Rapport d'activité

Analyse des données - Parcours Débutant

ALFARO Karla

Master 1 GAED, Parcours GéoSuds - sociétés, territoires, développement

Séance 1

Objectif Séance

- Comprendre le cadre du cours et ce qu'il vise à nous faire apprendre
- Prendre en main les outils mobilisés : Python, GitHub et Docker.
- Mettre en place l'environnement de travail et finaliser les installations nécessaires pour le semestre.

Outils utilisés:

- VS code
- Github
- ChatGPT et Grammarly - Pour traduction et explication des termes (étant étudiante étrangère, j'ai eu besoin de traduire à l'anglais et espagnol pour comprendre quelques questions et consignes à faire).
- Reddit Forum - Je me suis mise à la recherche des forum pour rétro alimenter et comprendre mes codages.

Apprentissages et sélection du parcours

J'ai choisi le niveau débutant afin de consolider les bases et de mieux maîtriser les notions à acquérir. Même si j'avais déjà une première expérience avec le code, j'ai préféré repartir d'un cadre plus fondamental pour comprendre les outils en profondeur, plutôt que d'avancer trop vite de manière superficielle.

Par ailleurs, comme il s'agissait de mon premier semestre à la Sorbonne, j'ai aussi fait le choix de m'aligner sur le niveau sélectionné par plusieurs camarades de mon parcours. Cela m'a permis de faciliter le travail en équipe, d'échanger sur différentes solutions possibles pour un même script, et de demander de l'aide plus facilement lorsque je bloquais.

J'ai bénéficié d'un soutien régulier de Zara Huston, qui m'a aidée à la fois sur la compréhension linguistique des consignes, la compréhension de rédaction de code et sur la rédaction de ce rapport, ce qui a été déterminant dans mon parcours.

Enfin, il est important de préciser que j'ai commencé cette séance avec un grand retard, car mon ordinateur personnel ne disposait pas de l'espace de stockage nécessaire. J'ai donc dû attendre de recevoir un ordinateur prêté par la Sorbonne pour pouvoir installer correctement l'environnement de travail. Malgré ce contretemps, la séance a été menée à terme et l'ensemble des outils requis a bien été installé.

Séance 2

Objectifs Séance

- Manipuler un fichier C.S.V.

- Faire des sorties graphiques
- Utiliser les bibliothèques Pandas (données) et Matplotlib (graphiques)
- Calculer des effectifs
- Calculer des fréquences
- Faire des graphiques (diagrammes en bâton et circulaires, et histogrammes)

Questions

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

A partir du texte il existe une relation très ambivalente entre la géographie et la statistique : à la fois une méfiance face à une dépendance pratique.

D'un côté, la géographie est décrite comme une discipline qui doute encore de son statut scientifique et qui garde une distance historique face aux mathématiques, chose qui conduit certains géographes à sous-estimer ou mal utiliser l'outil statistique.

D'un autre côté, elle travaille pourtant sur des quantités importantes de données, souvent massives, qui ne peuvent être analysées rigoureusement qu'à l'aide de méthodes statistiques.

A cause de cela, il y a donc un paradoxe : la géographie a besoin de la statistique pour étudier des phénomènes territoriaux, mais elle n'a pas toujours intégré cette exigence dans sa formation.

« La géographie est une discipline qui se cherche toujours. Il est fréquent qu'elle méprise les définitions mathématiques élémentaires de la statistique sous prétexte que cela n'entre traditionnellement pas dans son champ disciplinaire. Pourtant, elle produit des données massives que seul l'outil statistique permet d'étudier. Ainsi, les relations entre les deux disciplines sont très souvent tendues et complexes. » (p. 1, l. 3-8)

Le texte insiste sur le fait que les statistiques et la connaissance des lois du hasard ne sont pas à juste être considérées comme accessoires dans la géographie : elles sont présentées comme essentielles pour l'avenir de la géographie et pour faire de celle-ci une véritable science de l'information géographique et des échelles, plutôt qu'une simple « méthode » mise à disposition des autres disciplines.

« Les statistiques sont par conséquent essentielles pour l'avenir de la géographie. Connaître les lois du hasard est de fait une nécessité pour comprendre l'information (massive) géographique. » (p. 4)

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

Le texte ne répond pas directement à cette question mais montre que tout dépend du point de vue qu'on décide d'adopter.

Dans une perspective déterministe classique, inspirée et basée sur la physique, le hasard n'existe pas : chaque événement a une cause, même si elle n'est pas observable, et ce qui ressemble au hasard n'est qu'un « bruit » négligeable dans l'application des lois.

