NOURRIT
Philippine
M1 GAED
parcours « Géosuds »
Niveau « débutant »

RAPPORT

D’ACTIVITÉ

2025/2026

SOMMAIRE

Séance 2
Les principes généraux de la statistique

La géographie a longtemps entretenu, et entretient toujours un rapport paradoxal avec les statistiques. D'un coté elle se tient à distance des mathématiques et donc de ces dernières, en considérant que les questions soulevées par les statistiques ne sont pas dans son champs de recherche. D'ailleurs, une très grande quantité de géographes n'ont pas suivi de formation en mathématiques très poussée, ce qui les amènent souvent à se tourner vers des approches qualitatives.
Pourtant, la géographie est une discipline qui produit un nombre très important de données. Pour les analyser correctement, des outils statistiques sont nécessaires. Former les géographes aux outils de cette branche des mathématiques est donc essentiel. Cela permettrait à la fois d'élargir les possibilités d'analyse, d'ouvrir la géographie à des méthodes statistiques qui pourraient améliorer la recherche dans ce même domaine, mais aussi éviter les erreurs ou maladresses. Désormais, la géographie est donc forcée de se confronter aux statistiques.

Si l'on en croit la philosophie de Laplace, il existerait toujours des paramètres expliquant les phénomènes qui surgissent : chaque objet géographique serait donc déterminé par des facteurs naturels ou humains. Ainsi en science le hasard serait entendu comme ce que l'on ne connaît pas encore, ce que l'on a pas su expliquer, ce dont on a pas trouvé la loi. D'ailleurs dans les modélisations mathématiques on distingue deux types de hasard : le hasard bénin qui reste modéré et ne vient pas contredire la loi scientifique et le hasard sauvage qui lui peut sembler à première vue remettre en question une loi scientifique.
En géographie, ce hasard sauvage peut survenir surtout à des échelles plus petites : on ne peut pas savoir comment un acteur va agir individuellement, ou le territoire exact sur lequel un incendie risque de se produire par exemple. En revanche, il est possible de dégager des tendances globales, après une analyse des différents facteurs potentiellement explicatifs d'un phénomène. Elles seront certes inexactes à une petite échelle mais visibles à une grande échelle. C'est cette capacité à dégager des lois qui s'appliquent sur le territoire qui fait de la géographie une science.

Ainsi, il existe deux types d'information géographique. La première est celle qui s'intéresse à ce qui est dans le territoire, ce qui est contenu dans cet espace. Il peut s'agir de géographie humaine (densité, population humaine, activité économique) comme de géographie physique (température, précipitations etc). On appelle ce type d'information base attributaire dans un SIG (système d'information géographique).
La seconde concerne le territoire lui même. On relève alors des informations concernant sa « morphologie » c'est à dire sa forme (taille, structure etc).

Pour pouvoir comprendre un espace, la géographie a besoin de produire et/ou de collecter des données. Le géographe, la plupart du temps de les récolte pas lui même mais fait appel à des organismes extérieurs. Pour réaliser une étude statistique, il va donc devoir établir une nomenclature servant à définir des catégories le plus clairement possible tout en accompagnant les données de méta-données afin d'éviter des erreurs d'interprétation. Il s'agit ensuite d'analyser les données elles-mêmes, leur structure, afin d'identifier les types de variables, localiser les valeurs extrêmes et aberrantes, noter les différents types de distribution. On va chercher à utiliser le bon outils statistique, la bonne méthode pour analyser un panel de données et finir par analyser les résultats. Les statistiques, grâce aux outils et méthodes diverses qui ont permis de dégager des tendances, vont être interprétées. L'analyse statistique va donc finalement permettre de mieux comprendre le territoire. Du fait de l'abondance et du caractère multidimensionnel des données spatiales, la géographie a donc des besoins nombreux en analyse de données. L'objectif est d'aboutir à des interprétations les plus justes possible.

Néanmoins, les statistiques comme branche des mathématiques ne sont pas un bloc uni et homogène. Il existe des différences. On distingue la statistique descriptive de la statistique explicative par exemple.
En effet, la statistique descriptive permet de représenter de manière plus simple la réalité observée (avec « le minimum de mots, de paramètres et de graphiques »). Elle va pouvoir décrire les données, permettre de faire par la suite des comparaisons, détecter des valeurs extrêmes ou aberrantes, reconnaître les lois de probabilité. Elle permet aussi de produire des représentations simplifiées et synthétiques des données via des graphiques. Elle décrit la réalité grâce au traitement des données mais ne va pas expliquer ou tenter d'anticiper certains phénomènes comme le fait la statistique explicative.
Cette dernière, en revanche, cherche à comprendre, expliquer comme le nom l'indique des phénomènes. Elle va expliquer une variable à partir d'autres variables. Pour cela elle a recours à différentes méthodes qui lui permettront de faire des liens entre les variables, d'émettre des hypothèses voire même d'essayer d'anticiper certains phénomènes.

Face à l'abondance de données, il est nécessaire d'avoir des outils pour les représenter.
La géographie utilise différents types de visualisation de données : les histogrammes, les représentations sectorielles ou diagrammes circulaires, les diagrammes en bâtons, les polygones de fréquences, les courbes cumulatives, les diagrammes à rectangles horizontaux, les diagrammes à secteurs circulaires.
Pour choisir parmi ces différents types de visualisation, il faut se référer au type de variable. Les données qualitatives montrent des catégories donc on opte pour des graphiques par secteurs ou bâtons. Les quantitatives discrètes nécessitent souvent des bâtons puisqu'il s'agit de valeurs isolées, tandis que les quantitatives continues sont visualisées grâce à des histogrammes. Pour être plus précise : les variables qualitatives nominales ont besoin d'un graphique sectoriel appelé camembert ou un diagramme en bâtons. Les variables qualitatives ordinales nécessitent un histogramme avec des barres séparées (disjoint) tandis que es variables quantitatives discrètes nécessitent un diagramme en bâtons. Pour les variables quantitatives continues, le choix est plus vaste. On a le choix entre l'histogramme, le polygone des fréquences, la courbe cumulative et le diagramme à rectangles horizontaux.

Ensuite, concernant l'analyse de données elle-même, il existe différents types d'analyse de données possibles. On en identifie deux types : les méthodes descriptives et les méthodes explicatives.
Parmi les méthodes descriptives, il existe différentes façons d'analyser les données. D'abord, l'analyse factorielle en composantes principales permet de mieux visualiser les liens entre plusieurs variables quantitatives. Elle permet de mieux se représenter les informations. L'analyse factorielle des correspondances s'intéresse aux données qualitatives en permettant, elle aussi, de visualiser les liens entre les différentes données. Lorsque l'on a plus de deux variables qualitatives, on utilise l'analyse factorielle des correspondances multiples. La classification ascendante hiérarchique permet de rassembler/regrouper des données qui ont des ressemblances. Elles seront classées et hiérarchisées par la suite.
Les méthodes explicatives analysent, elles aussi, les données de différentes manières. On cherche à expliquer une donnée grâce à d'autres : « On cherche à relier une variable, à expliquer Y à des variables explicatives X1 ». Donc, on va expliquer Y à partir de X. Lorsque Y est quantitative, on a différentes méthodes, dont la première est la régression simple : on va expliquer Y à partir dune seule variable X. La seconde est la régression multiple qui fonctionne selon le même principe sauf que comme son nom l'indique, on va expliquer Y à partir de plusieurs variables explicatives que l'on peut noter comme étant X1, X2 etc. Il existe aussi l'analyse de la variance, lorsque les variables explicatives sont des catégories ou le modèle linéaire général qui combine les méthodes citées en utilisant plusieurs types de variables explicatives.
Lorsque Y est une donnée qualitative, on peut utiliser l'analyse discriminante, qui sert surtout à ranger les individus dans des catégories en fonction des variables explicatives, mais aussi la régression logistique qui sert à expliquer lY à partir de variables explicatives X, ou enfin la segmentation qui regroupe les individus selon une variable qualitative. Finalement, les modes de prévision proposent des modèles chronologiques, qui vont relier le présent au passé.
Il existe donc de multiples méthodes d'analyse de données, que l'on choisit en fonction de ses besoins et du type de données.

Mais, l'analyse de données n'a de sens que si elle se fonde sur des définitions précises. Nous allons donc faire un point vocabulaire.
En premier lieu, nous allons définir la population statistique. Elle correspond à un ensemble mathématiquement parlant, mais peut être définie comme un ensemble d'éléments présentant des caractéristiques similaires sur lesquels une étude statistique est destinée à être réalisée. Le cours propose des exemples concrets : le nombre d'habitants d'un territoire ou bien le personnel d'entreprise.
L'individu statistique, quant à lui, correspond à une unité statistique. Il n'est qu'un élément de l'ensemble, ici un élément de la population statistique. On peut distinguer deux types d'unités spatiales, des unités primaires qui sont non-agrégées ou des unités secondaires qui sont agrégées puisque issues des unités primaires. En géographie, les individus statistiques sont appelés unités spatiales puisqu'ils peuvent être localisés et cartographiés. Ils peuvent être composés d'éléments de niveau inférieurs c'est à dire de « sous unité », qu'il s'agisse de personnes, de parties de réseaux de secteurs particuliers etc. Par exemple, cela peut être une ville parmi un réseau inter-urbain ou bien un salarié parmi les autres.
Les caractères statistiques désignent les attribuent particuliers de l'individu statistique que l'on cherche à étudier dans l'analyse statistique. Ils sont l'objet de l'analyse statistique. On attribue des noms différents aux séries en fonction du nombre de caractères étudiés. Lorsqu'un seul

caractère est étudié, on parle de série numérique à une dimension. Puis, série numérique à deux dimensions quand il y a deux caractères, et série numérique multidimensionnelle s'il y en a plus. Ainsi, on parle de caractéristique statistique pour désigner une collection d'éléments sur un territoire, ou pour parler des particularités d'un salarié comme le salaire, la taille etc.
Les modalités statistiques correspondent aux « valeurs prises par un caractère » c'est à dire à une valeur possible d'un caractère, en sachant que chaque individu ne peut être associé qu'à une seule et unique modalité. Pour illustrer, si l'on a comme caractère la couleur des cheveux, le brun correspondra à une modalité, tout comme le blond ou le roux. La personne ne peut être associée qu'à une modalité car elle ne peut pas être à la fois brune, blonde, rousse. Cet exemple fonctionne aussi avec la couleur des yeux. Si l'on connaît les modalités du caractère individu par individu, il devient une « variable statistique » .
Il existe différents types de caractères : il y a les variables qualitatives qui sont difficilement mesurées incluant les qualitatives nominales qui décrivent des états et les qualitatives ordinales qui permettent une médiane et donc de crée un classement, ou un ordre. Puis, les variables quantitatives qui comme leur nom l'indique désigne une quantité et peuvent donc être mesurées (chiffrées). Elles sont soit discrètes et donc indépendantes ou « isolées » les unes des autres. Ce sont souvent des choix en oui/non. Ces données peuvent être comptées, c'est justement leur intérêt. Enfin les variables quantitatives continues sont comprises dans un intervalle (ex : salaire). On va donc pouvoir mesurer quelque chose grâce à elles. Il en existe deux types : les variables d'intervalle dans lesquelles le zéro fait parti intégrante de la mesure (ex : température) et les variables de rapport ou inversement le zéro correspond à « l'absence de la valeur mesurée » (ex : volume de vente).
Entre ces différents types de caractères, ce sont bien les variables quantitatives qui sont les plus utiles dans le cadre d'une analyse puisque comme indiqué précédemment, elles sont « associables à des lois de probabilités » et peuvent donc être comptées, servent de mesure etc. Les variables qualitatives sont difficilement soumises à des calculs permettant de créer une hiérarchie ou un ordre. Il est donc plus difficile de manipuler les données via des outils statistiques avec ces catégories là.

Utiliser une variable quantitative continue peut impliquer de la découper en classes c'est à dire en intervalles. C'est ce qu'on appelle la discrétisation. Mesurer l'amplitude signifie mesurer la largeur de la classe et uniquement à une classe particulière. Elle ne peut pas s'appliquer à la série entière. Elle peut donc être définie comme la valeur de b – a en sachant que a est la valeur minimale et b la valeur maximale.
La densité quant à elle, est le « rapport entre l'effectif ni et l'amplitude de la classe » ce qui donne « d = ni / (b – a) ». La densité s'avère utile lorsque les classes n'ont pas toutes la même amplitude. Cela permet de rendre les différentes colonnes de l'histogramme comparables, afin d'analyser plus simplement les données même si les classes n'ont pas la même largeur.

Certaines formules existantes peuvent nous aider dans l'analyse des données.
Les formules de Sturges et Yule en font parties. Elles servent toutes les deux à calculer le nombre de classes optimal.
En effet, la formule de Sturges définit le nombre de classes à utiliser en fonction de l'effectif total, notamment si l'on veut faire un histogramme par exemple. Cela évite d'avoir un nombre de classes insuffisant qui rendrait le travail imprécis, ou au contraire d'en avoir trop, le risque étant que le travail soit abscons à cause de l'illisibilité.

La formule de Yule quant à elle utilise la racine quatrième du nombre d'observations, toujours dans l'optique d'obtenir le nombre de classes optimal.
Néanmoins le nombre de classes reste un choix. Si l'utilisation des formules permet de donner une fourchette, c'est bien le chercheur qui va décider combien de classes utiliser en fonction de ses besoins, du degré de précision qu'il souhaite avoir, de la lisibilité, ou encore de la distribution des données. Ainsi l'objectif des formules est surtout de donner un ordre de grandeur afin que le graphique ne soit ni illisible, ni trop imprécis, en laissant le chercheur faire son choix dans la fourchette conseillée.

