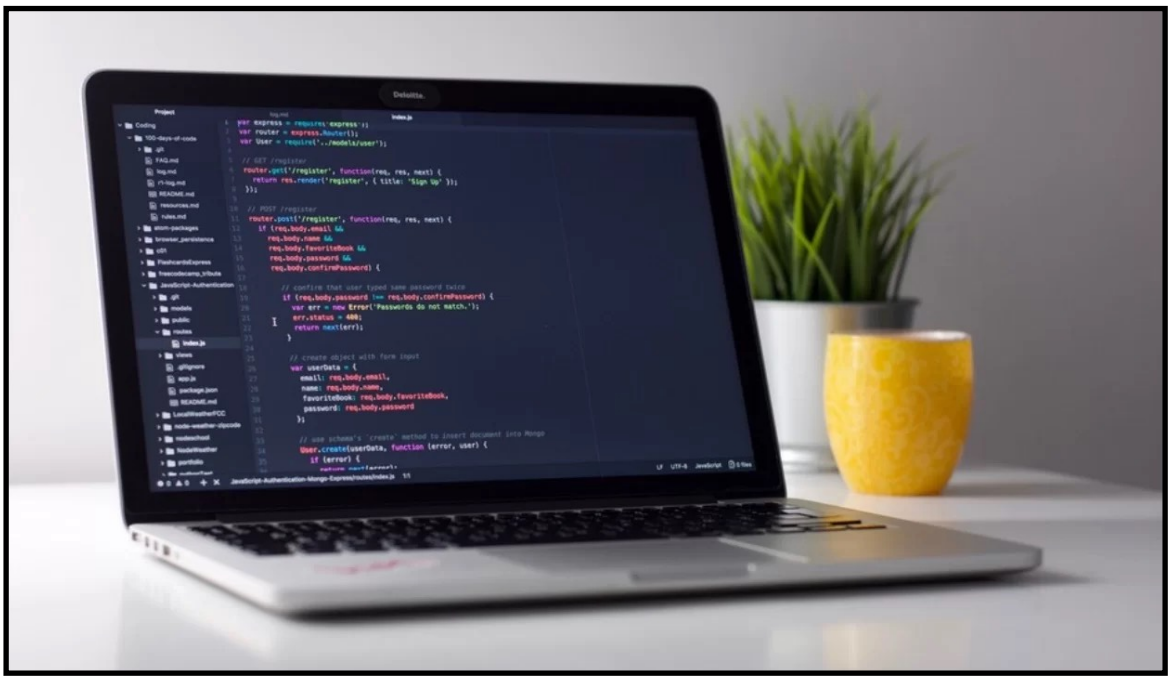
Roxane CELESTINE
21502680
M1 GAED-GéoSuds
Année 2025-2026
Niveau Débutants

Rapport final pour le cours Analyse de données en géographie



Séance 2

1. Quel est le positionnement de la géographie par rapport aux statistiques ?

Le positionnement de la géographie semble se montrer assez méfiante des statistiques car elles semblent n'avoir aucun lien. Pourtant, il est expliqué que c'est une discipline très utile dans les recherches de géographes : elle produit des données massives qui sont pourtant cruciale dans la compréhension de phénomènes.

2. Le hasard existe-t-il en géographie ?

Si l'on prend la définition comme tel du déterminisme « doctrine philosophique suivant laquelle tous les événements, et en particulier les actions humaines, sont liés et déterminés par la chaine des évènements antérieurs. » ; qui définit en partie la géographie, non.

Cependant deux notions sont discutées, celle de la nécessité et celle de la contingence : le premier dit qu'il est impossible que quelque chose se passe autrement, tandis que le second émet la possibilité que ça se fasse ou pas.

3. Quels sont les types d'information géographique.

Il y a deux séries statistiques possibles :

- Pour caractériser l'ensemble délimité par des éléments géographiques humaine ou physique, par la base attributaire
- Pour étudier la morphologie des ensembles délimités, par les données géométriques des SIG

Le géographe a surtout besoin de collecter des données pour ensuite l'interpréter et la mettre en forme.

4. Quels sont les besoins de la géographie au niveau de l'analyse de données ?

La géographie est utilisé pour diverses besoins dans l'analyse de données, notamment pour étudier la structure interne des données : rendant les comparaisons possibles ainsi que résumer ces données. De plus, elle permet de comparer les théories au pratique sur le terrain, ainsi que de la fiabilité des observations.

