite de lire les informations du fichier CSV.

Ensuite nous utilisons la méthode `DataFrame()` de la bibliothèque `Pandas`. Celle-ci permet de structurer l'affichage des données *(Data)* - de façon structurée, comme un tableau *(Frame)*.

On obtient en effet, comme schématisé sur la *Table 1*, toutes les informations du fichier dans une variable que nous nommons `table`.

Code du département	Libellé du département	Inscrits	...	Voix.11
01	Ain	438109	...	8998.0
02	Aisne	373544	...	5790.0
...	...	...	...	...
ZZ	Français établis hors de France	1435746	...	7074.0

Table 1 – Extraction des valeurs du fichier dans la variable `table`.

(Voir l'affichage du terminal lors de l'exécution du fichier `main.py`)

Question 6 : Mesure du nombre de lignes et de colonnes

La fonction native *len()* nous renvoie la longueur d'une liste. Or, un tableau étant une liste de listes, on obtient le nombre de lignes en codant :

```
nb_lignes = len(table)
```

Afin d'obtenir le nombre de colonnes, on s'intéresse alors à la liste des titres des colonnes, en codant :

```
nb_lignes = len(table.columns)
```

On trouve alors que le tableau (la variable `table`) contient 107 lignes et 56 colonnes.

(Voir l'affichage du terminal lors de l'exécution du fichier `main.py`)

Question 7 : Différents types de variable

Les variables peuvent être de différents types. Parmi les nombres, on distingue par exemple les nombres entiers de type `int` et les nombres décimaux de type `float`. Mais une variable n'est pas forcément un nombre, elle peut représenter une chaîne de caractères. Son type sera alors `str`. On peut également rencontrer le type `bool` qui désigne une variable booléenne - ne pouvant que valoir `True` ou `False`. On attribura enfin le type `None` lorsque le type est inconnu ou manquant.

Munis de ces considérations, nous pouvons désormais dresser la liste qui renseigne le type de variable de chaque colonne du tableau. Pour ce faire, nous isolons chaque colonne et analysons le type de son premier élément (après le titre, bien sûr).

(Voir l'affichage du terminal lors de l'exécution du fichier `main.py`)

Pour reprendre l'exemple de la *Table 1*, on découvre alors que les colonnes "Code du département" et "Libellé du département" sont de type `str`, mais que les colonnes "Inscrits" et "Voix.11" sont respectivement de type `int` et de type `float`. La distinction trouvera toute son importance lorsque nous voudrons effectuer des opérations mathématiques sur les colonnes.

Question 8 : Noms des colonnes

On affiche dans le terminal les noms des colonnes avec la méthode `head()`.
Notons qu'elle s'utilise ainsi : `table.head(n)` (avec `n` le nombre de lignes à afficher, en partant d'en haut).

(Voir l'affichage du terminal lors de l'exécution du fichier `main.py`)

Question 9 : Isoler la colonne des *Inscrits*

Pour isoler la colonne "Inscrits" du tableau général, on écrit `table["Inscrits"]`. On obtient alors une seule colonne, comme schématisée dans la *Table 2* :

438109
373544
...
1435746

TABLE 2 – Colonne "Inscrits" isolée.

(Voir l'affichage du terminal lors de l'exécution du fichier `main.py`)

Question 10 : Calcul des effectifs de chaque colonne

Nous allons calculer l'effectif de chaque colonne. Pour cela, il est impératif que le type de la colonne soit un nombre (de type `int` ou de type `float`).

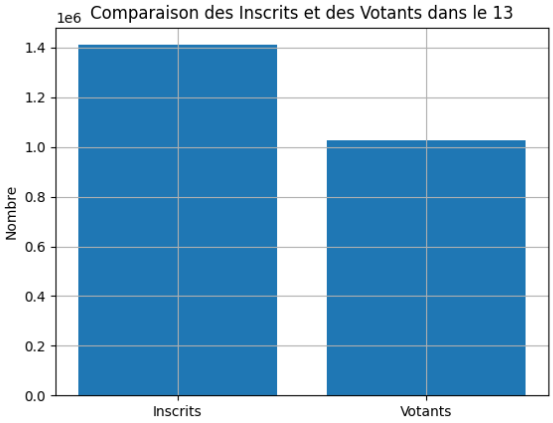
Pour chaque nom de colonne (obtenu à la question 8), on isole cette dernière. On en vérifie alors le type en se référant à la liste des types de chaque colonne (obtenue à la question 7). Si la colonne et de type `int` ou `float`, on en calcule la somme par la méthode `sum()`.

(Voir l'affichage du terminal lors de l'exécution du fichier `main.py`)

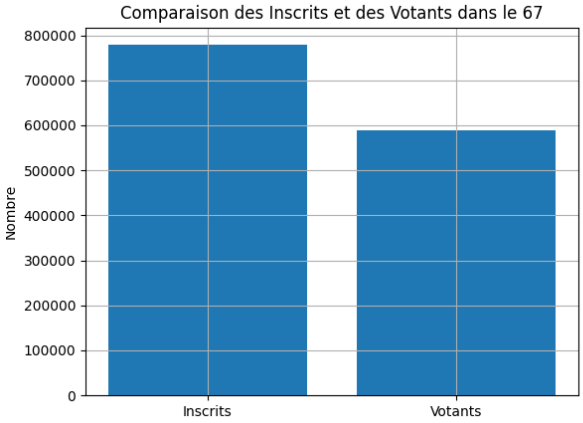
Question 11 : Diagrammes en barres

Après avoir isolé les colonnes *Code du département*, *Inscrits* et *Votants*, nous avons généré un diagramme en barres - comparant le nombre d'inscrits et le nombre de votants - pour chaque département français.

En voici 2 exemples en *Figure 1*. Les figures *(a)* et *(b)* comparent le nombre d'inscrits et le nombre de votants, respectivement pour les départements des Bouches-du-Rhône et du Bas-Rhin.



(a) Comparaison dans les Bouches-du-Rhône



(b) Comparaison dans le Bas-Rhin

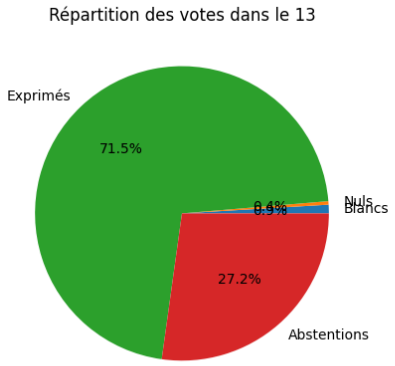
FIGURE 1 – Diagrammes en barres obtenus

En analysant l'intégralité des diagrammes en barres générés, on constate que dans la très grande majorité des départements français, moins de 50% des citoyens inscrits sont allé voter.

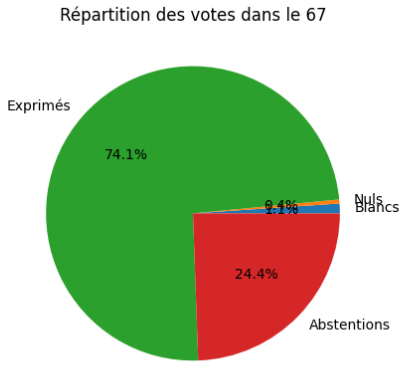
Question 12 : Diagrammes circulaires

Après avoir isolé les colonnes *Code du département*, *Blancs*, *Nuls*, *Exprimés* et *Abstentions*, nous avons généré un diagramme circulaire - montrant la répartition de la "validité" des votes - pour chaque département français.

En voici 2 exemples en *Figure 2*. Les figures *(a)* et *(b)* montrent la répartition de "validité" des votes, respectivement pour les départements des Bouches-du-Rhône et du Bas-Rhin.



(a) Répartition dans les Bouches-du-Rhône



(b) Répartition dans le Bas-Rhin

FIGURE 2 – Diagrammes circulaires obtenus

Cette représentation nous permet d'apprécier d'un seul coup d'œil les différents taux. En effet, pour la grande majorité des départements, on remarque que le taux d'abstention est d'environ 25%, alors que les taux de votes blancs ou

de votes nuls sont quasiment négligeables (de l'ordre de 1 à 2%).

Question 13 : Histogrammes

Grâce à la méthode `hist` de la bibliothèque `Matplotlib`, on peut tracer l'histogramme de la distribution des inscrits. Le voici en *Figure 3* :

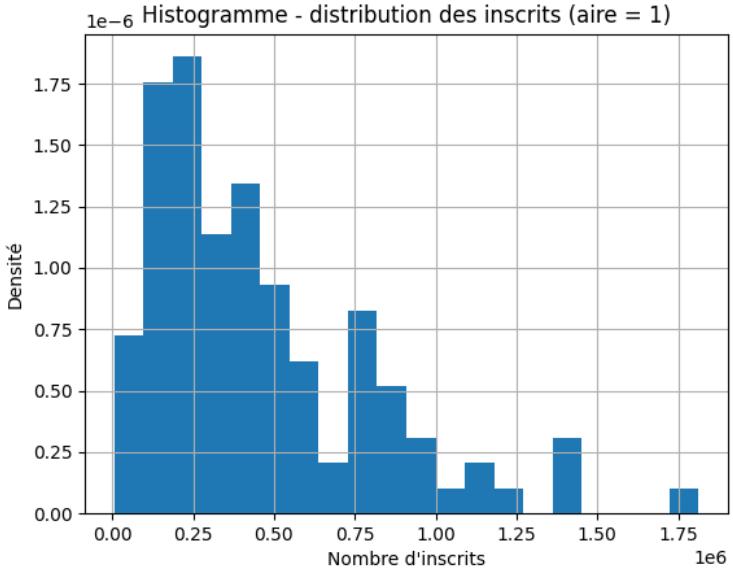
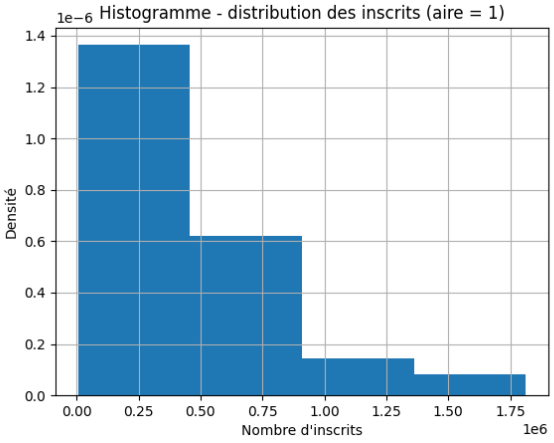


FIGURE 3 – Histogramme de la distribution des inscrits

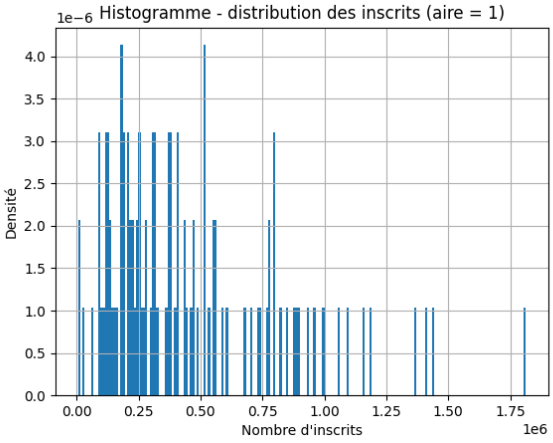
Nous avons ici choisit de décomposer l'intervalle des abscisses en 20 parties. C'est un choix parfaitement arbitraire, mais il convient de le déterminer avec soin.