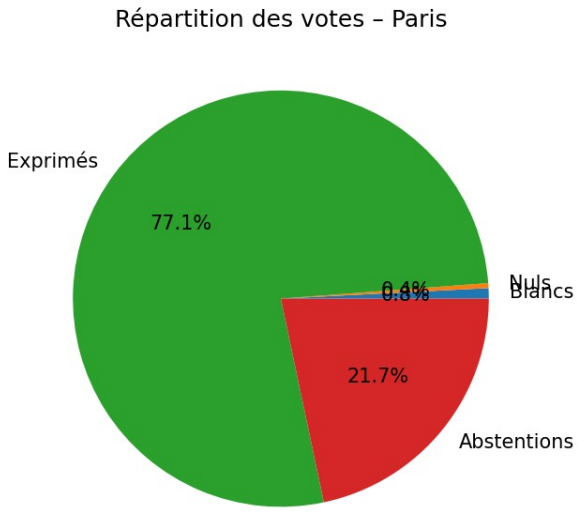
Finalement, l'étude des effectifs et fréquence permet de relier les probabilités et les statistiques.
L'effectif appelé aussi fréquence absolue correspond au nombre de fois ou la valeur apparaît au sein de la population statistique. Par exemple, si l'on a une population de vingts habitants dont dix ont les yeux marrons, cinq ont les yeux bleus, trois ont les yeux verts et les deux derniers les yeux noisette, l'effectif sera de 20.
La fréquence au contraire correspond à une part de la population. C'est une proportion. Si l'on reprend l'exemple précédent, la fréquence des yeux bleus au sein de la population est de 5/20 personnes, ce qui correspond à 25 % de la population statistique. Grace à la fréquence, on peut comparer des variables sans prendre en compte la population statistique.
La fréquence cumulée quant à elle, est calculée lorsque les valeurs sont rangées dans un ordre croissant, ce qui est le cas des variables quantitatives. C'est une fréquence qui monte étape par étape : elle s'obtient en « additionnant les valeurs inférieures ou égales à k ». Ainsi, la dernière fréquence cumulée fait toujours 1.
Finalement, une distribution statistique désigne l'organisation des données et le mode de répartition des valeurs dans une série de données. Elle s'intéresse à la structuration des valeurs. La fréquence sert à constituer une distribution statistique, à partir de données empiriques c'est à dire issues des données observées.

Mise en œuvre avec python

Nous allons maintenant nous intéresser aux résultats des manipulations de la séance 2. Pour précision, j'ai utilisé vs code durant toutes les séances. Je n'avais pas réussi à mettre correctement python sur docker, j'ai donc installé directement python « en dur ».

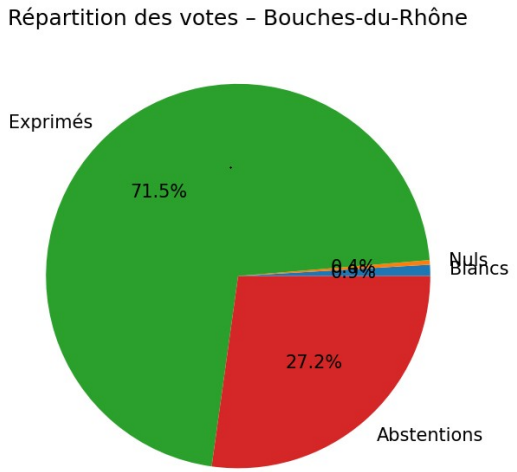
La séance 2 aboutit à l'élaboration de très nombreux diagrammes. La liste que je vais mettre ci dessous n'est donc pas exhaustive, j'en ai choisi certains pour les commenter. Je vais également comparer les résultats des diagrammes entre eux.

Le traitement des différentes données sur python permet d'établir des diagrammes qui ont pour objectif de montrer la répartition du nombre de votes « exprimés », « nuls », « blanc » ou bien d'« abstention » par département lors du premier tour des élections présidentielles de 2022.



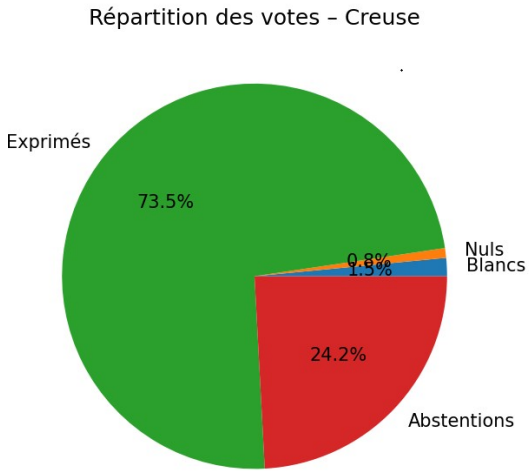
Ce graphique montre cette répartition dans Paris « intra-muros ». Tout d'abord, il est important de constater que les suffrages exprimés sont importants. En effet, ils sont estimés à 77,1 %, tandis que l'abstention est aux alentours de 21 %. Si l'on veut connaître la part des parisiens qui se sont rendus aux bureaux de vote, il faut ajouter les votes blancs et nuls dans les suffrages exprimés, ce qui permet d'arriver presque à 80 % de taux de participation à ces élections.
Nous pouvons interpréter ici les votes blancs comme étant un rejet des propositions politiques formulés par les partis. L'abstention quant à elle peut exprimer un désintérêt pour les questions politiques de la part des habitants, des difficultés pour se rendre au bureau de vote, ou un désaccord avec le fonctionnement du système politique actuel. Paris étant bien aménagée en termes de transport, on imagine que cette explication-ci n'est pas la plus importante à retenir.

Si j'ai mis l'accent sur la participation importante des habitants aux élections à Paris, c'est parce que j'ai sélectionné les graphiques au préalable, et que j'ai donc pu voir les suffrages exprimés et l'abstention dans les autres départements ou collectivités d'Outre mer. En effet, lorsque l'on compare la participation des habitants aux élections de 2022 entre Paris et la plupart des autres départements français, on s'aperçoit que la participation aux élections est presque toujours plus importante à Paris qu'ailleurs. L'exemple des Bouches-du-Rhône et de la Creuse ci dessous en témoignent.



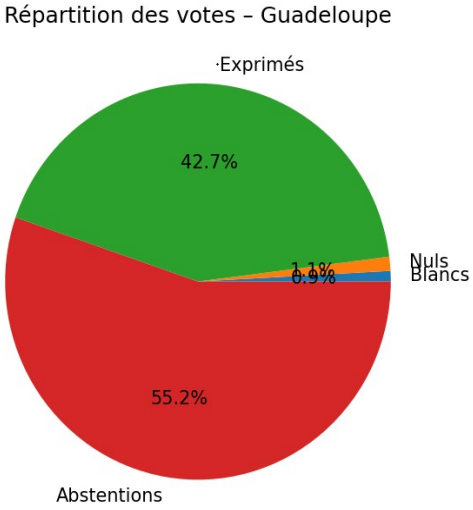
En effet, le diagramme des Bouches-du-Rhônes indique une abstention plus importante qu'à Paris. En regardant la répartition, on remarque que, si l'on grossit les traits, on est davantage sur environ 30 % d'abstention et 70 % de suffrages « exprimés ». Même si l'on ajoute les suffrages « blancs » et « nuls » aux exprimés pour montrer que les habitants se sont déplacés pour ces élections, on arrive vers 72,8 % de taux de participation. De plus, comptabiliser les suffrages « exprimés » avec les « blancs » et « nuls » n'est pas très honnête, ces derniers ne représentent pas du tout la même chose (on le fait ici pour pouvoir comparer). Néanmoins, les

suffrages exprimés restent malgré tout importants. La plupart des habitants de la région se déplacent pour voter. Je tiens à le souligner car ce n'est pas le cas pour tous les départements.

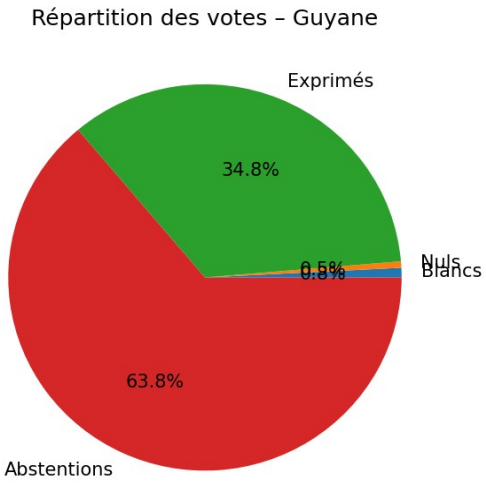


Concernant la Creuse, les suffrages exprimés restent importants également (73,5%). On observe toujours une moindre participation en comparaison avec Paris (24,2 % d'abstention), mais cette différence est très faible. Si l'on y ajoute les votes « nuls » et « blancs », elle est presque infime.

Désormais, nous allons nous pencher sur le cas des départements et collectivités d'Outre mer. Ici, l'abstention prend des proportions bien plus importantes.

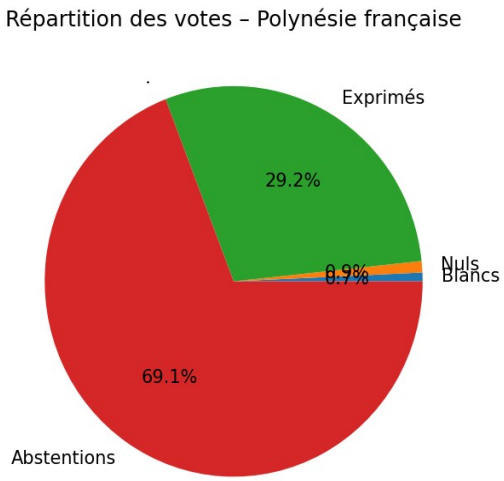


En effet, en Guadeloupe, les votes « exprimés » représentent à peine 42,7 %. Les habitants qui se sont abstenus sont donc très nombreux : ils sont même majoritaires avec 55,2 % d'abstention. Les votes nuls sont aussi légèrement plus importants que pour les départements métropolitains (néanmoins, la différence est infime et je doute qu'elle soit vraiment significative). Ceci est certainement révélateur du sentiment des habitants de Guadeloupe, qui ne se reconnaissent pas dans les propositions politiques des partis, et qui ne se sentent peut-être pas inclus au sein du fonctionnement de la démocratie représentative française.



Ce constat fait sens pour la Guyane également, qui se retrouve avec un pourcentage de suffrages exprimés dérisoire, de 34,8 %, alors que l'abstention atteint 63,9 %. Ainsi, on constate une énorme différence de participation aux élections entre les habitants de métropole et ceux des départements d'Outre mer. Cela peut nous

amener à nous questionner sur les propositions politiques des partis, mais aussi sur le sentiment d'appartenance de ces espaces et de leur habitants à la France. Ainsi les politiques menées par les anciens gouvernements ont pu peut être négliger ou laisser les départements d'Outre mer de coté, augmentant le sentiment d'indifférence de leurs habitants à l'égard des propositions politiques de tous bords.



La Polynésie française est une collectivité d'Outre mer, l'organisation politique et le lien avec la métropole est donc différent. L'autonomie de ces espaces est plus grande. L'abstention est certainement la plus élevée en comparaison avec les autres départements et collectivités. Seulement, il faut s'intéresser au contexte politique de l'île. En effet, les indépendantistes dirigés par Oscar Temaru ont une place importante dans la vie politique polynésienne et milite pour organiser un référendum d'autodétermination. La volonté de couper le lien avec la France est donc présente au sein de cette collectivité : en 2013 ces derniers avaient milité pour faire figurer la Polynésie française sur la liste de l'ONU des territoires non autonomes à décoloniser (Mike Leyral, le monde, 2025).

La réalisation de la séance deux a été assez compliquée, d'abord parce que je n'avais jamais codé. Je n'avais pas fait l'option numérique et sciences informatique pendant mon lycée, j'ai arrêté les mathématiques après la seconde dans le cadre de la nouvelle réforme du baccalauréat et l'initiation au langage de programmation obligatoire au lycée est tombée pendant le confinement : elle n'a donc pas été faite. De plus, après mon année de terminale, je suis allée en classe préparatoire littéraire pendant les trois ans de licence. Je n'ai donc jamais vraiment utilisé un ordinateur pour coder, et je n'ai plus eu de mathématiques dans mon parcours scolaire pendant presque 6 ans. La reprise de ce type de raisonnement n'était donc pas évidente, même si je n'étais pas particulièrement mauvaise à l'époque dans ces matières.

Néanmoins, les raisonnements mathématiques n'étaient pas tellement en cause, car contrairement à beaucoup d'élèves, je n'ai pas souffert de grandes difficultés d'apprentissage en mathématiques durant ma scolarité. J'ai surtout arrêté cette matière car je préférais les autres et non car je détestais celle-ci. Ainsi, c'est le langage de programmation qui a été compliqué à maîtriser (et l'installation de docker, mais ces problèmes ont déjà fait l'objet de discussions durant les cours, je ne vais donc pas m'y attarder). Pour maîtriser les bases afin de commencer à faire les manipulations des séances, j'ai utilisé des ressources disponibles sur youtube, notamment la vidéo « démarrer python sans rien installer en 30 sec-programmation-tutoriel-lycée » de jaicompris Maths (voici le lien : https://www.youtube.com/watch?v=pE11tsQo4dA&list=PL_ZtK1TB2InpOSbk7_EoteopTz5hqptRR). Il a fait plusieurs autres vidéos sur sa chaîne qui expliquent les variables, les différentes instructions et certaines erreurs à ne pas faire. Je me suis également procurée le livre *Python pour les débutants absolus, Un*

guide complet pour apprendre la programmation Python en 5 jours de Laurentine K.Masson. Ces ressources m'ont aidé à comprendre les bases du code, pour savoir ce que je faisais. Mais elles ne pouvaient pas répondre à mes questions sur des erreurs que j'avais dans mon code et que je ne comprenais pas.