5. Quelles sont les différences entre la statistique descriptive et la statistique explicative ?

La statistique descriptive a pour but de permettre la visualisation et les classements des données, tandis que la statistique explicative va ajuster les données disponibles avec la modélisation de relations de causalité et prédire de potentiels scénarios.

6. Quelles sont les types de visualisation de données en géographie ? Comment choisir celles-ci ?

Selon la nature de la variable étudiée, on peut avoir une variable quantitative continue sous forme d'histogramme, de la représentation sectorielle pour une variable qualitative.

7. Quelles sont les méthodes d'analyse de données possibles ?

Il existe trois grandes méthodes, le descriptif, qui permet de visualiser les données et qui les classent; l'explicatif où l'on « cherche à relier une variable à expliquer à des variables explicatives; et celui de prévision qui analyse et prévoit une série chronologique.

8. Comment définiriez-vous : (a) population statistique ? (b) individu statistique ? (c) caractères statistiques ? (d) modalités statistiques ? Quels sont les types de caractères ? Existe-t-il une hiérarchie entre eux ?

a) La population statistique est un ensemble de données à partir d'un élément général (quantitatif);
b) L'individu comme étant une donnée parmi cet ensemble (qualitatif);
c) L'individu statistique comme étant les distinctions/similitudes de cette donnée (qualitatif);
d) Les modalités statistiques seraient la valeur que prend ce caractère.

Il y aurait une hiérarchie entre eux les cratères car les caractères quantitatifs permettent plus de traitements que les caractères qualitatifs.

9. Comment mesurer une amplitude et une densité ?

L'amplitude est le calcul de la valeur maximale soustrait à la valeur minimale, alors que la densité est le calcul de l'effectif divisé par l'amplitude.

10. À quoi servent les formules de Sturges et de Yule ?

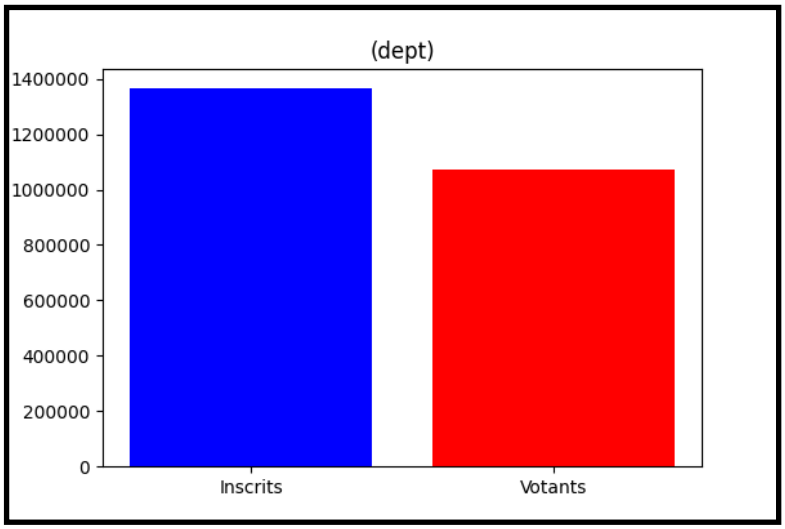
Les deux formules permettent de calculer la valeur approximative du nombre k de classes, cela permet lors de la construction de graphique, diagramme d'apporter un maximum d'informations et de le rendre lisible.

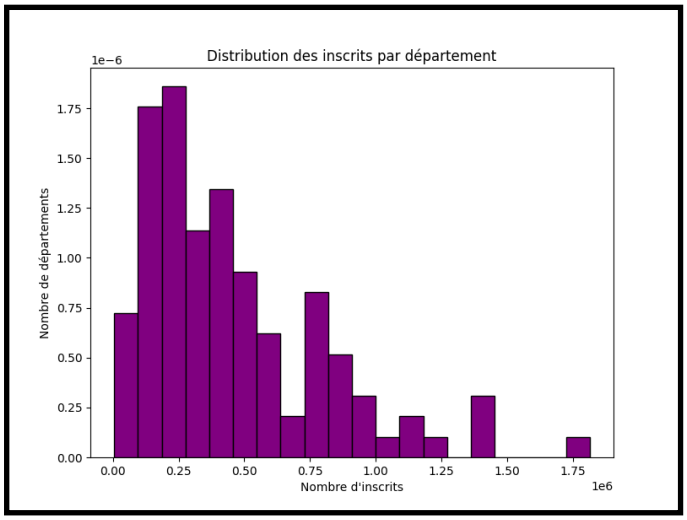
11. Comment définir un effectif ? Comment calculer une fréquence et une fréquence cumulée ? Qu'est-ce qu'une distribution statistique ?

