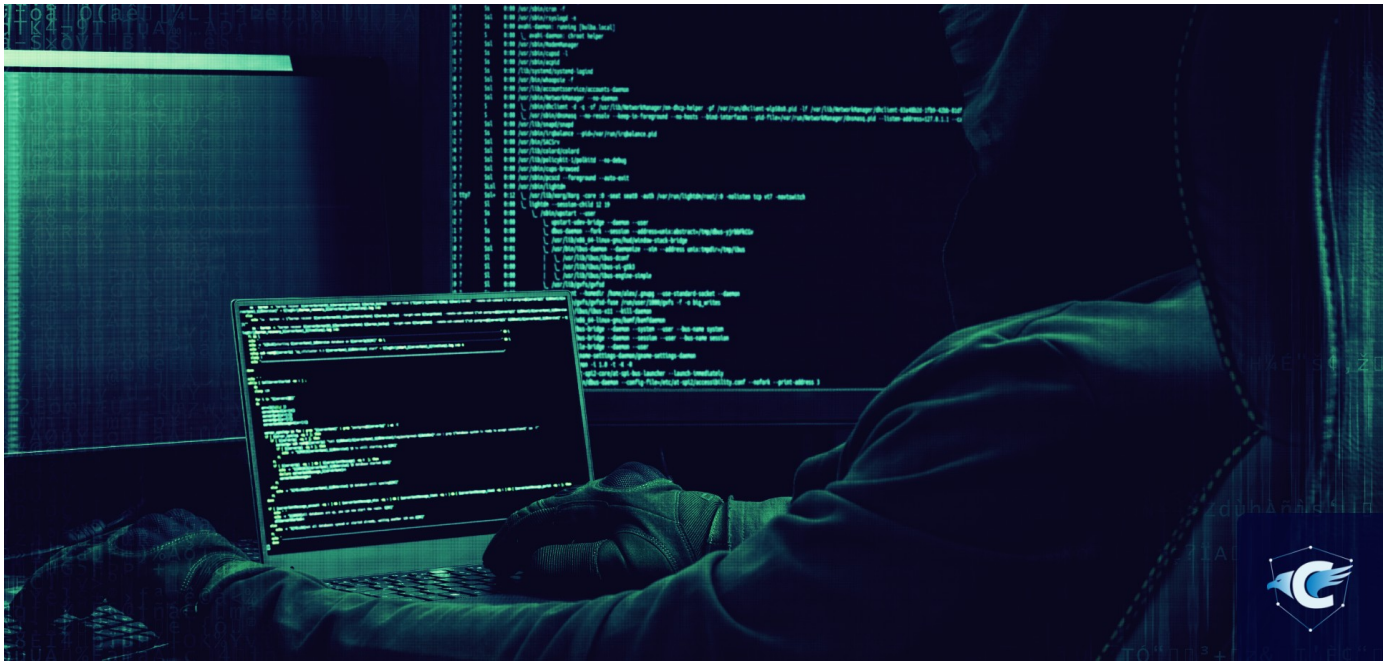
Rapport final d'analyse de données



Matthieu ELIZEON

MASTER 1 GAED : GEOSUDS

Parcours débutant

2025-2026

Remarques et difficultés :

Ce cours a été très difficile pour moi. La partie question était assez simple et compréhensible grâce à votre cours. Cependant les questions m'ont pris beaucoup de temps. En revanche, le codage était vraiment difficile. J'ai fait tout ce que j'ai pu afin de pouvoir finir toutes les séances. Mes principales difficultés concernaient le codage. La première était de pouvoir ouvrir les fichiers. En effet, *Python* avait du mal à trouver le dossier *data* alors qu'il était à l'emplacement prévu. J'ai dû ainsi utiliser la fonction du terminal « cd » pour montrer le chemin vers le dossier. Après plusieurs tentatives cela a fini par marcher notamment pour la séance 3. J'ai ensuite essayé d'écrire les lignes de code moi-même grâce à votre cours mais je prenais vraiment du temps et j'avais énormément de messages d'erreur. Des camarades et des vidéos ont alors été d'un grand soutien. Je dois avouer que pour le codage, j'ai utilisé l'intelligence artificielle pour m'aider sinon cela aurait été impossible pour moi notamment pour les lignes les plus complexes et pour forcer le chemin vers quelques fichiers que le logiciel ne trouvait pas. Ainsi je suis conscient des gros problèmes qu'il pourrait y avoir dans mon code, mais cela m'a quand même permis d'avoir des résultats de visualisation notamment. Certaines tâches étaient plus faciles que d'autres. Je ne compte plus les messages d'erreur que j'ai reçus. J'ai donc fait ce travail sérieusement durant tout le semestre afin de produire un travail assez convenable. J'ai évidemment mis qu'une infime partie des images que j'ai obtenues.

Je tiens particulièrement à remercier Zara Huston en Géosuds qui a installé *Python* et *VS Code* pour moi. Zara a plusieurs fois réservé des salles à la bibliothèque pour nous aider sur les différentes séances. Ainsi nous pouvions avancer tous ensemble en s'aidant mutuellement. Sans Zara, je n'aurais pas pu finir ce travail. Je remercie aussi Myriam Ménard qui m'a aidé à de multiples reprises également.

Séance 2

Question de cours :

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

La géographie reste dans une croyance où l'outil mathématique, l'outil statistique est dispensable. Alors que dans la géographie, sont présentes différents types de données que seule une approche statistique permet de traiter, analyser et interpréter. L'outil statistique malgré ses avantages et ses atouts évidents, restent très sous-estimés au regard de la géographie

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

Le hasard peut se définir comme un événement inattendu, imprévu et imprévisible, qui se passe par un concours de circonstances inexplicables. Affirmer que rien n'est dû au simple hasard serait adopter une vision déterministe du monde. Le déterminisme est une position philosophique qui défend que chaque évènement, action qui se passe dans le monde a pour origine une cause : c'est donc une relation de cause à effet. Ces évènements seraient donc nécessaires, c'est-à-dire qu'il se déroulerait toujours de la même façon. Au contraire, un évènement qui ne se produirait pas de la même façon à chaque fois serait contingent. Les scientifiques et géographes par le biais de la géographie physique peuvent savoir et démontrer ce que des causes vont produire comme effet. Cependant il est difficile de prévoir des cas particuliers. Par exemple on peut expliquer les causes qui vont amener à une chute de pierre (précipitations, humidité, roche fragile) mais il est plus difficile de prévoir quand et où il y aura une chute de pierre. De la même façon, on pourrait théoriquement prévoir des actions, déduire une tendance mais il y a nécessairement une part de hasard dans l'équation notamment quand l'on change d'échelle. Ainsi le hasard existe en géographie, mais l'étude des statistiques, d'analyse spatiale permettent de comprendre et même rationaliser ce hasard.

3. Quels sont les types d'information géographique ?

L'information géographique peut soit désigner « l'ensemble délimité par des éléments de la géographie humaine ou physique » soit « étudier la morphologie même des ensembles délimités ».

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

Les géographes ne produisent que très peu de données mais ils les étudient. De fait, les organismes publics produisent des données utiles à la géographie humaine. Les mesures sur le terrain viennent des géologues ou topographes. Le géographe doit cependant donner un sens à ces données en les analysant.

Le géographe doit ensuite produire une nomenclature et des métadonnées pour pouvoir mettre en œuvre une étude statistique.

5. Différences entre la statistique descriptive et la statistique explicative ?

La statistique descriptive étudie les données. Elle permet de mettre en lumière les « unités remarquables dans une distribution théorique connue ». Cela permet d’avoir une image claire de la réalité. On peut alors visualiser et classer plus aisément les données. La statistique explicative quant à elle doit expliquer les données, les interpréter afin d’y donner un sens en mettant en relation différentes données et de pouvoir trouver des tendances, des liens de causalité, des corrélations.