En effet, si le nombre de sous-intervalles est trop petit, la largeur des sous-intervalles sera trop grande, et ne laissera pas entrevoir l'aspect de la distribution. C'est le cas par exemple de la *Figure 4 (a)*, pour laquelle on a choisi 4 parties.

Il en va de même lorsque le nombre de sous-parties est trop grand, car la largeur des sous-intervalles deviens alors trop fine et ne recense alors que de rares "valeurs candidates". C'est le cas de la *Figure 4 (b)*, pour laquelle on a choisi 200 parties.



(a) Découpage en 4 sous-intervalles



(b) Découpage en 200 sous-intervalles

FIGURE 4 – Histogrammes obtenus mais non retenus

3 Séance 3 - Paramètres élémentaires

3.1 Questions de cours

Le caractère quantitatif peut être considéré comme le plus général par rapport au caractère qualitatif, dans la mesure où il permet une mesure numérique des phénomènes et autorise l'ensemble des traitements statistiques, alors que le caractère qualitatif se limite à une catégorisation des individus. Tout caractère qualitatif peut, dans certains cas, être transformé en caractère quantitatif par codage ou comptage, tandis que l'inverse n'est pas toujours possible sans perte d'information. Le caractère quantitatif offre ainsi une richesse analytique supérieure, car il permet le calcul de paramètres de position, de dispersion et de forme, ainsi que la modélisation statistique des phénomènes étudiés.

Les caractères quantitatifs se divisent en deux grandes catégories : discrets et continus. Les caractères quantitatifs discrets correspondent à des valeurs issues d'un dénombrement, prenant des valeurs isolées et distinctes, comme le nombre d'habitants ou le nombre de logements. Les caractères quantitatifs continus, quant à eux, résultent d'une mesure et peuvent prendre théoriquement une infinité de valeurs dans un intervalle donné, comme la température, la superficie ou la durée. Cette distinction est fondamentale car elle conditionne les méthodes d'analyse et de représentation : les variables continues nécessitent souvent un regroupement en classes et permettent l'utilisation d'outils comme les intégrales, tandis que les variables discrètes reposent sur des sommes et des effectifs précis.

Les paramètres de position visent à résumer une distribution autour d'une valeur centrale. Il existe plusieurs types de moyennes car aucune moyenne ne peut, à elle seule, rendre compte de toutes les structures possibles des données. La moyenne arithmétique est la plus courante, mais elle est sensible aux valeurs extrêmes ; d'autres moyennes, comme la moyenne harmonique, géométrique ou quadratique, sont plus adaptées à certains contextes spécifiques, notamment lorsqu'il s'agit de vitesses, de taux ou de phénomènes multiplicatifs. La médiane est quant à elle calculée afin de disposer d'un indicateur de tendance centrale robuste, peu sensible aux valeurs aberrantes, et particulièrement pertinent pour les distributions dissymétriques. Le mode, enfin, n'est calculable que lorsque certaines modalités ou classes présentent un effectif ou une densité maximale ; il n'existe pas toujours et peut être multiple, révélant alors la coexistence de plusieurs sous-populations au sein des données.

Les paramètres de concentration permettent d'analyser la manière dont une variable est répartie entre les individus. La médiale présente un intérêt majeur car elle partage la masse totale de la variable en deux parties égales, et non simplement les effectifs. Elle permet ainsi d'identifier des situations de concentration économique, sociale ou spatiale. L'indice de concentration de C. Gini, fondé sur la comparaison entre la médiale et la médiane et représenté à l'aide de la courbe de Lorenz, permet de mesurer le degré d'inégalité dans la distribution d'une variable. Plus la courbe s'éloigne de la diagonale d'égalité parfaite, plus la concentration est forte, ce qui en fait un outil central pour l'analyse des inégalités territoriales.

Les paramètres de dispersion mesurent la variabilité des données autour de leur valeur centrale. La variance est préférée à l'écart à la moyenne car elle évite les compensations entre écarts positifs et négatifs grâce à l'élévation au carré, tout en disposant de propriétés algébriques fondamentales. Elle est cependant exprimée dans une unité au carré, ce qui rend son interprétation moins intuitive ; c'est pourquoi on lui substitue souvent l'écart type, qui correspond à la racine carrée de la variance et s'exprime dans la même unité que la variable étudiée. L'étendue est calculée pour fournir une mesure simple de la dispersion globale, correspondant à la différence entre la valeur maximale et la valeur minimale, bien qu'elle soit très sensible aux valeurs extrêmes. Les quantiles sont construits afin de découper la distribution en parts égales, ce qui permet d'analyser la structure interne des données ; les plus utilisés sont les quartiles, les déciles et les centiles, le deuxième quartile correspondant à la médiane. La boîte de dispersion, ou boîte à moustaches, synthétise graphiquement ces informations en représentant la médiane, les quartiles et les valeurs extrêmes ; elle permet de comparer rapidement plusieurs distributions et d'identifier les asymétries et les valeurs atypiques.

Les paramètres de forme décrivent la structure globale de la distribution. Les moments absolus sont calculés par rapport à une origine donnée et permettent de caractériser l'ordre de grandeur des valeurs, tandis que les moments centrés sont calculés par rapport à la moyenne et servent à analyser la dispersion, la symétrie et l'aplatissement de la distribution. Ils sont utilisés car ils offrent une description mathématique complète des distributions statistiques. Vérifier la symétrie d'une distribution est essentiel pour choisir les indicateurs et les modèles appropriés : une distribution symétrique justifie l'usage privilégié de la moyenne, tandis qu'une distribution dissymétrique rend la médiane plus pertinente. La symétrie peut être évaluée en comparant moyenne, médiane et mode, ou à l'aide des moments centrés d'ordre supérieur, notamment le moment d'ordre trois, qui permet de mesurer l'asymétrie. Cette analyse de la forme complète l'étude de la position et de la dispersion, offrant une compréhension globale des données.

3.2 Manipulations avec Python

Question 1 à 4 : Extraction des données

Après avoir bien créé le dossier src/data/, et après y avoir introduit le ficher *resultats-elections-presidentielles-2022-1er-tour.csv*, on introduit le fichier `main.py` et on l'ouvre dans *VS Code*. On en extrait les données de la même manière que dans la séance précédente, au moyen de l'instruction `with` et des méthodes `read_csv()` et `DataFrame()`.

(Voir l'affichage du terminal lors de l'exécution du fichier `main.py`)

Question 5 et 6 : Calcul de paramètres statistiques

En réutilisant le travail de la séance 2, pour déterminer le type des données de chaque colonne, nous avons identifié les colonnes dites quantifiables. Il s'agit des colonnes de type `int` et de type `float`.

Une fois ce tri effectué, nous calculons - pour chaque colonne - certains paramètres statistiques élémentaires :
- La moyenne, avec la méthode `mean()` ;
- La médiane, avec la méthode `median()` ;
- Les modes, avec la méthode `mode()` ;
- L'écart-type, avec la méthode `std()` ;
- L'écart-absolu à la moyenne, avec la méthode `mad()` ; (nous reviendrons dessus)
- Et l'étendue, avec les méthodes `min()` et `max()` ;

On obtient ainsi des résultats, dont l'extrait suivant en *Table 3* :

Colonnes quantitatives	'Inscrits'	'Abstentions'	'Votants'	...
Moyennes	455587.63	119852.05	335735.58	...
Médianes	366859.0	95369.0	274372.0	...
Écarts-types	351003.78	117017.8	258393.81	...
Écarts-absolus	272240.72	74959.07	201517.17	...
Étendues	1808861.0	929183.0	1297100.0	...

TABLE 3 – Paramètres statistiques calculés pour chaque colonne quantitative

On fait le choix de ne pas afficher le résultat de la méthode `mode()`, qui donne les modes d'une liste de données. En effet, les modes sont les valeurs qui ont le nombre d'occurrences le plus élevé. Or, sur la centaine de départements étudiés, il est très peu probable que deux départements aient le même nombre d'inscrits, à l'unité près. *Idem* pour le nombre d'abstentions, *etc.* Ainsi, la méthode *mode()* des colonnes quantifiables renvoie la quasi intégralité de ces colonnes.

De plus, la méthode `mad()` n'est plus d'actualité. Certaines versions de *Python* l'acceptent encore, mais dans un souci de sécurité, nous en avons écrit cette version :

```
somme_ecarts_abs = 0
moyenne = mean(colonne)

for j in range(len(colonne)):    # Pour chaque élément de la colonne:
    # On somme les distances (écarts absolus) à la moyenne
    somme_ecart_abs += np.abs(moyenne - colonne[j])

ecart_abs = somme_ecarts_abs / len(colonne)  # On divise cette somme par le nombre d'élements
                                             # pour avoir la distance moyenne : l'écart-absolu.
```

(Voir les lignes de codes l.110 à l.114 du fichier `main.py`)

Évoquons maintenant l'interprétation que l'on peut faire des résultats de la *Table 3* :

La liste des moyennes nous indique par exemple que le nombre moyen d'inscrits par département est de 455 588.

Toujours pour cette colonne, la liste des médianes nous enseigne que la moitié des départements ont un nombre d'inscrits inférieur à 366 859. (Et l'autre moitié en a un nombre supérieur à cette même valeur.)

L'écart-type est de 351 004, ce qui est du même ordre de grandeur que la moyenne. Cela indique une forte dispersion des données autour de la moyenne. En d'autres termes, les valeurs sont très étalées, donc la moyenne est peu représentative de l'ensemble des données, car le nombre d'inscrits varie beaucoup d'un département à l'autre.

Ce même constat peut être appuyé par la valeur de l'étendue qui montre que la différence entre le département comptabilisant le plus d'inscrits et celui comptabilisant le moins d'inscrits, vaut 1 808 861, soit presque 2 millions d'individus, ce qui vaut environ 4 fois la valeur moyenne. Le nombre d'inscrits par département est donc très dispersé.

En revanche, un autre indicateur vient modérer ce propos : l'écart-absolu vaut ici 272 241, soit deux fois moins que la valeur de la moyenne. Cela indique plutôt une dispersion modérée des données. En effet, lorsque l'écart absolu vaut 50% de la moyenne, les valeurs sont jugées assez dispersées. Toutefois, elles restent du même ordre de grandeur que la moyenne. Donc la moyenne reste plutôt représentative, sans être pour autant très précise.

Question 7 et 8 : Boîtes à moustaches

Pour calculer la distance interquartile d'une colonne, on calcule la distance entre son premier quartile (donné par la méthode `quantile(0.25)` et son troisième quartile (donné par la méthode `quantile(0.75)`.

De même, pour calculer la distance interdécile d'une colonne, on calcule la distance entre son premier décile (donné par la méthode `quantile(0.10)` et son neuvième décile (donné par la méthode `quantile(0.90)`.