L'effectif correspond au nombre de fois qu'apparaît cette variable dans la population donnée. La fréquence est le calcul du rapport entre l'effectif ni par l'effectif total n, soit fi = ni / n, tandis que la fréquence cumulée est la somme des effectifs associés aux valeurs du caractère.

La distribution statistique correspond à l'ensemble de ces fréquences.

Diagramme en barre pour la Ville de Paris





Séance 3

1. Quel caractère est le plus général : le caractère quantitatif ou le caractère qualitatif?
   Justifier pourquoi.

Le caractère quantitatif est un caractère plus riche car il possède des informations mathématiques plus importantes. Le quantitatif peut se transformer en qualitatif mais pas l'inverse, il faudrait inventer les données et on perdrait alors des informations cruciales.

2. Que sont les caractères quantitatifs discrets et caractères quantitatifs continus? Pourquoi les distinguer ?

Les caractères quantitatifs discrets sont lorsque l'on classe les n valeurs de la série en ordre croisssant, elles comprennent deux cas possibles (pair ou impair pour n). Tandis que les caractères quantitatifs continus sont lorsque la médiane est le nombre me tel que la fréquence cumulée jusqu'à ce que la valeur m, soit égale à 1/2.

Il est très important de les distinguer car selon les données prises le résultat (sous forme de représentation graphique, ne donnera pas le même résultat.

3. Paramètres de position
   - Pourquoi existe-t-il plusieurs types de moyenne?

Il existe plusieurs moyennes (arithmétique, géométrique, harmonique, quadratique, mobile et focntionnelle), qui s'adaptent à la nature de la variable: la moyenne arithmétique convient aux séries simples, et assez utilisée, tandis que la géométrique est préférée pour les taux de croissance ou les ratios.

   - Pourquoi calculer une médiane ?

La médiane est calculée car elle n'est pas influencée par des valeurs extrêmes, contrairement à la moyenne, mais par le nombre de données. De plus, elle permet de couper un ensemble de données en deux groupes égaux, offrant ainsi une vision plus équilibrée de la série.

   - Quand est-il possible de calculer un mode?

Le mode agit comme une moyenne de fréquence, et pouvant être calculé pour tous les types de caractères, dont les qualitatifs. C'est la valeur la plus fréquente.

4. Paramètres de concentration
   - Quel est l'intérêt de la médiale et de l'indice de C. Gini?

La médiale vient partager de façon très égale deux parties, pour que leur « poids » conserve la même valeur des deux côtés. L'indice de Gini est très lié à elle, elle va décrire les effets de concentration d'ne population statistique. Ensemble, elles peuvent mettre en avant les possibles inégalités émergeantes dans les données.

5. Paramètres de dispersion
   - Pourquoi calculer une variance à la place de l'écart à la moyenne? Pourquoi la remplacer par l'écart type?

Il est plus intéressant de calculer une variance car, en prenant compte de toutes les données fournis, elle caractérise le mieux de la dispersion. L'écart-type est la racine carré de la variance, elle peut remplacer la variance car elle caractérise la dispersion d'une série de valeurs. C'est-à-dire, c'est le meilleur moyen de se rapprocher des données fournies.

   - Pourquoi calculer l'étendue?

L'étendue permet de mesurer l'étendue total d'une série entre la valeur minimale et la valeur maximale, elle est simple à calculer et ne prend que les valeurs extrêmes.

   - À quoi sert-il de créer un quantile? Quel(s) est (sont) le(s) quantile(s) le(s) plus utilisé(s)?

Un quantile sert à diviser une série ordonnée en parts égales pour ensuite en analyser la distribution. Les plus utilisés sont les quartiles (division en 4), les déciles (en 10) et les centiles (en 100).

   - Pourquoi construire une boîte de dispersion ? Comment l'interpréter?

La boîte de dispersion (ou boîte à moustaches) permet de visualiser les principales caractéristiques d'une distribution, à savoir : la dispersion, la symétrie et les valeurs aberrantes d'une série.

La ligne centrale de la boîte représente la médiane, le corps de la boîte les quartiles et valeurs extrêmes sont visibles avec les moustaches.