« Toute science recherche à réduire les éléments hasardeux au maximum. Cette position est défendue par le mouvement philosophique du déterminisme (...) Le hasard n'existe pas : il existe une cause à tout. » (p. 2)

À l'inverse, une grande partie des non-statisticiens et de nombreux géographes considèrent que le hasard est à l'origine de tout. Ce qui, poussé à l'extrême, rendrait impossible l'ambition de faire de la géographie une science.

« Avant de poursuivre, il est important de souligner que la majorité des non statisticiens affirment haut et fort que le hasard reste à l'origine de tout chose. Beaucoup de géographes défendent cette position. Un tel hasard aboutit au fait qu'il est impossible de faire de la géographie une science à proprement dit. » (p. 3)

Le hasard, dans ce sens à partir du texte, n'est pas tant une "réalité métaphysique" qu'une façon de penser l'incertitude : on ne renonce pas à connaître, mais on considère la connaissance des faits vers des régularités statistiques plutôt qu'une prédiction exacte de chaque cas.

3. Quels sont les types d'information géographique ?

En géographie, nous parlons souvent d'information géographique pour désigner un ensemble plus large que les simples statistiques. L'objectif est de regrouper les données en fonction des objets géographiques étudiés. Cette information géographique se décompose en deux grands types complémentaires.

Le premier type correspond aux données qui décrivent les caractéristiques d'un territoire donné : ce sont les informations thématiques sur la population, les structures sociales, les activités économiques, ou encore les paramètres physiques comme la température ou les précipitations. Ce sont ces données qui, dans un Système d'Information Géographique (SIG), constituent la base attributaire, c'est à dire les attributs attachés à chaque entité spatiale.

« L'information géographique se décompose en deux séries statistiques possibles. D'une part, il peut s'agir pour une entrée territoriale claire et précise d'étudier tout ce qui peut caractériser l'ensemble délimité par des éléments de géographie humaine (population humaine, caractéristiques sociales, caractéristiques économiques, etc.), ou de géographie physique (température, volume des précipitations, etc.). » (p. 4)

Le second type concerne la forme même des objets géographiques : la morphologie et la géométrie des ensembles délimités (limites, surfaces, formes, réseaux, etc.). Ces informations géométriques peuvent elles aussi faire l'objet d'analyses statistiques.

« D'autre part, il peut s'agir d'étudier la morphologie même des ensembles délimités. De fait, la géométrie des ensembles géographiques peut faire l'objet d'une étude statistique. » (p. 4)

D'une autre manière, l'information géographique comprend à la fois ce que l'on sait sur un lieu (ses propriétés) et la manière dont ce lieu est dessiné et organisé dans l'espace (sa géométrie), et c'est l'articulation des deux qui permet une véritable analyse spatiale.

« Dans le cadre d'un système d'information géographique (S.I.G.), elles définissent la base attributaire (ou les attributs). Les secondes caractérisent les données géométriques du S.I.G. » (p. 4)

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

Pour la géographie, l'analyse de données n'est pas une rareté technique mais une nécessité pour ne pas se laisser déborder par la masse d'informations produites sur les territoires.

Le texte rappelle que le géographe ne peut « rester passif face à l'abondance et la massification des données » : il doit être capable de les organiser, de les critiquer et de les interpréter, notamment en travaillant sur les nomenclatures et les métadonnées qui encadrent ces données.

Cela implique d'abord une compétence sur tout ce qui entoure la production des données (comment elles ont été construites, selon quelles définitions, avec quels biais possibles), puis une compétence proprement mathématique pour « étudier la structure interne des données analysées ».

L'analyse de données, fondée sur les probabilités et les statistiques, est décrite comme « le moment mathématique » du travail géographique, celui où l'on passe du simple constat à la mise en évidence de structures, de régularités ou de anomalies.

Enfin, pour que cette analyse ait du sens, elle doit toujours être confrontée à la méthodologie de production des données et aux connaissances substantives sur le phénomène étudié : la statistique ne remplace pas la réflexion géographique, elle la soutient. C'est de cette manière que le texte insiste sur le fait que les statistiques sont essentielles pour l'avenir de la géographie et que connaître les lois du hasard devient une condition pour comprendre l'information géographique massive contemporaine.