Lorsque des erreurs que je ne comprenais pas du tout s'affichaient, je posais la question dans un groupe whatsapp de classe que notre promo « Géosuds » avait crée spécialement pour l'analyse de données. Zara était très réactive sur ce groupe en plus du discord et a répondu à beaucoup de nos questions notamment lors de l'installation de vs code, utilisé par beaucoup d'entre nous suite à ses conseils. A partir de ce groupe, elle nous a proposé d'organiser des réunions dans des salles de la bibliothèque de Clignancourt. Cela nous a permis d'avancer sur les séances en groupe. Là encore, Zara nous a beaucoup aidé. J'ai également utilisé l'intelligence artificielle pour répondre à certaines erreurs de code (dont beaucoup d'erreurs d'écriture et des oublis d'espace qui faussaient le code, des fautes bêtes, je dois l'avouer).

Au bout d'un moment, le code est devenu plus fluide, mais je dois avouer que pendant les dernières séances, un problème est apparu : le fichier de code n'arrivait pas à trouver les données sur mon ordinateur, alors que je suivais les instructions comme il fallait. J'ai demandé à un de mes amis qui codait à l'université aussi en licence de mathématiques et il n'a pas réussi à arranger le problème. Zara m'a proposé de faire comme ceci : cd : copier le chemin d'accès du main.py, de tout enlever après séance-02, remettre les guillemets et cliquer sur entrer. Dans le code lui même, mettre with open(« data/resultats….tourcsv », encoding = « utf-8 ») as fichier : contenu = pd.read_csv(fichier) puis print(contenu). Ces modifications ont fonctionné, j'ai procédé de cette manière pour les dernière séances. En revanche, j'ai réalisé la séance 2 et 3 avec mon ancien ordinateur. Je n'ai donc pas eu ce problème.
Néanmoins, ce problème est très étrange car j'ai essayé sur un ordinateur différent pour voir si j'arrivais à faire les séances 5 et - sans problème pour trouver le fichier, et le code fonctionnait.

Séance 3
Les paramètres statistiques élémentaires

La séance 2 sur les paramètres statistiques élémentaires s'attarde principalement sur le caractère quantitatif, qui est le plus général. Cette information nous est donnée dès le début du chapitre dans l'introduction : « les paramètres statistiques concernent principalement les variables quantitatives, et ponctuellement qualitatives ».
En effet, le caractère quantitatif a l'avantage de pouvoir bénéficier de nombreuses méthodes de calculs, et donc de pouvoir faire l'objet de plus de manipulations qui seront éclairantes pour pour faire une analyse des données. La totalité presque des méthodes présentes dans le cours ne sont utilisables qu'avec des nombres comme valeurs, que l'on parle de paramètres de position, ou de dispersion.
Le caractère qualitatif ne correspond pas à ces outils. En effet, le principe même du caractère qualitatif est qu'il n'est pas forcément mesurable. Il n'est même pas forcément un nombre surtout lorsqu'on s'intéresse au qualitatif nominal par exemple. Les outils du cours ne sont donc pas utilisables pour ce type de données.

Le chapitre présente toutes les approches, tous les outils et tous les paramètres statistiques existant qui s'appliquent au caractère quantitatif. On compte trois grandes catégories de paramètres, que je vais présenter ci dessous.

D'abord, les paramètres de position comprennent la moyenne, la médiane, la moyenne quadratique, la moyenne harmonique. Sur la base des données recueillies dans un échantillon, les indicateurs de position fournissent des informations sur l'emplacement du « centre » de la distribution. Cela permet de savoir ou se situe l'ensemble des données. Ils sont très utiles pour pour résumer ou décrire une liste de données avec un seul paramètre. Je vais expliquer quelques unes des notions.
La moyenne se calcule en additionnant toutes les valeurs et en divisant par le nombre de valeurs. Elle permet d'obtenir la valeur centrale de toutes les données, elle donne le résultat typique mais pas le plus fréquent. Cependant, elle est très sensible aux valeurs extrêmes.
La médiane sert à obtenir la valeur centrale lorsque toutes les données sont dans l'ordre. Elle divise l'ensemble des données en deux : 50 % des valeurs sont plus petites et 50 % sont plus grandes. Elle permet de montrer la valeur du milieu. Son grande avantage : elle n'est pas sensible aux valeurs extrêmes.
La moyenne quadratique est calculée à partir des carrés des valeurs. Il faut que les valeurs soient positives pour que cela fonctionne. Dans ce cas, il en résulte une moyenne qui accorde plus d'importance aux grandes valeurs.
Au contraire la moyenne harmonique est plus utile lorsqu'on veut accentuer sur l'importance des petites valeurs. Elle va calculer des grandeurs par unité.

Puis, nous allons désormais nous attarder sur les paramètres de dispersion qui comprennent la variance, l'écart type, le coefficient de variation, l'étendue, l'écart interquartile, et l'écart moyen. Il décrit la dispersion des valeurs sur un échantillon autour d'un indicateur de position. Ils sont une mesure de l'ampleur des fluctuations d'un échantillon autour d'une valeur moyenne.
La variance mesure à quel point les valeurs s'écartent de la moyenne, en utilisant l'écart de chaque valeur à la moyenne, en élevant ces écarts au carré et en faisant la moyenne de ces carrés. L'objectif est d'éviter que les écarts positifs et négatifs s'annulent. La variance donne donc plus de poids aux grands écarts.
L' écart-type est la racine carré de la variance. On ramène la mesure dans la même unité que les données, ce qui est plus simple à analyser ensuite. Cela sert à comparer la variabilité de deux séries, détecter les données anormales.
L'étendue, quant à elle désigne la différence entre la valeur maximale et la valeur minimale. Elle permet d'obtenir une idée de l'amplitude mais est sensible aux valeurs extrêmes.
Le coefficient de variation se calcule en divisant l'écart type par la moyenne. Il s'utilise pour comparer la dispersion relative de deux séries de données même si elles n'ont pas la même unité ou la même échelle. On peut comparer des variabilités incomparables autrement. Plus le coefficient de variation est faible, plus le processus est fiable.
L'écart interquartile mesure l'écart de 50 % des données centrales. Il indique comment sont étalées les valeurs centrales, en ignorant les valeurs extrêmes. Il n'est donc pas influencé par les valeurs aberrantes, ce qui peut être un grand atout.
L'écart moyen permet de donner une variabilité, de savoir « de combien » les données s'éloignent de la moyenne. Il s'agit de « la moyenne des écarts absolus entre chaque valeur et la moyenne ». Il est plus solide que la variance, plus raisonnable vis à vis des valeurs extrêmes, mais pas forcément toujours utilisé.

Enfin, les paramètres de forme comprennent le coefficient d'asymétrie β 1 et le coefficient d'aplatissement β 2. Ces paramètres servent à connaître la forme de la distribution, pour savoir si elle est symétrique ou non ou bien s'il y a un aplatissement ou une concentration autour de la moyenne.
Pour comprendre ces deux coefficients, le cours explique qu'il faut comprendre ce qu'est un moment. Néanmoins, je ne vais pas me lancer dans une explication de ce dernier, il s'agit simplement ici de commenter dans les grandes lignes, de comprendre à quoi servent les outils mentionnés comme étant des paramètres de forme.
Le coefficient d'asymétrie β 1 permet de constater, comme son nom l'indique, si la distribution est symétrique ou non : si elle est plutôt à droite, plutôt à gauche ou bien complètement symétrique. Lorsque β 1 est inférieur à 0, alors l'asymétrie est positive et la distribution est étirée vers la droite. Au contraire, si β 1 est inférieur à 0, l'asymétrie est négative et la distribution est étirée vers la gauche. Si β 1 est égal à 0, la distribution est symétrique.
Le coefficient d'aplatissement β 2 sert à savoir si la distribution est pointue, aplatie ou normale. Si β 2 est supérieur à 0 c'est que la distribution est plus aplatie donc plus éloignée de la moyenne, si β 2 est inférieur à 0, c'est que la distribution est plus pointue c'est à dire plus rassemblée autour de la moyenne. Enfin, lorsque β 2 est égal à 0, c'est que nous avons en face de nous la loi normale.

Puis, pour finir, il existe les paramètres de concentration qui comprennent la médiale et le coefficient de concentration C

On compte aussi les paramètres de concentration qui comprennent la médiale et l'indice de Gini dont nous parlerons plus tard.

Chacun de ces paramètres, chacune de ces approches est utilisée uniquement pour la variable quantitative, et puisqu'ils constituent la quasi-intégralité du système statistique, on peut en conclure que le caractère quantitatif est le plus général.

Le cours distingue deux types de caractères quantitatifs, discrets et continus. Les caractères quantitatifs discrets prennent des valeurs isolées et sont composés par des modalités, associées à leurs effectifs. Les paramètres se calculent alors sous forme de sommes (finies). Les caractères quantitatifs continus sont décrits par une fonction de densité f (x) . Les paramètres se définissent alors à l'aide d'intégrales. Donc, si l'on prend l'exemple de la moyenne, le cours énonce : « La moyenne ayant été définie par la somme pour une variable discrète, devient une intégrale pour une variable continue. » On distingue les deux critères parce que les outils utilisés sont différents. D'un côté, on utilise des sommes et de l'autre des intégrales. D'un côté, on utilise des effectifs, et de l'autre une densité. L'utilisation du discret ou du continu change la manière de calculer, les formules changent complètement. Cela modifie le calcul des moyennes, de la variance et de l'écart type, de la médiane, la définition des quantiles et la théorie des moments.

Nous allons nous attarder quelques temps sur les paramètres de position.

Dans le cours, on nous présente un tableau des moyennes. Il y a six types de moyenne : arithmétique, quadratique, harmonique, géométrique, mobile/glissante, et fonctionnelle. Chacune de ces moyennes a ses spécificités et ses emplois. Il existe autant de contextes statistiques qu'il y a de moyennes différentes. Pour illustrer les différents emplois des moyennes, on peut se référer au tableau : la moyenne arithmétique est « sensible aux valeurs extrêmes », il peut donc être nécessaire de les supprimer pour ne pas affecter la moyenne ; la moyenne quadratique est utilisée pour la géométrie, la moyenne harmonique pour la vitesse, la moyenne géométrique pour les produits successifs. Il faut déduire l'utilité des moyennes mobiles et fonctionnelles d'après les formules dans le tableau. La moyenne fonctionnelle est utilisée pour les fonctions, et la moyenne mobile semble être utilisée pour calculer une moyenne sur un groupe de données qui fluctue, afin d'en donner une tendance.

La médiane « partage la série de données en deux parties en comprenant le même nombre de données d'une part et de l'autre. » De cette manière, elle n'est pas influencée par les valeurs extrêmes et ne change pas lorsqu'une valeur est très élevée ou très basse, contrairement à la moyenne. Elle est au centre exact et permet ainsi de rendre compte des distributions dissymétriques (elle distribue selon le classement, pas selon l'amplitude, donc on peut déduire l'influence plus importante de tel ou tel pôle). Cependant, la médiane a ses défauts : elle ne peut être combinée, on ne peut pas tirer une médiane globale à partir de sous-médianes, et elle ne peut pas non plus être utilisée pour des estimations, puisqu'elle est influencée par le nombre de données et non pas par leur valeur.

Le mode est défini comme la valeur qui correspond à l'effectif maximal pour une variable discrète ou la densité maximale de probabilité pour une variable continue. Il peut donc être calculé lorsque dans la chaîne il y a une valeur qui revient plus fréquemment que les autres. On

l’appelle donc la « moyenne de fréquence ». Donc, le mode n’existe pas, il ne peut être calculé dans une série où toutes les valeurs apparaissent à la même fréquence. Par ailleurs, il peut y avoir plusieurs modes (distribution bimodale ou plurimodale). Ainsi, « il faut considérer que deux ou plusieurs populations distinctes ayant chacune leurs caractéristiques propres sont en présence », des sous-populations en quelque sorte. Le mode n’existe donc que s’il y a une ou plusieurs valeurs dominantes. Il n’est pas toujours présent, ni toujours unique.

Nous allons maintenant nous attarder sur les paramètres de concentration.

La médiale divise la masse totale eu deux parties à 50%, pas les effectifs comme la médiane ; c’est une médiane des valeurs globales notées *nixi*. Pour l’obtenir, il faut calculer les valeurs globales relatives, notées *qi*. Il faut ensuite s’attarder sur les valeurs globales cumulées notée Qi. Pour cela, il faut cumuler les qi(Qi = q1,Q2 = q1 + q2), ...).Il faut chercher le plus petit Qi qui est supérieur à 0,5. C’est ce qu’on appelle la médiale par interpolation linéaire. La médiale se trouve dans l’intervalle où on remarque pour la première fois que les valeurs globales cumulées vont au-delà de 50%. La médiale sert surtout à analyser les phénomènes économiques en mesurant la répartition réelle. Cela est possible grâce à l’indice de concentration, ou indice de Gini, mesuré par la comparaison entre la médiale et la médiane. On la représente en graphique sur la courbe de Lorenz. Plus la médiale (en ordonnée) est supérieure à la médiane (en abscisse), plus la concentration est forte. Cela permet d’analyser, par exemple, les inégalités.