6. Types de visualisation de données en géographie et comment choisir ?

Il y a plusieurs types de visualisation de données en géographie. Ils cherchent à représenter les données dans un espace réduit. Le type de visualisation dépend du type de données traitées. Par exemple pour des données quantitatives absolues de la population de Paris par arrondissement il faudrait une carte en symboles proportionnels. Pour des déplacements, des migrations il faut opérer une carte de flux. Les cartes ne sont pas les seules types de visualisation, nous pouvons également utiliser des graphiques avec une courbe chronologique pour mesurer l’évolution d’un phénomène dans le temps. Par exemple l’évolution du taux de natalité de 2000 à 2025. Il existe de nombreux autres types de visualisation comme les cartes choroplèthes par exemple pour permettre la visualisation d’un taux ou d’un pourcentage. Il faut donc choisir le type de visualisation en fonction du type de données.

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

Il y a 3 types de méthode d’analyse de données possibles :

-Les méthodes descriptives qui visent à décrire, visualiser et synthétiser. Il ne s’agit pas encore d’expliquer la donnée mais plus de la mettre en ordre de façon claire et structurée.

-Les méthodes explicatives permettent de traiter les données, de les interpréter, de fournir un rapport. Elles permettent d’expliquer une variable en établissant un lien avec des variables explicatives.

-Enfin les méthodes de prévision qui permettent « la prévision d’une série chronologique ». En reliant le présent et le passé, nous sommes en capacité de construire un modèle. Prévoir l’évolution possible d’une variable grâce à ses valeurs passées.

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. Population statistique = un ensemble au sens mathématique. Il est donc quantifiable.

b. Individu statistique = élément de la population statistique. Ils sont alors localisables et cartographiables (unités spatiales). Les individus statistiques sont aussi composés d’un ensemble d’attributs.

c. Caractère statistique = caractéristique de l’individu pris dans la population statistique.

d. Modalité statistique = valeur prise par un caractère. « L’objectif est de caractériser l’appartenance ou la non appartenance, d’un individu à une modalité ».

Les caractères sont qualitatifs ou quantitatifs. Les caractères qualitatifs sont soit nominal ou ordinal et les caractères quantitatifs sont discret ou continu. Il existe une hiérarchie entre les caractères qui est déterminée par la quantité d’information et de données dont ils disposent.

9. Comment mesurer une amplitude et une densité ?

L’amplitude se mesure en calculant b-a c’est-à-dire la valeur maximale – la valeur minimale. La densité est le rapport entre l’effectif de la classe et son amplitude.

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

Les formules de Sturges et de Yule servent à estimer un nombre de classes approprié lors de la discrétisation. Il faut éviter que le nombre de classes soit trop petit, ou trop grand.

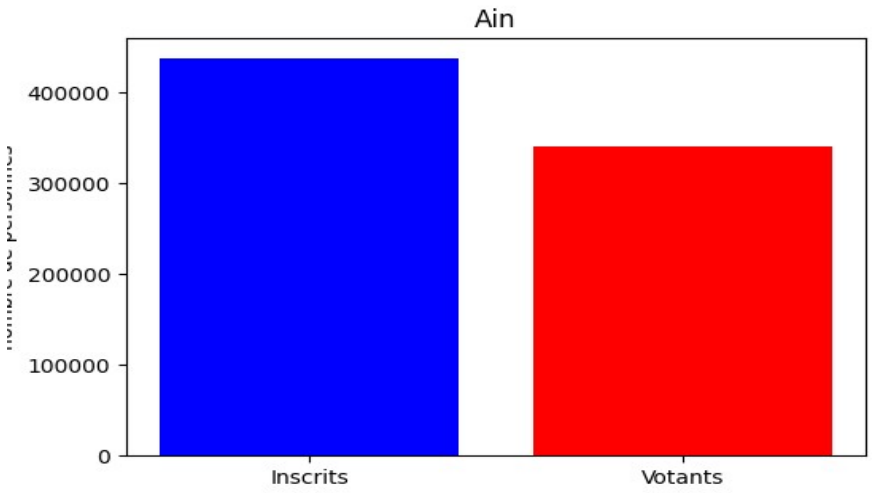
11. Comment définir un effectif ? Comment calculer une fréquence et une fréquence cumulée ? Qu'est-ce qu'une distribution statistique ?

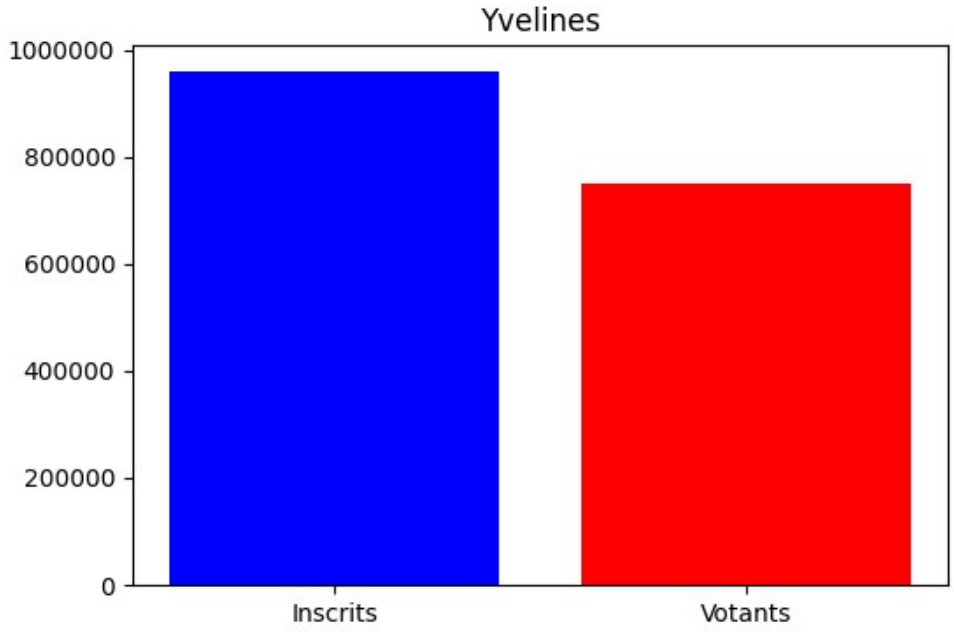
L'effectif « correspond au nombre d'apparitions de cette variable dans la population. » La fréquence c'est le rapport entre l'effectif d'une modalité et l'effectif total et la fréquence cumulée c'est la somme des effectifs associés aux valeurs du caractère qui sont inférieures ou égales à k. Enfin, une distribution statistique représente la répartition des individus selon les valeurs ou les classes d'une variable. Enfin, la distribution statistique permet de « conclure sur le type de loi de probabilité utilisée. »

Données, graphiques, tableaux

	Code du département	Libellé du département	Inscrits	Abstentions	...	Sexe.11	Nom.11	Prénom.11	oix.11
0	01	Ain	438109	97541.0	...	M	DUPONT-AIGNAN	Nicolas	8998.0
1	02	Aisne	373544	101089.0	...	M	DUPONT-AIGNAN	Nicolas	5790.0
2	03	Allier	249991	58497.0	...	M	DUPONT-AIGNAN	Nicolas	4216.0
3	04	Alpes-de-Haute-Provence	128075	29290.0	...	M	DUPONT-AIGNAN	Nicolas	2504.0
4	05	Hautes-Alpes	113519	25357.0	...	M	DUPONT-AIGNAN	Nicolas	2142.0
..	...	...	...	...	...	...	...	...	...
102	ZP	Polynésie française	205576	142121.0	...	M	DUPONT-AIGNAN	Nicolas	1969.0
103	ZS	Saint-Pierre-et-Miquelon	5045	2272.0	...	M	DUPONT-AIGNAN	Nicolas	82.0

