Rapport d'activité

Analyse de données - Parcours débutants

Par Calista KLEYN,

Étudiante en première année du Master GAED - EnviTERR, Faculté des Lettres, Sorbonne Université, Paris IV

TABLE DES MATIÈRES

SÉANCE 01

I. Objectifs

Notre objectif dans cette première séance était d'installer tous les outils utiles à l'utilisation de Python afin de pouvoir coder au mieux par la suite. Nous avons ainsi pu mettre notre environnement Python et Docker. Ainsi, tel que le dit le polycopié. Notre environnement devrait permettre plusieurs choses.
Installer python de façon automatique mais aussi charger les bibliothèques nécessaire à l'analyse de données (via requirements.txt) présent dans chaque séance de mon dossier en dur sur mon ordinateur.

II. Installation des outils cruciaux

Le plus important dans l'utilisation de python et dans sa compréhension reste l'installation de celui-ci. Pour l'installer, il faut utiliser plusieurs autres outils, *GitHub* et *Docker*. De plus, il fallait joindre mon GitHub sur ordinateur à mon espace de travail local. J'ai réussi à utiliser une méthode plus simple que celle de Monsieur Forriez afin de mettre les dossiers sur GitHub sans passer par les étapes de codage mais simplement en le transférant par le GitHub en dur sur l'ordinateur.
En revanche, pour ce qui est de l'installation de Docker et GitHub, cela a été plus compliqué que ce que je pensais. J'ai trouvé l'installation des deux et de python assez longue et laborieuse avec l'obligation de regarder des tutos youtube en parallèle. Sachant que j'ai fait l'erreur de mettre un docker falsifié et que cela ma apporté des virus par la suite (minime mais quand même présent désormais sur mon ordinateur).
Après tout cela, et après avoir compris à l'aide de mes camarades de classe toujours présent au séance de cours et hors des séances de cours, j'ai pu comprendre comment fonctionnent différentes applications et leurs importances dans l'utilisation de Python. À la suite des séances, je n'ai eu aucun problème technique sur de potentiel problème d'installation.

Néanmoins, ne comprenant pas NotePad ++, j'ai décidé sur conseil de mes camarades d'utiliser plutôt VSCode. Cela à été beaucoup plus simple et pédagogique à comprendre pour la manipulation et le bon fonctionnement des séances. En revanche, l'installation est un peu longue et doit être aidée d'un tuto internet.

SÉANCE 02

I. Objectifs

➔ Manipuler un fichier C.S.V.
➔ Faire des sorties graphiques
➔ Utiliser les bibliothèques Pandas (données) et Matplotlib (graphiques) N.B. pd et plt sont des alias qui remplacent respectivement pandas et matplotlib.pyplot.
➔ Calculer des effectifs
➔ Calculer des fréquences
➔ Faire des graphiques (diagrammes en bâton et circulaires, et histogrammes)

II. Question de cours

La géographie entretient depuis longtemps un rapport ambigu aux statistiques. Le cours souligne qu'elle est une discipline qui "se cherche toujours" et qui, malgré une méfiance persistante envers les outils mathématiques, produit des données massives que seules les méthodes statistiques permettent réellement d'analyser. La statistique (au singulier) correspond à une science, un ensemble de méthodes permettant de prendre des décisions en contexte d'incertitude, tandis que les "statistiques" désignent l'ensemble des données mobilisées en géographie. L'information géographique est donc indissociable de la statistique, qu'il s'agisse de données socio-économiques, environnementales ou morphologiques.

La question du hasard occupe une place centrale dans la réflexion géographique. Le cours rappelle les origines philosophiques du déterminisme, pour lequel le hasard n'est qu'une apparence masquant des causes encore inconnues. La géographie se situe entre deux pôles : d'un côté, une conception déterministe où les phénomènes ont des causes mesurables ; de l'autre, une vision "contingente", plus ouverte, dans laquelle certains événements peuvent se produire ou non sans nécessité. Pour la statistique, le hasard n'est pas une absence d'explication, mais un cadre permettant de dégager une "certitude globale" malgré l'impossibilité

de prévoir chaque réalisation individuelle. C'est précisément cette logique (penser des tendances globales face à des situations locales variées) qui fait de la géographie une science des échelles.

L'information géographique, pour être analysée, doit d'abord être produite ou collectée. Cela implique la mise en place de nomenclatures, qui définissent ce qui doit être observé, et l'usage de métadonnées, qui documentent les sources, définitions et conditions de collecte. Ces étapes sont essentielles car toute analyse statistique doit tenir compte de la qualité et de la fiabilité des données utilisées. Une fois ces bases établies, l'analyse statistique peut s'appuyer sur trois grands piliers : les probabilités (cadre théorique), les statistiques (méthodes descriptives et mathématiques) et l'analyse de données.

La statistique descriptive constitue la première étape du traitement. Elle permet de résumer les distributions observées, d'identifier les valeurs extrêmes, de visualiser les données et de préparer des comparaisons. Elle s'appuie sur le type de variable étudiée : qualitative ou quantitative, nominale, ordinale, discrète ou continue. Ce classement est fondamental, car il détermine les traitements possibles, les représentations graphiques et les lois de probabilité mobilisables. Par exemple, les variables qualitatives ne peuvent être décrites que par des fréquences, tandis que les variables quantitatives permettent le calcul de moyennes, d'écarts-types ou encore la recherche de distributions associées.