Désormais, nous nous intéressons aux paramètres de dispersion.

L’écart à la moyenne est la différence entre une valeur et la moyenne . Elle est utile pour commencer à mesurer la dispersion d’une série, les écarts entre les valeurs. Mais si l’on additionne tous les écarts, le résultat sera toujours 0. Donc, pour mesurer la dispersion, si les données sont regroupées ou étalées, si la série est homogène ou hétérogène, on ne peut pas faire une moyenne des écarts. C’est pour cette raison qu’on calcule une variance plutôt qu’un écart à la moyenne. La variance c’est le fait d’élever les écarts au carré. « Elle tient compte de toutes les données ». Cela permet de rendre les calculs possibles, de donner des propriétés mathématiques utiles, elle permet des opérations algébriques et est compatible avec la théorie des moments. Mais puisque la variance est exprimée au carré, il est plus difficile de l’interpréter de manière concrète. On utilise alors l’écart type, noté , puisqu’il correspond à « la racine carrée de la variance ». Cela facilite l’interprétation et conserve les propriétés du carré.

L’étendue est définie dans le cours comme la différence entre la valeur maximale et la valeur minimale. Elle est donc très facile à calculer et permet de donner une idée de la dispersion globale de la série. C’est un indicateur simple et direct. Cependant, elle est peu fiable car elle ne dépend que de deux valeurs extrêmes et ne permet pas d’entrer dans le détail des données, surtout s’il y en a beaucoup.

Un quantile sert à partager la série en plusieurs parties égales, notées k. Il sert donc à analyser la distribution de manière plus fine que la moyenne ou la médiane. Donc, si l’on divise la série en quatre parties égales, on introduit trois quartiles, la distribution sur laquelle le cours s’appuie le plus, (25%, 50%, 75% de la série), notés Q1,Q2,Q3. Le deuxième correspond donc à la

médiane, et l'écart entre le premier et le troisième contient donc 50% des valeurs. Les quantiles permettent donc d'isoler certaines parties de la série pour les étudier plus en profondeur. Les autres quantiles les plus utilisés sont les quintiles (k = 5), les déciles (k = 10), et les centiles (k = 100)…

La boîte de dispersion, ou boîte à moustache, permet de représenter la distribution en graphique, afin de visualiser les caractéristiques de la distribution : les quartiles, la médiane, les valeurs minimales et maximales, et la dissymétrie de la distribution surtout ! Entre Q1 et Q3, on trace un rectangle, et un trait marque la médiane. Les moustaches sont les segments qui ponctuent chacun la valeur minimale et la valeur maximale. La boîte à moustaches permet de comparer plusieurs séries, et de constater la symétrie ou la dissymétrie par rapport à la médiane : si le rectangle n'est pas axé sur le centre, sur la médiane.

Nous nous intéressons désormais aux paramètres de forme.

Les moments sont des outils, ils servent à décrire la forme d'une distribution. Les moments centrés sont calculés autour de la moyenne, on s'intéresse à comment les valeurs s'écartent de la moyenne, puis on élève ces écarts à la puissance r. Le moment centré d'ordre mesure donc la forme de la distribution par rapport à la moyenne. Les moments centrés sont divisés en plusieurs ordres : le moment centré d'ordre 1 vaut toujours 0, car la somme des écarts à la moyenne est toujours égale à 0. Le moment centré d'ordre 2 correspond à la variance. Le 3 correspond au coefficient d'asymétrie, et le 4 au coefficient d'aplatissement. Donc, les moments caractérisent la forme globale de la distribution.
Les moments absolus servent surtout à calculer l'écart moyen ; ils ajoutent simplement un outil de plus, mais basé sur la valeur absolue.

La symétrie d'une distribution permet de comprendre la structure et à déterminer les paramètres pour la décrire. Elle sert à l'analyse statistique, selon que la structure de la distribution est symétrique ou asymétrique, on peut en conclure différentes affirmations. Vérifier la symétrie est indispensable. Pour vérifier la symétrie d'une distribution, on utilise le coefficient d'asymétrie. Si , la distribution est symétrique. Si est supérieur à 0, il y a une dissymétrie positive (à droite), si est inférieur à 0, il y a une dissymétrie négative (à gauche). Une distribution symétrique a une moyenne, un mode et une médiane égales.

Mise en œuvre avec Python

Le code vise à sélectionner les caractères quantitatifs, calculer les paramètres de position et de dispersion, représenter graphiquement les paramètres avec une boite à moustaches et appliquer la catégorisation d'une variable quantitative continue.

D'abord, on sélectionne les variables quantitatives :
code : quant_cols = df.select_dtypes(include=[np.number]).columns.tolist()

On s’intéresse ici surtout aux caractères quantitatifs. Cette ligne permet donc de faire la distinction entre le caractère quantitatif et qualitatif. C’est une ligne qui prépare les calculs qui vont suivre.

Le calcul des paramètres de position :

Moyenne :
code : moyennes = df[quant_cols].mean()
La moyenne arithmétique est un paramètre de position. Le code va ici calculer l’espérance empirique de chaque variable quantitative

Médiane :
code : medians = df[quant_cols].median()
La médiane va partager la population en deux selon les indications vues dans la partie questions de cours. La médiane est moins sensible aux valeurs extrêmes que la moyenne

Mode :
code : modes_df = df[quant_cols].mode()
Le mode est la valeur la plus fréquente. Ce code sert à identifier les valeurs dominantes (ou la valeur dominante). Il conserve le premier mode, ce qui est cohérent avec le cas de distributions éventuellement plurimodales.

Le calcul des paramètres de dispersion :

L’écart type :
code : stds = df[quant_cols].std()
L’écart-type est la racine carrée de la variance. Il mesure la dispersion des valeurs autour de la moyenne.

L’écart absolu moyen :
code : abs_dev_mean = df[quant_cols].apply
(lambda x: np.mean(np.abs(x – x.mean())))
Le code va calculer les écarts à la moyenne, va prendre la valeur absolue et faire la moyenne de ces écarts. Il s’agit d’obtenir une alternative à la variance

L’étendue
code : etendue = df[quant_cols].max() - df[quant_cols].min()
Le code va mesurer des dispersions et dépend seulement des valeurs extrêmes.

Quantiles et distances interquantiles

Distance interquartile
code : iqr = df[quant_cols].quantile(0.75) - df[quant_cols].quantile(0.25)
L’écart interquartile contient 50 % de la population centrale normalement

Distance interdécile
code : idr = df[quant_cols].quantile(0.9) – df[quant_cols].quantile(0.1)

Les déciles sont mentionnés comme quantiles usuels. On va mesurer la distribution des 80 % au centre de la distribution.

Boite à moustache
code : df.boxplot(column=[col])
La boite à moustache permet de comparer visuellement les distributions. Le graphique va synthétiser les paramètres de position et de position calculés précédemment.

Enfin, la catégorisation d'une variable quantitative continue:
code : surface[(surface > 0) & (surface <= 10)]
Cette ligne permet de créer des classes d'intervalles et de dénombrer des effectifs par classe. Il s'agit d'une étape de préparation de distribution statistique regroupée.

Séance 4
Les distributions statistiques

Cette séance s'intéresse aux distributions statistiques. Elle est accompagnée de captures d'écran des graphiques du code réalisé dans le cadre de la section « manipulations ».

Tout d'abord, il existe différents type de distribution statistique selon le type de variables. Ainsi, pour choisir entre une distribution statistique avec des variables discrètes et une avec des variables continues, il faut d'abord s'intéresser à la « nature du phénomène étudié », c'est à dire à la nature de la variable.

Les lois discrètes s'appliquent lorsque la variable représente un comptage : elles sont utilisées pour modéliser. Elle ne prend que des valeurs particulières, surtout des nombres entiers et il n'y a pas de valeur entre les deux valeurs possibles. Le texte cite quelques exemples éclairants : résultat d'un jeu de hasard, réponse d'un sondage, nombre de personnes dans une file d'attente etc. On choisit une loi discrète parce qu'on peut comptabiliser quelque chose, notamment grâce au fait qu'il s'agit de nombre entier et pas de « demi-chose » et donc des chiffres décimaux.
En revanche, les variables continues peuvent être n'importe quelles valeurs dans un intervalle. Le nombre de valeurs possibles est infini, raison pour laquelle on doit avoir des intervalles. Ainsi, une variable continue peut prendre une multitude de valeurs (donc des décimaux également, contrairement aux variables discrètes). Ainsi, l'objectif est de mesurer quelque chose. Le texte cite encore une fois certains exemples comme la taille de quelqu'un, la pluie etc. La différence majeure entre les deux se situe donc ici : la distribution statistique avec des variables discrètes marche avec une « fonction de répartition en escalier ». Cela signifie que l'on peut avoir 10 et 11 en variable mais pas 10,01, ni 10,02. IL n'y aura rien entre les deux. Ce n'est pas le cas pour la distribution statistique des variables continues qui sont denses et peuvent être calculées avec des intégrales. La variable se situera n'importe ou dans un intervalle : elle pourra être 10,01 ou 10,02 si l'on reprend l'exemple précédent.

L'attention doit désormais être portée sur la forme de la distribution empirique. Il faut regarder la forme de toutes les données qu'il s'agisse d'histogramme ou autres avant de choisir. Ce critère s'additionne à celui de la nature des données. Concernant les variables discrètes, il faut prendre en compte la nature de ces données. Comme nous l'avons vu précédemment, les variables discrètes produisent des sauts. Or, on pourra constater ces derniers sur les représentations qui sont en escalier. Cela signifie qu'on verra des barres séparées, des valeurs qui seront spécifiques avec des nombres entiers, des paliers. Ainsi, on optera pour une loi discrète. En revanche si on ne voit pas de ruptures, si la courbe est continue, comme les courbes exponentielles par exemple, ce sont donc des variables continues.
Il est aussi possible de comparer les distributions que l'on observe avec une loi théorique, notamment une courbe normale qui a une forme de cloche, une courbe log-normale qui est très

asymétrique, une courbe de Pareto qui est plus élevée à droite, et une courbe Zipf qui a des lignes décroissantes par rang.

Ensuite, il faut aussi faire attention à la temporalité : il faut regarder moyenne, qui centralise les données, la variance qui mesure la dispersion, l'asymétrie.
Par exemple, la loi poisson est utile pour modéliser les données car la moyenne équivaut à peut près à la variance. Il faut aussi faire attention à la symétrie : voir si la distribution est symétrique, donc représente une loi normale, ou décalée à droite ce qui permet d'identifier la log-normale et Pareto, ou bien asymétrique pour Zipf ou Pareto également. Chaque type de distribution a ses propriétés, donc la variance permet elle aussi d'en déterminer certaines. Zipf et Pareto ont par exemple une très grande variance.

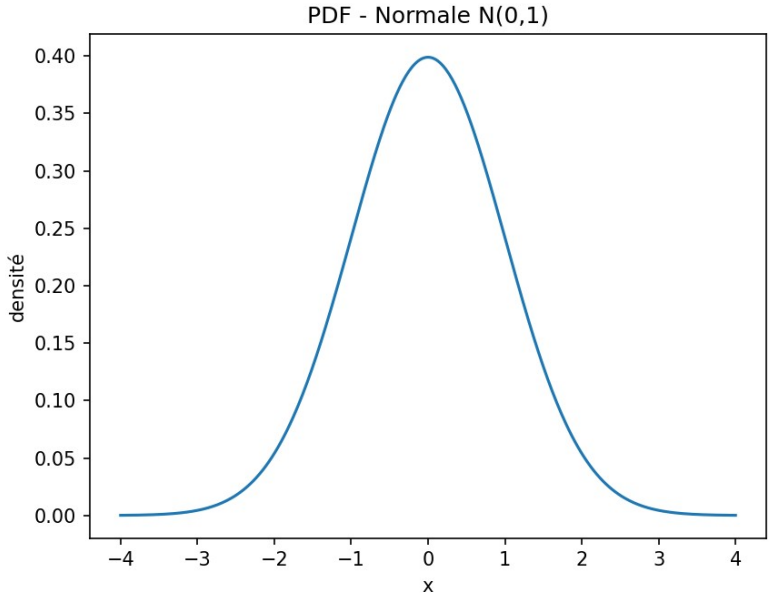
Il est également conseillé de faire attention au nombre de paramètres des lois. En effet, les paramètres permettent d'adapter la position, la largeur, la forme ou encore la symétrie de la distribution. Ainsi, cette dernière peut mieux s'adapter aux données.
Si l'on prend l'exemple de la loi normale avec deux paramètres, on va rapidement remarquer son manque de flexibilité. Elle reste complètement symétrique : elle peut aller plus vers la gauche ou vers la droite, être plus condensée mais elle ne peut pas montrer des représentations asymétriques.
Autre exemple, la loi de Pareto n'a qu'un seul paramètre, ce qui l'empêche d'avoir une bonne flexibilité. A l'inverse, elle représente une tige et n'est donc pas du tout symétrique. Avoir plus de paramètres permet de bouger la courbe, la rendre plus épaisse, modifier son apparence etc.