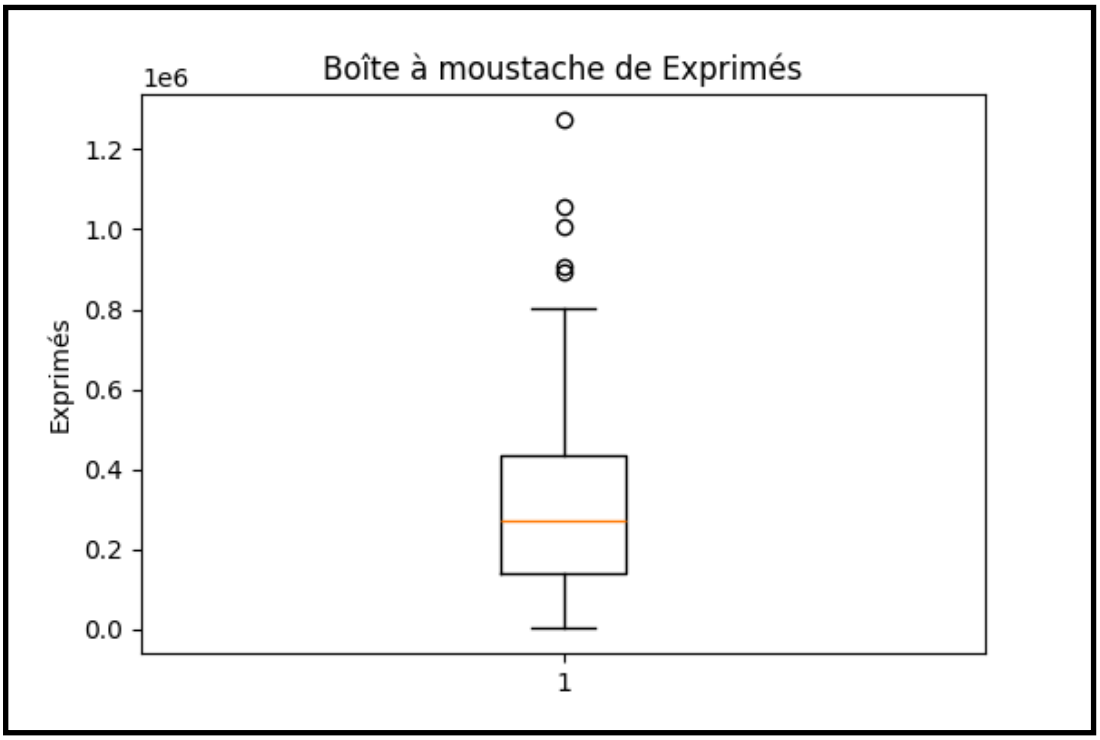
6. Paramètres de forme
   - Quelle différence faites-vous entre les moments centrés et les moments absolus ?

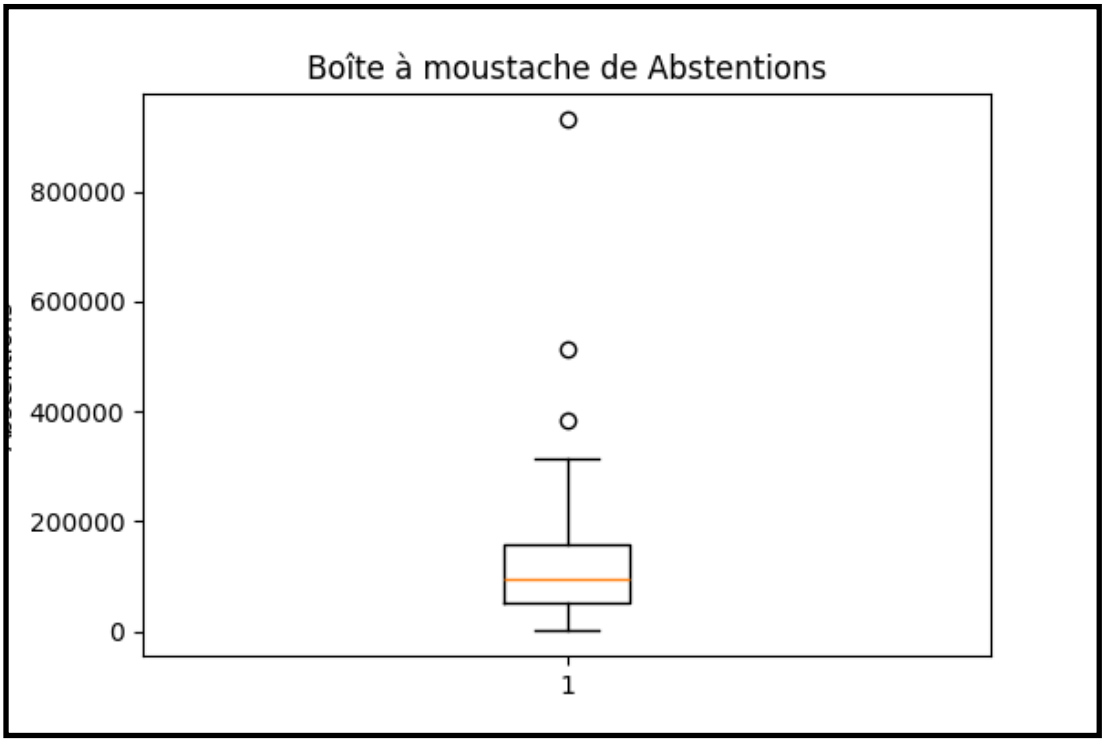
Les moments centrés mesurent les écarts par rapport à la moyenne, tandis que les moments absolus s'intéressent aux valeurs brutes. Ils permettent de caractériser une distribution.