Les besoins de la géographie en analyse de données sont ainsi multiples : décrire, classer, expliquer et modéliser. Les méthodes de visualisation (histogrammes, représentations sectorielles, cartes factorielles issues de l'ACP, de l'AFC ou de l'ACM) permettent de rendre lisibles des jeux de données complexes. Le choix d'une méthode dépend de la nature des variables (qualitatives, quantitatives, ou mixtes) et de l'objectif poursuivi : réduire la dimension d'un tableau, visualiser des proximités entre individus ou catégories, classer des unités spatiales, etc.

Au-delà de la description, les méthodes explicatives visent à comprendre les relations entre une variable à expliquer et des variables explicatives. Il

peut s’agir de modèles linéaires, de régressions, d'analyses discriminantes ou de régressions logistiques selon la nature de la variable étudiée. Ces méthodes cherchent à ajuster un modèle aux données pour dégager des relations significatives et interprétables. La statistique mathématique, quant à elle, vise à prévoir, par exemple à travers l’analyse des séries chronologiques.

Le cours insiste également sur la nécessité de bien définir les notions de population statistique, d’individu statistique, de caractère et de modalité. La population est l’ensemble des unités d’observation, un individu est une unité spatiale ou une entité élémentaire, un caractère est une propriété mesurée, une modalité est la valeur prise par ce caractère. Ces notions s’imbriquent hiérarchiquement et permettent de structurer toute analyse statistique.

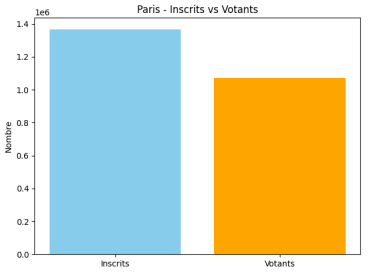
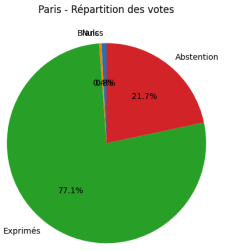
Pour les variables quantitatives, la discrétisation joue un rôle important : elle consiste à regrouper des valeurs continues en classes. Deux paramètres en découlent : l’amplitude de classe (b−a) et la densité (effectif divisé par amplitude). Le choix du nombre de classes peut être guidé par les formules de Sturges ou de Yule, qui donnent une estimation optimale en fonction de la taille de l’échantillon. Cette étape est nécessaire pour construire histogrammes, polygones de fréquence ou courbes cumulatives.

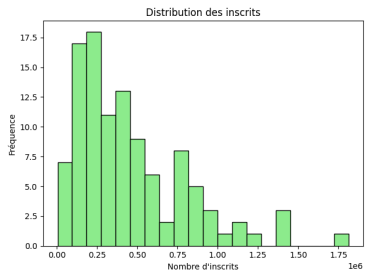
Enfin, l’analyse statistique repose sur les notions d’effectif, de fréquence et de fréquence cumulée. L’effectif correspond au nombre d’individus dans une modalité, la fréquence à son poids relatif dans la population totale, et la fréquence cumulée à l’accumulation progressive des fréquences. Ces notions permettent de construire une distribution statistique empirique, étape indispensable pour relier les observations réelles aux lois théoriques du hasard et pour interpréter les phénomènes géographiques.

Ainsi, la statistique apparaît comme un outil indispensable à la géographie : elle organise les données, met en évidence les tendances, éclaire la structure interne des phénomènes et permet de confronter les théories à la réalité observée. Elle ne remplace ni le raisonnement explicatif ni les

connaissances disciplinaires, mais elle les complète en donnant accès aux régularités cachées derrière la complexité des territoires.

III. Résultats obtenus et piste de réflexion

Question n°	Résultat(s)	Illustration(s)
5	Lecture du fichier CSV, ouverture du terminal avec le contenu.	
6	Calcule du nombre de lignes et de colonnes du fichier CSV	
7	Affichage par colonne du type de données	
8	Affichage du nom des colonnes	
9	Sélection du nombre des inscrits	
10	Calcule les effectifs de chaque colonne, placement dans une liste	
11	Diagrammes en barres avec nombres inscrits et votants pour chaque département	
12	Diagrammes circulaires avec les votes blancs, nuls, exprimés et l'abstention pour chaque département.	

13	Histogramme de la distribution des inscrits	

Je n'ai pas eu de problème réellement lors de la réalisation de cette séance. Néanmoins, cela à été compliqué de rentrer vraiment dans ce nouveau domaine, dans la compréhension de celui et dans la production des dossiers et son utilisation.

J'ai eu quelques soucis avec GitHub afin de transférer cette première séance de code sur Github en ligne. Mais grâce à mes camarades, nous avons trouvé une méthode plus simple que Monsieur Forriez pour tout connecté ensemble.