Les critères sont donc variés pour choisir entre une distribution statistique avec des variables discrètes ou continues. Néanmoins la nature des données reste le critère déterminant.

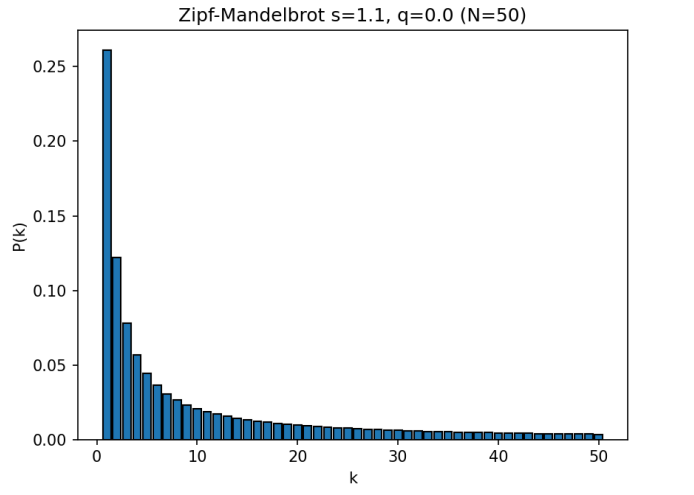
Néanmoins, s'il existe de nombreuses lois statistiques liées à la géographie, certaines sont malgré tout, plus utilisées que d'autres. Nous allons donc évoquer par la suite certaines lois particulièrement importantes en géographie.

Tout d'abord, la loi normale est peut-être celle qui revient le plus souvent. On l'utilise régulièrement pour tout ce qui concerne les phénomènes naturels comme la température, la hauteur, les précipitations, mais aussi les densités par exemple. Elle est très proche des moyennes arithmétiques et utilise l'écart type pour représenter les dispersions ou l'éparpillement des données. Elle est donc très utile pour les mesures. Statistiquement parlant, les variables sont souvent continues et symétriques.

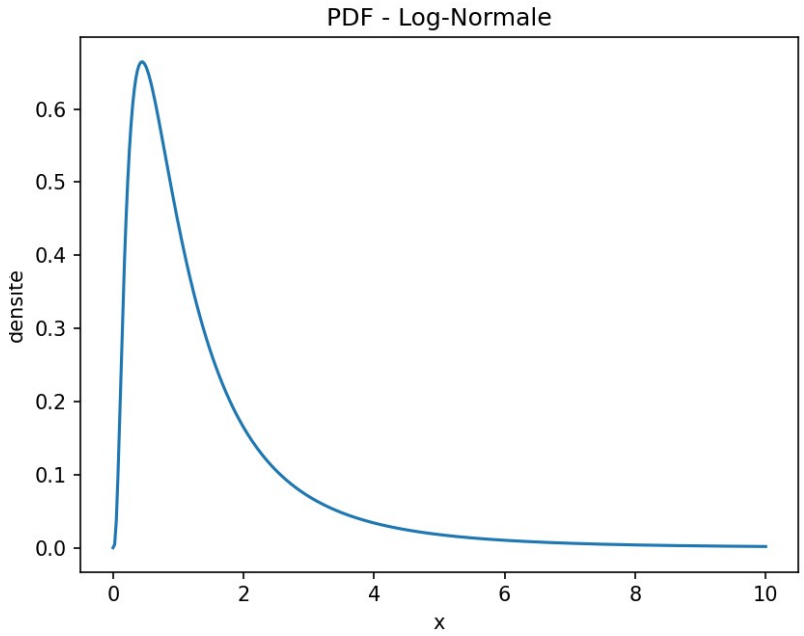
La présence récurrente de cette loi est tout à fait normale. En effet, les phénomènes géographiques sont souvent dû à une multitude de causes qui s'accumulent. Or la loi normale permet de modéliser tout cela, de « donner sens » à la moyenne, aboutissant à une forme en cloche. Elle est donc particulièrement utilisée dans ce type de cas.De plus, la loi normale est une « distribution limite à l'infini d'autres distributions statistiques ». Ainsi, comme elle est une distribution limite, beaucoup de lois deviennent normales. Lorsqu'on additionne les différents éléments qu'il s'agisse de phénomènes climatiques, de mesures légèrement différentes, ou bien d'autres cause, l'ensemble va tendre vers une représentation en cloche donc une loi normale. Cela correspond aux déclarations de C.F.Gauss. Ainsi, comme la géographie accumule beaucoup de données et tout particulièrement des variables continues (en géographie physiques notamment), la loi normale n'est que d'autant plus utilisée.



Ensuite, on utilise la loi de Zipf pour les lois rang-taille. Cette dernière est une règle de distribution rang/taille appliquée aux villes. Elles sont classées en fonction du rang de leur population. Ce classement par nombre d'habitants se veut comme une règle : la première ville est plus peuplée que la deuxième dans un rapport qui serait universel dans le temps et l'espace, tout comme le rapport entre la deuxième et la troisième. Cette règle sert surtout à faire des comparaisons entre les systèmes urbains. Cette loi dit donc que la ville classée première est beaucoup plus peuplée que la ville classée deuxième, de même pour la troisième qui est plus peuplée que la quatrième etc.
Statistiquement parlant, on a donc des variables discrètes correspondant au nombre d'habitants et, puisqu'on a des distributions asymétriques, on va utiliser une loi qui convient, adaptée aux tailles extrêmes. La loi de Zipf est donc très utile en géographie des villes, dans le cadre de la loi rang-taille.

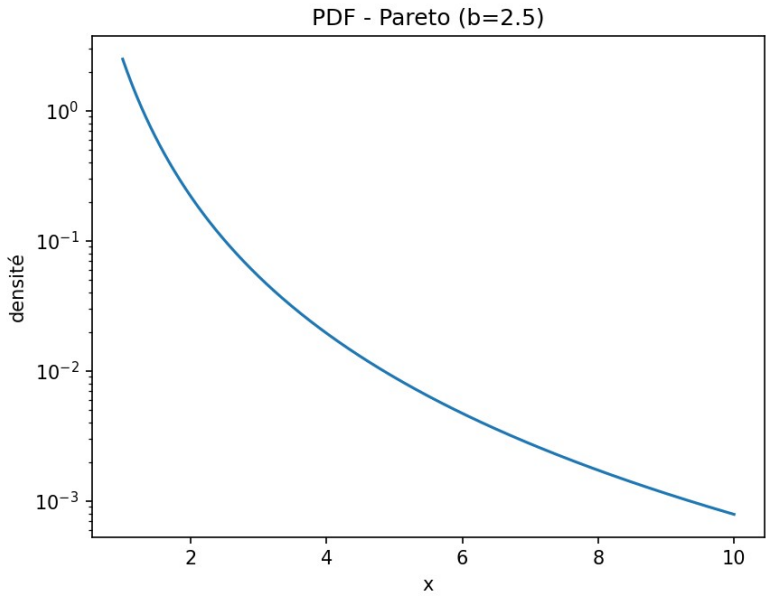


La Loi log normale est elle aussi utilisée en géographie.
Elle produit des distributions asymétriques, voire très asymétriques : on a presque un pic à gauche de la représentation graphique (voir ci dessous). Ceci est dû à la présence de nombreuses petites valeurs et quelques très grandes valeurs. En effet, la loi log normale amplifie les grandes valeurs, limite les petites et fait disparaître les valeurs négatives puisqu'elle ne comptabilise que des phénomènes positifs. Ces caractéristiques correspondent néanmoins à de nombreux cas en géographie.
Les fortes différences entre les valeurs sont récurrentes dans cette discipline, c'est un cas typique. On va donc pouvoir utiliser la loi log normale pour s'intéresser aux revenus d'un territoire par exemple, ou concernant des précipitations très différentes rapprochées dans le temps ou non.
La loi log normale peut aussi être rencontrée lorsqu'on s'intéresse à la loi rang-taille et à la hiérarchie des villes. Il est parfois difficile de la distinguer de la loi Zipf. Cependant, comme je l'ai mentionné longuement dans le paragraphe précédent, on associe surtout la loi rang-taille à Zipf.



Finalement, la loi de Pareto est utilisée en géographie au début du XXème siècle dans le cadre de travaux sur les grandeurs. On peut prendre l'exemple des travaux de Lotka (1925) et Korčák (1940). Elle est donc utilisée pour analyser certains phénomènes en géographie.
C'est une distribution « scalante », ce qui signifie qu'elle prend l'échelle de l'observateur, puisqu'elle n'a pas d'échelle propre. Donc, si l'on change d'échelle, les formes restent similaires, ce qui peut être intéressant pour analyser les reliefs ou même les villes par exemple.
Il s'agit également d'une loi qui crée des modélisations asymétriques, ce qui signifie qu'il y a encore une fois beaucoup de petites valeurs et peu de grandes valeurs. Comme dit précédemment, il est très courant d'avoir des données très irrégulières et extrêmes en géographie.
Finalement, les grandes valeurs (ou extrêmes) sont mieux présentées grâce à cette loi. Elles ont une probabilité plus élevée de se réaliser par rapport à la loi normale.

Cette loi a donc été utilisée dans le cadre de travaux sur la distribution de la taille de certains phénomènes géographiques.



Pour conclure le bilan de la séance 4, j'ai veillé à faire en sorte que mon code sorte les modélisations des différentes lois et non seulement les formules qui sont, je trouve plus abstraites. En effet, ces représentations me servent à nourrir le rapport et donnent un sentiment d'accomplissement à la fin du code. Seulement, je ne souhaitais pas non plus tous les enregistrer sur mon ordinateur (on a déjà beaucoup de graphiques en stock avec les autres codes).

Mise en œuvre avec python

→ Loi uniforme discrète

Formule : $P(X=k)=n/1, k \in \{1,\ldots,n\}$

Résultat du code : Uniforme discrète sur {1,...,6}
Moyenne théorique : 3.5000
Écart-type théorique : 1.7078
Simulation (10000 tirages) -> moyenne emp = 3.4771, std emp = 1.7148

On remarque que les valeurs empiriques sont très proches des valeurs théoriques. On peut supposer que le léger écart est du au fait que la simulation soit aléatoire ou à la taille définie de l'échantillon. La loi uniforme est bien symétrique ce que confirme la stabilité de la moyenne autour du centre.

→ Loi binomiale

Formule :
espérance : $E(X)=np$
variance : $V(X)=np(1-p)$

Résultat du code : Binomiale B(n=10, p=0.3)
  Moyenne théorique : 3.0000
  Écart-type théorique : 1.4491
  Simulation (10000) -> moyenne emp = 2.9707, std emp = 1.4595

On observe une concordance très forte entre la théorie et la simulation. L'écart sur la moyenne est normal c'est à dire inférieur à 1 %. La loi reste néanmoins légèrement asymétrique puisque p n'est pas égal à 0,5.

→ Loi de poisson

Propriété : $E(X)=V(X)=\lambda$

Résultat du code : Poisson(λ=3)
Moyenne théorique = 3.0000,
écart-type théorique = 1.7321
  Simulation (10000) -> moyenne emp = 2.9796, std emp = 1.7175

Ici la moyenne est à peu près égal à la variance ce qui est typique de la loi de poisson. Les petits écarts sont due au fait que la taille de l'échantillon soit finie, mais les résultats théoriques et empiriques sont sinon assez bon.

→ Loi de Zipf-Mandelbrot

Formule (loi de rang-fréquence) : $p(k)\propto(k+q)s/1$
La distribution est normalement très asymétrique

Résultat : Zipf-Mandelbrot (s=1.1, q=0.0, N=50)
  Moyenne (sur 1..N) = 9.7431, écart-type = 11.9110
  Simulation (10000) -> moyenne emp = 9.8663, std emp = 11.9226

L'écart type est plus grand que la moyenne, on observe donc une très forte dispersion, ce qui est typique de cette loi. En effet, il y a normalement un grand nombre de petites valeurs accompagnées de quelques très grandes. N=50 est nécessaire pour éviter que certains moments ne divergent. Contrairement aux lois précédentes, dans ce cas, la convergence est plus lente. En effet ces lois sont mal décrites par la moyenne seule. Comme le mentionne le cours, on peut utiliser des échelles logarithmiques.

→ Loi normale

Formule (loi symétrique) : $E(X)=\mu, V(X)=\sigma 2$

Résultats : Loi normale N(0,1)
 moyenne théorique = 0, écart-type = 1
  Simulation (10000) -> moyenne emp = -0.0025, std emp = 1.0081

Ici, la concordance est presque parfaite et la moyenne est très proche de 0. L'écart type est aussi proche de 1, ce qui confirme une bonne normalisation, si on peut le dire de cette manière. En tant que modèle central en statistique, la loi normale est ici bien illustrée.

→ Loi log-normale

Mentionné dans le cours : Si ln(X)~N, alors X est log-normale.

Résultats : Loi log-normale (s=0.9, scale=1.0)
moyenne théorique ≈ 1.4993, std théorique ≈ 1.6749
Simulation (10000) -> moyenne emp = 1.4815, std emp = 1.6589

On observe également ici une bonne concordance. Contrairement à la loi normale, la moyenne n'est pas égale à la médiane qui n'est pas égale au mode. On observe donc une forte dissymétrie.

→ Loi uniforme continue

Formule : $E(X)=(a+b)/2$, $V(X)=(b-a)2/12$

Résultats : Uniforme continue U(0,1)
moyenne = 0.5000, std = 0.2887
Simulation (10000) -> moyenne emp = 0.5027, std emp = 0.2882
La moyenne est très proche de la valeur théorique. L'écart-type empirique correspond presque parfaitement à la valeur analytique. Cela confirme la validité de la simulation.