7. Pourquoi les utiliser?

- Pourquoi vérifier la symétrie d'une distribution et comment faire ?

La vérification de la symétrie permet de savoir comment les données se répartissent autour de la moyenne. De cette manière, on peut voir si elle est dissymétrique à droite (donc positive) ou bien dissymétrique à gauche (donc négative)


Boîte à moustache de Exprimés
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Exprimés
1


Boîte à moustache de Abstentions
800000
600000
400000
200000
0
1

Séance 4

1. Quels critères mettriez-vous en avant pour choisir entre une distribution statistique avec des variables discrètes et une distribution statistique avec des variables continues?

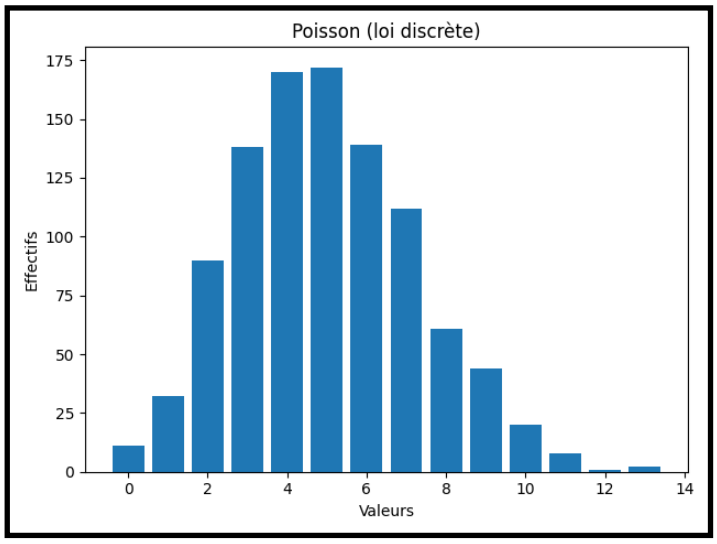
Les distributions statistiques sont rattachées aux lois de probabilité, le choix entre les deux disruptions reposerait entre 4 principes:

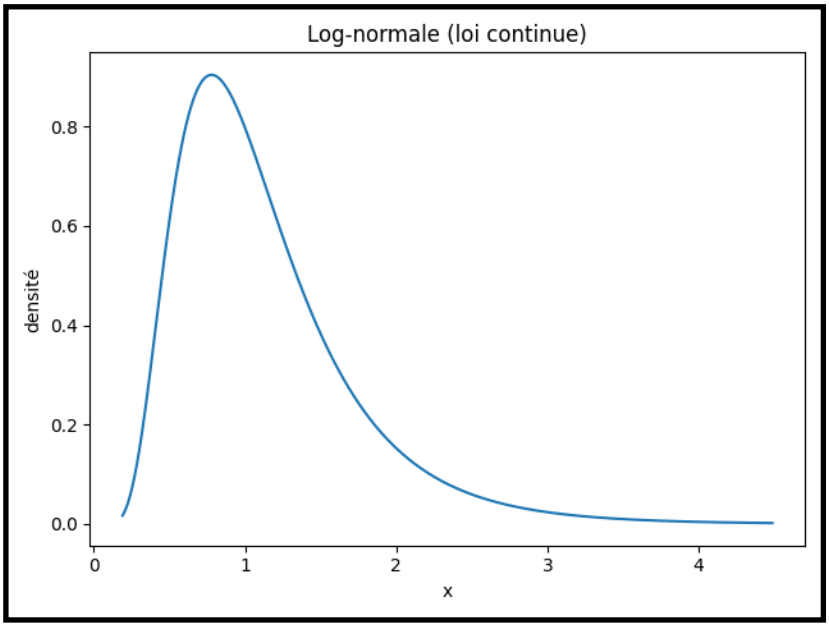
- À la nature du phénomène étudié afin de choisir entre loi discrète et loi continue ;
- À la forme de la distribution empirique;
- À la connaissance et à l'interprétation des principales caractéristiques de l'ensemble de données : espérance, médiane, variance, écart type, coefficients d'asymétrie et de dissymétrie, etc.;
- Au nombre de paramètres des lois, une loi dépendant de plusieurs paramètres pouvant s'adapter plus facilement à une distribution.

