Cours d’analyse de données – Questions de cours

Séance 2

La géographie est une discipline qui cherche à comprendre comment les sociétés s’organisent dans l’espace, comment les milieux évoluent et comment les territoires fonctionnent. Pour répondre à toutes ces questions, les données jouent un rôle essentiel, et c’est là que les statistiques entrent en jeu. Elles sont plutôt utilisés comme un outil parmi d’autres. Elles permettent au géographe de passer de simples observations à des mesures plus solides et comparables. On pourrait dire qu’elles apportent une forme de rigueur indispensable lorsqu’il s’agit de décrire ou d’expliquer un phénomène spatial. En somme, la géographie s’appuie sur les statistiques non pas pour tout réduire à des chiffres, mais pour donner un fondement plus objectif aux analyses qu’elle mène.

Mais lorsqu’on parle d’objectivité, on peut se demander si le hasard occupe une place en géographie. À première vue, certains phénomènes semblent distribués au hasard sur le territoire, comme l’emplacement de certaines maisons isolées, des accidents de la route, ou encore l’ouverture de petits commerces. Pourtant, quand on creuse un peu, on se rend compte que le hasard est assez rare. La plupart du temps, ce qui paraît aléatoire cache en réalité une multitude de facteurs économiques, historiques, sociaux ou même environnementaux ; ce qui influencent la répartition dans l’espace. Le hasard existe, mais il est souvent « apparent ». Le travail du géographe, consiste justement à déterminer ce qui relève de vraies logiques spatiales et ce qui relève de l’aléatoire.

Pour étudier ces phénomènes, les géographes utilisent différents types d’informations. Certaines sont qualitatives : par exemple le type d’occupation du sol, le genre d’habitat ou encore une catégorie socioprofessionnelle. D’autres sont ordinales : elles introduisent une notion de classement ou de hiérarchie, comme des niveaux de risque ou des classes de revenus. Enfin, il existe les données quantitatives, sans doute les plus utilisées : elles peuvent être discrètes (le nombre d’habitants, de logements, de commerces) ou continues (l’altitude, la température, la distance). Comprendre la nature de ces données est fondamental, car cela conditionne le type d’analyse et les outils statistiques que l’on peut utiliser.

La géographie a donc de nombreux besoins du point de vue de l’analyse des données. Elle doit pouvoir décrire des situations spatiales, comparer des territoires, repérer des inégalités ou des dynamiques, identifier des ruptures ou des tendances. Elle a aussi besoin de synthétiser l’information lorsqu’elle travaille avec des bases de données très volumineuses. Et bien sûr, elle doit pouvoir représenter visuellement les phénomènes pour faciliter leur compréhension, telles des graphiques, cartes, diagrammes… Les statistiques permettent tout cela, en donnant de la cohérence et de la structure aux données brutes.

Dans cette perspective, il est essentiel de distinguer la statistique descriptive de la statistique explicative. La première se contente de présenter et de résumer l’information : calculer des moyennes, représenter des valeurs sur un graphique, construire une carte thématique… Elle répond à la question « que voit-on ? ». La statistique explicative, essaye d’aller plus loin : elle cherche à comprendre pourquoi on observe ces valeurs-là, quelles variables influencent un phénomène, quelles relations existent entre elles. C’est là que l’on utilise des régressions, des corrélations ou des

analyses multivariées. Les deux approches ne s’opposent pas, elles s’enchaînent : d’abord on décrit, ensuite on explique.

Pour rendre ces analyses plus lisibles, les visualisations de données sont essentielles. On utilise des diagrammes en barres pour comparer des catégories, des diagrammes circulaires pour montrer des proportions, des histogrammes pour représenter une distribution, des cartes pour localiser les phénomènes… Le choix du type de graphique dépend de plusieurs choses : le type de variable, le but de l’analyse, ce que l’on veut mettre en avant, mais aussi le public auquel on s’adresse. La géographie, parce qu’elle travaille sur l’espace, donne une place particulièrement importante aux cartes.

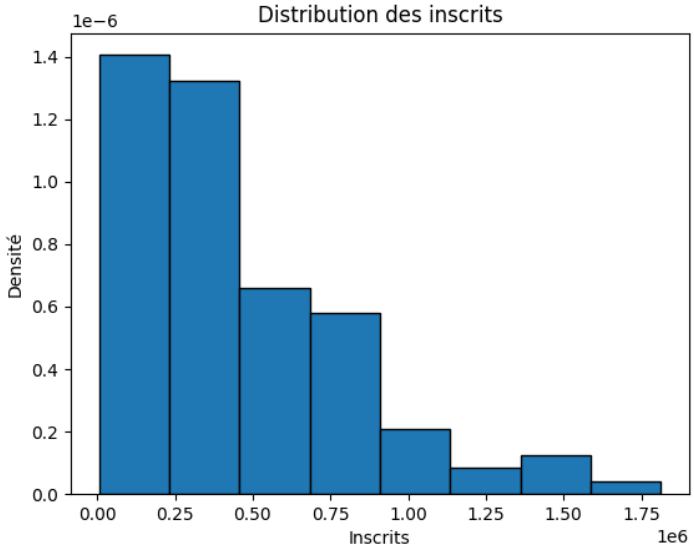
Les méthodes d’analyse de données utilisées en géographie sont nombreuses. Certaines se concentrent sur une seule variable, d’autres comparent deux variables, et d’autres encore permettent d’examiner plusieurs dimensions en même temps. Les analyses multivariées, comme l’ACP ou la CAH, sont particulièrement utilisées lorsqu’on veut dégager de grandes tendances dans des territoires complexes. Elles permettent de simplifier sans déformer, et de faire apparaître des structures que l’on ne verrait pas à l’œil nu.