→ Loi du Chi²

Résultats : Chi2(df=4)
moyenne = 4.0000, std = 2.8284
Simulation (10000) -> moyenne emp = 3.9985, std emp = 2.8388

Les valeurs empiriques sont toujours très proches des valeurs théoriques. L'écart-type empirique est légèrement supérieur ce qui est un effet normal de la simulation finie et liée au fait que la variance élevée est logique étant donné l'asymétrie de la loi.
La loi Chi2 est toujours positive et fortement asymétrique pour les faibles degrés de liberté. Lorsque k augmente, la loi devient plus symétrique et elle tend vers une loi normale.

→ Loi de Pareto

Résultats : Pareto(b=2.5, scale=1.0)
moyenne théorique = 1.6667, std théorique = 1.4907
Simulation (10000) -> moyenne emp = 1.6736, std emp = 1.2362

Encore une fois, la moyenne empiriques est proche de la valeur théorique. C'est cohérent car b=2.5>1. L'écart type empirique est en revanche beaucoup plus instable : il est très sensible aux valeurs rares mais très grandes.

Séance 5
Les statistiques inférentielles

L'échantillonnage peut se définir comme une méthode consistant dans le fait de prélever au hasard un petit groupe au sein d'une population délimitée. Cela permet d'étudier un petit nombre d'individus pour en déduire des informations sur toute la population mère. On part du principe que le groupe est représentatif de l'ensemble. Chose inhabituelle en mathématiques : on étudie presque toujours un ensemble duquel on déduit des informations afin de déterminer des ensembles plus petits. Le fait de choisir un petit groupe à l'ensemble de la population s'explique car il est impossible d'étudier une population entière lorsqu'elle est trop grande. Donc, on utilise surtout l'échantillonnage lorsqu'il est impossible d'avoir l'information totale sur une population (« l'ensemble des poissons contenus dans l'océan » par exemple). Le prix d'une telle étude serait également trop élevé et le tout serait inutilisable pour certaines enquêtes : d'opinions, de comportements… Le mieux est de prélever plusieurs échantillons, la différence entre chaque résultat de chacun des échantillons est appelée la fluctuation de l'échantillonnage.

On distingue deux types de méthodes : les échantillons non biaisés, équiprobables, et les échantillons biaisés, non équiprobables. L'échantillon non biaisé est celui qui a été prélevé au hasard, chaque individu a autant de chance que les autres d'être tiré au sort. L'échantillon biaisé correspond à un individu qui a été sélectionné, qui n'est pas pris au hasard. La plupart des méthodes requièrent des enquêtes, c'est-à-dire questionner ou observer les individus de la sous population isolée. Parmi les méthodes aléatoires, on peut citer le tirage avec remise. Cela consiste à tirer un individu au sort, à noter son numéro, et à le remettre dans la population. C'est une méthode très simple, les calculs qu'on en tirera seront simplifiés, mais un même individu peut être tiré au sort à nouveau son numéro apparaître plusieurs fois… C'est un modèle irréaliste, qui trompe la réalité. Il est surtout utilisé en simulation. Il existe ensuite le tirage sans remise : le principe est le même, mais une fois l'individu tiré, il est retiré de la liste et son numéro est barré. C'est une méthode plus représentative, où il n'existe aucun doublon. Cependant, on introduit des calculs plus complexes où intervient le taux de sondage n/N. Il est utilisé pour les sondages, les enquêtes de terrain, l'échantillonnage classique.

Ensuite, nous pouvons énumérer les méthodes non aléatoires, ou biaisées. L'échantillonnage systématique consiste à choisir la taille n de l'échantillon, calculer le pas de sondage (k = N/n), choisir le numéro de départ, puis les individus. On va prendre régulièrement une personne, toutes les k personnes : il s'agit d'une sorte de règle où on choisit l'individu placé à tel intervalle, et on ignore tous les autres. Par exemple, sur 100 personnes, on déciderait de prendre la 10 ème , la 20 ème , la 30 ème , la 40 ème , etc. N est la population totale, n est le nombre d'individus que je veux dans mon échantillon. Le pas de sondage permet de déterminer où fixer l'intervalle, en fonction du nombre d'individus total et le nombre que l'on veut échantillonner. Le point de départ est aléatoire, mais il est situé entre 1 et k. Puis, on lui rajoute k. Grossièrement, si on continue notre intervalle tous les 10, on peut choisir entre 1 et 10, prenons 4, auquel on rajoute k (= 10). Puis on poursuit, donc on sélectionne la personne 14, puis 24, puis

34, etc. Cette méthode couvre toute la population de façon équilibrée, mais elle n'est pas représentative si la population a une structure répétitive. Il existe aussi la méthode des quotas : on impose des proportions à respecter, et l'échantillon doit avoir les mêmes proportions que la population mère sur certaines variables, comme l'âge, le sexe ou la profession. Par exemple, une population avec 40% de personnes de plus de soixante ans, et 60% de personnes de moins de 60 ans, l'échantillon doit représenter les mêmes proportions. On utilise cette méthode pour les populations trop grandes, lorsqu'une base de sondage est incomplète, ou lorsqu'on veut un échantillon qui représente la population sans avoir à faire de tirage au sort, simple et rapide à réaliser, sans un coût trop élevé. Cette méthode est très utilisée dans les sondages d'opinion. Malheureusement, même si les quotas sont respectés, l'échantillon peut être influencé par le lieu d'enquête, l'enquêteur, et la représentativité de la population n'est pas garantie, du fait des variables qui n'ont pas été prises en compte.

Enfin, il existe la méthode de Monte-Carlo, une méthode de simulation utilisée pour reproduire (virtuellement) des phénomènes aléatoires, que l'échantillonnage direct est irréalisable parce que la population est trop grande ou inconnue, et pour observer, grâce à de nombreux tirages aléatoires, comment varie une statistique. Cela permet de créer artificiellement des tirages et de voir la dispersion des résultats. On choisit une distribution réelle ou inventée (par exemple, le nombre d'absences de 100 employés), on simule des tirages aléatoires, on transforme ces tirages en observations, et on répète le tout plusieurs fois. On peut étudier la distribution des résultats, les moyennes, et en déduire une estimation du paramètre étudié au départ, noté μ. Afin de choisir l'une de ces méthodes, l'estimateur fait face à plusieurs facteurs déterminants. Ces facteurs l'amèneront à sélectionner telle ou telle méthode, biaisée ou non biaisée, que ce soit l'existence d'une base de sondage, le degré de représentativité, le coût et la simplicité, la taille de la population, et surtout l'objectif de l'étude.

La théorie de l'estimation pose une question : comment estimer un paramètre inconnu de la population à partir d'un seul échantillon. La théorie de l'estimation mobilise des outils, nommés estimateurs ; un estimateur est une fonction, une variable aléatoire Yn dépendant des Xi (les données de l'échantillon). Il est donc conçu à partir de l'échantillon, et doit pouvoir se rapprocher du paramètre inconnu θ de la population mère. L'échantillon change à chaque prélèvement, et par conséquent l'estimateur aussi (la moyenne des tailles de dix personnes ne sera jamais la même d'un tirage au sort à l'autre). Ainsi, ce sont sa variance et son biais qui sont étudiés. L'estimateur est une fonction qui s'écrit ainsi : , où y est la fonction et Xn le nombre d'observations de l'échantillon. L'estimateur est donc, avant tout, une formule mathématique. On peut en citer plusieurs comme la moyenne empirique, notée : , où on estime μ (la moyenne réelle), ou la proportion observée, notée : , où on estime p (la proportion réelle).
Il faut distinguer l'estimateur Yn de l'estimation, yn, qui est une valeur obtenue à partir des données réelles. C'est le résultat de l'estimateur, de la formule, à partir des données. On remplace simplement les Xi par les valeurs observées dans l'échantillon. L'estimateur donne une valeur concrète, qui est donc l'approximation du paramètre θ. L'estimateur existe donc avant de connaître les données et l'estimation après les avoir observées.

Désormais, nous allons nous intéresser à l'intervalle de fluctuation et l'intervalle de confiance qu'il s'agira de distinguer.

L’intervalle de fluctuation est utile lorsque la proportion réelle p est connue. C’est un intervalle dans lequel la fréquence f doit se trouver, si l’échantillon est compatible avec la proportion réelle de la population. Il indique les valeurs possibles que peut prendre la fréquence f. Il sert à vérifier si un échantillon est cohérent avec la population mère, avec un risque d’erreur a %. On utilise alors la formule : . C’est un outil d’échantillonnage, pas d’estimation.
L’intervalle de confiance, en revanche est un outil d’estimation. Il est utilisé lorsque la proportion réelle est inconnue. On essaie alors de l’estimer à partir de l’échantillon, grâce à la formule : . Le cours énonce clairement : « l’intervalle de confiance doit avoir de grandes chances de contenir la vraie valeur du paramètre. » On ne teste plus la fréquence, on veut donner un cadre à la proportion réelle, en se demandant quelle fourchette de valeurs est le plus cohérent pour le paramètre recherché (la proportion réelle par exemple). C’est une méthode d’estimation pour le paramètre inconnu. Donc, le premier sert à décider si l’échantillon est vraisemblable et cohérent, et le second à estimer un paramètre.

Le cours nous permet de conclure qu’il existe un biais dans la théorie d’estimation. Ce biais est aussi appelé « erreur d’estimation ». Il est défini comme la différence entre le point où l’estimateur estime la moyenne et la vraie valeur du paramètre θ. Autrement dit : si les deux valeurs coïncident, il n’y a pas de biais. Cela signifie que l’estimateur vise souvent juste, et qu’il donne, en moyenne, la bonne valeur.
Dans le cours, la moyenne est définie comme un estimateur sans biais : $E(\hat{\mu}) - \mu = 0$. Dès lors qu’elles ne coïncident pas, l’estimateur est dit biaisé : il existe un biais. Un estimateur biaisé va systématiquement donner des valeurs erronées, soit trop grandes soit trop petites, qui s’écartent systématiquement de θ. Il existe aussi ce qu’on appelle un estimateur asymptotiquement sans biais, noté . Autrement dit, il peut être biaisé pour des échantillons de petite taille, mais l’« erreur d’estimation » s’efface dès que l’échantillon devient plus grand.

On appelle recensement, une statistique travaillant sur la population totale. Il consiste à observer tous les individus de la population, sans échantillonnage. Ainsi, c’est une « enquête exhaustive » qui englobe toute la population mère, sans tirage au sort, ni estimation. Mais il faut noter qu’il est rare, car quasi impossible, d’obtenir l’information totale d’une population mère car cela coûte trop de temps et d’argent, et qu’une telle population représente un trop grand nombre d’individus. Le recensement est peu usité, on lui préfère, comme vu plus haut, les sondages ou les échantillonnages. Pour parvenir à une information totale, ou presque, on introduit la notion de données massives, ou big data. Les données massives permettent d’obtenir de milliers, des centaines de milliers voire des millions d’informations, en provenance directe de la population dans son ensemble (on peut utiliser les bases administratives). Les données massives peuvent servir à sur la population quasi-complète plutôt que sur des échantillons représentatifs. Les données massives sont des quasi-recensements. C’est un traitement quasi-exhaustif de la population qui remplace les notions indirectes et représentatives par une méthode plus directe.

L’estimateur est très important, son choix soulève différents enjeux.
En effet, les estimateurs concentrent plusieurs facteurs importants à prendre en compte : leur biais,leur variance, leur ERQM (l’erreur quadratique moyenne, ou l’erreur moyenne au carré entre l’estimateur et la vraie valeur de l’estimation yn), la consistance, c’est-à-dire le fait qu’ils soient de plus en plus précis si la taille de l’échantillon augmente, la convergence, qui

correspond au fait que, lorsque l'échantillon grandit, la probabilité que l'estimateur s'éloigne du paramètre recherché devient nulle, l'efficacité (le fait d'avoir un faible biais et une faible variance), et la robustesse, c'est-à-dire le fait que l'estimateur ne change pas ou peu lorsqu'une valeur est extrême, lorsqu'il y a une erreur, c'est le fait d'être peu sensible aux dérèglements dans les données. Le premier enjeu dans le choix d'un estimateur est d'éviter le biais dont nous avons parlé précédemment : l'erreur systématique introduit nécessairement une mauvaise estimation, et va orienter les recherches et les données dans un mauvais sens. Il faut privilégier un estimateur sans biais.
Ensuite, il faut limiter la dispersion. Un estimateur sans biais peut toujours être impertinent si, d'un échantillon à l'autre, les résultats varient trop. La dispersion doit être la plus faible possible, l'estimateur doit être stable. Il faut un estimateur précis, ce qui peut être mesuré par la formule suivante : . Plus l'ERQM est bas, plus l'estimateur est précis (car le biais et la variance sont faibles). De même, il est nécessaire de vérifier la consistance de l'estimateur : « l'erreur quadratique moyenne doit tendre vers 0 », il faut être certain que plus l'échantillon augmente, plus l'estimateur devient exact. Autrement dit, comme il est écrit dans le cours : « si la distribution se concentre autour de la valeur à estimer quand n est ∞. » Si l'estimateur reste imprécis alors que l'échantillon devient plus grand, l'estimateur est à abandonner. Cela rejoint la convergence, l'estimateur doit converger vers la vraie valeur au fur et à mesure que grandit l'échantillon.
Enfin, le dernier enjeu du choix de l'estimateur est celui de choisir l'estimateur le plus efficace : il faut pouvoir comparer plusieurs estimateurs et déterminer celui qui est le plus rapide, le plus précis et qui prend en compte les plus grandes plages de données. L'information de Fisher indique la quantité d'information disponible à partir d'un échantillon, une donnée essentielle afin de choisir un estimateur capable de les prendre toutes en compte.