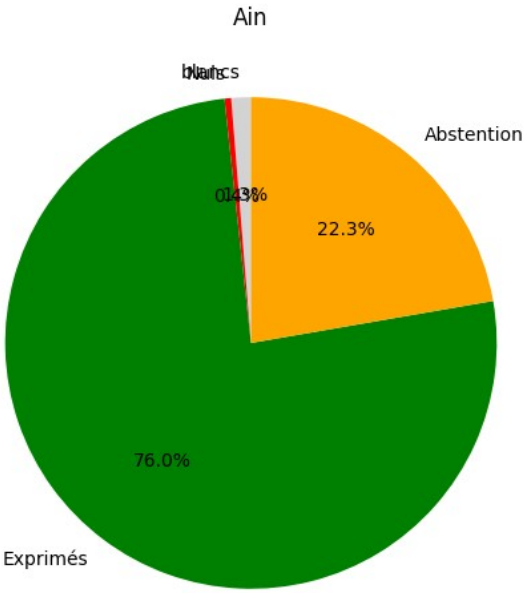
Commentaire de résultat graphique : J'ai pu lors de cette première séance me familiariser avec les commandes python. J'ai notamment appris à ouvrir un fichier à l'aide de la bibliothèque pandas. J'ai néanmoins été confronté à un problème : python n'arrivait pas à trouver le chemin vers mon fichier. Je me suis alors aperçu que c'était parce que j'ai téléchargé deux fois le fichier il y avait donc un (1). J'ai également utilisé la commande cd pour indiquer dans quel dossier se trouvait le fichier. J'ai calculé le nombre de colonnes, de lignes et déterminé le type de variables. Ensuite grâce à Matplotib j'ai affiché le tableau.



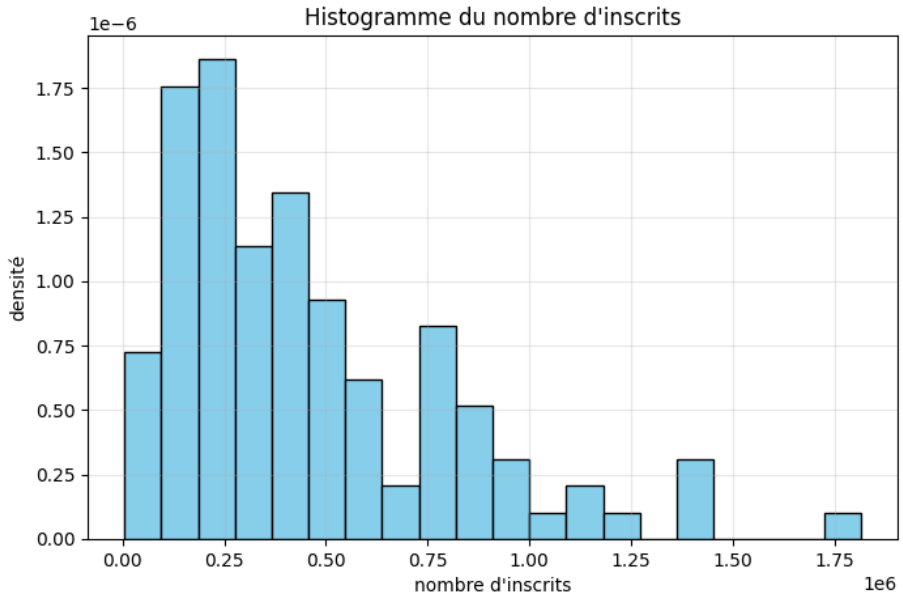


Commentaire de résultats graphiques : Ces deux graphiques en barre pour le département de l'Ain et des Yvelines le nombre d'inscrits et de votants. Pour le département de l'Ain, on observe que le nombre d'inscrits est d'environ 438 000, tandis que le nombre de votants est d'environ 340 000. L'écart entre les deux barres représente les abstentions car une partie des électeurs inscrits ne s'est donc pas rendue aux urnes. Pour les Yvelines, les effectifs sont plus élevés, avec environ 960

000 inscrits et 750 000 votants. Ces graphiques en barre nous permettent donc d'analyser le nombre d'abstention en faisant la différence entre le nombre d'inscrits et de votants, et ce de manière synthétique.



Commentaire de résultats graphiques : Ce diagramme circulaire représente la répartition des votes lors du premier tour de l'élection présidentielle de 2022 dans le département de l'Ain. J'ai réussi à l'afficher grâce à Matplotlib. Ici, la part des votes exprimés est largement majoritaire. Elle représente environ 76 % des inscrits, ce qui montre que la majorité des électeurs ayant participé au vote ont exprimé un choix valide pour un candidat. Il y a quand même 22,3 % d'abstentions. Il y a peu de votes blancs et nuls.



Commentaire de résultats graphiques : J’ai créé grâce au code un histogramme du nombre d’inscrits, ce qui nous permet de voir plusieurs choses. Cet histogramme représente la distribution du nombre d’inscrits par département lors du premier tour de l’élection présidentielle de 2022. Il est différent des précédentes figures car il démontre la dispersion et l’asymétrie de la distribution du nombre d’inscrits.

Séance 3

Questions de cours :

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

Selon moi le caractère qualitatif est le plus général car il peut prendre plusieurs types très variés de données alors que le caractère quantitatif prend exclusivement les nombres comme données.

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

Les caractères quantitatifs discrets prennent seulement en compte les valeurs entières alors que les caractères quantitatifs continus prennent en compte les nombres réels, c'est-à-dire des entiers, des nombres décimaux, des fractions. Dans python il faut les distinguer, car on ne peut pas diviser un nombre entier comme 2 (int) et les nombres réels, décimaux 3,6 (float). Nous pouvons faire une opération avec ces 2 nombres que s'ils sont du même type. A l'intérieur de Python, il est possible de changer le type de la valeur.

3. Paramètres de position

— *Pourquoi existe-t-il plusieurs types de moyenne?*

Il existe plusieurs types de moyenne pour avoir une représentation des données différentes, n'exprimant pas les mêmes tendances. La moyenne diffère en fonction du besoin de l'information recherchée. Les données sont mises en rapport différemment.

— *Pourquoi calculer une médiane?*

La médiane partage en deux parties égales un ensemble de données, elle sépare la première moitié inférieure, de la seconde moitié supérieure. Elle est utile car dans certains cas elle fournit une vision plus claire de l'ensemble statistique que la moyenne car elle n'est pas influencée par les extrêmes.

— *Quand est-il possible de calculer un mode?*

Le mode est une modalité correspondant à l'effectif maximal. Nous pouvons calculer le mode lorsqu'une valeur de la série statistique se répète au moins une fois. S'il n'y a pas de répétitions de la donnée, alors nous ne pouvons pas calculer le mode car il fournit l'élément dominant dans la série statistique. Il peut être considéré comme une mesure de tendance centrale. Il fournit l'élément dominant dans la série statistique.

4. Paramètres de concentration

— Quel est l'intérêt de la médiale et de l'indice de C. Gini?

La médiale contrairement à la médiane considère la valeur de la variable elle-même. On s'intéresse donc à la valeur médiane de la distribution et non à l'effectif. Elle partage également les valeurs globales en deux parties égales. L'intérêt de la médiale est de pouvoir déterminer l'indice de concentration dans une série statistique.

L'intérêt de l'indice de GINI quant à lui est de pouvoir rendre compte de la répartition d'une variable à l'intérieur d'un groupe. Elle décrit aussi la concentration dans une population statistique. Les valeurs données varient entre 0 et 1. Il est souvent utilisé pour analyser le niveau d'inégalité d'un pays, le 0 étant l'égalité parfaite et le 1 étant la situation la plus inégalitaire qui soit. Cet indice prend en compte les revenus, les salaires et niveau de vie. L'intérêt est alors d'avoir un indice hiérarchisé qui permet démontrer une concentration des 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?

La variance est « la moyenne de la somme des carrés des écarts par rapport à la moyenne arithmétique ». Il est dans certains cas plus utile de calculer une variance car elle démontre le rapport qu'entretient les valeurs d'un ensemble de données les unes par rapport aux autres. Alors que l'écart à la moyenne permet de montrer seulement la distance et le lien des valeurs par rapport à la moyenne, afin d'identifier par exemple des valeurs aberrantes. La variance prend toutes les données en compte.