Ces analyses reposent sur quelques notions fondamentales : la population statistique, qui désigne l’ensemble des unités étudiées ; l’individu statistique, qui est chaque élément de cette population (une commune, un ménage, un lieu ...) ; le caractère statistique, qui correspond à la propriété observée ; et la modalité, qui représente les différentes valeurs possibles de ce caractère. Les caractères peuvent être qualitatifs ou quantitatifs, et il existe effectivement une hiérarchie entre eux. Les variables quantitatives permettent plus de calculs et d’analyses, alors que les variables qualitatives offrent surtout des regroupements.

Lorsque l’on étudie une série statistique, il est souvent utile de mesurer l’amplitude ; c’est-à-dire l’écart entre les valeurs minimale et maximale car cela donne une idée de la dispersion des données. On calcule également la densité dans le cas des histogrammes, ce qui permet de représenter correctement les classes en tenant compte de leur largeur. Ces détails techniques jouent un rôle clé pour éviter de fausses interprétations.

Les formules de Sturges et de Yule sont justement là pour aider à déterminer le nombre optimal de classes dans un histogramme. Trop de classes rendent le graphique difficile à lire et trop peu le simplifient excessivement. Les formules proposent un compromis basé sur le nombre d’observations, ce qui permet d’obtenir une représentation équilibrée.

Enfin, pour construire une distribution statistique, on a besoin de comprendre ce que sont les effectifs, les fréquences et les fréquences cumulées. L’effectif représente le nombre d’individus correspondant à une modalité donnée. La fréquence exprime la proportion que cet effectif représente dans l’ensemble. Et la fréquence cumulée additionne ces proportions les unes après les autres. L’ensemble forme ce qu’on appelle la distribution statistique, qui est indispensable pour analyser comment les valeurs se répartissent dans une population.



L’objectif principal est de prendre en main un jeu de données réel à l’aide du langage Python, en particulier à travers les bibliothèques Pandas pour la manipulation des données et Matplotlib pour leur représentation graphique.
La première étape a consisté à importer correctement le fichier CSV. Cette opération permet de transformer les données brutes en un tableau structuré, appelé DataFrame, qui facilite grandement les traitements statistiques. Une fois les données chargées, l’analyse s’est concentrée sur les effectifs, notamment le nombre total d’inscrits et de votants, calculés à l’aide des méthodes de Pandas. Une attention particulière a été portée à la sélection des variables quantitatives, afin d’éviter des erreurs de calcul sur des colonnes qualitatives comme les noms de départements. Cette étape montre l’importance de bien connaître la structure d’un jeu de données avant d’effectuer des traitements automatiques.

Enfin, nous avons mis l’accent sur la visualisation des données. Des diagrammes en barres ont été réalisés pour comparer le nombre d’inscrits et de votants par département, tandis que des diagrammes circulaires (qui auraient dû apparaitre avec le codage) permettent de représenter la répartition des votes (blancs, nuls, exprimés, abstention). Un histogramme (image ci dessus) a également été construit afin d’observer la distribution du nombre d’inscrits. Ces représentations graphiques permettent de passer d’une lecture purement numérique à une lecture visuelle, beaucoup plus intuitive et adaptée à l’analyse géographique.

Les résultats obtenus lors de cette séance montrent clairement l’intérêt de Python pour explorer rapidement un jeu de données volumineux. Les calculs d’effectifs permettent d’avoir une vision globale de la participation électorale, tandis que les graphiques mettent en évidence des différences importantes entre départements. L’histogramme, en particulier, permet d’observer la dispersion du nombre d’inscrits et de repérer des départements très peuplés par rapport à la majorité.

Séance 3

Lorsqu'on parle des caractères statistiques, la première distinction à faire est celle entre les caractères qualitatifs et les caractères quantitatifs. Le caractère qualitatif est généralement considéré comme le plus général, dans le sens où il permet de classer les individus en catégories, même quand aucune information chiffrée n'est disponible. Il se contente de décrire une propriété ou une appartenance comme le type d'habitat, le mode de transport utilisé, la situation familiale… .Le caractère quantitatif, lui, va plus loin en exprimant une valeur numérique mesurable. Autrement dit, on peut toujours commencer par qualifier un phénomène, mais pouvoir le quantifier nécessite une information précise supplémentaire. C'est pour cela que le qualitatif est souvent vu comme plus général car il constitue la première étape dans la description d'un phénomène.

Parmi les caractères quantitatifs, on distingue les caractères discrets et les caractères continus. Un caractère discret prend des valeurs entières dénombrables (nombre d'enfants, nombre de logements, nombre de véhicules…). Un caractère continu, au contraire, peut prendre n'importe quelle valeur réelle dans un intervalle (la température, l'altitude, une distance, un revenu exprimé à l'euro près…). La distinction est importante car elle conditionne le type d'analyse ou de représentation à utiliser. Par exemple, un caractère discret peut se représenter avec un diagramme en barres, alors qu'un caractère continu nécessite souvent un histogramme puisqu'il s'agit d'une distribution.

Les statistiques utilisent ensuite plusieurs paramètres dit « de position », car ils permettent de situer une série de données (la moyenne, la médiane, le mode). Il existe plusieurs types de moyennes (arithmétique, géométrique, harmonique) parce que les phénomènes ne se comportent pas tous de la même manière. Une moyenne arithmétique convient à des valeurs additives, une moyenne géométrique est utilisée dans les phénomènes multiplicatifs (comme des taux d'évolution), et la moyenne harmonique est utile lorsqu'on travaille avec des vitesses, des ratios ou des taux.