SÉANCE 03

I. Objectifs

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

II. Questions de cours

L'analyse statistique repose sur l'étude de caractères décrivant une population. Le caractère qualitatif est le plus général, car il ne nécessite aucune structure numérique : il permet simplement de distinguer des catégories ou modalités. À l'inverse, le caractère quantitatif impose une mesure et donc des traitements davantage contraints. Parmi les caractères quantitatifs, on distingue les variables discrètes, qui ne prennent qu'un ensemble dénombrable de valeurs, et les variables continues qui peuvent prendre toute valeur d'un intervalle. Cette distinction est essentielle, car elle conditionne les méthodes de calcul : les variables discrètes se traitent par sommes tandis que les variables continues nécessitent des intégrales.

Lorsque l'on souhaite résumer une distribution, les paramètres de position constituent un premier ensemble d'indicateurs. La moyenne, sous ses différentes formes (arithmétique, harmonique, géométrique…), existe en plusieurs versions car chaque type répond à des situations précises et à une manière particulière d'agréger les valeurs. Toutefois, la moyenne est sensible aux valeurs extrêmes, ce qui justifie l'usage complémentaire de la médiane. Cette dernière partage la population en deux ensembles de même effectif et n'est pas influencée par les valeurs aberrantes, ce qui la rend particulièrement utile pour décrire des distributions dissymétriques. Le mode, quant à lui, correspond à la valeur la plus fréquente ; il n'existe pas toujours, peut être multiple et sert surtout à identifier des groupes ou des pics de fréquence.

Certains paramètres cherchent non pas à exprimer une valeur centrale, mais à mesurer la concentration des valeurs au sein de la distribution. La

médiale, par exemple, ne partage plus la population en effectifs égaux, mais en deux parts de masse totale égales : chacune représente 50 % du total cumulé des valeurs. Le décalage entre la médiale et la médiane indique le degré de concentration, et plus cet écart est grand relativement à l'étendue, plus la distribution est inégalitaire. Cette logique de concentration se retrouve dans l'indice de Gini ou dans la courbe de Lorenz, outils permettant d'appréhender visuellement et numériquement la distribution d'une masse, par exemple les revenus.

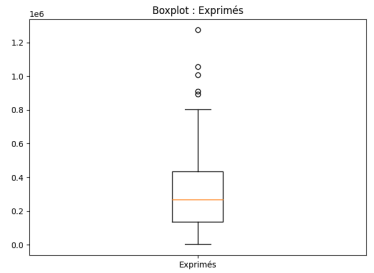
Les paramètres de dispersion complètent l'analyse en mesurant l'écart des données autour d'une valeur de référence. La variance est la mesure de dispersion la plus classique : elle correspond à la moyenne des carrés des écarts à la moyenne. L'usage du carré est nécessaire, car la moyenne des écarts simples est toujours nulle et ne donne aucune information. Toutefois, la variance s'exprime dans des unités au carré, ce qui la rend difficile à interpréter directement ; l'écart-type est alors employé pour ramener cette mesure à l'unité initiale. D'autres indicateurs existent : l'étendue, particulièrement simple, mesure la différence entre la plus grande et la plus petite valeur, mais ne repose que sur des données extrêmes et devient peu informative lorsque la série est longue. Les quantiles, en revanche, permettent de découper la série en parties égales et d'étudier la répartition interne des valeurs. L'écart interquartile, qui contient 50 % des valeurs, constitue un outil robuste pour évaluer la dispersion. La boîte de dispersion, ou box-plot, permet alors une représentation graphique synthétique, faisant apparaître quartiles, médiane et valeurs extrêmes.

Enfin, les paramètres de forme décrivent la structure générale de la distribution. Ils s'appuient sur les moments, qui généralisent la notion de moyenne. Les moments centrés, notamment d'ordre 2, 3 et 4, permettent de mesurer respectivement la variance, l'asymétrie et l'aplatissement de la distribution. La différence entre moments centrés et moments absolus repose sur le fait que les seconds utilisent la valeur absolue, ce qui empêche les compensations entre valeurs positives et négatives. Ces mesures sont utiles pour comprendre si une distribution est symétrique ou non. L'asymétrie est caractérisée par le coefficient β_1 : s'il est positif, la

distribution est étirée vers les valeurs élevées ; s'il est négatif, elle est étirée vers les valeurs faibles. L'aplatissement, mesuré par β_2, indique si la distribution est plus ou moins « pointue » que la loi normale. Vérifier la symétrie d'une distribution est crucial, car dans une distribution parfaitement symétrique, la moyenne, la médiane et le mode coïncident, et les paramètres de forme proches de zéro signalent cette symétrie.

Ainsi, les différents paramètres statistiques (de position, de dispersion, de concentration et de forme) fournissent un ensemble cohérent d'outils permettant de décrire et de comprendre la structure d'une distribution. Leur utilisation conjointe est indispensable pour analyser pleinement les caractéristiques d'un ensemble de données et pour en saisir à la fois les tendances centrales, les inégalités, la variabilité et la forme générale.

III. Résultats obtenus et piste de réflexion

Question n°	Résultat(s)	Illustration(s)
5	Sélection et calcul des caractères quantitatifs : moyennes, médianes, modes écart type, écart absolu, étendu	
6	Affichage de la liste des paramètres sur le terminal	
7	Calcule de la distance interquartile et interdécile de chaque colonne	
8	Boîte à moustache de chaque colonne quantitative	
10	Catégorisation et dénombrement du nombre d'îles ayant une surface comprise dans chaque intervalle + création d'un organigramme	Ouverture du fichier

Aucun problème lors de la production de cette séance.