La première dépendrait de ce que l'on étudie, si elle se concentre sur des données plus quantitatives, on prend la loi discrète mais si ce sont des données plus qualitatives, une lozi continue.
La deuxième comprend que le rendu visuel des données doit correspondre à la théorie.
La troisième évoque la concordance entre les indicateurs et les propriétés de la loi utilisée.
Enfin, la quatrième évoque la facilité d'un modèle face à une loi, plus elle a de paramètres, mieux elle s'adaptera à des formes complexes.

2. Expliquez selon vous quelles sont les lois les plus utilisées en géographie?

D'après moi, les lois les plus utilisés en géographie sont la loi de Zipf-Mandelbrot qui permet de confronter le nombre d'habitants d'une ville avec son rang au sein d'un territoire, avec la loi rang-taille. Par ailleurs, c'est une loi que j'ai pu utilisé lors de la manipulation avec Python.
Aussi la loi de Benford, qui se concentre sur la longueur des fleuves du globe et à la superficie des pays. Ce sont des loi qui se concentrent directement sur les données utiles dans le domaine de la géographie.




Log-normale (loi continue)
densité
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
x

Séance 5

1. Comment définir l'échantillonnage. Pourquoi ne pas utiliser la population en entier? Quelles sont les méthodes d'échantillonnage? Comment les choisir ?

L'échantillonnage correspond à la sélection d'un échantillon (d'individus) au sein d'une population, on n'utilise pas une population entière car celle-ci est bien trop grande pour être fait, cela prendrait trop de temps. Bien que ça ne lui enlève pas de pertinence, bien au contraire.

Les méthodes d'échantillonnage sont soit aléatoire (non-biaisées), soit non-aléatoire(biaisées). Le premier procède par tirage au sort, où l'on prend au hasard des individus dans la population mère, ce ce qui implique une base de songe au préalable.

La seconde se construit sur des modèles réduits de la population mère, et en prenant d'autres procédés que le tirage au sort. On peut retrouver en « sous-base », l'échantillonnage systématique, qui implique de déterminer la taille de l'échantillon et une numérotation des individus choisis.

2. Comment définir un estimateur et une estimation?

Un estimateur est une variable aléatoire, une fonction de données, qui permet d'observer une caractéristique, qui est proche de la valeur du paramètre. C'est un outil pour l'estimation.

L'estimation est une valeur numérique précise qui s'obtient une fois le calcul fait sur les données étudiées, elle permet d'estimer les paramètres d'une loi de probabilité et d'observer des caractéristiques générales.

3. Comment distingueriez-vous l'intervalle de fluctuation et l'intervalle de confiance ?

L'intervalle de fluctuation est utilisé lorsque l'on fait de l'échantillonnage, elle a pour but de « prendre une décision avec un certain risque d'erreur selon l'appartenance ou non de la fréquence observée à l'intervalle de confiance asymtotique. » Elle suppose qu'il y a une connaissance sur la population et que l'on parte d'un résultat pour estimer quelque chose. À l'inverse, l'intervalle de confiance est utilisé lorsqu'on ne connaît pas la population.

4. Qu'est-ce qu'un biais dans la théorie de l'estimation?

Une biais est une erreur dans la théorie de l'estimation, si le résultat de l'estimateur est loin de la valeur réelle du paramètre, il sera biaisé.

5. Comment appelle-t-on une statistique travaillant sur la population totale? Faites le lien avec la notion de données massives 1 ?

Il existe plusieurs statistiques travaillant sur la population totale, tout d'abord le recensement, qui se constitue un dénombrement total de la population, c'est une donnée exhaustive. Toutes les informations nécessaires sont présentes dans un ensemble de données, sous forme de variable quantitative.

Les données massives, ou aussi connus sous le nom de « Big data », permettent d'accumuler des données très importantes sur une ou plusieurs domaines. cependant, elles peuvent être biaisées, car elles peuvent représenter qu'une partie d'un ensemble.