— Pourquoi calculer l'étendue ?

L'étendue d'une série statistique est l'écart entre la plus petite valeur et la plus grande. Calculer l'étendue peut être utile si on s'intéresse aux extrêmes d'une série statistique, elle ne dépend pas de d'autres valeurs. Elle est utile si la série de données est courte. De plus le calcul est très rapide et très simple.

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

Un quantile est utile en statistique car il permet de mettre en évidence les données du centre mais aussi celles autour du centre. Il permet de découper la série statistique en plusieurs parties. Les quantiles les plus utilisés sont les quartiles donc division en 4 parties, les déciles en 10, et les centiles en 100.

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

La boite de dispersion ou boîte à moustaches est un outil statistique puissant. Elle permet d'avoir une visualisation claire d'une série statistique en montrant par exemple la médiane, les quartiles ou encore les extrêmes. La médiane est représentée par un trait au milieu, on voit les valeurs extrêmes à gauche (minimum) et à droite (maximum) du graphique. La boite quant à elle représente l'intervalle interquartile qui contient 50 % des observations. Les points isolés sont les valeurs aberrantes et la taille de la boite indique la dispersion des valeurs centrales. Son utilisation est très pertinente en statistique car elle montre en un seul graphique plusieurs mesures de position ce qui permet de se faire une idée concrète et nette d'une série statistique.

6. Paramètres de forme

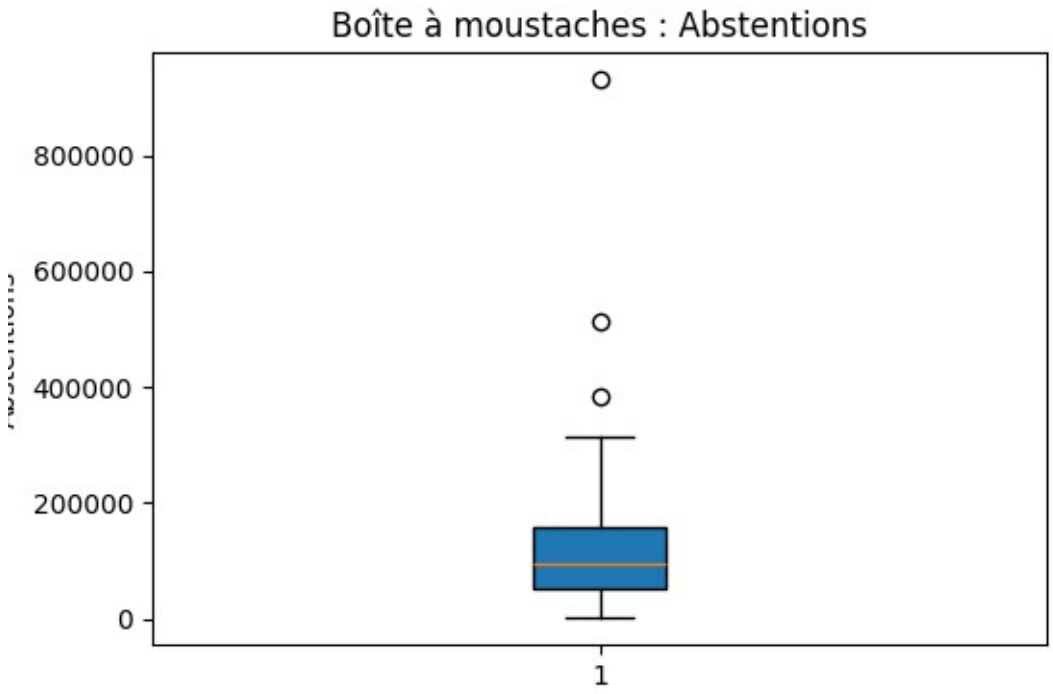
— Quelle différence faites-vous entre les moments centrés et les moments absolus? Pourquoi les utiliser?

Le moment centré analyse une distribution statistique qui reste autour de la moyenne en tenant compte du signe. Le moment absolu quant à lui mesure la distance à un point de référence sans tenir compte du signe. Ils sont utilisés pour analyser la dispersion et la forme d'une série statistique.

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

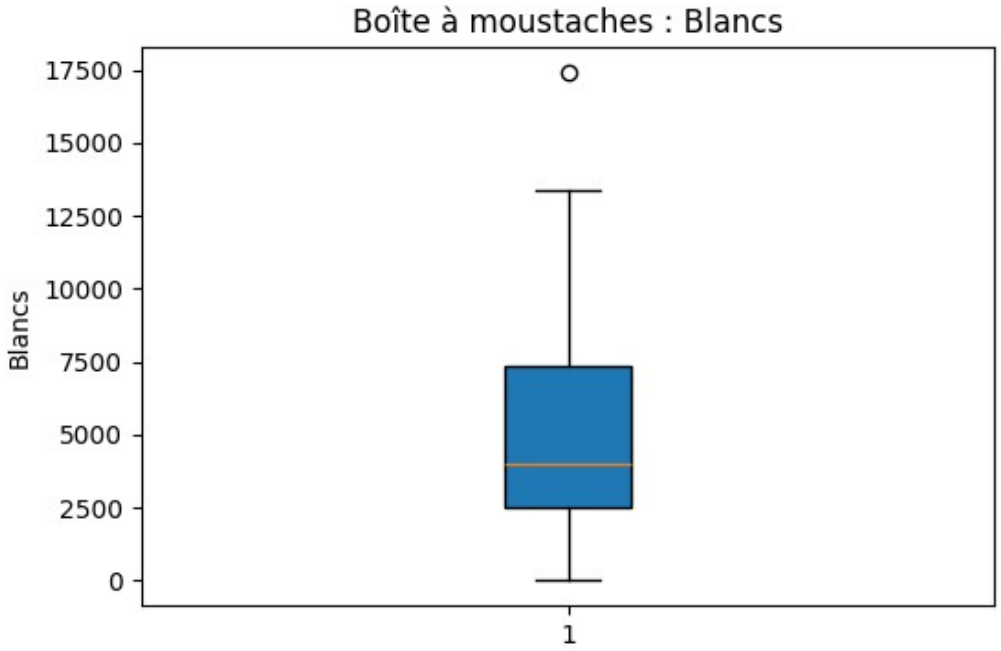
Il faut vérifier la symétrie d'une distribution pour choisir les indicateurs (de tendance centrale, de dispersion, de forme…) adaptés et comprendre la forme de la distribution. On peut le vérifier en s'intéressant à l'asymétrie. Si elle est proche de 0, cela signifie que la distribution est symétrique. Si l'asymétrie est significativement éloignée de 0, la distribution est alors asymétrique.

Données, graphiques, tableaux



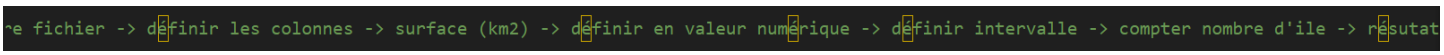
Commentaire de résultats graphiques : Ce graphique représente la boite à moustaches du nombre d'abstentions par département lors du premier tour de l'élection présidentielle de 2022. La médiane

se situe dans la partie inférieure de la boîte, ce qui indique que plus de la moitié des départements comptent un nombre d'abstentions relativement modéré, elle se situe aux alentours des 80 000 à 100 000 abstentions. On voit également que le taux d'abstentions varie d'un endroit à un autre. Il y également une forte asymétrie, comme on le voit les moustaches vont vers les valeurs élevées, elles s'étendent de quelques milliers à 150 000. La boite à moustache résume de façon synthétique la distribution en identifiant la variabilité, et en repérant les valeurs extrêmes.



Commentaire de résultat graphique : Pour la seconde boite à moustaches, celle des votes blancs par département, la médiane se situe autour des 3 500 à 4 000. Il y a une grande variabilité selon la taille du département comme le montre le grand écart entre les deux moustaches

Organigramme.