La médiane, va servir à repérer le point où une population se divise en deux groupes égaux. On la calcule lorsqu'on veut éviter l'influence des valeurs extrêmes, ce qui est parfois essentiel. Par exemple, pour décrire les revenus d'une population très inégale, la moyenne est souvent trompeuse, alors que la médiane donne une idée beaucoup plus réaliste de la situation. Quant au mode, il n'a de sens que lorsqu'une valeur revient plus fréquemment que les autres, c'est un paramètre intéressant pour des données qualitatives ou fortement regroupées, mais il n'est pas toujours utile pour des données continues.

Les paramètres de concentration permettent d'évaluer comment les valeurs se répartissent au sein d'une série. La médiale (ou médiane de Lorenz) et l'indice de Gini sont des outils conçus pour mesurer les inégalités. Ils sont particulièrement utilisés pour des revenus, des richesses ou toute variable dont on veut étudier la répartition. L'indice de Gini, par exemple, résume en un seul chiffre le degré d'inégalité : plus il est proche de 1, plus la concentration est forte. Ces outils sont précieux

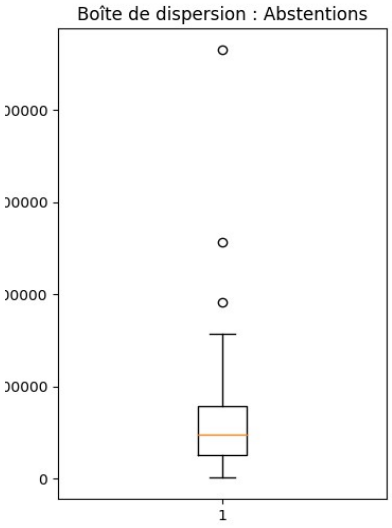
car ils permettent de dépasser les simples moyennes, qui masquent souvent la manière dont les valeurs sont réellement réparties.

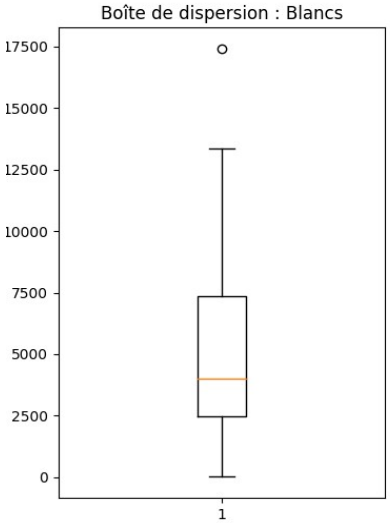
Les paramètres de dispersion servent quant à eux à mesurer l'écart entre les valeurs d'une série. La variance est préférée à l'écart moyen à la moyenne parce qu'elle prend en compte l'ensemble des écarts en les élevant au carré, ce qui évite que les écarts positifs et négatifs s'annulent. Cependant, comme la variance produit une unité de mesure différente de celle de la variable d'origine (par exemple des m^2 pour une variable exprimée en mètres), on utilise souvent l'écart-type, qui est simplement la racine de la variance et qui revient donc dans l'unité initiale. L'étendue est la différence entre la plus grande et la plus petite valeur, c'est une mesure simple mais informative car elle donne immédiatement une idée de la largeur de la distribution.

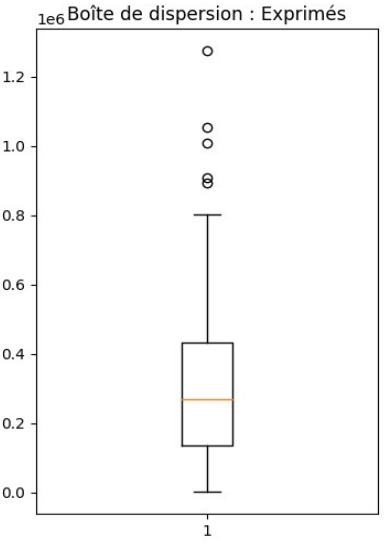
Les quantiles quant à eux, servent à découper la population en parts égales. Ils permettent de voir comment sont réparties les valeurs et de repérer des seuils importants. Les quantiles les plus utilisés sont les quartiles (qui découpent en quatre parts) et les déciles (dix parts). Le quantile le plus célèbre reste la médiane, qui correspond au 50^e percentile. Ce découpage en quantiles est à la base de la construction d'une boîte de dispersion. Cette boîte résume en un seul graphique la valeur minimale, les quartiles, la médiane et les valeurs extrêmes. Elle permet d'un simple coup d'œil d'apprécier la dispersion, la position centrale, et même d'identifier des valeurs atypiques. C'est un outil très efficace pour comparer plusieurs distributions entre elles.

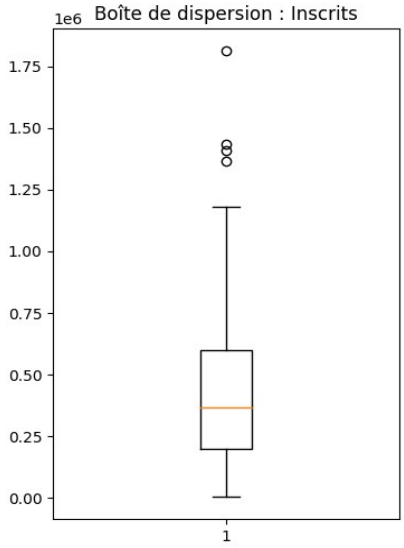
Enfin, les paramètres de forme s'intéressent à la manière dont une distribution se comporte. On distingue les moments centrés et les moments absolus. Les moments centrés se calculent à partir de la moyenne et permettent de mesurer des propriétés comme l'asymétrie ou l'aplatissement. Les moments absolus, utilisent donc des valeurs absolues, ce qui les rend plus robustes lorsqu'il y a des valeurs extrêmes importantes. On les utilise pour mieux décrire la structure réelle d'une distribution, notamment quand elle n'est pas normale.