SÉANCE 04

I. Objectifs

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

II. Questions de cours

Le choix entre une distribution statistique à variables discrètes ou continues repose avant tout sur la nature du phénomène étudié. Lorsqu'une variable ne peut prendre que des valeurs entières et dénombrables, comme un nombre d'habitants, d'événements ou d'objets, elle est dite discrète. À l'inverse, lorsqu'une variable est mesurable sur un continuum et peut prendre n'importe quelle valeur dans un intervalle donné (par exemple l'altitude, la température, la distance, la durée ou le revenu) elle est considérée comme continue. Dans ce cas, aucune probabilité n'est associée à une valeur ponctuelle, mais uniquement à un intervalle de valeurs, selon le principe fondamental des variables continues pour lesquelles Pr(X=x)=0\text{Pr}(X = x) = 0Pr(X=x)=0.

Un second critère essentiel réside dans la forme empirique de la distribution observée. Une distribution présentant des valeurs isolées ou une structure en « marches » suggère un processus discret, tandis qu'une distribution lisse décrite par une densité continue oriente vers un modèle continu. Ce critère est particulièrement important en pratique, car les données observées peuvent parfois masquer la nature réelle de la variable, notamment lorsqu'une variable continue est mesurée de manière grossière ou arrondie, donnant l'illusion d'un comportement discret.

Le choix de la loi statistique dépend également du processus générateur du phénomène étudié. Les phénomènes résultant de l'addition d'un grand nombre de facteurs indépendants suivent fréquemment une loi normale, tandis que les processus de croissance multiplicative conduisent plutôt à des distributions log-normales. Les comptages d'événements rares et indépendants dans l'espace ou dans le temps sont quant à eux bien

modélisés par une loi de Poisson, à laquelle est souvent associée la loi exponentielle pour décrire les durées entre événements. Ainsi, l'interprétation du mécanisme sous-jacent constitue un élément déterminant dans le choix du modèle probabiliste.

Les paramètres descriptifs disponibles constituent un autre critère de décision important. L'analyse de l'espérance, de la variance, de la médiane, de l'asymétrie ou encore de l'aplatissement permet d'orienter le choix entre plusieurs lois possibles. Par exemple, une moyenne proche de la variance peut suggérer une loi de Poisson, tandis qu'un phénomène symétrique avec une variance bien définie est compatible avec une loi normale. Lorsque plusieurs lois sont envisageables, ces indicateurs statistiques aident à sélectionner le modèle le plus pertinent.

Un dernier critère méthodologique concerne le nombre de paramètres de la loi choisie. Les lois simples, comportant peu de paramètres, offrent une interprétation plus directe mais une flexibilité limitée. À l'inverse, les lois paramétriques plus complexes peuvent mieux épouser la forme d'une distribution empirique, au prix d'une interprétation parfois plus délicate. Le choix final doit donc trouver un équilibre entre simplicité, adéquation empirique et cohérence théorique.

En géographie, certaines lois statistiques occupent une place centrale car elles permettent d'analyser les structures spatiales, les hiérarchies et les dynamiques territoriales. La loi de Zipf, ou loi rang-taille, est largement utilisée pour étudier la hiérarchie urbaine et la distribution des tailles de villes, en mettant en évidence les fortes inégalités entre grandes et petites agglomérations. Sa généralisation par le modèle de Zipf-Mandelbrot permet de décrire des systèmes urbains plus complexes ou s'écartant du schéma idéal.

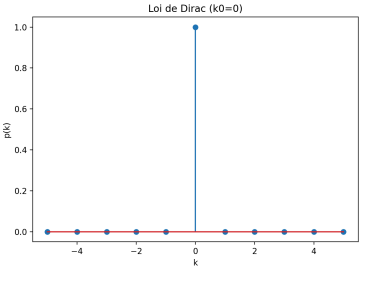
La loi normale intervient fréquemment dans l'analyse de grandeurs continues telles que les altitudes, les températures ou les résidus de modèles statistiques. Elle joue également un rôle central dans la modélisation des erreurs de mesure et dans la construction des intervalles de confiance. La loi log-normale, aussi appelée loi de Gibrat, est