Séance 4

Question de cours : séance 4

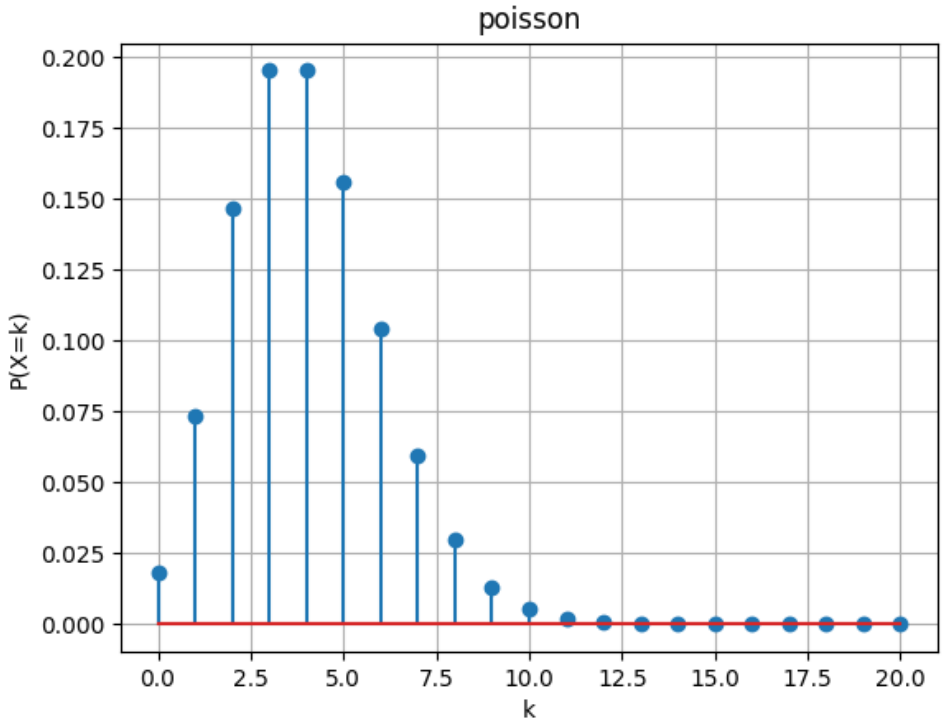
1. Quels critères mettriez-vous en avant pour choisir entre une distribution statistique avec des variables discrètes et une distribution statistique avec des variables continues?

La variable discrète a une valeur finie et entière alors que la variable continue a des valeurs relatives. Les critères qui permettent de choisir une distribution statistique aux variables discrètes et continues sont multiples. En effet ce choix dépend de l'ensemble que l'on veut étudier. Si nous devons dénombrer les élèves d'une classe on utilise des variables discrètes car un élève ne peut être coupé en deux. Si l'on parle du poids de ces élèves par exemple, il est préférable d'utiliser des variables continues. Si les valeurs sont entières on utilise alors les variables discrètes, si elles sont fractionnaires on utilise la distribution continue. L'utilisation de variables continues ou discrètes dans les diagrammes, les histogrammes ou les courbes dépend alors du type de données étudiées. Il faut alors se demander s'il est plus pertinent d'étudier avec des valeurs discrètes ou continues.

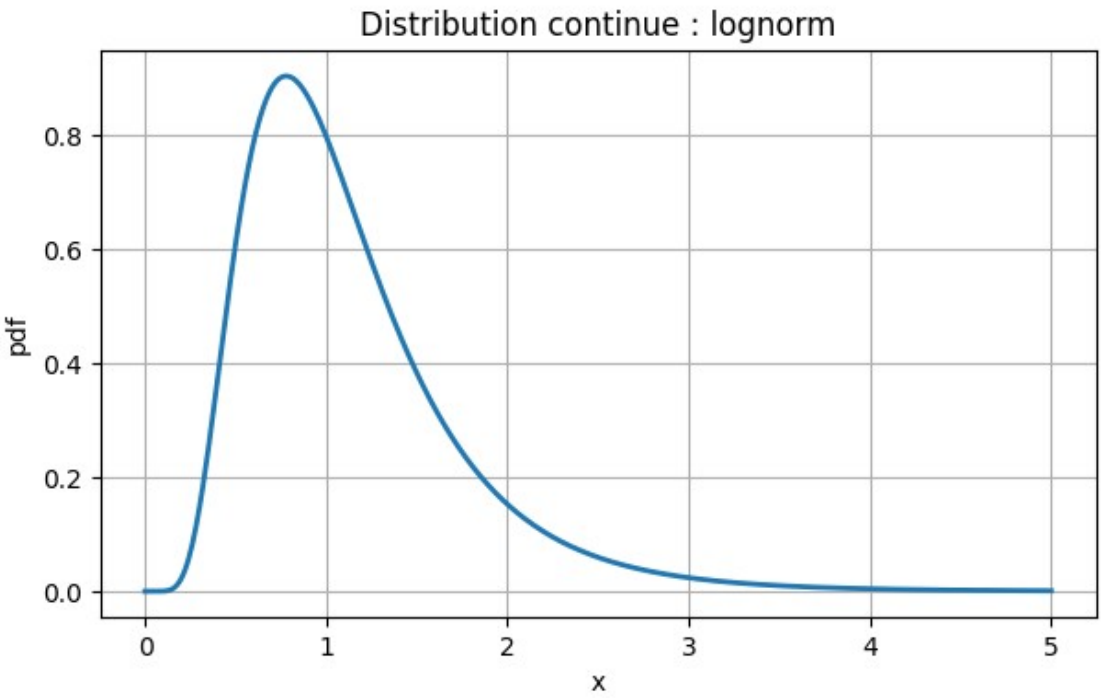
2. Expliquez selon vous quelles sont les lois les plus utilisées en géographie?

Selon moi les lois statistiques les plus spécifiquement liées à la géographie sont la loi de Zipf et sa généralisation, rencontrées dans les lois rang-taille qui confrontent le nombre d'habitants d'une ville avec son rang au sein d'un territoire. Une autre loi testée sur des grandeurs géographiques est la loi de Benford qui s'applique à la longueur des fleuves du globe et à la superficie des pays. Les variables aléatoires concernant le dénombrement de certains objets géographiques comme les lacs et les montagnes ne suivaient pas une loi normale ce qui montre la spécificité des distributions géographiques. Ce sont les lois qui semblent les plus utilisées en géographie et qui s'y appliquent directement.

Données, graphiques, tableaux



Commentaire de résultats graphiques : Ce graphique représente la loi de Poisson, qui est une distribution statistique discrète. Elle modélise le nombre d'occurrences d'un événement aléatoire sur un intervalle donné, lorsque ces événements sont rares et indépendants. On remarque que la probabilité augmente pour les petites valeurs de k. Elle atteint un maximum autour de k = 3 ou k = 4, puis diminue progressivement. Cela indique que le paramètre λ de la loi de Poisson est proche de 4, puisque la moyenne et la variance d'une loi de Poisson sont toutes deux égales à λ. La distribution est asymétrique à droite. Les valeurs élevées de k ont des probabilités de plus en plus faibles, mais elles ne sont pas vraiment nulles. Ainsi, on voit que les barres verticales montrent que la probabilité est concentrée sur un nombre fini de valeurs.



Commentaire de résultats graphiques : Ce graphique représente la loi log-normale, qui est une distribution statistique continue. Une variable suit une loi log-normale lorsque son logarithme suit une loi normale. La distribution est fortement asymétrique à droite. La queue de distribution s'étend vers les grandes valeurs de x, ce qui indique que des valeurs élevées sont possibles mais rares. Ainsi on voit qu'il est différent du premier graphique, il montre que les valeurs proches de 1 sont plus probables.