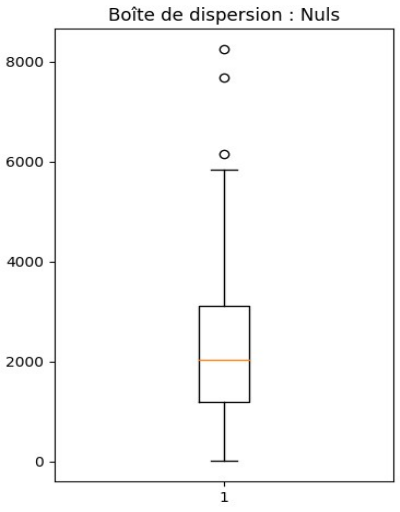
Vérifier la symétrie d'une distribution est utile parce qu'une distribution symétrique est souvent plus facile à analyser : moyenne, médiane et mode coïncident, et les modèles statistiques classiques s'appliquent mieux. Pour identifier la symétrie ou l'asymétrie, on compare souvent la moyenne et la médiane : si la moyenne est plus grande, la distribution est probablement asymétrique vers la droite ; si elle est plus petite, elle est asymétrique vers la gauche. On peut également regarder l'aspect du histogramme ou calculer un coefficient d'asymétrie. Ces informations aident à comprendre comment les valeurs se répartissent réellement, ce qui influence fortement les choix méthodologiques.

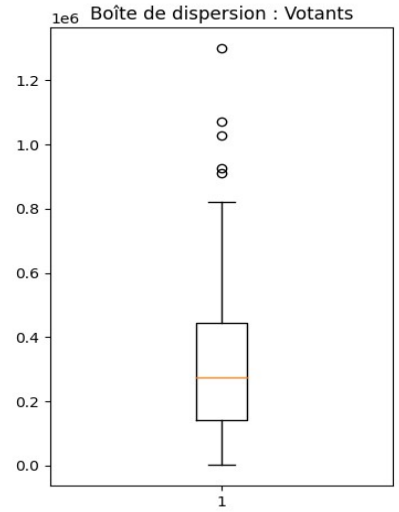

Boîte de dispersion : Abstentions
1


Boîte de dispersion : Blancs
17500
15000
12500
10000
7500
5000
2500
0
1


1e6
Boîte de dispersion : Exprimés
1.2
1.0
0.8
0.6
0.4
0.2
0.0
1


1e6
Boîte de dispersion : Inscrits
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
1


Boîte de dispersion : Nuls
8000
6000
4000
2000
0
1


1e6
Boîte de dispersion : Votants
1.2
1.0
0.8
0.6
0.4
0.2
0.0
1

La séance 3 s’inscrit dans la continuité directe de la séance 2, mais avec un objectif plus analytique. Il ne s’agit plus seulement de décrire et de visualiser les données, mais de calculer des paramètres statistiques précis afin de caractériser les distributions.

À partir des colonnes quantitatives du même fichier électoral, plusieurs indicateurs ont été calculés (moyennes, médianes, modes, écarts types, écarts absolus à la moyenne et étendues). Ces paramètres ont été obtenus grâce aux méthodes intégrées de Pandas, ce qui permet de garantir à la fois rapidité et fiabilité des calculs. L’utilisation de la fonction « `round()`» a permis d’harmoniser les résultats et d’améliorer leur lisibilité.

La séance a également introduit le calcul des quantiles, en particulier les quartiles et les déciles, afin de mieux comprendre la dispersion des valeurs. Ces quantiles ont ensuite été mobilisés pour tracer des boîtes à moustaches, qui offrent une synthèse graphique très efficace de la distribution d’une variable. Les boxplots permettent notamment de repérer les valeurs extrêmes et de comparer visuellement plusieurs variables entre elles.

Enfin, un second jeu de données portant sur la surface des îles a été utilisé pour apprendre à catégoriser une variable quantitative continue. Cette étape a permis de passer d’une logique purement descriptive à une logique de classification, essentielle en géographie pour construire des typologies territoriales.

Les résultats permettent donc d’aller plus loin dans la compréhension des données. Les moyennes donnent une première idée des valeurs typiques, mais ce sont surtout la médiane et les quantiles qui révèlent la structure réelle des distributions, souvent asymétriques. Les écarts types et les étendues montrent que certaines variables présentent une forte dispersion, ce qui confirme l’existence de fortes inégalités territoriales. Donc pour conclure les boîtes à moustaches sont particulièrement intéressantes, car elles permettent de visualiser rapidement ces écarts et d’identifier des valeurs atypiques.

Séance 4

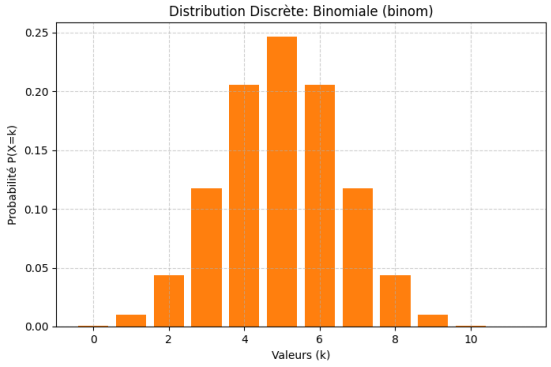
Le point de départ de toute analyse quantitative est la compréhension de la variable étudiée, qui impose le choix entre un modèle discret ou continu.
La distinction entre les variables statistiques repose sur leur capacité à prendre des valeurs. Pour choisir entre une distribution à variables discrètes et une distribution à variables continues, on va donc se baser sur la nature de la mesure.
Les variables discrètes sont utilisées lorsque la variable est le résultat d'un comptage et ne peut prendre qu'un nombre fini ou dénombrable de valeurs entières (le nombre d'habitants, le nombre de séismes…). Leurs probabilités sont définies par une Fonction de Masse de Probabilité (PMF).
Les variables continues quant à elles, modélisent les phénomènes résultant d'une mesure et peuvent prendre n'importe quelle valeur réelle dans un intervalle donné (la température, le revenu, l'altitude…). Leurs probabilités sont définies par une Fonction de Densité de Probabilité (PDF).
La pertinence d'une distribution dépend donc entièrement de la variable géographique à modéliser.
En géographie, certaines lois se distinguent par leur capacité à décrire les phénomènes spatiaux complexes, c'est le cas de 3 loi :