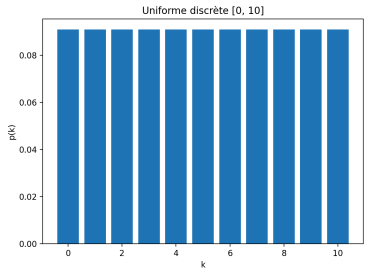
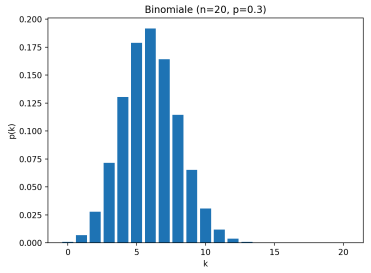
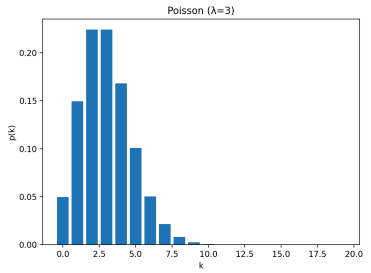
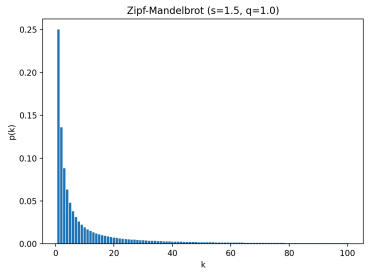
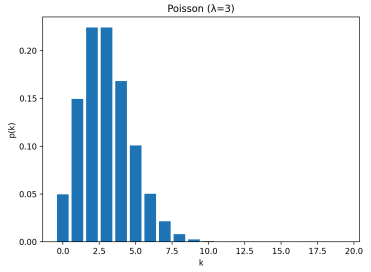
particulièrement adaptée à l'étude des phénomènes de croissance proportionnelle ou cumulative, comme les populations, les revenus ou les surfaces de bassins versants, et rend compte de distributions fortement asymétriques.

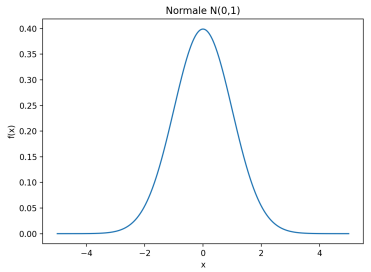
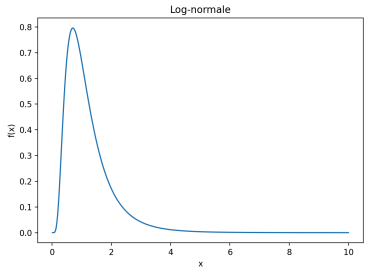
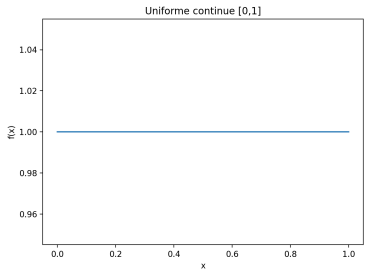
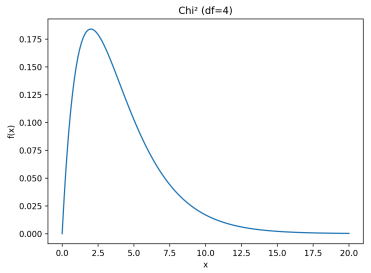
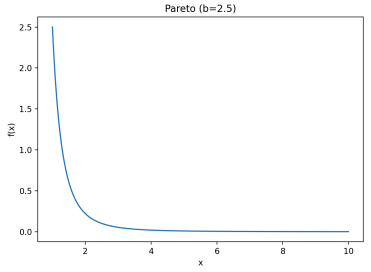
La loi de Poisson et la loi exponentielle sont mobilisées pour analyser la répartition spatiale ou temporelle d'événements rares et indépendants, tels que des accidents, des catastrophes naturelles ou certaines occurrences environnementales. Enfin, la loi uniforme continue occupe une place spécifique en géographie théorique et en simulation spatiale. Elle traduit l'hypothèse d'équirépartition et constitue un modèle de référence pour les espaces isotropes, la génération aléatoire de points ou certaines approches bayésiennes fondées sur une absence d'information préalable.

Ainsi, le choix d'une distribution statistique ne relève pas d'une simple préférence technique, mais d'une démarche méthodologique intégrant la nature de la variable, la forme empirique des données, les mécanismes générateurs du phénomène et les objectifs de l'analyse. En géographie, ces lois permettent de relier les formes observées dans l'espace aux processus statistiques et spatiaux qui les produisent, offrant ainsi une lecture rigoureuse des structures territoriales.

III. Résultats obtenus et piste de réflexion

Question n°1	Résultat(s)	Illustration(s)
Distributions statistiques de variable discrètes :	Loi de Dirac	

	Loi uniforme discrète	
	Loi binomiale	
	Loi de Poisson	
	Loi de Zipf-Mandelbrot	
Distributions statistiques de variables continues :	Loi de Poisson (Continue)	

	Loi normale	
	Loi log-normale	
	Loi uniforme (continue)	
	Loi du Chi2	
	Loi de Pareto	
Question n°2	Calculer la moyenne, l'écart-type	

Aucun problème sur cette séance. En revanche, j'ai trouvé que l'énoncé n'était pas assez complet et manquait énormément d'information. J'ai

énormément réfléchir à comment produire ces lois grâce à python, mais une fois trouvé c'était facile à reproduire sur les autres lois

SÉANCE 05

I. Objectifs

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

II. Questions de cours

L'échantillonnage consiste à étudier un sous-ensemble d'une population mère afin d'en tirer des conclusions sur l'ensemble. On y a recours lorsque l'observation exhaustive de la population est impossible, trop coûteuse ou inutile, notamment lorsque celle-ci est très vaste ou difficilement accessible. L'enjeu central de l'échantillonnage est donc de constituer un échantillon suffisamment représentatif afin de limiter l'erreur liée à la fluctuation d'échantillonnage et de garantir la validité des inférences produites.