Séance 5

Questions de cours :

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

L'échantillonnage désigne une sélection d'individus au sein d'une population. C'est un sous-ensemble de la population visée par l'étude. L'intérêt de l'échantillonnage est de pouvoir à l'aide d'un nombre plus faible d'individus déterminer des caractéristiques généraux sur la population ce qui représente un gain de temps, d'accessibilité (on ne peut interroger une population entière) considérable et d'un coût souvent moindre. Il y a plusieurs méthodes d'échantillonnage. Le choix de la méthode est fait de sorte à ce que l'échantillon soit représentatif de la population. Il faut alors choisir deux méthodes : les méthodes probabilistes ou aléatoires et les méthodes non probabilistes qui sont non aléatoires. L'intérêt est de parvenir à un échantillon non biaisé. Le choix de la méthode d'échantillonnage est alors nécessaire car l'intérêt d'un échantillon est de pouvoir obtenir un résultat fiable pour toute la population.

Ainsi les méthodes sont de deux natures : aléatoires (non-biaisées) et non aléatoires (biaisées). Les méthodes aléatoires se font sans aucun calcul, chaque individu a la même chance de faire partie de l'étude. La sélection se fait au hasard. Par exemple tirer au sort des personnes avec un logiciel, ou aller voir chaque personne qui passe dans une zone d'étude déterminée à l'avance dans le cadre d'un questionnaire. Il peut être systématique, par exemple interroger un individu sur 5. L'échantillonnage peut également se faire de manière stratifiée, donc diviser en groupes distincts les individus pour en sélectionner certains au hasard dans chaque groupe. Il existe d'autres méthodes d'échantillonnage probabilistes.

Ensuite, la méthode non aléatoire ne se font pas au hasard. L'échantillonnage peut se faire par quotas, séparer la population en pourcentage équivalent de sorte à avoir une parfaite égalité (autant d'hommes que femmes). Elle peut se faire à l'aide de contacts, interroger des personnes ciblées donc par cooptation.

Ainsi l'échantillonnage est nécessaire dans une étude statistique. Il rend l'étude possible. On peut vérifier les résultats de l'échantillon avec le principe de vraisemblance.

2. Comment définir un estimateur et une estimation?

Un estimateur concerne une variable aléatoire, c'est une fonction de données. Il permet d'observer et d'estimer une caractéristique, en étant aussi proche que possible de la valeur du paramètre. Son intérêt est de permettre de déterminer des estimations ponctuelles. C'est un outil pour l'estimation.

L'estimation permet d'estimer les paramètres d'une loi de probabilité, autrement dit utiliser des données étudiées pour observer des caractéristiques générales.

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

L'intervalle de fluctuation est utilisé dans une démarche d'échantillonnage. Le paramètre de fluctuation est utile quand on connaît la population pour vérifier la façon dont l'échantillon peut varier autour du paramètre connu. L'intervalle de confiance quant à lui n'est pas utilisé quand on connaît toute la population, il s'intéresse particulièrement à l'échantillon. À partir de cet échantillon on essaye d'estimer les paramètres d'une population. Ainsi, le premier suit une démarche déductive car il examine la position de l'échantillon par rapport à une population dont le paramètre est connu. Le second relève d'une démarche inductive car il utilise les informations issues de l'échantillon pour estimer le paramètre de la population entière, il sert à relever l'erreur probable de l'estimation.

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

Un biais est une erreur dans la théorie de l'estimation ($Biais=E(\hat{\theta})-\theta$). Ainsi si le résultat de l'estimateur est loin de la valeur réelle du paramètre, on peut affirmer qu'il est biaisé. L'estimation est alors sans biais quand l'estimateur donne en moyenne la bonne valeur du paramètre.

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

Il y a plusieurs statistiques qui travaillent de façon générale sur la population totale. Il y a notamment le recensement qui est une opération statistique de dénombrement général d'une population (Géoconfluences), c'est une statistique exhaustive. Il permet de dénombrer la population

d’un pays par exemple. On étudie cette population grâce à un paramètre qui sont des nombres qui doivent résumer l’information d’une variable statistique quantitative. L’enquête est alors exhaustive. On peut faire le lien avec la notion de données massives, car elles fonctionnent comme un recensement massif de données qui permettent de visionner les paramètres. ON peut calculer le paramètre réel grâce à la quantité de données. Les données massives permettent une compréhension comme le recensement des tendances et des modèles. Cependant, il est difficile d’étudier toute une population totale, ainsi l’échantillonnage est privilégié.

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

L’intérêt d’un estimateur est qu’il soit le plus proche possible de la valeur du paramètre, il faut alors sélectionner le meilleur estimateur pour que l’estimation ne soit pas biaisée. L’échantillon fournit une information partielle, l’estimateur doit pouvoir minimiser les erreurs. L’estimateur doit être sans biais sinon il introduit une erreur systématique. Cependant l’absence de biais ne veut pas dire que l’estimateur est pertinent, il faut également qu’il ait une variance minimale parmi tous les estimateurs qui ne sont pas biaisés. On peut calculer la précision d’un estimateur avec l’« Erreur Quadratique Moyenne » qui «correspond à la somme de la variance de l'estimateur et du carré de son biais ». L’estimateur doit aussi être convergent, robuste. Ainsi le choix de l’estimateur ne se fait pas au hasard, il est entouré de plusieurs conditions. Si un estimateur ne réunit pas ces nombreuses conditions, le résultat de l’estimation de l’échantillon sera loin des paramètres réels de la population. L’estimation sera parsemée d’erreurs.

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

Il existe plusieurs méthodes d’estimation d’un paramètre. Il y a d’abord la méthode des moindres carrés qui est utile lorsque l’on souhaite étudier plusieurs variables. Elle consiste à choisir les paramètres qui minimisent la somme des carrés des résidus. Ensuite, la méthode du maximum de vraisemblance est une approche générale qui trie les différentes valeurs du paramètre selon leur probabilité. Cela permet de maximiser la fonction de vraisemblance. D’autres méthodes plus classiques sont utilisées comme l’estimation ponctuelle qui consiste à fournir une valeur unique ou l’estimation par intervalle de confiance en fournissant un intervalle avec une forte probabilité de contenir un paramètre réel.

On sélectionne l’une de ces méthodes en sélectionnant en réalité le meilleur estimateur. Il faut alors soit chercher des statistiques qui résument l’information de l’échantillon sans perte en évitant ainsi les biais et les variances. On peut également utiliser l’étude de la quantité d’information de Fisher qui mesure la quantité d’information contenue dans l’échantillon sur le paramètre. Les choix qui guident vers le meilleur estimateur sont dans la réponse précédente.

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

Les analyses statistiques permettent de déterminer la signification d'un effet, d'une différence ou de l'adéquation à une loi à partir d'un échantillon. On identifie diverses sortes de tests : paramétriques (t de Student, F de Fisher), non paramétriques (Mann-Whitney, Wilcoxon), d'ajustement (Khi², Shapiro-Wilk), et d'indépendance (Khi²).

Pour concevoir un test, on énonce une hypothèse nulle et une alternative. Il faut déterminer un seuil de risque, choisir une statistique de test, on établit la dimension critique, puis il faut calculer et comparer pour décider si l'on doit accepter ou rejeter l'hypothèse.

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

La statistique inférentielle, et en particulier les tests d’hypothèses, est souvent critiquée pour plusieurs raisons. Les principales critiques concernent les hypothèses nulles souvent irréalistes, l’influence de la taille de l’échantillon sur la significativité, le manque de puissance avec des petits échantillons, la mauvaise interprétation du non-rejet de l’hypothèse nulle, la confusion entre valeur et probabilité que l’hypothèse nulle soit vraie, la difficulté de réaliser des méta-analyses ou encore la fixation excessive sur la significativité statistique.

Je pense que ces critiques peuvent s’entendre dans le sens où la statistique inférentielle produit des résultats qui ne sont pas fiables entièrement car il y a une marge d’erreur dans l’estimation. Cependant la statistique inférentielle est selon moi nécessaire car elle permet la multiplication d’études possibles. Nous avons pu démontrer grâce aux questions qu’elle est nécessaire, car il est rare de nos jours de faire des études d’une population entière. L’échantillonnage permet de rendre les études pratiques, rapides et moins coûteuses, tout en donnant des résultats proches de la réalité grâce à des méthodes fiables.