On a ainsi obtenu la liste de toutes les distances interdéciles et de toutes les distances interquartiles, dont on montre ici un extrait en *Table 4* :

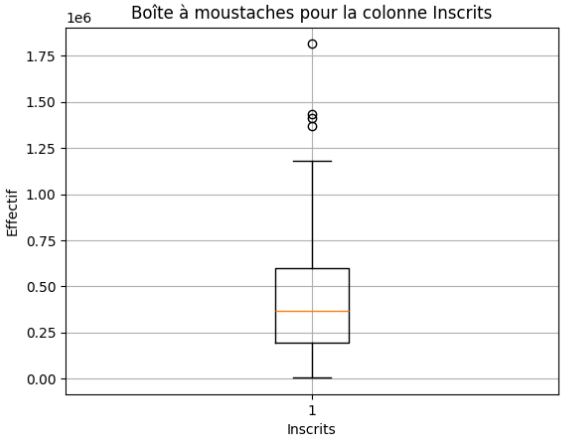
Colonnes quantitatives	'Inscrits'	'Abstentions'	'Votants'	...
Distances interquartiles	401050.0	106489.0	301770.5	...
Distances interdéciles	793988.8	193676.2	602687.2	...

TABLE 4 – Distances interquartiles et interdéciles calculées pour chaque colonne quantitative

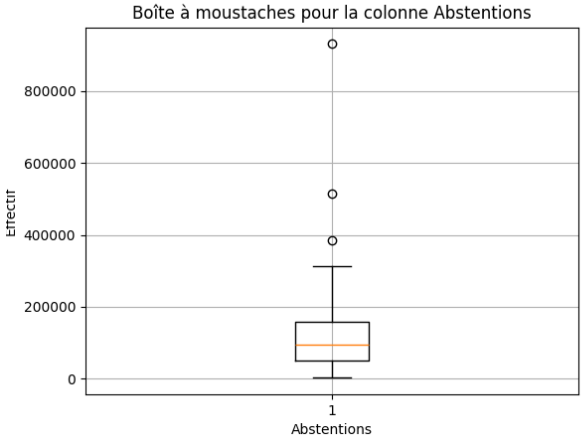
(Voir l'affichage du terminal lors de l'exécution du fichier `main.py`)

Ces listes nous ont permis de générer une boîte à moustaches pour chaque colonne.

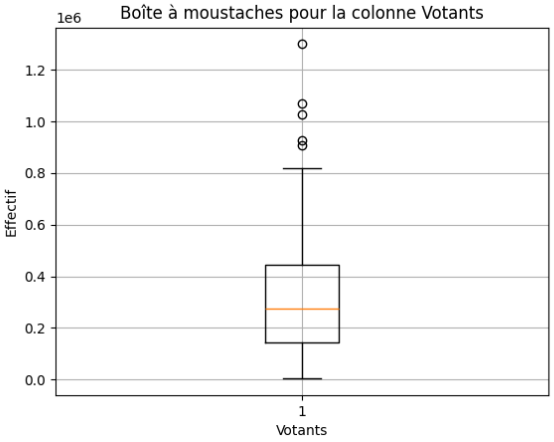
Voici en *Figure 5 (a), (b), (c) et (d)*, respectivement les boîtes à moustaches obtenues pour la répartition des valeurs du nombre d'inscrits, du nombre d'abstentions, du nombre de votes blancs et du nombre de votes nuls.



(a) Pour les votes exprimés



(b) Pour les Abstentions



(c) Pour les votes Blancs

(d) Pour les votes Nuls

FIGURE 5 – Boîtes à moustaches obtenues

Dans une boîte à moustaches, la longueur du segment noir délimité par des tirets représente la distance interdécile. La hauteur du rectangle noir représente la distance interquartile. Enfin la position du tiret orange représente la moyenne.

Les différentes boîtes à moustaches de la *Figure 5* nous offrent ainsi la possibilité de comparer d'un seul coup d'oeil les différentes étendues des séries d'observations. Il faut en revanche bien faire attention aux valeurs indiquées sur l'axe des ordonnées, car celles-ci peuvent varier énormément d'une boîte (ou graphe) à l'autre.

Question 9 et 10 : Catégorisation des variables quantitatives

On s'intéresse maintenant à catégoriser des îles en fonction de leur surface (donnée en km^2). Ainsi, après avoir extrait les informations du fichier *island-index.csv*, et après avoir isolé la colonne intitulée "Surface (km²)", nous opérons la logique de catégorisation suivante (les chiffres étant données en km^2) :

La surface est-elle inférieure à 10 ?
Si OUI : On la comptabilise dans l'intervalle]0,10].
Si NON : On passe au test suivant :

La surface est-elle inférieure à 25 ?
Si OUI : On la comptabilise dans l'intervalle]10,25].
Si NON : On passe au test suivant :

La surface est-elle inférieure à 50 ?
Si OUI : On la comptabilise dans l'intervalle]25,50].
Si NON : On passe au test suivant :

La surface est-elle inférieure à 100 ?
Si OUI : On la comptabilise dans l'intervalle]50,100].
Si NON : On passe au test suivant :

Et ainsi de suite...

Avec cette méthode, nous avons obtenu les résultats suivants, dans la *Table 5* :

Entre 0 et 10km²	78423 îles
Entre 10 et 25km²	2327 îles
Entre 25 et 50km²	1164 îles
Entre 50 et 100km²	788 îles
Entre 100 et 2 500km²	1346 îles
Entre 2 500 et 5 000km²	60 îles
Entre 5 000 et 10 000km²	40 îles
Au-delà de 10 000km²	71 îles

TABLE 5 – Catégorisation des îles terrestres en fonction de leur surface

Cette catégorisation nous permet de constater que la grande majorité des îles sont de surface inférieure à 10 km². Cela concerne environ 7 îles sur 8 dans le Monde. Ensuite, il semble se dessiner la tendance générale selon laquelle les îles se raréfient à mesure que leur taille augmente.

On constate donc que cette représentation est parlante, mais qu'elle n'est pas très informative. Ce constat nous invite donc à nous intéresser aux différentes distributions statistiques.

4 Séance 4 - Distributions statistiques

4.1 Questions de cours

Le choix entre une distribution statistique fondée sur des variables discrètes ou sur des variables continues repose sur plusieurs critères méthodologiques étroitement liés à la nature du phénomène étudié, aux objectifs de l'analyse et aux traitements statistiques envisagés. Le premier critère est la nature intrinsèque du caractère observé : lorsqu'un phénomène résulte d'un dénombrement et ne peut prendre que des valeurs entières distinctes, comme le nombre d'habitants, d'équipements ou d'événements, il relève naturellement d'une variable quantitative discrète. À l'inverse, lorsqu'un phénomène est issu d'une mesure et peut théoriquement prendre une infinité de valeurs dans un intervalle, comme la température, la superficie ou la durée, il est plus pertinent de le considérer comme une variable quantitative continue. Un second critère tient à l'échelle de mesure et à la précision des données : certaines variables discrètes peuvent être assimilées à des variables continues lorsque leur effectif est élevé et que l'analyse porte sur des tendances globales. Le choix dépend également des objectifs analytiques, car les distributions continues permettent l'ajustement à des lois théoriques, le calcul de densités et l'utilisation d'outils analytiques plus avancés. Enfin, les contraintes de représentation graphique jouent un rôle important : les variables continues nécessitent souvent un regroupement en classes afin de faciliter la lecture et l'interprétation, tandis que les variables discrètes peuvent être analysées directement à partir des effectifs observés.

Dans le cadre de l'analyse géographique, plusieurs lois statistiques sont particulièrement mobilisées en raison de leur capacité à rendre compte de phénomènes spatiaux et socio-spatiaux récurrents. La loi normale occupe une place centrale, car de nombreux phénomènes géographiques, tels que certaines variables climatiques ou des caractéristiques physiques, tendent à se distribuer de manière symétrique autour d'une moyenne. Elle constitue également une référence théorique essentielle pour de nombreuses méthodes statistiques. La loi log-normale est fréquemment utilisée pour décrire des phénomènes caractérisés par une forte dissymétrie, comme les distributions de revenus, de tailles de villes ou de surfaces agricoles, où les petites valeurs sont très nombreuses et les grandes valeurs rares mais déterminantes. La loi de Poisson est particulièrement adaptée à l'étude des phénomènes discrets et rares dans l'espace, tels que la localisation d'accidents, d'événements naturels ou d'équipements, lorsque ceux-ci sont indépendants et répartis de manière aléatoire. En géographie urbaine et économique, la loi de Pareto et, plus largement, les lois de puissance sont souvent mobilisées pour analyser les hiérarchies spatiales, notamment la distribution des tailles de villes ou des revenus, mettant en évidence des processus de concentration et d'inégalités. Enfin, certaines analyses spatiales reposent sur des lois exponentielles ou binomiales selon la nature des processus étudiés. Le recours à ces lois ne vise pas seulement à ajuster des données empiriques, mais permet aussi d'interpréter les mécanismes géographiques sous-jacents, en reliant les formes de distribution observées aux dynamiques sociales, économiques et spatiales.

4.2 Manipulations avec Python

Visualisons chacune des distributions statistiques suivantes, à l'aide d'un exemple.

4.2.1 Distributions statistiques de variables discrètes

Loi de Dirac

On a une loi de Dirac lorsqu'on peut dire qu'une seule issue est probable à 100%.
Sur la *Figure 6*, l'issue X=7 est certaine, alors que les autres issues sont impossibles.

FIGURE 6 – Distribution de la loi de Dirac - discrète

Loi Uniforme discrète

On a une loi Uniforme discrète lorsqu'on peut dire que toutes les issues sont équiprobables.
Sur la *Figure 7*, chacune des 20 issues à la même probabilité de 1/20 d'advenir.

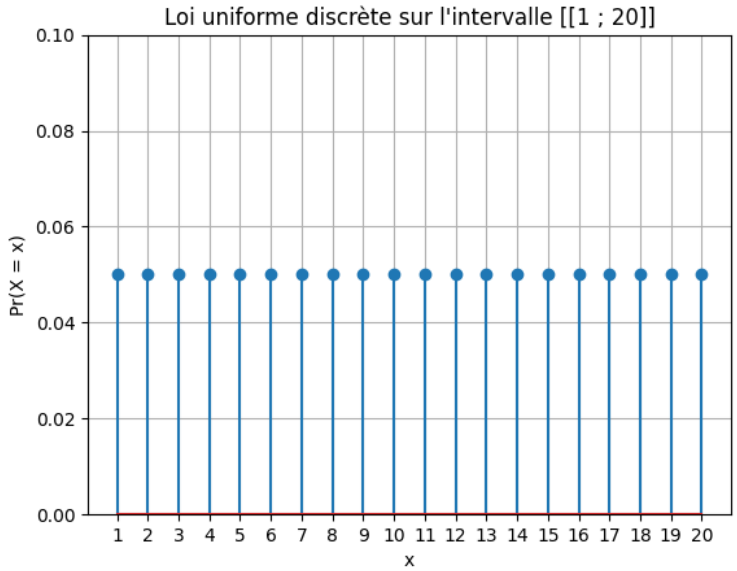


FIGURE 7 – Distribution de la loi Uniforme - discrète

Loi Binomiale

On a une loi Binomiale par exemple dans le cas suivant :

"Je joue à pile ou face n fois. Combien ai-je de chance de remporter au total : 0 manche ? 1 manche ? 2 manches ? ... n manches ?"

Ainsi la *Figure 8* en livre un exemple. C'est une loi binomiale où l'expérience est répétée n =20 fois, avec une probabilité de succès à chacune, de $p = 1/2$.