Les statistiques inférentielles reposent sur ce principe fondamental : il est généralement plus efficace d'analyser un échantillon limité mais bien construit plutôt que l'ensemble de la population. Dans de nombreuses situations, comme l'étude des intentions de vote à l'échelle d'un pays, l'analyse d'un échantillon représentatif permet d'obtenir une approximation fiable des caractéristiques de la population mère tout en réduisant fortement les coûts de collecte et de traitement des données.

Il existe deux grandes familles de méthodes d'échantillonnage : les méthodes aléatoires et les méthodes non aléatoires. Les méthodes aléatoires reposent sur un tirage au sort, avec ou sans remise, et nécessitent l'existence d'une base de sondage recensant l'ensemble des individus de la population. Elles garantissent théoriquement l'équiprobabilité des tirages et offrent de bonnes propriétés statistiques de représentativité. Les méthodes non aléatoires, quant à elles, visent à reconstituer un « modèle réduit » de la population sans recourir au hasard pur. Parmi elles figurent

l'échantillonnage systématique, où les individus sont sélectionnés à intervalles réguliers après un premier tirage aléatoire, et l'échantillonnage par quotas, largement utilisé en sciences sociales. Ce dernier reproduit la structure de la population selon certains critères jugés pertinents, mais nécessite des moyens importants et une connaissance fine de la population étudiée. Le choix de la méthode dépend des objectifs de l'étude, de l'accès aux données, du niveau de précision recherché et des contraintes pratiques.

À partir des échantillons constitués, la statistique inférentielle vise à estimer des paramètres inconnus de la population, tels qu'une moyenne ou une proportion. Cette estimation repose sur l'utilisation d'estimateurs, définis comme des fonctions mathématiques des observations issues de l'échantillon. L'estimation correspond à la valeur numérique obtenue à partir de l'estimateur. La qualité d'un estimateur est évaluée à l'aide de plusieurs propriétés théoriques : l'absence de biais, la précision (variance faible), la convergence vers la valeur réelle lorsque la taille de l'échantillon augmente, la robustesse face aux valeurs aberrantes et l'efficacité. Par exemple, la moyenne empirique est un estimateur sans biais et convergent de la moyenne de la population, tandis que la variance empirique non corrigée est biaisée, ce qui justifie l'application d'un facteur correctif.

La statistique inférentielle mobilise également deux outils fondamentaux : l'intervalle de fluctuation et l'intervalle de confiance. L'intervalle de fluctuation s'utilise lorsque la valeur du paramètre est connue. Il décrit les variations attendues d'une fréquence observée dans un échantillon de taille donnée et permet d'évaluer la compatibilité d'une observation avec un paramètre théorique. À l'inverse, l'intervalle de confiance est utilisé lorsque le paramètre est inconnu. Construit à partir d'un estimateur et de son erreur standard, il fournit une plage de valeurs dans laquelle le paramètre réel a une probabilité élevée de se situer, généralement fixée à 95 %. Plus la taille de l'échantillon augmente, plus l'intervalle de confiance est étroit et précis.

Lorsque l'analyse porte sur l'ensemble de la population, on parle de statistique exhaustive. Dans ce cas, les paramètres sont directement

observables et l'inférence perd sa fonction traditionnelle, puisque la variabilité liée à l'échantillonnage disparaît. Avec le développement des données massives, certaines analyses se rapprochent de cette logique, même si les données peuvent rester biaisées, incomplètes ou de qualité inégale, ce qui justifie le maintien d'outils inférentiels.

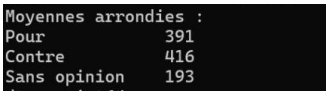
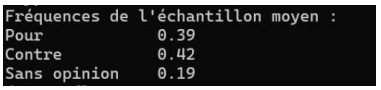
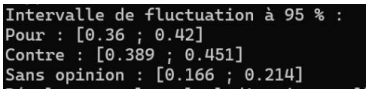
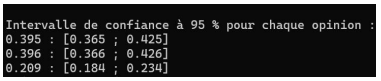
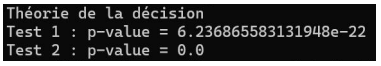
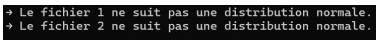
Les méthodes d'estimation se divisent principalement entre l'estimation ponctuelle, qui fournit une valeur unique du paramètre, et l'estimation par intervalle. Elles peuvent être mises en œuvre à l'aide de différentes approches, telles que la méthode des moments, le maximum de vraisemblance, les moindres carrés, le bootstrap ou encore les approches bayésiennes, qui intègrent une information a priori. Le choix de la méthode dépend de la taille de l'échantillon, de la loi supposée des données, du paramètre à estimer et du niveau de précision recherché.