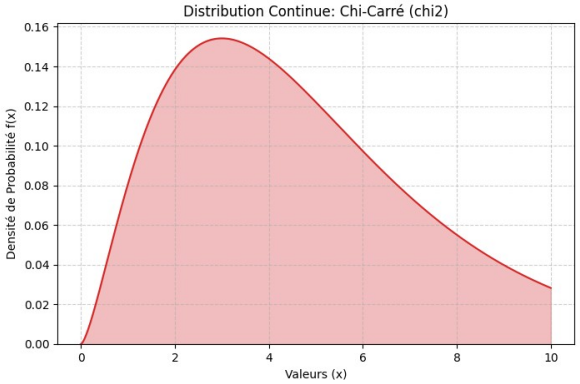
- La **Loi Normale** est omniprésente, souvent utilisée pour les erreurs ou les caractéristiques humaines agrégées.
- La **Loi de Poisson** est essentielle pour les événements rares sur un territoire (comptage de points), comme les foyers d'une maladie ou l'occurrence de feux de forêt.
- La **Loi de Pareto** est fondamentale dans l'étude des inégalités et des hiérarchies (ex : la taille des villes, la concentration des richesses…).

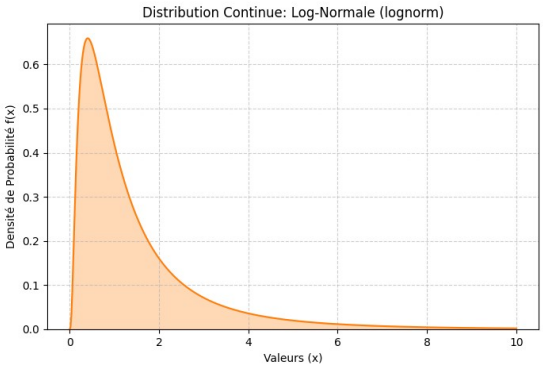
Les manipulations Python ont concrétisé ces concepts en calculant la moyenne et l'écart type pour divers modèles (Binomiale, Normale, Pareto…), montrant comment chaque loi est caractérisée par ses propres paramètres analytiques.

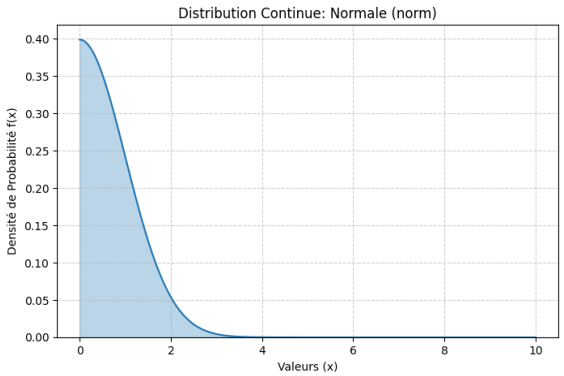
Nom de la loi	Nature de la variable	Moyenne	Ecart-type	Objectif de la manipulation
Dirac	Discrète	5,00	0,00	Modélisation d'une probabilité ponctuelle unique
Uniforme		3,50	1,7078	Distribution équiprobable de valeurs entières
Binomiale		5,00	1,5811	Succès cumulés d'épreuves de Bernoulli indépendantes
Poisson		3,00	1,7321	Modélisation des occurrences d'évènement rares
Zipf-Mandelbrot		1,8899	1,6973	Étude des hiérarchies (ex: rang-taille des villes)
Normale	Continue	0,00	1,00	Référence pour la distribution des mesures

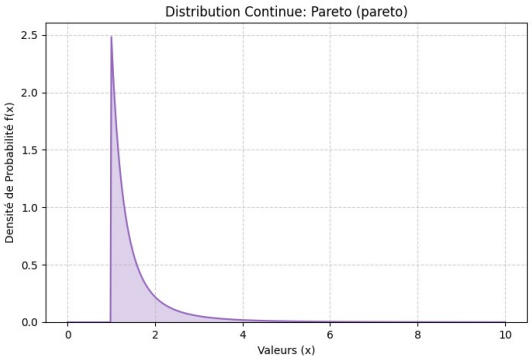
				physiques
Log-normale		1,5703	1,9011	Analyse de variables à forte croissance (revenus)
Uniforme		5,00	2,8868	Densité de probabilité constante sur un intervalle
Chi-carré		5,00	3,1623	Préparation aux tests d'ajustement et de dépendance
Pareto		1,6667	1,4907	Modélisation des fortes disparités et inégalités