6. Quels sont les enjeux autour du choix d'un estimateur ?

Il y a des enjeux autour du choix, dont celui d'être au plus proche de la valeur de paramètre, pour éviter que ce soit biaisé, il faut choisir le meilleur estimateur. Ce dans quoi l'estimateur puisse ses informations ne lui donne qu'une part minime ou pas totale. Il doit donc être le plus précis possible.

7. Quelles sont les méthodes d'estimation d'un paramètre? Comment en sélectionner une ?

Deux méthodes se distinguent avec la méthode des moindre carrés, qui étudie plusieurs variables, et choisi les paramètres qui minimisent la somme des carrés des résidus.

La méthode du maximum de vraisemblance trie les valeurs du paramètre selon leur probabilité.

Il y a aussi l'estimation ponctuelle, qui se concentre sur une seule valeur et l'estimation par intervalle qui prend une plage de valeurs. On va sélectionner selon le besoin de précision, qui prend le plus d'informations pour éviter des variances.

8. Quels sont les tests statistiques existants ? À quoi servent-ils ? Comment créer un test ?

Il existe des tests, dont les paramétriques avec t de Student, F de Fisher, les non-paramétriques avec Mann-Whitney et Wilcoxon, d'ajustement avec Khi² et Shapiro-Wilk, ainsi que d'indépendance (Khi²). Pour créer un test, on définit une hypothèse nulle, puis on choisit un seuil de risque pour enfin calculer une statistique de test pour voir si le résultat est significatif ou non.

9. Que pensez-vous des critiques de la statistique inférentielle ?

Cette statistique est souvent critiqué et les chercheurs soulignent que les hypothèses nulles sont irréalistes, avec l'influence de la taille qui peut rendre n'importe quel

détail significatif. Aussi le manque de puissance de plus petits échantillons qui peuvent masquer des phénomènes réels.

Ce sont des critiques que j'entends et pour lesquels je comprend, surtout en tant que géographe. Avec les statistiques que j'ai pu étudier plus en profondeur depuis quelques années, j'ai bien remarqué que les échantillons pouvaient vite se confronter à des blocages, de ce que critique les chercheurs. Cependant, je pense qu'il est nécessaire de poursuivre la statistique inférentielle car elle permet de produire des études. Et ce, malgré les marges d'erreur possibles.

Séance 6

1. Qu'est-ce qu'une statistique ordinale? À quel autre statistique catégorielle s'oppose-t-elle? Quel type de variables utilise-t-elle ? En quoi cela peut matérialiser une hiérarchie spatiale?

D'après l'ESRI, une statistique ordinale est « une méthode d'organisation des valeurs de données en une hiérarchie classée sans intervalle fixe entre les niveaux hiérarchiques. ». Elle s'appuie sur du classement de données, de manière croissante. C'est un outil très puissant qui permet d'évaluer la force d'une impression, ou bien de montrer l'évolution d'entités.
Elle s'oppose à la statistique nominale, qui ne suivent pas d'ordre logique, elle utilise des variables qualitatives ordinales.
Elle peut matérialiser une hiérarchie spatiale en classant les entités, comme établir des classements de villes par rapport aux nombres d'habitants par ordre croissnat, aussi pour suivre leur évolution et les comparer entre elles. C'est ce qui la rend essentiel en géographie.

2. Quel ordre est à privilégier dans les classifications?

L'ordre a privilégié est celui dans la classification par ordre croissant pour des raison de logique mais aussi par défaut.

3. Quelle est la différence entre une corrélation des rangs et une concordance de classements?

La corrélation permet la mesure entre deux classements distincts d'un ensemble d'individus ou d'objets, tandis que la concordance de classements est un outil statistique qui permet d'observer la concordance entre plusieurs classements, elle utilise le coefficient W de Kendall, qui mesure la dispersion des sommes des colonnes, par rapport à leur moyenne.