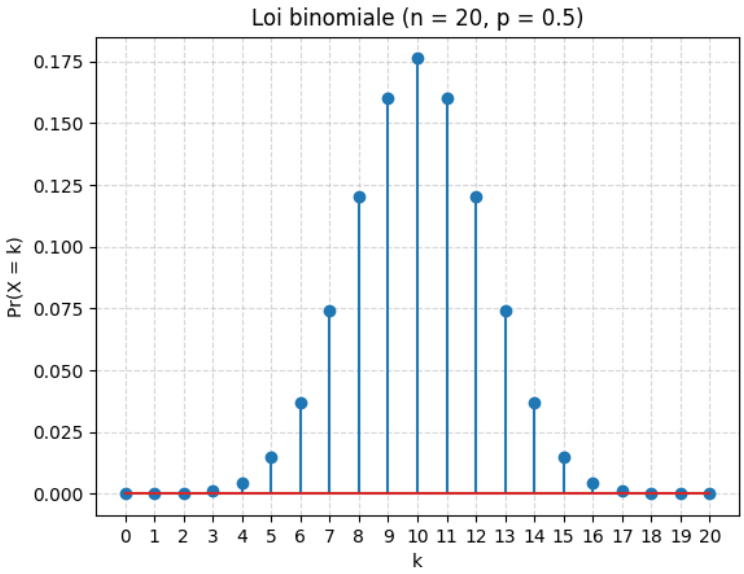


FIGURE 8 – Distribution de la loi Binomiale - discrète

Loi de Poisson

La loi de Poisson suit la même philosophie que la loi Binomiale, mais quand l'événement est rare.
Par exemple, notre hypothèse est de dire : il y a en moyenne 2 étoiles filantes dans le ciel chaque 5 minutes. (On note lambda = 2).
Maintenant on choisit dans la soirée une tranche de 5 min. Alors la loi de Poisson donne la probabilité : qu'il y passe 0 comète / qu'il y passe 1 comète / qu'il y passe 2 comètes / ... / qu'il y passe 14 980 comètes / *etc...*

Dans la *Figure 9*, la loi de Poisson étudiée est celle de paramètre lambda = 5.

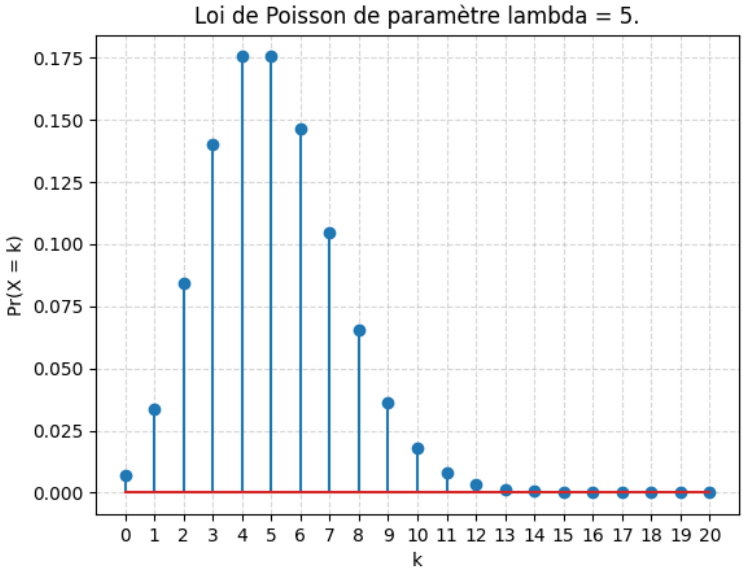


FIGURE 9 – Distribution de la loi de Poisson - discrète

Loi de Zipf-Mandelbrot

La Zipf nous disait : "Dans un livre, si le mot le plus fréquent apparaît 1000 fois, alors le 2ème mot apparaît 1000/2 fois. Le 3ème mot apparaît 1000/3 fois, *etc* ...

Mais la loi de Zipf-Mandelbrot est une généralisation de cette tendance, en apportant des paramètres correctifs :

$$s > 1 \text{ : exposant}$$
$$q >= 0 \text{ : décalage de Mandelbrot}$$
$$N \text{ : taille maximale du support}$$

De ces trois valeurs découle la constante de normalisation H qui vaut :

$$H = \frac{1}{(1+q)^s} + \frac{1}{(2+q)^s} + ... + \frac{1}{(N+q)^s}$$

Zipf-Mandelbrot se retrouve souvent en géographie. Par exemple dans le classement des populations des villes.

Voici en *Figure 10* une distribution de Zipf-Mandelbrot de paramètres s=1.5, q=2, et N=50.

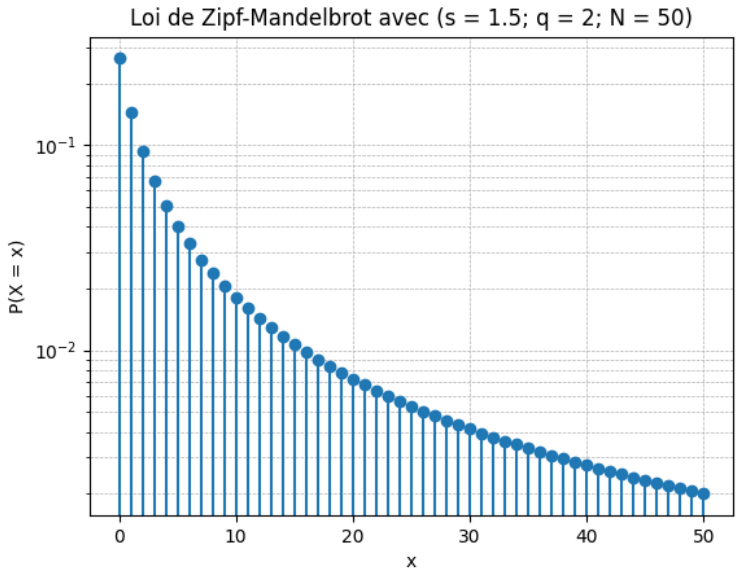


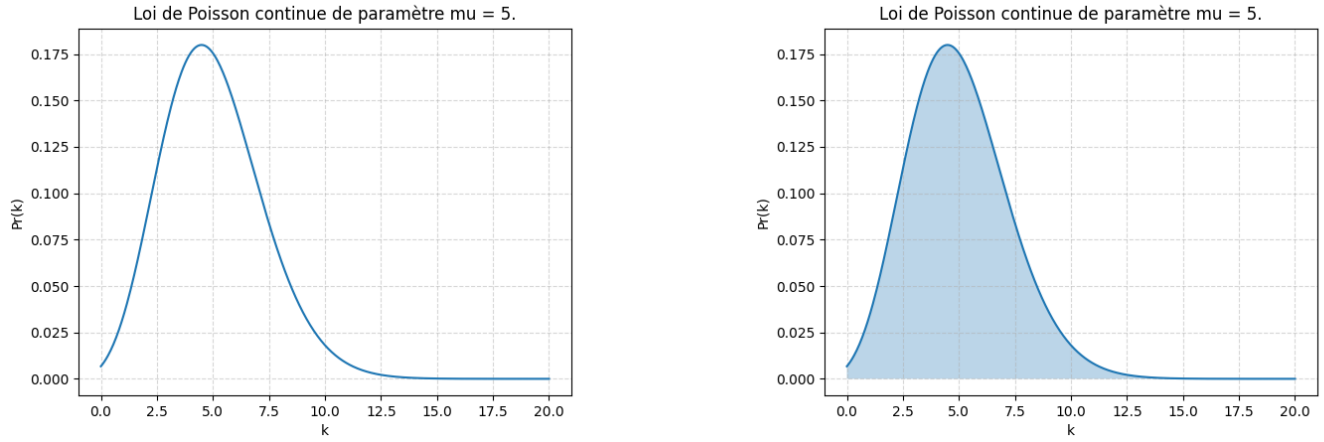
FIGURE 10 – Distribution de la loi de Zipf-Mandelbrot - discrète

4.2.2 Distributions statistiques de variables continues

Intéressons nous maintenant aux distributions statistiques de variables continues.

Loi de Poisson

On retrouve notre même loi de Poisson, en considérant cette fois-ci que l'on dénombre une quantité continue (par exemple une température exacte, ou une longueur exacte).On indique désormais le paramètre mu, et non lambda.
Voici en *Figure 11* deux représentations de la même distribution de Poisson, avec mu = 5.



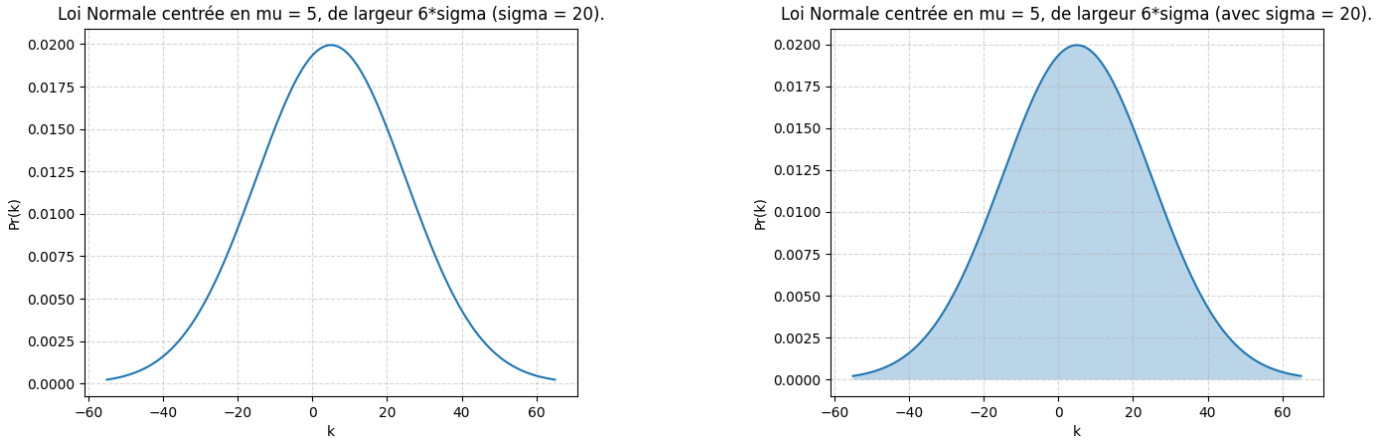
(a) Représentation courbe (b) Représentation intégrale

FIGURE 11 – Distribution de la loi de Poisson - continue

Loi Normale

Loi Normale est la "fameuse Gaussienne". Elle est centrée en la valeur μ, s'étend de - l'infini à + l'infini mais 99.7% de ses valeurs se trouvent entre $\mu - 3\sigma$ et $\mu + 3\sigma$. Avec σ son écart-type.