En effet, il y a certes une marge d'erreur dans ces statistiques, mais il y a des moyens qui permettent de réduire cette marge d'erreur comme on l'a montré en choisissant le meilleur estimateur à l'aide de calculs comme l'« Erreur Quadratique Moyenne ». Il faut cependant bien comprendre et interpréter correctement les résultats pour éviter les pièges liés à la p-valeur, au rejet de H_0 ou à la surinterprétation de la significativité comme le montre les critiques. Je pense qu'il ne faut pas penser de façon absolument dichotomique, la statistique inférentielle est très utile et permet à la science d'avancer, mais elle doit effectivement pouvoir réduire sa marge d'erreur au maximum. C'est pour cela qu'elle doit être le résultat de multiples vérifications rigoureuses.

Données, graphiques, tableaux

```
Moyennes arrondies des 3 opinions :
Pour            391.0
Contre          416.0
Sans opinion    193.0
dtype: float64

Fréquences estimées à partir des moyennes :
Pour            0.39
Contre          0.42
Sans opinion    0.19
dtype: float64

Fréquences de la population mère :
Pour            0.39
Contre          0.42
Sans opinion    0.19
dtype: float64
```

Commentaire de résultats graphiques : Ce tableau affiche les moyennes arrondies dérivées de 100 échantillons aléatoires représentant une enquête d'opinion, avec une moyenne de 391 individus « Pour », 416 « Contre » et 193 « Sans opinion ». Les fréquences associées estimées sont 0,39, 0,42 et 0,19 respectivement, ce qui s'aligne presque parfaitement sur les fréquences réelles de la population mère à deux décimales près. Ce résultat indique que la moyenne des échantillons offre une estimation remarquable de la structure véritable de la population.

```
Intervalle de fluctuation à 95% pour chaque opinion :
Pour : [0.36 ; 0.42]
Contre : [0.39 ; 0.45]
Sans opinion : [0.17 ; 0.21]
Résultat sur le calcul d'un intervalle de confiance
Théorie de la décision
Résultat sur le calcul d'un intervalle de confiance

Fréquences du premier échantillon :
Pour : 0.4
Contre : 0.4
Sans opinion : 0.21

Intervalle de confiance à 95% pour le premier échantillon :
Pour : [0.37 ; 0.43]
Contre : [0.37 ; 0.43]
Sans opinion : [0.18 ; 0.24]
```

Commentaire de résultat graphique : Dans cette distributions, les plages de variation à 95 % indiquent que les fréquences déduites des échantillons sont cohérentes avec celles de la population d'origine, confirmant ainsi la représentativité de l'échantillonnage. Les fréquences notées dans le premier échantillon se rapprochent des valeurs théoriques et leurs intervalles de confiance englobent les proportions effectives. On peut donc déduire que, malgré les fluctuations aléatoires, les résultats obtenus sont statistiquement fiables. Les techniques d'évaluation et de prise de décision employées garantissent une confiance dans les déductions basées sur ces échantillons.

```
Shapiro-Wilk for Loi-normale-Test-1.csv :
Statistic = 0.9639, p-value = 0.0000
-> On rejette H0 : distribution non normale.
Théorie de la décision
PS C:\Users\Matthieu\AppData\Local\Programs\Microsoft VS Code>
```

Commentaire de résultats graphiques : On a procédé à l'application du test de normalité de Shapiro-Wilk sur les deux fichiers pour déterminer s'ils obéissent à une distribution normale. Pour le fichier Loi-normale-Test-1.csv, la valeur p calculée est largement inférieure à 0,05, entraînant le rejet de l'hypothèse nulle de normalité : on conclut donc que la distribution n'est pas normale. Cependant, en ce qui concerne Loi-normale-Test-2.csv, la valeur p dépasse le seuil de 5 %, ce qui rend impossible de refuser l'hypothèse de normalité. Nous déduisons donc que Loi-normale-Test-2.csv suit une loi normale, contrairement à Loi-normale-Test-1.csv.

Séance 6

Questions de cours :

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?

Une statistique ordinale selon l’ESRI est « Méthode d'organisation des valeurs de données en une hiérarchie classée sans intervalle fixe entre les niveaux hiérarchiques. » Autrement dit c’est le classement de données qui doit se faire généralement par ordre croissant. C’est un outil statistique puissant puisqu’il peut permettre d’évaluer la force d’un sentiment ou d’une impression comme avec l’échelle de Likert ou de faire des classements en montrant l’évolution montante ou descendante des entités.

On oppose généralement la statistique ordinale à la statistique nominale, et non aux données qualitatives en général. En effet, une donnée qualitative peut très bien être de nature ordinale. Ainsi elle utilise des variables qualitatives ordinales, elles sont constituées de catégories qui sont classées par ordre logique, sans que cet écart soit mesurables contrairement aux données quantitatives.

Elle matérialise donc une hiérarchie spatiale en établissant par exemple des classements d’objets géographiques afin de suivre leur évolution, peut produire des classements de villes par rapport au nombre d’habitants par ordre croissant ou encore en établissant des liens de dépendance. La

statistique ordinale établit donc une hiérarchie spatiale de différentes entités grâce à des critères et caractéristiques mesurables, elle est donc nécessaire en géographie.

2. Quel ordre est à privilégier dans les classifications?

L'ordre qui doit être préféré dans les classifications est l'ordre croissant pour une raison logique mais aussi de visibilité des données, nous pouvons appliquer directement les statistiques de rang comme la médiane, ou les quantiles. Les entités sont souvent présentées en termes d'évolution. Cependant il existe des exceptions. Certaines lois ne s'inscrivent pas dans cette logique d'ordre croissant comme la loi rang-taille qui utilise volontairement un ordre décroissant car la hiérarchie est interprétée du plus important vers le moins important afin de mesurer la hiérarchie des villes.

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

La corrélation des rangs sert à comparer deux classements distincts d'un ensemble d'individus ou d'objets pour démontrer s'ils sont identiques ou non. Les deux tests pour la corrélation des rangs sont les tests de C. Spearman et M. G. Kendall. La concordance de classements quant à elle, est l'outil statistique utilisé pour évaluer l'accord entre plusieurs classements en utilisant 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 la distance au carré entre les rangs pour mesurer la corrélation entre deux classements d'un même ensemble d'objets, cela se fait donc de façon plus directe. Le test de Kendall quant à lui, analyse des paires concordantes et discordantes pour mesurer la concordance entre des classements ce qui garantit une généralisation à plusieurs classements. Ainsi pour plusieurs classements le test de Kendall est privilégié.

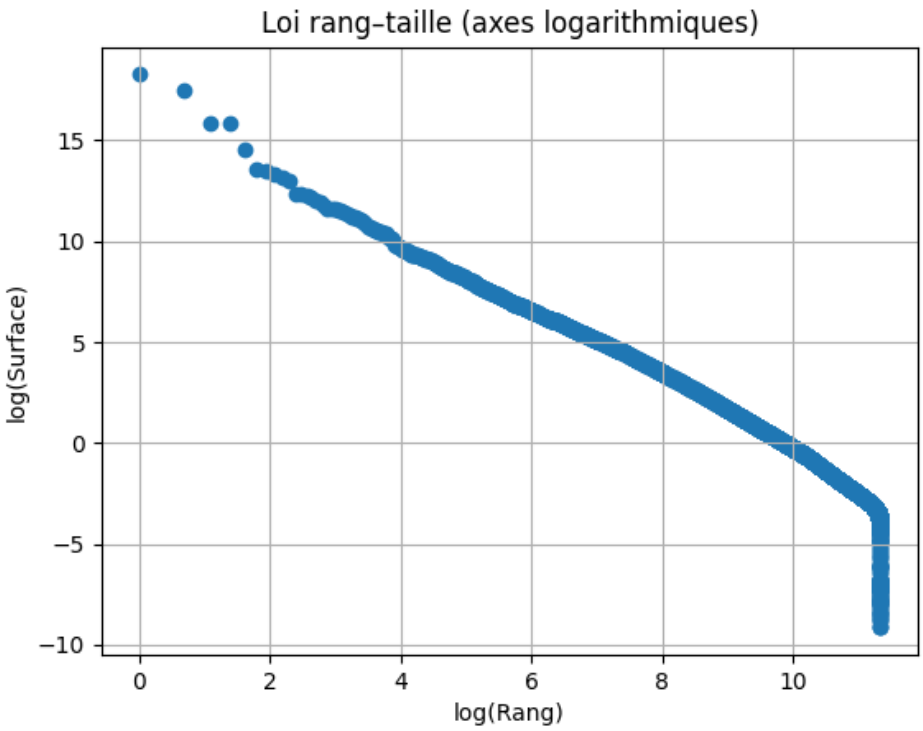
5. Àquoi servent les coefficients de Goodman-Kruskal et de Yule?