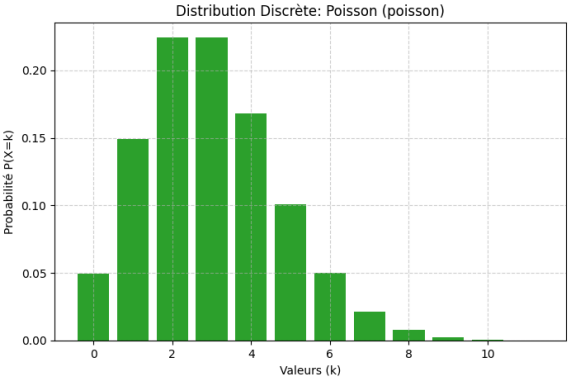












La Loi de Dirac : La figure obtenue est la plus singulière, car elle ne présente qu'une seule barre verticale de probabilité (P=1) sur une valeur précise, toutes les autres étant nulles. Elle modélise une certitude absolue ou une localisation ponctuelle sans aucune dispersion spatiale.
La Loi Uniforme Discrète : Cette figure se caractérise par une série de barres de hauteurs strictement identiques sur un intervalle défini. Elle illustre une situation d'équiprobabilité parfaite, où aucune valeur n'est privilégiée par rapport à une autre.
La Loi Binomiale : Le graphique affiche une distribution en forme de cloche, mais composée de barres distinctes. Elle représente le nombre de succès lors de répétitions d'expériences indépendantes. Plus le nombre d'essais augmente, plus sa silhouette se rapproche visuellement d'une courbe normale.
La Loi de Poisson : Pour de faibles moyennes, la figure montre une forte asymétrie vers la gauche, avec une décroissance rapide des probabilités pour les valeurs élevées. Elle est l'outil privilégié pour visualiser la répartition d'événements rares dans l'espace ou le temps.
La Loi de Zipf-Mandelbrot : La figure obtenue montre une chute brutale de la probabilité entre le premier rang et les suivants. C'est une courbe de hiérarchie qui explique, par exemple, pourquoi la ville principale d'un pays est souvent bien plus grande que ses subordonnées.
La Loi Normale : C'est la figure de référence, affichant une courbe en cloche parfaitement symétrique autour de sa moyenne. Elle symbolise l'équilibre et la convergence des phénomènes géographiques influencés par de multiples facteurs aléatoires.
La Loi Log-normale : Contrairement à la précédente, cette courbe présente une asymétrie marquée avec une « longue traîne » vers la droite. Elle permet de visualiser des distributions où une minorité d'individus ou d'entités possède des valeurs très élevées (comme la taille des entreprises).
La Loi Uniforme Continue : La figure prend la forme d'un rectangle parfait sur un intervalle donné, indiquant que la densité de probabilité est constante.
La Loi du chi2 : La forme de cette courbe évolue de manière significative selon les degrés de liberté injectés. Elle est visuellement asymétrique pour de petites valeurs et s'élargit au fur et à mesure que la complexité du système augmente.
La Loi de Pareto : Le graphique montre une densité qui décroît selon une loi de puissance. Elle est l'illustration visuelle de la règle des « 80/20 », où une petite portion de la population (le début de la courbe) concentre l'essentiel de la valeur étudiée.
Ces figures permettent donc de diagnostiquer la nature d'un phénomène spatial avant même de procéder à des tests statistiques complexes. Chaque forme (symétrie, asymétrie, étalement) raconte une organisation différente du territoire.

La démarche adoptée pour cette séance a consisté à transformer des concepts mathématiques abstraits en outils d'analyse concrets.
La première étape de notre travail a été la création de fonctions de visualisation distinctes pour les variables discrètes et continues. Pour les lois discrètes, comme la loi Binomiale ou celle de Zipf-Mandelbrot, nous avons privilégié la Fonction de Masse de Probabilité (PMF). Cette méthode permet d'attribuer une probabilité exacte à chaque valeur entière, illustrant parfaitement les phénomènes de comptage.
À l'inverse, pour les variables continues telles que la loi Normale ou la loi de Pareto, nous avons exploité la Fonction de Densité de Probabilité (PDF). Cette approche est indispensable pour modéliser des mesures réelles (températures, revenus, distances) où la variable peut prendre une infinité de valeurs dans un intervalle donné. Nous avons notamment noté que la loi de Poisson, bien

que classée par convention parmi les variables discrètes, a été examinée dans un cadre continu comme le suggérait l'énoncé.

Le cœur de notre analyse a résidé dans le calcul automatisé de la moyenne et de l'écart-type, deux valeurs statistiques qui résument à elles seules la structure d'une distribution. La moyenne nous a servi d'indicateur de position centrale. Par exemple, la moyenne de 3,5 obtenue pour un dé théorique (loi uniforme discrète) confirme que les valeurs sont parfaitement équilibrées. Pour la loi de Poisson, nous avons mathématiquement validé que la moyenne (3,0) est la fondation même de la distribution, dictant la fréquence d'occurrence des événements rares. Puis l'écart-type, quant à lui, a révélé le degré de dispersion ou d'incertitude. Un résultat frappant est celui de la loi de Dirac, où l'écart-type est de 0,0, traduisant une absence totale d'aléa, car toute la probabilité est concentrée sur un point unique, ce qui est utile pour modéliser une localisation géographique fixe sans erreur de mesure. À l'opposé, la loi Uniforme continue présente un écart-type élevé (2,8868), montrant un étalement maximal des probabilités sur l'ensemble de l'intervalle [0, 10].

En définitive, cette séance nous a permis de comprendre que derrière chaque graphique se cache une réalité numérique précise. En géographie, savoir que la loi de Pareto présente une moyenne faible (1,66) mais un écart-type non négligeable permet de quantifier les inégalités spatiales, où une majorité de petites entités coexiste avec quelques pôles dominants. Cette maîtrise technique est le préalable indispensable à toute démarche d'inférence statistique que nous aborderons par la suite.

Séance 5

La statistique inférentielle permet, à partir d'informations partielles (l'échantillon), de tirer des conclusions robustes sur l'ensemble (la population), constituant la pierre angulaire de la recherche en sciences humaines et sociales.