Tout ceci étant dit, comment sélectionner la méthode d'estimation de paramètre qui convient ? Tout d'abord, il existe trois familles de méthodes d'estimation d'un paramètre inconnu : la méthode des moments, le principe de vraisemblance et un rééchantillonnage. La première est celle des moindres carrés, elle repose sur le modèle $y_i = \mu_i + \varepsilon_i$. μ_i dépend des paramètres à estimer, ε_i est une erreur aléatoire qui permet de vérifier le bon fonctionnement de l'estimateur, en vérifiant : $E(\varepsilon_i) = 0$, $V(\varepsilon_i) = 0$, $cov(\varepsilon_i, \varepsilon_j) = 0$ si $i \neq j$. Cette méthode générale, et puissante, repose sur le fait d'utiliser un modèle probabiliste connu, qui s'ajuste aux données, surtout quand les quantités à estimer sont des espérances. Cette méthode permet de trouver les paramètres qui minimisent les erreurs. On peut l'utiliser lorsque plusieurs variables aléatoires sont présentées.
Ensuite, il existe la méthode du maximum de vraisemblance. Elle consiste à prendre, la valeur, parmi toutes les valeurs possibles de θ, celle pour laquelle l'échantillon avait le plus de probabilité d'apparaître. La vraisemblance est définie à partir de la densité de l'échantillon $L(x, \theta) = \prod f(x_i , \theta)$. Cela permet de chercher la valeur θ la plus plausible. Formulé autrement, on maximise $V(\theta)$ car cela permet de rendre les données observées les plus plausibles possibles.
Enfin, il existe une troisième méthode, dite de rééchantillonnage, ou méthode du bootstrap. On tire, avec remise, au sein de l'échantillon initial, pour créer de nouveaux échantillons. Chacun donne une estimation, et on répète le processus. Cela permet de donner un grand nombre d'échantillons, et d'obtenir une distribution empirique de l'estimateur. On l'utilise lorsqu'il n'y a aucune formule ou aucun modèle utilisable.
Le choix d'une méthode d'estimation d'un paramètre repose sur plusieurs facteurs : le type de modèle statistique utilisé. Par exemple, la méthode du maximum de vraisemblance est utilisée

pour un modèle probabiliste très spécifique, car elle donne des estimateurs sans biais et une variance minimale. La nature de la relation entre les variables intervient aussi. De même lorsqu'il n'y a pas de possibilité d'utiliser une formule, on utilise le rééchantillonnage. Les propriétés recherchées par l'estimateur décident de la méthode à choisir, si l'estimateur doit être sans biais, la méthode du maximum de vraisemblance est préférable, de même s'il faut une variance faible. Bien sûr la nature des données influence le choix de la méthode.

Un test statistique permet de décider si les données observées répondent à l'hypothèse que l'on fait sur une population. Le test mobilise H0 (l'hypothèse qu'on va tester) et H1 (l'hypothèse alternative si H0 est réfutée). Les tests statistiques comparent le modèle théorique aux données (le test du χ^2 d'ajustement), deux ou plusieurs populations (tests de Student, Fisher, Mann-Whitney, Wilcoxon…), vérifient le lien ou l'indépendance entre deux variables (les tests non paramétriques), et enfin vérifient une valeur d'un paramètre dont on n'est pas sûr (les tests paramétriques). Il existe quatre principales catégories de tests : les tests de signification qui servent à mesurer si les données sont compatibles avec H0, les tests paramétriques (sur la moyenne, sur la variance, de comparaison entre deux ou plusieurs moyennes), les tests non paramétriques, et les tests d'ajustement (pour comparer une distribution empirique et une loi théorique).
Afin de construire un test statistique, le cours énumère 11 étapes : tout part d'une simple question, à laquelle on répond par une hypothèse H0, on formule une hypothèse alternative H1, on trouve la loi statistique qui peut répondre à H0, on choisit le seuil α, c'est-à-dire la probabilité de rejeter l'hypothèse alors qu'elle est vraie. On le définit à 0,01, soit 1%, qui implique un test strict, ou à 0,05, soit 5% de risque, qui implique un test plus permissif. Parfois, il peut s'élever à 10% pour des tests exploratoires. Ensuite, il faut vérifier les conditions d'application, par exemple si la variance est connue ou inconnue, quelle est la taille de l'échantillon, etc. Puis, on prélève l'échantillon représentatif. On choisit le test, on détermine la région où on rejette H0 et celle où le risque d'erreur, le seuil α, est défini. On finit par calculer la valeur observée de la statistique et enfin on compara la statistique à la zone critique, soit la valeur au seuil.

La statistique inférentielle regroupe plusieurs outils qui permettent de déduire des conclusions à partir d'échantillons d'une population. Elle soulève plusieurs critiques, dont les limites que nous avons pu parcourir au fil du cours. On peut déduire de nombreuses critiques implicites, comme la dépendance de la statistique inférentielle au choix de l'échantillon, au risque de biais et à la fluctuation d'échantillonnage. Les tests peuvent produire deux types d'erreur : réfuter H0 alors qu'elle est vraie, et l'infirmer alors qu'elle est fausse. Mais le risque d'erreur, bien que central, n'est pas forcément bien compris car un test ne donne jamais la probabilité que l'hypothèse soit vraie. Cela reste une probabilité. On peut aussi critiquer la dépendance au seuil de risque α, qu'on peut juger arbitraire et impliquer des décisions et des cadres artificiels. Mais le cours rappelle que c'est un seuil utilisé de manière mécanique, et fixé à 5% car plus raisonnable, servant simplement à définir la région critique. Le tout dépend de la taille de l'échantillon, car même si n augmente, β diminue. On peut reprocher aux modèles de ne pas correspondre parfaitement au réel, de l'idéaliser et de ne pas prendre en compte toutes les nuances et toutes les variables. Les distributions peuvent être biaisées, dissymétriques et surtout issues de processus réels qui ne s'inscrivent pas dans les lois statistiques classiques. La

nécessité d’ajustement graphique, de tests d’adéquation, de tests robustes vient corriger ces lacunes.
Les autres critiques sont celles de la sensibilité aux valeurs extrêmes qui nécessitent des estimateurs résistants, le fait de se reposer sur un modèle probabiliste qui peut être mal configuré, et enfin les tests multiples nécessaires, mais qui peuvent tout de même conduire à une mauvaise interprétation. La statistique inférentielle se caractérise par une fragilité et une rigueur obligatoire, ce qui n’empêche par l’erreur pour autant. Enfin et surtout, elle indique seulement le risque minimal, sans jamais diriger vers une décision claire. Elle sert d’outil dans que vient compléter les contextes économiques, politiques, et les enjeux sociaux.
La statistique inférentielle reste un outil indispensable pour mesurer l’inconnu et l’incertitude à partir d’échantillons, afin de présenter un cadre rationnel validé mathématiquement pour prendre des décisions. Il faut simplement avoir conscience qu’elle repose sur des hypothèses, des seuils arbitraires et que chaque résultat dépend d’une rigueur mathématique permanente à laquelle chaque être humain peut faillir. Ce sont des chiffres utiles auxquels il faut prêter un but : donner une vue d’ensemble.

Mise en œuvre avec python

Nous allons désormais nous attarder sur les résultats afin d’en produire une analyse au regard du cours.

1) Théorie de l’échantillonnage

Résultats :
Moyennes observées sur 100 échantillons :
Pour 391.0
Contre 416.0
Sans opinion 193.0
dtype: float64

Fréquences observées (moyennes) :
{'Pour': np.float64(0.39), 'Contre': np.float64(0.42), 'Sans opinion': np.float64(0.19)}

Fréquences de la population réelle :
{'Pour': np.float64(0.68), 'Contre': np.float64(0.32), 'Sans opinion': np.float64(0.0)}

Concernant les résultats de la théorie d’échantillonnage, lorsque l’on compare les fréquences moyennes observées sur les 100 échantillons avec les fréquences réelles dans la population mère, on remarque que les résultats des échantillons sont assez différents des proportions mises en avant dans la population réelle. En effet, tandis que l’on obtient 0,39 pour sur l’échantillon, nous avons 0,68 pour les fréquences réelles. On retrouve également une différence concernant les résultats contre : 0,42 pour l’échantillon et 0,32 pour l’échantillon mère. Cette différence est moins importante que la première mais reste notable.

Cela correspond à ce que le cours évoquait concernant les fluctuations d'échantillonnage. En effet, les échantillons présentent des fluctuations importantes par rapport aux vraies proportions. Ainsi, les valeurs issues des échantillons ne reproduisent pas forcément les proportions de la population mère. Ceci est tout à fait normal dans un échantillonnage aléatoire, d'autant plus si les échantillons ne sont pas représentatifs, ou s'il ne suivent pas la distribution théorique.

Résultats :

Intervalles de fluctuation (95%) :
Pour : [0.36 ; 0.42]
Contre : [0.389 ; 0.451]
Sans opinion : [0.166 ; 0.214]

Les intervalles de fluctuation nous permettent de conclure que les échantillons simulés ne sont pas cohérents avec la population mère réelle. On pouvait déjà le constater en comparant les fréquences observées aux fréquences réelles. En effet, normalement, si la fréquence réelle se situe en dehors de l'intervalle de de fluctuation, alors les échantillons ne sont soit pas représentatifs, soit ne proviennent pas de la bonne population mère, soit les tirages n'ont pas les mêmes paramètres que ceux énoncés

2) Théorie de l'estimation

Résultats :

Fréquences du premier échantillon :
{'Pour': 0.395, 'Contre': 0.396, 'Sans opinion': 0.209}

Intervalles de confiance (95%) :
Pour : [0.365 ; 0.425]
Contre : [0.366 ; 0.426]
Sans opinion : [0.184 ; 0.234]

La construction de l'intervalle de confiance se fait à partir d'un échantillon unique. C'est la différence avec l'intervalle de fluctuation qui nécessite un paramètre qui soit connu. Dans ce cas ci, l'estimateur est utilisé sans biais et convergent. Normalement les proportions du premier échantillon sont de bonnes estimations, mais il reste forcément beaucoup de doute et d'incertitude. Les intervalles de confiance obtenues sont logiques avec les moyennes des 100 échantillons, ils recouvrent presque les mêmes valeurs

L'intervalle de confiance révèle ici le doute ou l'incertitude de l'échantillon choisi. Il ne contient jamais vraiment les proportions réelles de la population mère. Cela vient encore confirmer que l'échantillon n'est pas représentatif de la population, et qu'une estimation sur cette base est biaisée, sûrement par un mauvais tirage initial.

3) Théorie de la décision

Test 1 :
W = 0.9639, p-value = 0.0000
Pas normal

Test 2 :
W = 0.2609, p-value = 0.0000
Pas normal

Le test de Shapiro-Wilk sert à vérifier que la distribution suit une loi normale. Ici les deux séries de données semblent incompatibles avec une loi normale selon ce test. La statistique W est éloigné de 1 pour le test 2. C'est un signe d'une distribution très peu normale. Ni la loi normale, ni ses propriétés ne sont compatibles avec les fichiers testés. Ici on ne peut donc pas utiliser de méthodes reposant sur la loi normale

Séance 6
La statistique d’ordre des variables qualitatives

Une statistique ordinale, ou statistique d’ordre, repose sur le classement d’observations en rangs. Elle sert à montrer quel individu ou entité monte, stagne ou descend dans le classement. L’idée est que l’on peut mettre en ordre une plage d’observations du plus petit au plus grand, ou inversement (mais on privilégie l’ordre croissant). Plus simplement, cela consiste à traduire des valeurs en rangs, méthode appelée statistique d’ordre associée à la série d’observations.
Elle s’oppose aux statistiques catégorielles sans ordre naturel, les variables qualitatives nominales. Les variables nominales, comme les couleurs, certaines catégories administratives, n’ont pas de classement. Elles ne permettent pas d’ordre. Les statistiques ordinales s’appliquent lorsque apparaissent des hiérarchies, naturellement.
La statistique ordinale utilise des variables qualitatives ordinales (des variables aux catégories ordonnées). Elles permettent de travailler sur les rangs (les rangs sont des nombres entiers issus du classement), la comparaison entre les classements, les valeurs aberrantes, l’ordre lui- même. Cela peut matérialiser une hiérarchie spatiale car elle révèle ou donne forme à une hiérarchie préexistante entre des objets géographie : « la statistique d’ordre est le cœur de la géographie humaine » après tout ; on peut classer des villes par population et révéler une hiérarchie urbaine, classer par richesse et révéler une hiérarchie économique, classer par zones sismiques en fonction de la magnitude, soumises aux tsunamis, aux cyclones ou aux crues pour révéler une hiérarchie des risques. On souligne alors l’importance et la domination de certains espaces sur d’autres. Cette statistique est donc un outil qui rend compte, analyse et compare les hiérarchies spatiales.