Enfin, les tests statistiques constituent un outil essentiel de l'inférence. Ils permettent de juger si les données observées sont compatibles avec une hypothèse donnée en contrôlant un risque d'erreur prédéfini. Leur construction repose sur la formulation d'une hypothèse nulle et d'une hypothèse alternative, le choix d'une statistique de test, la fixation d'un seuil de signification et une règle de décision fondée sur la valeur critique ou la p-value. On distingue les tests paramétriques, qui supposent certaines conditions sur la distribution des données, et les tests non paramétriques, utilisés lorsque ces conditions ne sont pas vérifiées.

Bien que la statistique inférentielle fasse l'objet de critiques (dépendance à des hypothèses parfois idéalisées, risques de biais liés à un mauvais échantillonnage ou interprétations abusives des résultats) elle demeure un ensemble d'outils indispensables pour produire des connaissances généralisables à partir de données partielles. Son utilisation requiert toutefois une approche rigoureuse, critique et éclairée.

III. Résultats obtenus et piste de réflexion

Question n°	Résultat(s)	Illustration(s)

1	Calcul de la moyenne pour chaque colonnes	
	Calcul des fréquences	
	Calcul de l'intervalle de fluctuation de chacune des fréquences	
2	Prendre premier échantillon de la liste	
	Calcul de la somme de la ligne, les fréquences avec l'effectif total	
	Calcul l'intervalle de confiance	
3	Théorie de la décision	
	Test de Shapiro	

J'ai trouvé cela un peu plus compliqué car les résultats étaient plutôt à chercher dans le programme. Je n'ai trouvé d'ailleurs que des distributions anormales et aucunes étaient normales. Or, nous nous sommes concertés avec mes camarades et nous sommes nombreux (voir la totalité) à avoir eu ce résultat.

SÉANCE 06

I. Objectifs

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

II. Questions de cours

La statistique ordinale regroupe l'ensemble des méthodes fondées sur le classement des individus, des objets ou des territoires, en privilégiant leur ordre relatif plutôt que leurs valeurs numériques exactes. Elle repose sur l'ordonnancement d'une série d'observations et sur l'attribution de rangs, notés $X(1) \leq X(2) \leq \cdots \leq X(n)$X(1) \leq X(2) \leq \dots \leq X(n)$X(1) \leq X(2) \leq \cdots \leq X(n)$. Elle se distingue ainsi des statistiques nominales, qui se limitent à répartir les individus en catégories sans relation d'ordre entre elles. Les variables mobilisées sont des variables qualitatives ordinales, pour lesquelles un ordre naturel peut être défini, le plus souvent croissant.

La statistique ordinale occupe une place centrale en géographie, car de nombreux phénomènes spatiaux s'organisent spontanément sous forme de hiérarchies. Classer des villes selon leur taille, des territoires selon leur dynamisme socio-économique ou des événements naturels selon leur intensité permet de rendre visibles des structures hiérarchiques et d'analyser la position relative des espaces. C'est pourquoi elle est souvent considérée comme un outil fondamental de la géographie humaine, mais aussi de la géographie physique.

Dans les démarches de classification, l'ordre croissant (également appelé ordre naturel) est généralement privilégié. Il facilite l'interprétation des rangs, l'identification des valeurs extrêmes et l'analyse globale des distributions. Cet ordre est notamment mobilisé dans l'étude de phénomènes tels que la loi rang-taille en géographie urbaine ou l'analyse des maxima en géographie physique.

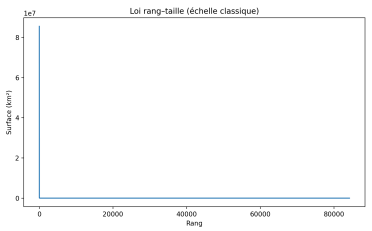
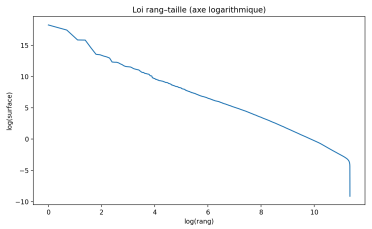
Lorsque plusieurs classements doivent être comparés, deux approches complémentaires sont mobilisées : la corrélation des rangs et la concordance des classements. La corrélation des rangs vise à mesurer la proximité globale entre deux classements en comparant les rangs attribués aux mêmes individus. Elle renseigne sur la force et le sens de la relation entre deux ordres. La concordance, quant à elle, s'intéresse à la cohérence interne des classements en examinant, paire par paire, si les individus sont classés dans le même ordre ou dans l'ordre inverse. Une concordance parfaite signifie que toutes les paires évoluent dans le même sens, tandis qu'une concordance nulle correspond à un équilibre entre paires concordantes et discordantes.

Ces différences conceptuelles se retrouvent dans les deux tests classiques de comparaison des classements : le test de Spearman et le test de Kendall. Le coefficient de Spearman repose sur le calcul des différences entre les rangs et s'inscrit dans une logique quantitative. Il mesure la proximité des rangs et varie entre −1, lorsque les classements sont inverses, et +1, en cas d'identité parfaite. Ce test est sensible à la présence d'ex æquo et, pour des effectifs suffisants, sa distribution peut être assimilée à une loi normale. Le test de Kendall, en revanche, repose uniquement sur le comptage des paires concordantes et discordantes. Conceptuellement plus simple, il compare directement l'ordre de chaque paire d'individus et se généralise plus facilement au cas de plusieurs classements simultanés. De manière générale, Spearman mesure la proximité globale des rangs, tandis que Kendall évalue la cohérence de l'ordre.