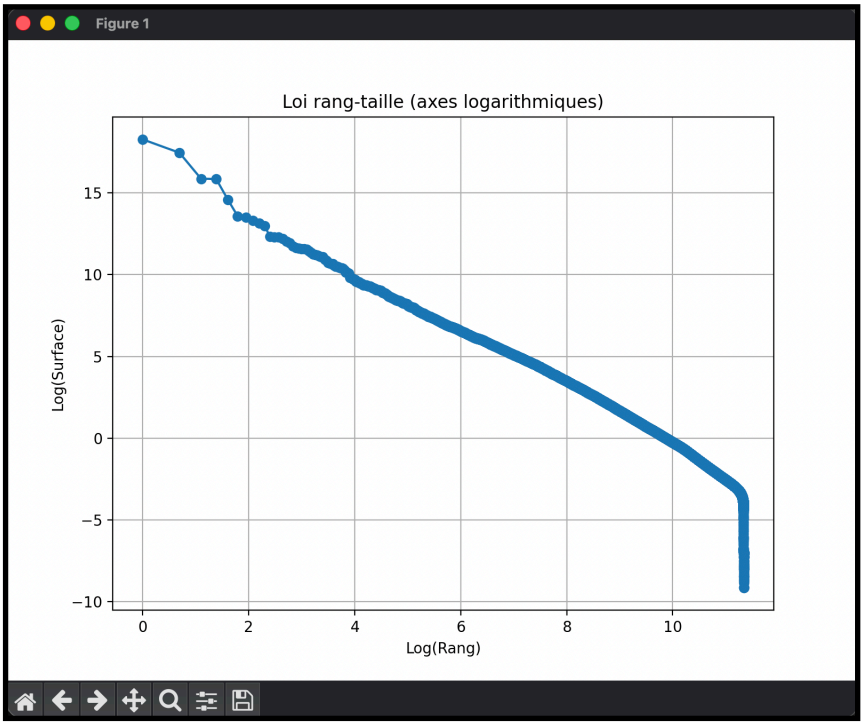
4. Quelle est la différence entre les tests de Spearman et de Kendal?

Les tests de Spearman mesurent au carré, la distance entre les rangs pour mesurer la corrélation entre deux classements d'un même ensemble.
Tandis que les tests de Kendall, est dans le calcul de paires concordantes et discordantes, permettant une généralisation à plusieurs classements.

5. À quoi servent les coefficients de Goodman-Kursdal et de Yule?

Les deux coefficients servent à mesurer l'association entre des variables qualitatives ordonnées. Le coefficient de Goodman-Krustal évalue la relation entre paire de données, concordantes ou discordantes.

Le coefficient de Yule se fonde sur les écarts des quartiles, pour déterminer l'intensitté de l'association entre deux variables.



Avis et ressenti du cours

Ce cours m'a beaucoup appris, notamment, découvrir de façon plus approfondie le langage de Python et à mélanger la géographie, ce que j'ai trouvé particulièrement pertinent.

Cependant, c'est un travail qui a demandé beaucoup de temps, car étant débutante dans le domaine, il me fut très difficile de comprendre ce que je faisais.

Tout d'abord, j'ai rencontré des difficultés en essayant d'ouvrir les fichiers, prendre du temps à les ouvrir me décourager par moment, mais lorsque je parvenait à comprendre, cela m'amusait un peu.

Bien qu'étant un cours très complexe, je l'ai trouvé extrêmement pertinent et utile. Dans la manière où vous nous avez expliqué les avantages de connaître le codage python à l'avenir et son importance . Cela a pu être un élément motivant.

J'ai tenté de faire ses cours au mieux, et je regrette de ne pas avoir eu plus de temps pour y consacrer du travail, mais je suis ravie de pouvoir dire que je possède des bases.

Je tiens à remercier particulièrement Zara HUSTON, qui m'a permis d'installer les fichiers, mais surtout VS code, sa connaissance du logiciel m'a beaucoup permis de progresser. De plus. Je remercie également. Inès GBALE, avec qui j'ai pu avancer tout au long du semestre, d'autant plus que nous avions toutes les deux des MacBook, ce qui était légèrement compliqué. J'ai également reçu l'aide de Myriam MÉNARD et d'Anaé DELLIÈRE.

Concernant le cours en lui-même

Dès le début, j'ai ressenti du découragement du fait que l'ordinateur que j'avais emprunté à la fac n'était pas assez performant pour me permettre de faire les exercices. À la vue des rendus à faire, cela me faisait très peur car il représentait une quantité de travail importante, et comme dit plutôt n'ayant pas de connaissance encodage, cela me semblait impossible.

Il aurait été pertinent de faire des cours plus interactifs où on apprenait les bases comme des leçons. Je note tout de même que ce n'est pas facile d'apprendre du langage python à des élèves. Au lycée, mes professeurs ont tenté de nous initier mais sans véritables succès, nous n'avons jamais vraiment poursuivi dans le codage.

C'est que j'ai tout de même apprécié, c'est l'entraide entre élèves qui a permis un véritable avancement pour chacun d'entre nous.