Ces coefficients servent à évaluer l’association ou l’indépendance des variables catégorielles.
Le coefficient de Goodman-Kruskal mesure l’association entre paires concordantes et discordantes, son calcul est fondé sur une différence possible entre le nombre de paires concordantes et le nombre de paires discordantes. Il calcule grâce à cela le « surplus » de paires concordantes par rapport aux paires discordantes. Le coefficient de Yule est « un cas particulier du Goodman-Kruskal ». Il se fonde sur les écarts des quartiles et détermine si deux variables binaires sont associées, indépendantes ou inversement associées.

Données, graphiques, tableaux

```
Corrélation des rangs (Spearman) : 0.9862, p-value = 1.0697e-134
Concordance des rangs (Kendall) : 0.9043, p-value = 2.0819e-69
```

Commentaire de résultats graphiques : Les coefficients de corrélation par rang ont été déterminés en utilisant les fonctions spearmanr() et kendalltau() de la bibliothèque Scipy.Stats, en faisant une comparaison entre le classement des pays basé sur leur population et celui basé sur leur densité de population. L'indice de Spearman calculé ($\rho = 0{,}986$) révèle une corrélation positive très marquée entre les deux classements, indiquant que les pays qui ont un classement élevé en termes de population ont également tendance à avoir un bon classement en matière de densité. Le coefficient de concordance de Kendall ($\tau = 0{,}904$) confirme cette constatation grâce à une forte concordance des rangs. Les p-values, qui sont très faibles, liées à ces résultats indiquent qu'ils ont une signification statistique. Il est donc possible de conclure qu'il y a un lien très solide et robuste entre les classements pris en compte.



Commentaire de résultats graphiques : Ce diagramme illustre la loi de rang-taille des surfaces suite à l'organisation des territoires du plus grand au plus petit, avec une transformation des axes en échelle logarithmique. L'axe horizontal illustre le classement, alors que l'axe vertical indique la superficie. On constate une baisse graduelle des valeurs à mesure que le rang s'élève : les premiers éléments possèdent des surfaces très étendues, qui ensuite régressent rapidement. Utiliser une échelle logarithmique facilite la lecture du graphique, étant donné l'importance significative de la différence entre les grandes et petites surfaces. L'aspect général du nuage de points indique une tendance à la baisse, sans toutefois être parfaitement en ligne, ce qui laisse penser que le classement a plus d'importance que les valeurs précises. Cette sorte de représentation souligne l'importance des rangs et justifie l'utilisation de tests basés sur les classements pour effectuer des comparaisons entre ces types de données.

Réflexion personnelle sur les sciences des données et les humanités numériques

Comme beaucoup de mes camarades je ne connaissais rien à Python. J'ai fait 3 ans de CPGE littéraire et je suis ensuite directement entré directement en Master. Les mathématiques n'ont jamais été faits pour moi. Ainsi ce semestre j'ai dû me confronter à python. Les sciences des données et humanités numériques ne sont pas simple de compréhension. Les nombreuses formules ont eu tendance à me fournir des maux de crânes, je ne suis pas sûr d'en retenir vraiment beaucoup mise à part les indicateurs de base que je ne maîtrisais pas auparavant comme la médiane, l'écart-type ou les variances. J'ai pu comprendre et approfondir les paramètres de dispersions ou les variances. J'ai même réussi à retenir certaines commandes et cela devenait plus simple au fur et à mesure même si je suis loin d'avoir obtenu un niveau convenable. Les autres formules étaient assez abstraites pour moi, j'en comprenais certaines grâce à des explications dans votre cours ou sur internet mais d'autres restaient obscures. Cependant, même si c'était fastidieux, j'ai vu le potentiel des sciences des données et des humanités numériques. Les sciences des données permettent d'analyser des phénomènes complexes, de comprendre des réalités sociales, ou politiques difficile à saisir. L'outil de visualisation permet d'avoir une vision synthétique des données qui s'étudie cette fois-ci assez facilement. J'ai aussi vu un sens nouveau à toutes ces données que les institutions statistiques fournissent elles sont en réalité toutes utilisables dans différents cas. On ne décrit plus, on analyse en profondeur diverses données, et on peut évidemment appliquer tout cela à la géographie. Les humanités numériques permettent de travailler sur un grand volume de données en valeur avec les sciences sociales, c'est donc un outil nécessaire pour nous géographes. Cependant je me demande si il faudrait obligatoirement une généralisation de ces compétences. En effet, la plupart des géographes sont capables d'analyser des tableaux, des graphiques, est-ce qu'il est réellement nécessaire qu'ils sachent en produire eux-mêmes? J'ai lu plusieurs articles et thèses en géographie et je n'ai jamais vu ou alors remarqué des calculs statistiques aussi poussés.

J'ai cependant été témoin de la puissance de l'outil python. Avec quelques (ou plusieurs) lignes de code, Python peut nous fournir une quantité de données infinie, elles sont analysées, triées avec une grande rapidité. Sa capacité à générer plusieurs images ou cartes est assez impressionnante (et satisfaisante par la même occasion). D'autant plus, je pense qu'en tant que débutant, je n'ai vu qu'une infime partie des capacités de Python. Je vais donc pour la suite essayer de continuer à utiliser un peu ce langage Python, il pourrait m'aider pour certaines tâches. Je pense cependant, que je ne deviendrai jamais *data analyst* ou développeur.

Avis personnel sur le cours

Le cours d'analyse de données a été une véritable épreuve pour moi et pour beaucoup de mes camarades. Je dois avouer qu'il était source de stress durant tout le semestre. Ce stress s'explique par la sensation de ne pas pouvoir le faire, et par la grande quantité de travail demandée. Les exercices de code étaient très complexes, loin selon moi d'un niveau débutant. Peut-être aurait-il fallu commencer par des exercices plus simples ? De plus nous étions assez livrés à nous-mêmes durant le semestre. Je sais que vous étiez disponibles pour répondre aux questions sur le Discord mais il est difficile de régler un problème à distance car dans mon cas je ne suis pas doué en informatique. Le problème en classe, était que vous étiez souvent occupés avec d'autres élèves, ce qui ralentissait ma progression car je n'arrivais pas à résoudre un problème. J'étais alors obligé de patienter. Je trouve que la matière m'a pris beaucoup trop de temps dans le semestre, même si c'est peut être dû à ma lenteur dans les exercices, elle a pris un peu trop de place sur mes autres matières qui correspondaient davantage à mon projet professionnel. Il aurait peut être fallu nous donner des cours directement afin de bien comprendre les commandes de base, et comprendre réellement le langage Python. Mais vous avez sûrement vos raisons d'avoir effectué ce choix de pédagogie inversée. Cela ne doit pas être évident d'enseigner directement la langage Python à des élèves ayant pour la plupart reçues une formation littéraire mais je pense que cela aurait été préférable. Aussi je sais que vous avez beaucoup donné dans la production des pages de cours et des exercices. Je ne suis finalement peut-être pas devenu développeur, mais j'ai sincèrement fait ce que j'ai pu pour parvenir aux exigences de la matière.