On en donne un représentation pour $\mu = 5$, et $\sigma = 20$, en *Figure 12* :

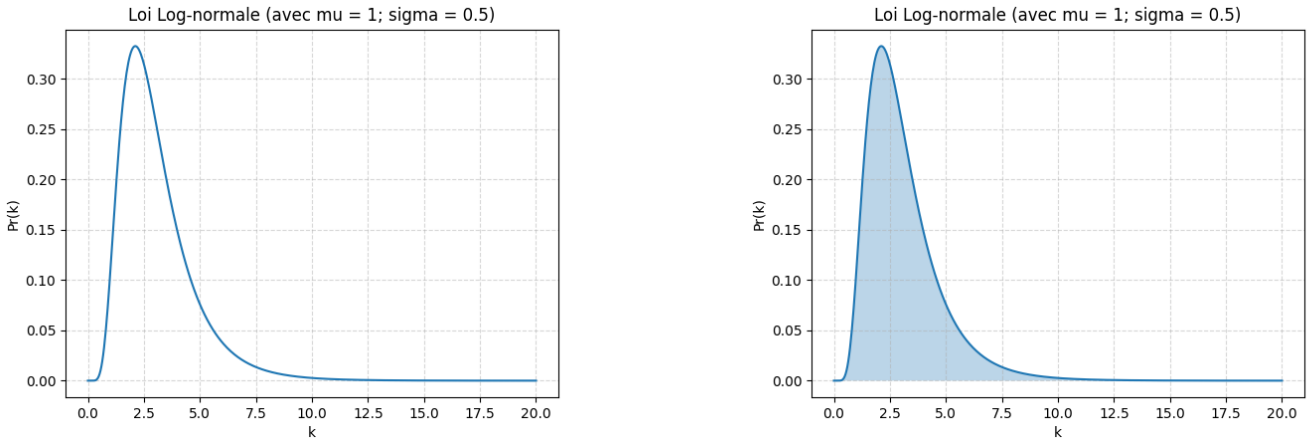


(a) Représentation courbe (b) Représentation intégrale

FIGURE 12 – Distribution de la loi Normale - continue

Loi log-normale

Voici en *Figure 13* deux représentations de la même distribution log-normale, avec $\mu = 1$ et $\sigma = 0.5$.



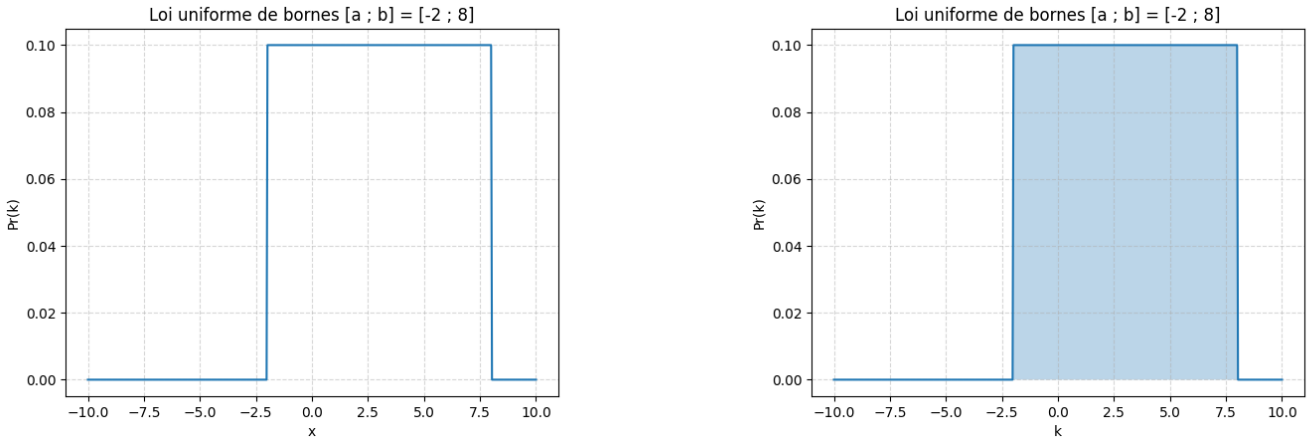
(a) Représentation courbe (b) Représentation intégrale

FIGURE 13 – Distribution de la loi log-normale - continue

Loi Uniforme

On retrouve la loi uniforme dans son aspect continu. Ici, c'est tout un intervalle qui a la même probabilité de se réaliser. La probabilité est alors de 1/(largeur de l'intervalle).

Voici en *Figure 14* deux représentations de la même distribution uniforme continue, de bornes [-2 ;8] (donc de largeur 10) et donc de probabilité $1/10 = 0.1$

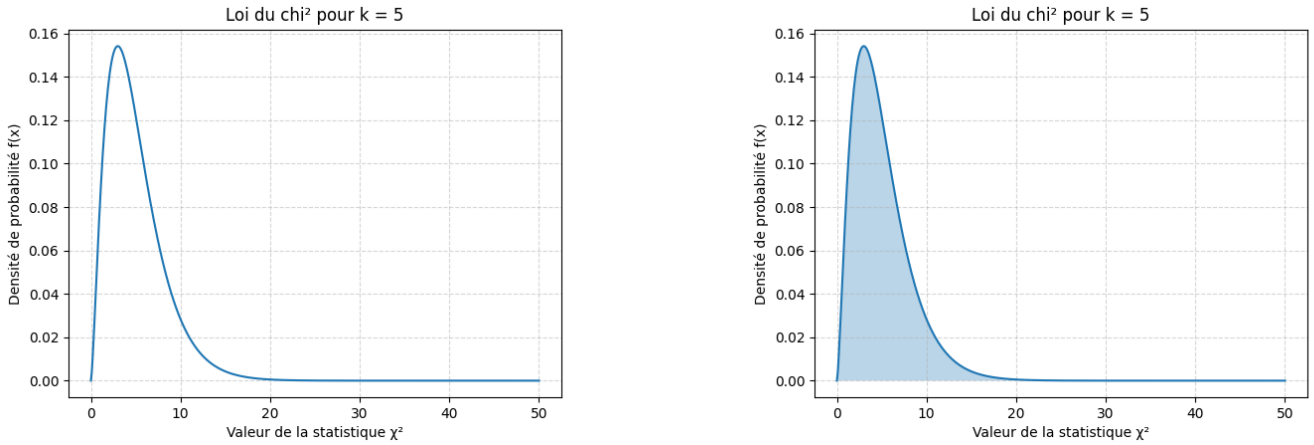


(a) Représentation courbe (b) Représentation intégrale

FIGURE 14 – Distribution de la loi Uniforme - continue

Loi du χ2

Voici en *Figure 15* deux représentations de la même distribution d'une loi de χ^2 de paramètre k=5.



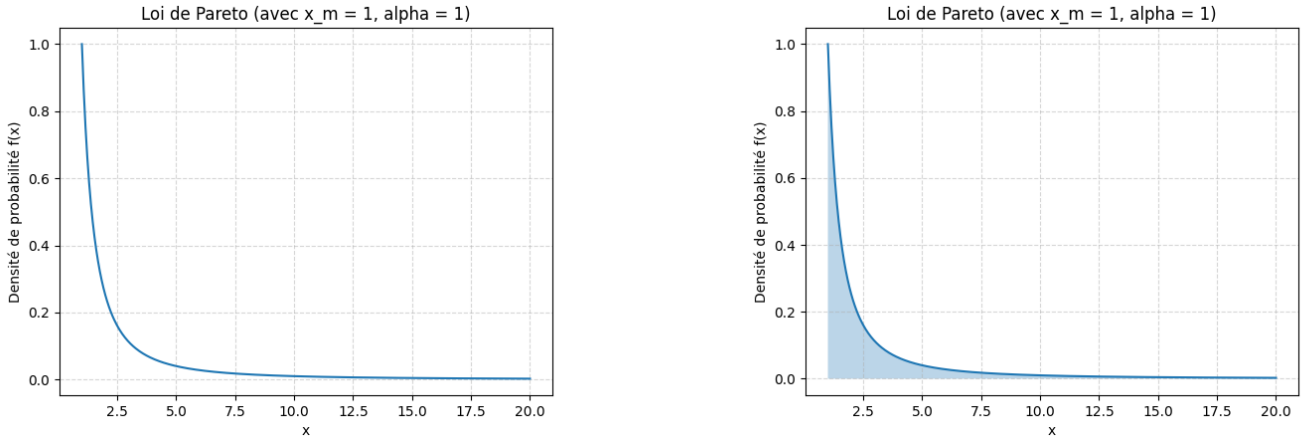
(a) Représentation courbe (b) Représentation intégrale

FIGURE 15 – Distribution de la loi du χ2 - continue

Loi de Pareto

Enfin, nous donnons ici un exemple de la loi de Pareto, aussi connue sous le nom de la loi des "80-20". Dans beaucoup de domaines indépendants, il est notable qu'environ 20% de l'investissement représente 80% des bénéfices. Ou encore 20% des semences donnent 80% des récoltes.
On généralise donc cette loi grâce à la distribution continue de Pareto.

En voici en *Figure 16* deux représentations, pour des paramètres $x_m = 1$ et $\alpha = 1$.



(a) Représentation courbe (b) Représentation intégrale

FIGURE 16 – Distribution de la loi de Pareto - continue

5 Séance 5 - Statistiques inférentielles

5.1 Questions de cours

L'étude statistique d'une population entière est souvent impraticable en raison d'un trop grand nombre d'individus à analyser ou de coûts prohibitifs, ce qui conduit à prélever un échantillon représentatif pour inférer des caractéristiques générales. L'échantillonnage désigne ainsi le processus de sélection d'un sous-ensemble d'une population mère, permettant d'étendre les observations à l'ensemble plus large grâce à des méthodes probabilistes. Au lieu d'examiner exhaustivement la population, comme dans un recensement, on opte pour cette approche afin d'obtenir des résultats fiables avec une fluctuation contrôlée, évitant des biais liés à une collecte totale impossible, par exemple pour estimer les intentions de vote de millions de personnes. Parmi les méthodes d'échantillonnage, on distingue les aléatoires, telles que le sondage aléatoire simple avec tirage au sort, avec ou sans remise, qui repose sur une base de sondage pour assurer l'équiprobabilité, et les non aléatoires, comme l'échantillonnage systématique basé sur un pas fixe ou la méthode des quotas qui respecte les proportions de variables corrélées pour minimiser les biais. Le choix d'une méthode dépend de facteurs comme la disponibilité d'une liste des individus, le budget, la représentativité requise et la nature de la population : les aléatoires sont idéales pour des inférences valides en cas de ressources suffisantes, tandis que les non aléatoires conviennent mieux à des contextes hétérogènes ou contraints, en évaluant toujours l'erreur potentielle pour garantir la généralisation.

Un estimateur se définit comme une fonction aléatoire appliquée à un échantillon pour approximer un paramètre inconnu de la population, tel que la moyenne ou la variance, et il est caractérisé par des propriétés comme la convergence vers la valeur vraie. L'estimation, en revanche, représente la valeur numérique spécifique obtenue à partir des données observées, constituant une réalisation ponctuelle de cet estimateur. Cette distinction est essentielle en analyse de données, où l'estimateur reste théorique et variable, tandis que l'estimation fournit une approximation concrète, souvent évaluée par son biais et sa variance pour assurer la fiabilité.

L'intervalle de fluctuation décrit la plage de variabilité attendue pour une statistique due au hasard de l'échantillonnage, calculée pour une probabilité donnée et liée à la distribution sous-jacente. À l'opposé, l'intervalle de confiance est une estimation intervalle d'un paramètre fixe, construit à partir d'un échantillon unique avec un niveau de confiance indiquant la probabilité que, sur de multiples répétitions, il capture la valeur vraie. Bien que les deux concepts soient interconnectés via des théorèmes comme celui central limite, le premier met l'accent sur la dispersion des statistiques, tandis que le second oriente vers l'inférence paramétrique, avec une interprétation méthodologique plutôt que probabiliste sur l'intervalle lui-même.