L'échantillonnage est l'art de sélectionner un sous-ensemble d'une population mère pour en faire l'étude. La raison est simple, l'utilisation de la population entière est souvent irréalisable en raison des coûts, du temps et de la logistique associés au recueil exhaustif des données (imaginons une enquête sur la France entière, cest irréalisable).

Le choix de la méthode d'échantillonnage (aléatoire simple, stratifié, par grappes…) est crucial. Pour l'inférence, on privilégie les méthodes probabilistes qui garantissent que chaque individu a une chance connue d'être sélectionné, permettant ainsi d'estimer la marge d'erreur.

Un estimateur est une formule mathématique (une statistique) calculée à partir de l'échantillon pour prédire un paramètre de la population (ex : la moyenne de l'échantillon est l'estimateur de la moyenne de la population). Alors qu'une estimation est la valeur numérique obtenue par l'estimateur pour un échantillon donné.

Ainsi un biais se produit lorsque l'estimateur surévalue ou sous-évalue systématiquement le vrai paramètre de la population. L'enjeu est de choisir un estimateur qui soit à la fois non biaisé et efficace (faible variance).

Les statistiques qui travaillent sur la population totale sont appelées des paramètres. L'essor des Données Massives (*Big Data*) nous rapproche d'une situation où les données sont quasiment des populations, réduisant la nécessité de l'inférence par échantillonnage classique, car nos statistiques (calculées sur l'intégralité des données disponibles) tendent vers les paramètres réels.

La distinction est essentielle, l'Intervalle de Fluctuation (IF) est utilisé dans la théorie de l'échantillonnage. Il part de la population (paramètre p connu) pour prédire l'intervalle dans lequel la fréquence f d'un échantillon représentatif doit "fluctuer". A l'inverse l'Intervalle de Confiance (IC) est utilisé dans la théorie de l'estimation. Il part de l'échantillon (fréquence f observée) pour encadrer la valeur inconnue du paramètre p de la population.

Les manipulations sur les échantillons d'opinion (Pour, Contre, Sans opinion) ont illustré ces concepts.

Nous avons calculé les fréquences moyennes observées sur 100 échantillons et les avons comparées à l'IF théorique (seuil de 95%, ZC=1,96) basé sur la population mère (N{pop}=2185).

L'Intervalle de Fluctuation (IF) est centré sur la fréquence réelle (p) de la population mère. Si la fréquence observée moyenne d'un grand nombre d'échantillons tombe dans cet intervalle, cela confirme que les échantillons utilisés sont, en moyenne, conformes à la théorie et représentatifs de la population. Donc pour chaque opinion, la fréquence observée moyennée est tombée dans l'IF. Ceci valide l'ensemble des 100 échantillons utilisés pour le calcul comme étant statistiquement conformes à la population mère au seuil de 95%.

L’intervalle de confiance (IC) a été calculé sur le premier échantillon isolé (utilisant sa propre taille n et sa fréquence f). On peut interpréter que l'IC encadre la fréquence réelle de la population mère avec un certain niveau de confiance. Par exemple, pour l'opinion "Pour", si l'IC était compris entre [0,3647 ; 0,4253], nous sommes confiants à 95% que la vraie fréquence p(Pour) de la population est comprise dans cet intervalle. En Comparaison avec le résultat précédent, l'IC est centré sur la valeur observée de l'échantillon (f), tandis que l'IF est centré sur la valeur réelle de la population p. Le fait que la fréquence réelle de la population mère tombe dans l'IC observé sur l'échantillon, valide l'échantillon isolé comme une bonne base d'estimation de la population.

La théorie de la décision intervient lorsque les intervalles de confiance ou de fluctuation ne suffisent plus à valider une hypothèse scientifique. L'enjeu ici était de déterminer laquelle de deux séries de données aléatoires, contenues dans les fichiers de test, suivait effectivement une loi normale. Pour trancher cette question de manière objective, nous avons mobilisé le test de Shapiro-Wilks, une méthode statistique spécifiquement conçue pour éprouver la normalité d'une distribution.

L'application de ce test repose sur la confrontation de deux hypothèses : l'hypothèse nulle (H0), postulant que les données sont issues d'une population normale, et l'hypothèse alternative (H1), suggérant un écart significatif par rapport à ce modèle. En observant les résultats obtenus sur nos fichiers, une distinction nette apparaît, bien que la lecture des statistiques demande une certaine finesse d'interprétation.

Pour le premier fichier, correspondant au Test-1, la statistique W s'élève à environ 0,96. Mathématiquement, plus cette valeur est proche de l'unité, plus la distribution ressemble à une courbe en cloche parfaite. Cependant, la p-value associée est extrêmement faible (6,29 * 10^{-22}), ce qui nous pousse techniquement à rejeter l'hypothèse de normalité au seuil classique de 5%. Ce phénomène est fréquent sur des échantillons de grande taille car la moindre irrégularité, même insignifiante en pratique, suffit à rendre le test statistiquement « significatif ». Malgré ce rejet formel, le score W très élevé désigne sans ambiguïté, comme étant le candidat modélisant la loi normale dans cet exercice