Le cours énonce clairement que l’ordre à privilégier dans les classifications est l’ordre croissant, donc du plus petit au plus grand, du moins au plus important, du plus faible au plus fort ; cet ordre est le référent universel de la statistique ordinale puisqu’il suit l’ordre naturel de la progression des valeurs. On préfère utiliser l’ordre croissant car la statistique d’ordre est fondée sur l’idée que les nombres se suivent, c’est une suite de nombre : $X(1) \leq X(2) \leq ... \leq X(n)$; donc, comme n’importe quelle suite, on part de 1 et on continue à l’infini vers les valeurs plus élevées. De même, cet ordre facilite l’analyse des étendues et des distributions (et donc des anomalies) parce qu’on peut étudier les valeurs extrêmes, qui sont exposées par cet ordre naturel : X(1) est la plus petite valeur et X(n) la plus grande. Les fonctions de répartition (Weibull, Hazen, Chegodaev, Tukey) sont construites autour de cet ordre également. On peut signaler cependant l’exception de la loi rang-taille, qui va à l’encontre de l’ordre croissant. On classe les valeurs en ordre décroissant, la plus grande valeur est au rang 1 et la plus petite le dernier rang. On utilise ce classement en géographie pour étudier les hiérarchies des villes (classées par population par exemple), ou des surfaces, ou des tailles d’îles…

Le cours s'assure que nous puissions bien faire la distinction entre une corrélation des rangs et une concordance de classements.
La corrélation des rangs mesure le lien statistique entre deux classements grâce à deux méthodes. On peut utiliser d'abord le coefficient de Spearman (rs) : on mesure la corrélation entre deux séries de rang A et B. Le coefficient vérifie si les classements sont identiques ou non, en se reposant sur les écarts entre les rangs, selon la formule suivante : $rs = 1 - 6 / [n(n^2 - 1)] \times \Sigma(ui - vi)^2$. Selon les résultats, on peut distinguer si les classements concordent ou non. Si rs est égal à 1, les classements sont identiques, si rs est égal à -1, les classements sont strictement inverses, si rs est égal à 0, ils n'ont rien à voir l'un avec l'autre. Ce premier coefficient permet de mesurer une relation globale entre deux classements. Le deuxième coefficient, celui de Kendall, noté τ, compare des paires d'objets. On observe si les classements ordonnent les paires d'objets dans le même sens, on note Na pour des paires concordantes et Nd pour des paires discordantes. On exécute la formule $\tau = 2Sc / [n(n - 1)]$, avec $Sc = Na – Nd$: si τ est égal à 1, la concordance est parfaite, si c'est égal à -1, ils sont strictement inverses, et si c'est égal à 0, les deux classements sont indépendants. La corrélation des rangs est donc une mesure d'un degré de relations entre deux classements et mesure à quel point ils sont similaires et vont dans la même direction. Elle crée deux coefficients statistiques, rs ou τ, et s'intéresse à la continuité de la relation à travers une analyse globale des rangs.
La concordance des classements analyse paire par paire. Une paire d'objets est concordante si les classements vont dans le même sens, et discordante s'ils vont dans le sens opposé. Si l'ordre est respecté, on note +1, sinon on note -1. On dit que la concordante est totale lorsque $ST = n(n-1)/2$. On mesure le nombre de paires qui sont ordonnées de l'exacte même manière, c'est un comptage : dans le tableau du cours, on note qu'il y a 61 concordances pour 5 discordances, donc le coefficient $\tau = 56 / 66 \approx 0{,}85$.
En somme, la corrélation mesure à quel point deux classements varient dans le même sens, et la concordance mesure si les classements ont le même ordre, paire après paire.

Les deux tests de Spearman et Kendall visent à comparer deux classements pour établir un coefficient. Cependant, le rs transforme les données en rangs et calcule le coefficient de corrélation entre ces rangs. Encore une fois, si rs est égal à 1, les classements sont identiques, si rs est égal à -1, les classements sont strictement inverses, si rs est égal à 0, ils n'ont rien à voir l'un avec l'autre. En principe, il ne peut y avoir d'ex aequo, et dans ce dernier cas il existe des formules de correction. Enfin, si n est supérieur à 30, on peut utiliser une loi normale.
Le coefficient de Kendall compare chaque paire d'objet, et toutes les paires, et vérifie pour chacune si l'ordre est similaire ou strictement inverse. On décompte donc le nombre de paires concordantes et discordantes. La formule du coefficient va de 1 à -1. On peut faire une loi normale à partir du moment où n est supérieur ou égal à 8. Cette dernière méthode est facile à appliquer à plusieurs classements en même temps, et facilite donc la comparaison.
Le choix entre ces deux méthodes dépend soit de la taille de l'échantillon, de la volonté d'avoir une mesure construite à partir d'écarts quadratiques ou à partir de paires concordantes et discordantes.

Finalement, nous allons nous attarder sur les coefficients de Goodman-Kursdal et de Yule.
Le coefficient de Goodman-Kursdal permet de mesurer l'association entre deux classements. Il repose sur la différence entre les paires concordantes notées Na et les paires discordantes notées Nd. Il va varier entre -1 et 1 : s'il est égal à 0, cela signifie qu'il y a une association, s'il

s’approche de +1 alors les classements sont concordants, au contraire s’il s’approche de -1 alors les classements sont discordants. Il permet d’évaluer l’importance et le sens de l’association entre deux variables qui sont ordinales, en mesurant le rapport de concordance ou discordance. Le coefficient de Yule doit être compris comme un cas particulier du coefficient de Goodman-kursdal. Il permet de mesurer l’association positive entre deux variables, l’absence d’association ou l’association négative. Ce coefficient s’intéresse à un cas plus particulier : c’est une mesure d’association ou l’on considère que deux modalités pour chaque variable. Q (coefficient de Yule) va donc indiquer si les évènements sont positivement associés, indépendants ou négativement associés.

Mise en œuvre avec python

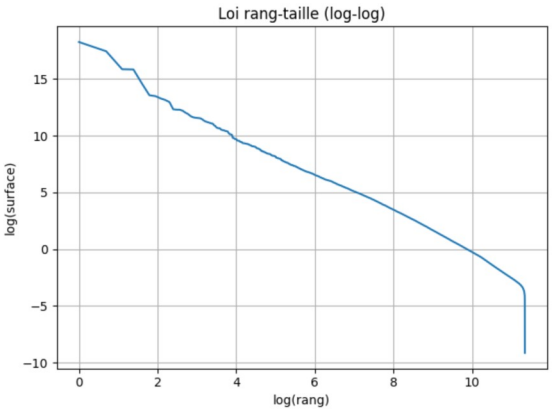
Nous allons encore une fois revenir étape par étape sur les résultats du code

Comme pour les fois précédentes nous utilisons la fonction *OuvrirUnfichier()*. L’objectif est ici de pouvoir avoir accès au fichier CSV et de renvoyer un Dataframe.

Puis l’exercice nous demande d’isoler la colonne « surface ». Cela se traduit par : *surface = list(iles["Surface (km²)"])* dans le code. On est en contact ici avec une variable quantitative que l’on va changer en valeur ordinale : le chapitre 5 traite en effet de l’ordination d’une variable quantitative pour créer une statistique d’ordre.

On va ajouter quatre surfaces continentales converties en float : le but est de rendre la série d’observation plus riche avant l’ordination. Cela correspond à l’augmentation de l’échantillon n qui modifie les statistiques d’ordre et la fonction de répartition empirique.

Puis on ordonne la liste obtenue avec la fonction locale *ordreDecroissant()*. En effet, normalement l’ordre naturel est croissant, mais comme nous l’avons vu précédemment, la loi rang-taille, quant à elle, nécessite un tri décroissant. Le rang 1 correspond alors à la plus grande valeur.



On crée un graphique pour visualiser la loi rang-taille. Le résultat obtenu est une courbe fortement décroissante, avec les grandes surfaces qui sont largement présentes. Justement, la loi rang-taille a une distribution asymétrique avec un petit nombre de valeurs très grandes et une majorité de petites valeurs. Le graphique illustre donc bien ces caractéristiques.

Pour plus de lisibilité, l’exercice nous demande d’utiliser la fonction *conversionLog()*. L’objectif est de transformer la courbe pour la lecture.

On va ensuite s’intéresser aux populations et densité.

En utilisant les données du fichier Le-Monde-HS-Etats-du-monde-2007-2025.csv, que l'on va ouvrir de la même manière qu'avant, on va commencer par isoler les colonnes « État », « Pop 2007 », « Pop 2025 », « Densité 2007 » et « Densité 2025 ». J'utilise la fonction *find_column()*. C'est à peu près de même processus que précédemment. Les statistiques d'ordre nécessitent un ordre comme leur nom l'indique, une sorte de classement préalable. L'objectif est de préparer le terrain pour obtenir des classements pour chaque année et pour chaque variable.

On utilise ensuite la fonction *Ordrepopulation()* pour trier les valeurs, associer les États et renvoyer une liste. Bref, il s'agit de la construction des classements ordonnés en utilisant l'ordre décroissant.

Pour effectuer la comparaison demandée, on va utiliser la fonction classementPays() afin d' aligner les états présents dans les années comparées (donc 2007-2025) et renvoyer une autre liste. On met ici en place les couples de rangs u et v pour calculer Spearman et Kendall que l'on a expliqué précédemment dans les questions de cours.

On va extraire les colonnes : il s'agit de la préparation des deux vecteurs mentionnés en lien avec Kendall et Spearman. On va par la suite utiliser s*pearmanr() et kendalltau()* pour faire les calculs qui correspondent.
Le code nous transmet un coefficient et une sorte de test d'hypothèse, ce qui correspond aux objectifs de Spearman et de kendall.

Voici les résultats obtenus pour la population :
Correlation Population 2007 vs 2025
Spearman : coef = 0.9863, p-value = 8.19e-136
Kendall : tau = 0.9052, p-value = 5.91e-70

Comme 0,9863 est très proche de 1, nous pouvons conclure que le classement mondial de la population des États en 2025 est presque identique à celui de 2007. Cela illustre une stabilité dans le classement avec les plus grands pays qui restent en haut, et les plus petits en bas. Si p s'était approché de -1, alors les classements auraient été inverses.
Ensuite, le test de Spearman permet de savoir si la corrélation est due au hasard. Ici le résultat obtenu grâce au code, ce que l'on nomme la p-value, est presque nul. Cela signifie que le classement de 2025 n'est pas hasardeux et est bien corrélé à celui de 2007.
Concernant l'application de Kendall en code, la valeur 0,9052 est assez haute et s'approche fortement de 1, ce qui laisse penser qu'il y aurait une immense majorité de couples concordants et à l'inverse très peu de paires discordantes entre 2007 et 2025. Les États conservent donc à peu près la même place, si l'on s'intéresse au classement des populations.
Ainsi, comme les pays ne changent presque pas de rang entre 2007 et 2025, on peut conclure que la démographie mondiale est stable.

Concernant la densité, voici les résultats obtenus :
Correlation Densite 2007 vs 2025
Spearman : coef = 0.9678, p-value = 2.20e-104
Kendall : tau = 0.8589, p-value = 3.73e-63

Le test Spearman obtient 0,9678 : le résultat est très proche de 1. Cela montre que les pays conservent globalement le même rang de densité entre 2007 et 2025. Il existe quelques

modifications qui restent néanmoins trop faible pour avoir une influence majeure. Comme ce que l'on a vu précédemment pour la population, la p-value est faible. Cela signifie donc que la corrélation n'est pas hasardeuse.
Pour Kendall, le résultat du code est encore une fois très élevé : on a donc beaucoup de concordances. La hiérarchie mondiale des densités reste donc très stable, même si elle l'est légèrement moins que celle de la population.

Ces résultats témoignent du constat fait dans le cours : les classements géographiques surtout démographiques, que l'on a analysé ici, « présentent des hiérarchies très stables dans le temps ».

CONCLUSION

Si la géographie est enseignée comme une matière « littéraire », les cinq séances d'analyse de données nous montrent l'intérêt de changer de perception de cette discipline. Les outils mathématiques sont nécessaires pour rassembler les nombreuses données géographiques afin d'en dresser une analyse pertinente. Il est donc essentiel d'avoir une certaine maîtrise de ces outils. Dans ce cadre, la programmation via python peut nous aider. Le code n'est pas un simple outil technique, mais un véritable instrument de raisonnement scientifique, indissociable de l'analyse statistique.

L'étude des principes généraux de la statistique nous a permis de comprendre les fondamentaux en statistiques et de commencer à manier les outils informatiques. Les statistiques univariées ont ensuite permis d'explorer la distribution et la variabilité des données. Les bibliothèques Python sont une grande aide pour le calcul et la visualisation. Grace à elles, il est possible de tester plusieurs hypothèses descriptives et de comparer différents indicateurs. C'est une possibilité formidable pour la géographie : on peut analyser des corpus volumineux ou hétérogènes, qu'il s'agisse de données spatiales, sociales ou autres. Ces fondamentaux nous ont permis de continuer à avancer.

Ainsi le code a permis de réaliser une meilleure analyse des phénomènes, plus complète. Il renforce la rigueur scientifique. Savoir le maîtriser est essentiel pour ne pas faire des erreurs de calculs qui fausseraient tout le raisonnement du géographe. Mais pas seulement. La maîtrise de ces outils permet aux géographes de continuer à être ceux qui vont étudier, comprendre et expliquer les phénomènes géographiques. Ne pas maîtriser ce type d'outils, c'est risque de laisser d'autres disciplines, notamment l'informatique, s'accaparer les données géographiques. Or pour bien utiliser des données spatiales et faire des manipulations pertinentes, il faut avoir compris ce qu'est un raisonnement sur l'espace.