Dans la théorie de l'estimation, un biais désigne l'écart systématique entre l'espérance d'un estimateur et le paramètre réel, pouvant provenir d'une sélection non aléatoire ou d'hypothèses erronées, menant à des sous-estimations ou surestimations persistantes. Un estimateur non biaisé, comme la moyenne d'un échantillon, aligne son espérance sur le paramètre, mais en présence de biais, l'erreur quadratique moyenne augmente, soulignant l'importance de méthodes correctives pour des inférences précises.

Une statistique opérant sur la population totale porte le nom de paramètre de population ou statistique descriptive exhaustive, fournissant des valeurs exactes sans incertitude d'échantillonnage, comme dans un recensement. Ce lien avec les données massives est évident : avec les volumes croissants de big data, issus de sources comme les capteurs ou les réseaux sociaux, on peut traiter quasi-intégralement la population, transformant l'inférence en description pure, bien que des défis comme les biais de sélection persistent, nécessitant des outils scalables pour exploiter pleinement cette exhaustivité.

Les enjeux du choix d'un estimateur résident dans sa capacité à équilibrer non-biais, efficacité et robustesse, en minimisant la variance tout en convergeant vers le paramètre vrai, comme illustré par des critères tels que l'inégalité de

Cramer-Rao. Un estimateur inadapté peut entraîner des décisions erronées, d'où l'évaluation via des simulations ou des validations croisées, particulièrement critique en analyse de données pour gérer des distributions complexes.

Les méthodes d'estimation d'un paramètre incluent les moments, qui égalent les statistiques empiriques aux théoriques pour une simplicité computationnelle, le maximum de vraisemblance, optimisant la probabilité des données observées pour des lois paramétriques, les moindres carrés pour des contextes de régression, et les approches bayésiennes intégrant des priors pour incorporer des connaissances antérieures. La sélection repose sur la distribution des données, les hypothèses sous-jacentes et les objectifs : le maximum de vraisemblance est souvent privilégié pour son efficacité asymptotique, évaluée par des techniques comme le bootstrap en cas d'incertitudes.

Parmi les tests statistiques, on recense les paramétriques comme le t-test pour comparer des moyennes ou l'ANOVA pour des groupes multiples, assumant normalité et homoscédasticité, et les non-paramétriques tels que Mann-Whitney pour des médianes sans hypothèses distributives, Wilcoxon pour des données appariées, ou Kruskal-Wallis pour plusieurs échantillons. Ces tests servent à valider des hypothèses nulles contre alternatives, en mesurant la signification via une p-value pour détecter des effets ou des différences. Pour créer un test, on formule les hypothèses, choisit une statistique sensible sous l'alternative, détermine sa distribution sous la nulle via des théorèmes limites, fixe un seuil alpha, et rejette si la p-value est inférieure, assurant ainsi une décision contrôlée en termes de risques d'erreur.

Les critiques de la statistique inférentielle, soulignant des hypothèses nulles souvent fausses, une dépendance excessive à la taille d'échantillon menant à des significativités artificielles, ou des confusions sur les p-values, sont fondées et appellent à une pratique plus nuancée. Elles encouragent l'usage complémentaire d'intervalles de confiance et de mesures d'ffet pour éviter les pièges, renforçant finalement la robustesse de l'inférence en analyse de données plutôt que de la discréditer.

5.2 Manipulations avec Python

5.2.1 Théorie de l'échantillonnage

Nous avons accès à un fichier qui présente 100 échantillons aléatoires. Pour chaque échantillon, il y a un nombre de votes "Pour", un nombre de vote "Contre" et un nombre de vote "Sans opinion".
Voici en *Table 6* les moyennes obtenues sur les 100 échantillons.

Vote	Moyenne
Pour	391
Contre	416
Sans opinion	193

TABLE 6 – Moyennes obtenues sur les 100 échantillons

Voici en *Table 7* les fréquences obtenues sur les 100 échantillons.

Vote	Fréquence (2 décimales)	Fréquence (3 décimales)
Pour	0.39	0.391
Contre	0.42	0.416
Sans opinion	0.19	0.193

TABLE 7 – Fréquences obtenues sur les 100 échantillons

Voici en *Table 8* les fréquences de la population mère.

Vote	Fréquence (2 décimales)	Fréquence (3 décimales)
Pour	0.39	0.390
Contre	0.42	0.417
Sans opinion	0.19	0.193

TABLE 8 – Fréquences de la population mère

D'après les *Tables 7* et *8* on pourrait penser que les fréquences sont égales, si on ne regardait qu'avec une précision de 2 décimales. Mais la troisième colonne des tableaux révèle que les fréquences calculées avec les 100 échantillons ne font qu'approcher celles de la population mère.

Il convient donc de calculer les 3 intervalles de fluctuation, pour déterminer si il est plausible que les échantillons soient bien représentatifs de la population mère.
Voici en *Table 9* les intervalles de fluctuation.

Vote	Intervalle Fluctuation	Fréquence (3 décimales)
Pour	[0.2943, 0.4855]	0.390
Contre	[0.3203, 0.5136]	0.417
Sans opinion	[0.1158, 0.2705]	0.193

TABLE 9 – Intervalles de fluctuation

Comme les 3 valeurs de fréquences sont comprises dans les intervalles de fluctuation, on en conclut que l'échantillon est représentatif de la population mère - avec 95% de chances d'avoir raison : du fait qu'on ait calculé les intervalles avec le facteur "d'exigence" $z_c = 1.96$.

5.2.2 Théorie de l'estimation

Dans cette sous-partie, familiarisons-nous avec la théorie de l'estimation.
Cette fois-ci nous ne disposons que d'un unique échantillon, que voici :

$$\text{Échantillon} = [395 \text{ votes } \textit{Pour} \text{ , } 396 \text{ votes } \textit{Contre} \text{ , } 209 \text{ votes } \textit{Sans opinion}]$$

Voici en *Table 10* les fréquences obtenues de l'échantillon isolé.

Vote	Fréquence (3 décimales)
Pour	0.395
Contre	0.396
Sans opinion	0.209

TABLE 10 – Fréquences obtenues de l'échantillon isolé

Il convient donc de calculer les 3 intervalles de confiance, pour déterminer si il est plausible que l'échantillon soit bien représentatif de la population mère.
Voici en *Table 11* les intervalles de confiance.

Vote	Intervalle Confiance	Fréquence (3 décimales)
Pour	[0.3647, 0.4253]	0.395
Contre	[0.3657, 0.4263]	0.396
Sans opinion	[0.1838, 0.2342]	0.209

TABLE 11 – Intervalles de confiance

Comme les 3 valeurs de fréquences sont comprises dans les intervalles de confiance, on en conclut à nouveau que l'échantillon est représentatif de la population mère - avec 95% de chances d'avoir raison : du fait qu'on ait aussi calculé les intervalles de confiance avec le facteur "d'exigence" $z_c = 1.96$.

Remarque : l'intervalle de confiance est plus mince, donc plus restrictif que l'intervalle de fluctuation. En effet, plus l'échantillon est grand, plus on a "d'assurance". A l'inverse si l'on manque d'échantillon comme dans le cas de l'échantillon unique, on est plus restrictif car on émet un doute plus grand.

5.2.3 Théorie de la décision

On dispose de deux tests, nommées test 1 et test 2. Il s'agit de deux suites aléatoires de nombres.

La théorie de la décision va nous permettre de déterminer laquelle d'entre elle correspond à une distribution normale.

On applique alors des tests de Shapiro, avec la méthode `scipy.stats.shapiro()`. El les résultats qu'on obtient sont les suivants :

Test	paramètre stat	p_value
n°1	0.96394	6.2369 e-22
n°2	0.26089	0.0

TABLE 12 – Résultats des tests de Shapiro

Les résultats indiquent ceci : le paramètre stat du test 1 est assez proche de 1 (O.96 n'est pas considéré comme fortement proche, mais il renseigne une bonne correspondance entre les valeurs du test 1 et une distribution de loi normale). C'est alors que la p_value vient confirmer qu'on peut faire confiance au fait que stat = 0.96, car p_value

vaut 6.2369 e-22, ce qui est extrêmement proche de zéro.

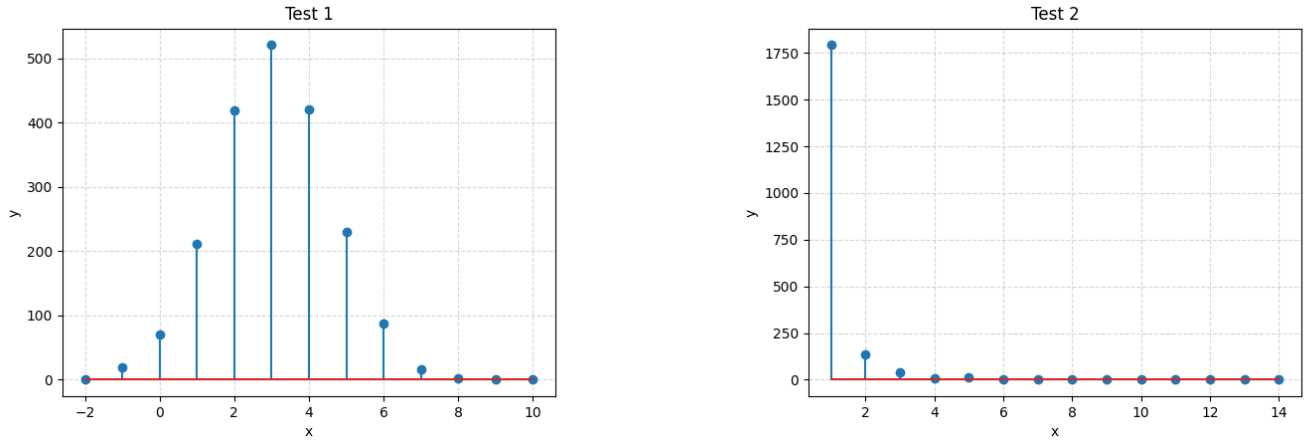
A l'inverse, bien que le p_value du test 2 soit lui aussi très proche de zéro (dans la réalité il n'est pas nul, mais si petit que la représentation informatique de la valeur vaut 0.0), celui-ci confirme que le paramètre stat vaut bien 0.26, ce qui est très loin de la valeur 1.

Donc ces deux tests de Shapiro nous montrent que le test 1 pourrait bien correspondre à une distribution normale, mais que le test 2 n'y correspond très probablement pas.

5.2.4 Bonus

Les deux figures suivantes permettent donc de déterminer la distribution de chaque test.

Voici en *Figure 17* les distributions discrètes des deux tests.