D'autres coefficients permettent d'analyser l'association entre variables ordinales. Le coefficient de Goodman-Kruskal repose sur la comparaison du nombre de paires concordantes et discordantes et fournit une mesure comprise entre −1 et +1, allant de l'inversion totale à la concordance parfaite. Il est proche, dans son esprit, du coefficient de Kendall, mais s'interprète comme un indice de concordance fondé sur des proportions. Le coefficient de Yule constitue un cas particulier de cette approche, limité aux tableaux de contingence 2×2. Il permet de mesurer l'association entre deux variables dichotomiques et s'interprète selon les mêmes bornes.

Ainsi, les statistiques ordinales offrent un ensemble d'outils rigoureux permettant d'analyser les hiérarchies spatiales, de comparer plusieurs systèmes de classement et d'évaluer la cohérence des ordres observés. En géographie comme dans l'ensemble des sciences sociales, elles constituent des instruments essentiels pour comprendre, comparer et interpréter les structures ordonnées des territoires, des populations et des dynamiques spatiales.

III. Résultats obtenus et piste de réflexion

Question n°	Résultat(s)	Illustration(s)
3	Isoler la colonne "surface (km2)" et ajouter la liste	
4	Ordonner la liste	
5	Visualiser la loi rang-taille	
	Loi rang-taille logarithmique	
7	Possibilité de faire un test sur les rangs ?	
10	Isoler les colonnes Etat, pop 2007, Densité 2007, Densité 2025.	
11	Les ordonner de manière décroissante	
12	Comparaison des listes sur la population et la densité	

13	Isoler les deux colonnes	
14	Calcul le coefficient de corrélation des rangs et la concordance des rangs.	

Aucun problème lors de cet séance.

Retour générale sur le cours

Pour être entièrement honnête, j'ai eu beaucoup de difficultés avec ce cours. C'est d'ailleurs celui avec lequel j'ai rencontré le plus de problèmes ce semestre.

Tout d'abord, j'ai eu le sentiment que nous n'étions pas tous logés à la même enseigne. Il y avait un réel manque d'équité dans les conditions de réussite, notamment entre les étudiants disposant d'un Mac et ceux ayant un ordinateur trop ancien ou inadapté. Certains ont même dû acheter un nouvel ordinateur ou en emprunter un auprès du service informatique de l'université. Cela témoigne, selon moi, d'un manque d'organisation et de considération envers les étudiants. J'espère sincèrement que ces aspects seront mieux anticipés l'année prochaine et pris en compte de manière plus équitable. En effet, nous n'avons pas tous pu réussir l'exercice de code, qui comptait pour 40 % de la note globale, alors que cela ne relevait pas nécessairement d'un manque de compétences, mais plutôt d'une prise en charge insuffisante des contraintes matérielles.

Au-delà de ces difficultés, et une fois la matière mieux comprise, j'ai trouvé que la manipulation de Python ainsi que le contenu des séances n'étaient finalement pas très complexes. Toutefois, certains problèmes informatiques, difficiles à résoudre lorsque l'on est novice, ont pu compliquer l'apprentissage. Néanmoins, l'application concrète des séances s'est révélée moins difficile que je ne l'avais initialement imaginé.

Pour conclure sur une note positive, je suis satisfaite d'avoir pu être initiée au travail sur Python. Ce type d'enseignement n'est pas proposé dans tous les cursus, et j'en suis finalement ravie.

Réflexion sur les humanités numériques

Les humanités numériques désignent un champ de recherche interdisciplinaire qui associe les sciences humaines et sociales aux outils et méthodes numériques. Elles ne se limitent pas à la numérisation des sources, mais interrogent la manière dont le numérique transforme la production, l'analyse et la diffusion des savoirs.

Cette réflexion m'a permis de mieux comprendre l'impact du numérique sur les pratiques de recherche en sciences humaines et sociales. Avant ce travail, je percevais surtout le numérique comme un outil technique. Les humanités numériques m'ont amenée à le considérer comme un véritable cadre de réflexion méthodologique, influençant les choix de recherche et les résultats produits.

L'un des apports essentiels concerne la place centrale des données. La numérisation massive des corpus ouvre de nouvelles possibilités d'analyse, notamment à grande échelle, mais pose aussi la question des biais, de la qualité des données et de la non-neutralité des outils numériques. Cette dimension critique m'a permis de comprendre son importance pour produire une recherche rigoureuse.

Les humanités numériques favorisent également une complémentarité entre approches quantitatives et qualitatives. Elles ne remplacent pas les méthodes traditionnelles, mais les enrichissent, notamment par la visualisation, la cartographie ou l'analyse de données, qui nécessitent toujours une interprétation contextualisée.

Enfin, ce travail m'a sensibilisée aux enjeux éthiques et sociaux liés au numérique, comme l'accessibilité des données et la diffusion des savoirs. En tant qu'étudiante en géographie, j'y vois un fort potentiel pour analyser les dynamiques territoriales, tout en restant attentive aux limites et aux responsabilités liées à l'usage du numérique.