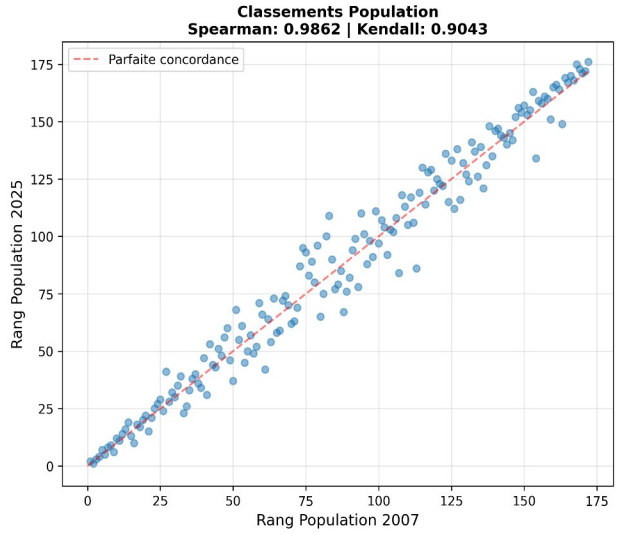
À l'opposé, le second fichier Test-2 présente des caractéristiques radicalement différentes. Sa statistique W s'effondre à 0,26, tandis que sa p-value est encore plus dérisoire (7,05 * 10^{-67}). Ici, l'écart par rapport à la loi normale n'est pas seulement statistique, il est structurel. La distribution est massivement asymétrique, rendant toute hypothèse de normalité totalement obselète.

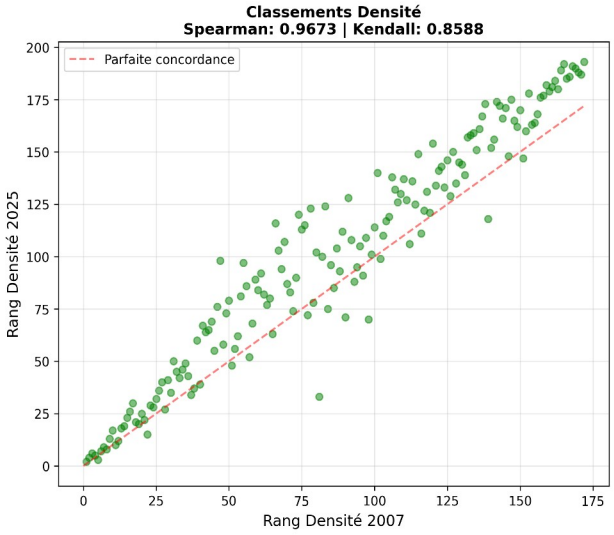
Ainsi l'examen visuel et statistique des données du Test-2 nous oriente vers une autre famille de distributions étudiée précédemment. Avec une forte concentration de valeurs faibles et une traîne

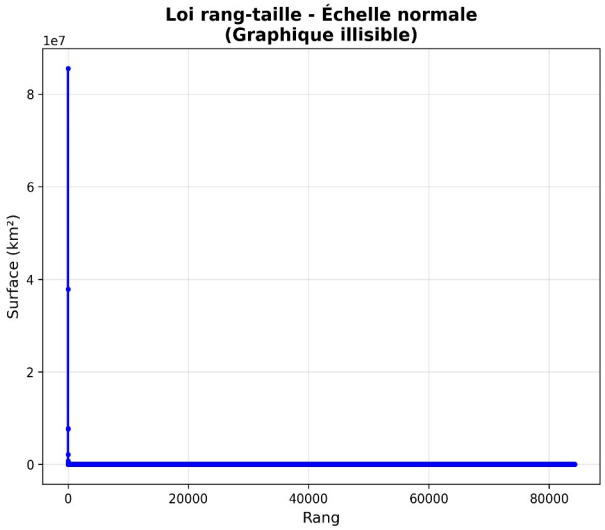
s'étirant vers des valeurs plus rares, cette série semble suivre une Loi de Poisson. Cela correspondrait typiquement à un processus de comptage d'événements isolés dans l'espace, bien loin de l'équilibre symétrique de la loi normale.

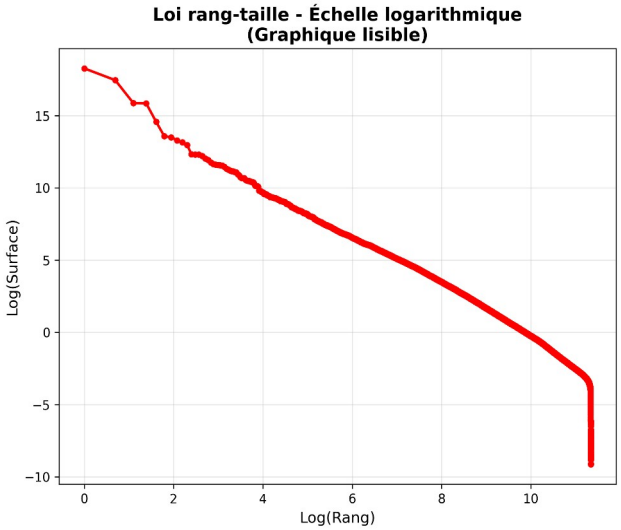
En conclusion, si la statistique inférentielle nous fournit des outils de décision puissants, elle nous rappelle également que la réalité des données de terrain (ou de simulations massives) s'écarte souvent des modèles théoriques parfaits, nous obligeant à interpréter les résultats au-delà du simple seuil de probabilité.

Séance 6


Classements Population
Spearman: 0.9862 | Kendall: 0.9043
Parfaite concordance
Rang Population 2025
Rang Population 2007
0
25
50
75
100
125
150
175


Classements Densité
Spearman: 0.9673 | Kendall: 0.8588
Parfaite concordance
Rang Densité 2025
Rang Densité 2007
0
25
50
75
100
125
150
175
200


Loi rang-taille - Échelle normale
(Graphique illisible)
1e7
Surface (km²)
Rang
0
2
4
6
8
0
20000
40000
60000
80000


Loi rang-taille - Échelle logarithmique
(Graphique lisible)
Log(Surface)
Log(Rang)
−10
−5
0
5
10
15
0
2
4
6
8
10