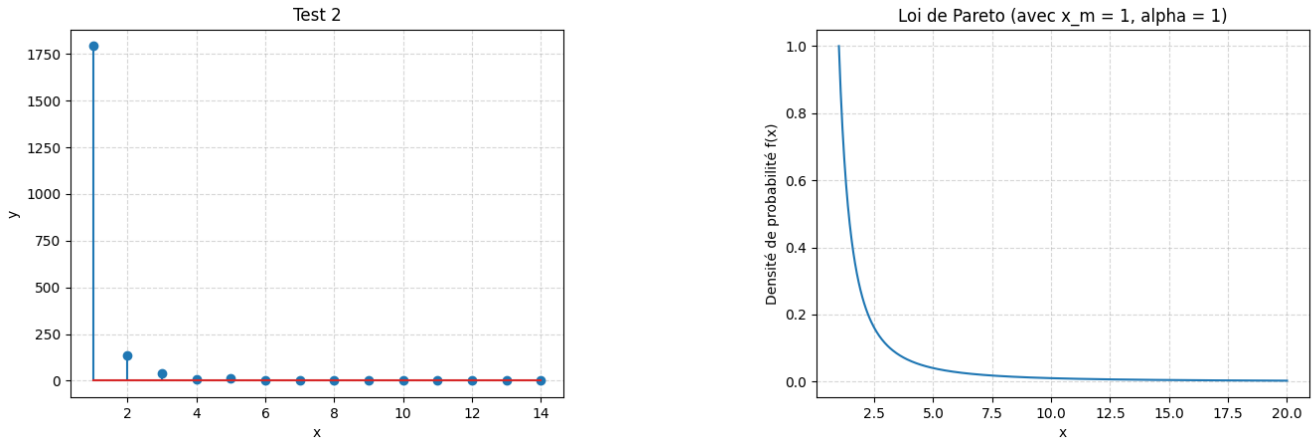


(a) Test 1 (b) Test 2

FIGURE 17 – Distribution discrète des deux tests

On voit en effet qu'une distribution normale correspondrait bien au comportement des valeurs du test 1. Cela conforte les résultats de la partie précédente.

En revanche, le test 2 correspondrait plutôt à une distribution de Pareto, comme on peut le constater en comparant les deux en *Figure 18* :



(a) Test 2 (b) Distribution de Pareto

FIGURE 18 – Comparaison des deux distribution

Il est d'ailleurs d'usage de privilégier la représentation visuelle, pour identifier une loi de Pareto.

6 Séance 6 - Statistique d'ordre des variables qualitatives

6.1 Questions de cours

Une statistique ordinale désigne une mesure appliquée à des données catégorielles où les catégories possèdent un ordre naturel intrinsèque, sans que les distances entre elles soient nécessairement égales ou quantifiables, permettant ainsi de classer les observations selon une hiérarchie implicite comme faible, moyen ou élevé. Elle s'oppose à la statistique nominale, qui traite des catégories sans ordre inhérent, telles que des couleurs ou des genres, où les étiquettes sont purement distinctives sans notion de rangement. Ce type de statistique utilise des variables ordinales, qui sont qualitatives mais ordonnées, issues d'observations indépendantes triées par rang croissant pour identifier des valeurs extrêmes ou des distributions. En géographie, cela matérialise une hiérarchie spatiale en révélant des structures comme les classements urbains ou les lois rang-taille, où des entités géographiques sont ordonnées par taille, importance ou intensité, soulignant des disparités territoriales et des dynamiques socio-économiques, par exemple dans l'analyse de crues fluviales ou de tremblements de terre pour anticiper des risques hiérarchisés.

Dans les classifications statistiques, l'ordre à privilégier est généralement l'ordre croissant, également appelé ordre naturel, qui facilite la détection de valeurs aberrantes trop petites ou trop grandes et permet d'étudier des lois comme celle de la plus grande valeur dans une série d'observations. Des exceptions existent en géographie, notamment pour des modèles comme la loi rang-taille où un ordre décroissant peut être appliqué pour mettre en évidence des hiérarchies inverses, mais l'approche croissante reste la norme pour assurer une cohérence dans l'analyse des fonctions de répartition et des densités de probabilité.

La corrélation des rangs mesure le degré d'association monotone entre deux classements ou variables ordinales, en évaluant si les rangs varient de manière similaire ou opposée, comme avec les coefficients de Spearman ou Kendall qui quantifient la force et la direction de cette relation pour deux séries. En revanche, la concordance de classements étend cette idée à plus de deux classements, en évaluant l'accord global entre plusieurs juges ou critères via des mesures comme le coefficient W de Kendall, qui normalise la dispersion des sommes de rangs pour déterminer si les classements multiples sont indépendants ou alignés, particulièrement utile pour des analyses multivariées où l'on teste une harmonie collective plutôt qu'une paire isolée.

Le test de Spearman, proposé en 1904, calcule un coefficient rs comme la corrélation de Pearson appliquée aux rangs, sensible aux écarts linéaires et plus adapté à des données continues converties en rangs, mais il peut être influencé par des erreurs ou des liens, rendant ses valeurs souvent plus élevées en absolu. À l'opposé, le test de Kendall, introduit en 1938, utilise tau basé sur le nombre de paires concordantes et discordantes, offrant une robustesse accrue aux outliers et une efficacité supérieure pour des distributions non normales, avec des calculs fondés sur des comparaisons binaires qui le rendent préférable pour des tests d'indépendance, bien que plus computationnellement intensifs pour de grands échantillons.

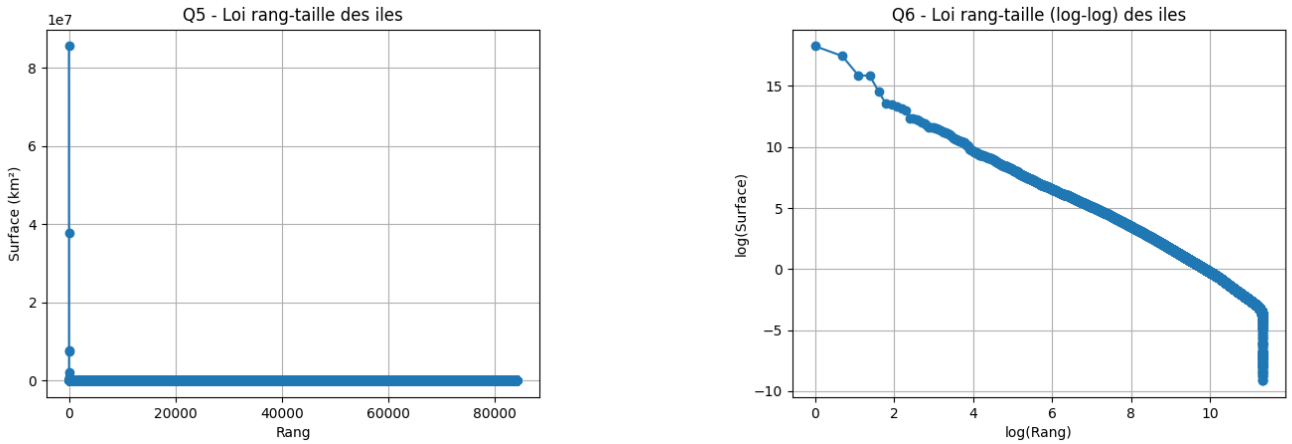
Les coefficients de Goodman-Kruskal, notamment gamma noté Γ, servent à mesurer l'association ordinale entre deux variables catégorielles en calculant le surplus de paires concordantes sur discordantes, variant de -1 à +1 pour indiquer la force et la direction de la relation, avec une interprétation probabiliste utile pour tester l'indépendance via un test de Student, tout en étant invariant sous transformations monotones. Le coefficient Q de Yule, un cas particulier de gamma pour des tableaux de contingence 2x2, évalue l'association entre deux variables dichotomiques via un rapport de cotes, allant de -1 (association négative totale) à +1 (positive parfaite), et est employé pour quantifier des liens binaires comme oui/non ou positif/négatif, facilitant l'analyse de dépendances dans des données qualitatives simples sans nécessiter d'hypothèses distributives.

6.2 Manipulations avec Python

On souhaite maintenant travailler sur la statistique d'ordre.

6.2.1 Partie Iles

Le premier aboutissement de cette séance à été de générer le graphe suivant, présenté en *Figure 19 (a)*. Puisque sa lecture est impossible, on a éditer le graphe en *Figure 19 (b)* :



(a) Loi rang-taille

(b) Loi rang-taille (log-log)

FIGURE 19 – Loi rang-taille des îles terrestres

Le graphe (b), étant corrigé par une représentation log-log nous permet ainsi de voir un lien évidente entre le rang et la taille. Mais ce lien ne peut pas être étudier d'un point de vue statique, car le "rang" dont on parle, a été imposé par nous de façon choisie.

Question 7

En effet, la réponse est : Non, on ne peut pas faire un test (statistique) sur les rangs.
Les rangs ne sont pas issus d'un échantillon aléatoire, mais ils sont générés par une simple numérotation ordonnée (1er, 2e, 3e, ...).

En effet, nous avons créé la liste des rangs après avoir trié la liste des surfaces. La corrélation entre le rang et la taille est donc évidente, car forcée "par construction". Le rang et la taille ne sont donc pas des variables indépendantes issues de mesures distinctes. Cela implique qu'on ne peut pas appliquer un test de corrélation (au sens statistique) entre le rang et la taille.

6.2.2 Partie Pays

Nous avons obtenus les classements suivants :
ord_pop_2007 = [[1, 'Chine'], [2, 'Inde'], [3, 'États-Unis'], [4, 'Indonésie'],...]
ord_pop_2025 = [[1, 'Inde'], [2, 'Chine'], [3, 'États-Unis'], [4, 'Indonésie'],...]
ord_densite_2007 = [[1, 'Singapour'], [2, 'Malte'], [3, 'Bangladesh'], [4, 'Maldives'],...]
ord_densite_2025 = [[1, 'Singapour'], [2, 'Malte'], [3, 'Bangladesh'], [4, 'Maldives'],...]

(Voir l'affichage du terminal lors de l'exécution du fichier `main.py`)

```
On fait alors le calcul du coefficient de corrélation rho entre
les rangs selon la population en 2007, et les rangs selon la densité en 2007.\\


La fonction \texttt{spearmanr()} renseigne deux valeurs:\\

#    rho : mesure la corrélation entre les deux classements.
#              - si rho proche de +1 : les classements sont très similaires
#              - si rho proche de -1 : les classements sont inversés
#              - si rho proche de  0 : les classements ne corrèlent pas
#
#     p-value : c'est la probabilité qu'il n'y ait aucune corrélation entre les classements.
#               (On note H0 l'hypothèse "nulle": H0 = "il n'y a aucune corrélation entre les classements")
#
#              - si p-value  < 0.05 (soit  < 5%): --> on rejette l'hypothèse H0
#                                                     (avec un risque de se tromper < 5%)
#                                                 --> la corrélation observée est donc
#                                                                          très probablement réelle
#                                                     (corrélation sûre à plus de 95%)
#
#              - si p-value >= 0.05 (soit >= 5%): --> on ne peut pas rejeter l'hypothèse H0
#                                                     (car on aurait un risque de se tromper >= 5%)
#                                                 --> aucune corrélation n'est donc prouvée.
#                                               Ainsi, si rho montre une corrélation,
#                                               celle-ci est peut-être dûe au hasard
#                                                     (corrélation sûre à moins de 95%)

Calcul du coefficient de concordance tau entre les rangs selon la population en 2007
 et les rangs selon la densité en 2007
#La fonction kendalltau() renseigne deux valeurs:
#     tau : mesure la concordance entre les deux classements.
#              - si tau proche de +1 : les classements concordent fortement
#              - si tau proche de -1 : les classements discordent fortement
#              - si tau proche de  0 : les classements n'ont pas de lien
#
#     p-value : c'est la probabilité qu'il n'y ait aucune concordance entre les classements.
#               (On note H0 l'hypothèse "nulle": H0 = "il n'y a aucune concordance entre les classements")
```

```
#
#              - si p-value  < 0.05 (soit  < 5%): --> on rejette l'hypothèse H0
#                                                     (avec un risque de se tromper < 5%)
#                                                 --> la concordance observée est donc
                                                                très probablement réelle
#                                                     (concordance sûre à plus de 95%)
#
#              - si p-value >= 0.05 (soit >= 5%): --> on ne peut pas rejeter l'hypothèse H0
#                                                     (car on aurait un risque de se tromper >= 5%)
#                                                 --> aucune concordance n'est donc prouvée.
# Ainsi, si tau montre une concordance, celle-ci est peut-être dûe au hasard
#                                                     (concordance sûre à moins de 95%)


Interprétation
# Pour l'année 2007, on obtient rho = 0.093 et tau = 0.067. Ces valeurs sont proches de 0, ce qui
# signifie qu'il n'y a ni corrélation ni concordance entre les classements selon la population et
# selon la densité.
#
# De plus, les 'p-value' des tests de Spearman et de Kendall sont respectivement de 0.224 et 0.192, ce qui
# est bien supérieur à 0.050 (5%). On ne peut donc pas rejeter l'hypothèse "nulle", selon laquelle "il n'y
# a pas de relation entre les classements".
#
# Conclusion : D'après les données de 2007, il n'y a donc pas de corrélation - statistiquement
# significative - entre
# la population d'un État et sa densité de population : ce sont deux grandeurs variant indépendamment.


#Pour l'année 2025

#Interprétation
# Pour l'année 2025, on obtient rho = -0.027 et tau = -0.007. Ces valeurs sont quasiment nulles, ce qui
# signifie qu'il n'y a ni corrélation ni concordance entre les classements selon la population et
# selon la densité.
#
# De plus, les 'p-value' des tests de Spearman et de Kendall sont respectivement de 0.709 et 0.877, ce qui
# est largement supérieur à 0.050 (5%). On accepte donc l'hypothèse "nulle", selon laquelle "il n'y
# a pas de relation entre les classements".
#
# Conclusion : D'après les données de 2025, il n'y a donc pas de corrélation - statistiquement
# significative - entre
# la population d'un État et sa densité de population : ce sont deux grandeurs variant indépendamment.
#
# N.B.:L'analyse des données de 2007 et l'analyse des données de 2025 aboutissent à cette même conclusion.
```

6.2.3 Partie Bonus

(Veuillez l'affichage du terminal lors de l'exécution du fichier `main.py`)

7 Conclusion : Réflexion sur les sciences des données et les humanités numériques

Au terme de ce parcours débutant en analyse de données avec Python, il apparaît clairement que les sciences des données et les humanités numériques constituent aujourd'hui un champ en pleine expansion, marqué par une hybridation croissante entre technologie, traitement automatisé de l'information et traditions intellectuelles issues des sciences humaines. Mais que sont les humanités numériques ? Les humanités numériques se définissent comme l'ensemble des connaissances textuelles ayant subi une numérisation, et plus largement, l'intégration du numérique dans les pratiques de recherche en lettres, sciences humaines et sociales. Cette perspective s'inscrit dans une continuité historique : dès le XIX^{e} siècle, l'automatisation du traitement textuel accompagnait le progrès industriel, réintroduisant les lettres au sein du domaine des sciences sous une forme renouvelée.

Aujourd'hui, le numérique n'est plus seulement un outil technique, mais un véritable objet de recherche et un instrument de communication, capable de rapprocher les connaissances scientifiques et de recomposer les projets des humanités du XVI^{e} au XX^{e} siècle. Il s'impose comme un facteur commun aux sciences et aux lettres, ouvrant de nouvelles perspectives interdisciplinaires. Les humanités numériques poursuivent ainsi plusieurs objectifs : revitaliser les filières des humanités et l'écriture scientifique, transformer les pratiques de recherche, encourager les collaborations entre disciplines et articuler approches qualitatives et quantitatives. Elles s'organisent également autour d'enjeux majeurs, tels que l'inscription des humanités dans la modernité technologique, la diversité des formes d'édition ou encore l'analyse des représentations et des usages du numérique.

Cependant, malgré leurs apports, les humanités numériques comportent certains risques : celui de privilégier les méthodes quantitatives au détriment de l'interprétation, de sous-estimer le contexte social de production des données, d'accentuer l'hégémonie de l'informatique dans les sciences humaines et sociales ou encore, de confondre le traitement des données avec la compréhension du sens. Ces limites mettent en évidence la nécessité d'une vigilance éthique et méthodologique : il revient au chercheur de connaître les outils, d'en maîtriser les limites et de maintenir un regard critique sur les pratiques employées.

De plus, l'évolution des liens entre humanités et informatique, des premières métadonnées des années 1970 jusqu'à l'avènement des humanités numériques, témoigne d'un processus structurant. L'usage croissant des bases de données a permis de cataloguer et relier les documents ; la numérisation du patrimoine a rendu accessibles des corpus volumineux ; le Web sémantique a offert aux machines la possibilité de comprendre et relier les données selon leur sens. Ces transformations ont largement contribué à la structuration des humanités numériques et à leurs enjeux contemporains.

Les cinq étapes de leur méthodologie (trouver l'information, modéliser les données, numériser les sources, analyser le contenu et valoriser les résultats), trouvent un écho direct dans ma pratique débutante de Python. Qu'il s'agisse de nettoyer un jeu de données, d'écrire des scripts d'automatisation ou de produire des visualisations (tableaux, graphiques, boîtes à moustaches, Khi², etc.), chaque étape rappelle que le numérique n'est pertinent que s'il est articulé autour d'une réflexion sur la structure, le sens et les usages des données.

Ainsi, ce parcours débutant trouve également des points d'ancrage concrets avec d'autres domaines de mon Master 1 GAED Géopolitique-GEOINT, notamment les Systèmes d'Information Géographique (SIG) et le Geospatial Intelligence (GEOINT). Dans le cours de SIG dispensé par M. de Matos-Machado, je manipule quotidiennement de vastes ensembles de données spatialisées (rasters, vecteurs, données attributaires), accompagnées de métadonnées fournies par des institutions comme l'Institut National de l'information Géographique et Forestière (IGN). Leur prise en compte est indispensable pour garantir la validité des analyses spatiales. L'apprentissage de Python m'a ainsi permis de mieux comprendre ces enjeux, notamment en automatisant le nettoyage, en détectant les incohérences ainsi que les données aberrantes et en structurant des bases complexes, ce qui rejoint directement les problématiques de rigueur et de norma-

lisation propres aux SIG.

En GEOINT, champ dédié à la fusion de données géolocalisées multi-sources et multi- capteurs, la gestion de données hétérogènes soulève des défis supplémentaires : compatibilité, interopérabilité, sécurité, protection de l'information. La difficulté à faire dialoguer des bases issues de services différents illustre parfaitement les enjeux des humanités numériques : comment croiser et interpréter des données produites dans des contextes variés ? La fusion de données satellitaires, terrestres, institutionnelles ou ouvertes met en lumière la nécessité d'une méthodologie rigoureuse et d'une compréhension fine des contraintes techniques et sémantiques.

Même des outils plus simples comme Excel (lien avec le cours "Méthodes quantitatives" de Madame Hugueninn-Richard) montrent combien la manipulation des données repose sur une compréhension fine de leur structure : trier, filtrer, structurer un tableau ou vérifier la cohérence des valeurs constituent une première approche de la logique de la donnée. La transition entre Excel et Python n'est donc pas une rupture mais plutôt une montée en complexité, permettant une automatisation poussée, une reproductibilité accrue et une capacité à traiter des volumes bien plus importants. Ces compétences se retrouvent au cœur de la cultural analytics, des digital methods ou encore de la distant reading, qui renouvellent l'étude des textes, des images, des discours et des comportements numériques. Elles montrent que la maîtrise des outils ne peut être dissociée d'une réflexion critique sur la nature des données et leurs analyses.

Ainsi, cette initiation à l'analyse de données avec Python m'a permis de développer des compétences techniques essentielles pour les sciences humaines et sociales. Elle m'a également fait prendre conscience de la place désormais centrale du numérique dans la production de connaissances, en particulier dans la pratique de la géographie. Les disciplines des sciences humaines et sociales mobilisent aujourd'hui une grande diversité de jeux de données (spatiales, textuelles, statistiques ou multimédias). La maîtrise d'outils tels que Python devient alors indispensable pour les structurer, les analyser et en tirer des interprétations pertinentes.

Ce cours m'a également permis de comprendre que les humanités numériques ne constituent pas une rupture avec les traditions intellectuelles, mais bien un prolongement des méthodes d'analyse et un accélérateur des traitements et des opérations complexes. Elles permettent à la fois de revisiter des questionnements anciens grâce à de nouvelles capacités techniques, tout en renforçant la complémentarité entre rigueur scientifique, interprétation qualitative et analyse quantitative.

Cependant, j'ai pris conscience que cette évolution rapide des outils impose une vigilance éthique. À l'avenir, les humanités numériques et l'analyse de données seront donc confrontées à des défis majeurs qui redéfiniront profondément nos manières de produire, d'organiser et d'interpréter le savoir. L'augmentation continue des volumes d'information, leur diversification et leur circulation mondiale imposeront des capacités techniques et théoriques toujours plus pointues, notamment en matière de gestion des big data, de standardisation, de sécurité et de transparence des méthodes. Les frontières entre disciplines continueront de s'estomper, tandis que les questions de souveraineté numérique, de protection des données, de biais algorithmiques et de traçabilité des modèles deviendront centrales dans les sciences humaines et sociales. L'arrivée d'IA génératives, capables d'interagir avec les corpus, les images ou les données spatiales, posera également de nouveaux dilemmes : Comment garantir la fiabilité des analyses ? Comment préserver l'intégrité du sens lorsque les traitements sont automatisés ? Quelle place accorder à l'interprétation humaine dans des environnements où l'automatisation devient la norme ? Fort de ce constat, je dois ainsi veiller à prendre du recul sur les technologies utilisées, en particulier sur les intelligences artificielles qui s'intègrent progressivement aux workflows d'analyse de données. Leur développement transforme en profondeur les pratiques : automatisation accrue, assistance aux traitements, génération de contenus. Plus que jamais, l'avenir des humanités numériques dépendra de notre capacité à concilier innovation technologique, exigence scientifique et responsabilité éthique, afin de construire une recherche qui demeure à la fois rigoureuse, critique et profondément humaine.

Notes de bas de page

Cultural Analytics : Approche qui utilise des méthodes computationnelles (statistiques, vision par ordinateur, NLP) pour analyser de grands ensembles de données culturelles (images, textes, vidéos, plateformes sociales) afin de repérer des tendances, formes, styles ou dynamiques culturelles.

Digital Methods : Ensemble de méthodes qui exploitent les traces natives du web (liens, hashtags, likes, moteurs de recherche, réseaux sociaux) pour analyser des phénomènes sociaux et culturels à partir de données produites en ligne.

Distant Reading : Méthode d'analyse textuelle qui observe de grandes collections de textes grâce à l'informatique (modèles de topics, statistiques lexicales, embeddings, etc.) plutôt qu'une lecture rapprochée traditionnelle. L'objectif est d'identifier des motifs généraux ou macro-tendances.

Workflows : Ensemble structuré d'étapes permettant de collecter, nettoyer, analyser et visualiser des données. Dans l'analyse numérique, un workflow formalise la chaîne complète de traitement, souvent reproductible et automatisée.

- FIN DU COMPTE RENDU -