« Les statistiques sont par conséquent essentielles pour l'avenir de la géographie. Connaître les lois du hasard est de fait une nécessité pour comprendre l'information (massive) géographique. » (p. 4)

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

Le texte fait distinction entre deux grandes manières d'utiliser les statistiques, qui correspondent à l'opposition entre statistique descriptive et statistique mathématique:
« Les statistiques se divisent en deux branches : 1. la statistique descriptive qui consiste à étudier des données. L'objectif est de dégager des propriétés remarquables par rapport à une distribution théorique connue. Cela permet ainsi d'obtenir une image simplifiée de la réalité en mettant de l'ordre dans les données (caractéristiques numériques ou graphiques). 2. la statistique mathématique dont l'objectif est de prédire à partir des statistiques descriptives, et surtout de la distribution de probabilités théorique que l'on a établie, des scénarios possibles.» (p. 6–7)

La statistique descriptive a pour objet la description d'un ensemble de données : elle cherche à étudier des données, à en dégager des propriétés remarquables et à produire une image simplifiée de la réalité en mettant de l'ordre dans les distributions à l'aide de paramètres numériques et de représentations graphiques. Elle s'intéresse à l'ensemble des variables d'un tableau sans qu'il y ait "variable à expliquer", son objectif est résumer, visualiser et de classer les données, et aussi comme préparer le terrain pour des comparaisons ou des prédictions finales.

D'un autre côté, la statistique mathématique vise explicitement à modéliser une relation entre une variable à expliquer (Y) et une ou plusieurs variables explicatives. On utilise alors d'un tableau individus × variables, et il s'agit « d'ajuster les données disponibles » à un modèle dont la forme dépend de la nature de la variable réponse : numérique ou qualitative. On passe comme ça de la simple description à la recherche de relations causales ou de dépendance, à travers des outils comme la régression, l'analyse de la variance ou les méthodes discriminantes.

En conclusion, la statistique descriptive répond surtout à la question "à quoi ressemblent mes données ?", alors que la statistique explicative répond à comment une variable Y dépend-elle d'autres variables X.

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

Par le texte, la visualisation des données est pensée à plusieurs niveaux, des graphiques simples ou d'analyse multidimensionnelle. Le choix du type de représentation dépend beaucoup de la nature des variables étudiées.

Au niveau le plus simple, la description statistique propose des "visualisations graphiques" adaptées au type de variable

- L'histogramme est recommandé pour les variables continues, il permet de représenter la distribution d'une quantité le long d'un axe
- La représentation sectorielle (diagramme en secteurs ou camembert) est utilisée pour les variables qualitatives afin de visualiser chaque partie.
- Les tableaux de synthèse accompagnent les graphiques en rassemblant les paramètres essentiels et les intervalles de confiance.

Quand on travaille sur plusieurs variables à la fois, ils existent des méthodes de visualisation plus avancées:

- L'analyse factorielle en composantes principales (ACP) pour visualiser des données quantitatives,
- L'analyse factorielle des correspondances (AFC) pour représenter un tableau de contingence entre variables qualitatives
- L'analyse factorielle des correspondances multiples (ACM) quand il y a plus de deux variables qualitatives.

Pour choisir quel type de graphique il faut considérer le type de variables (quantitatives, qualitatives, mixtes), le nombre de variables (une, deux ou plusieurs) et l'objectif. Cela peut être décrire une distribution, comparer des catégories, ou explorer la structure d'un phénomène complexe. Avec cela c'est qu'on choisit le type de visualisation à représenter l'information.

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

Il y a trois méthodes: les méthodes descriptives; les méthodes explicatives et les méthodes d'explication.

- **Méthodes descriptives:** servent uniquement à décrire les données. Elles s'appliquent à des tableaux où toutes les variables jouent le même rôle et il n'y a pas de variable à expliquer. Son objectif est de visualiser et de classer les données avec des caractéristiques numériques ou graphiques pour donner une image simplifiée de la réalité.
- **Méthodes explicatives:** Il y a une *variable à expliquer* et des *variables explicatives*. L'objectif n'est pas seulement décrire, mais aussi à mettre en évidence une relation entre les données.
- **Méthodes de prévision:** Ces méthodes travaillent avec des données dans le temps. On cherche à prévoir une valeur future avec la construction d'un modèle en relevant des séries à travers le passé et les répétitions des données.

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 ?

- La population statistique est l'ensemble des individus sur lesquels porte l'étude.
- L'individu statistique est une unité de cet ensemble.
- Le caractère statistique, qui représente les propriétés observées.

Il y a deux grands types de caractères: les caractères qualitatifs, qui expriment une qualité ou une catégorie, et les caractères quantitatifs, qui expriment une grandeur mesurable.
Les quantitatifs se divisent encore en discrets (valeurs entières) ou continus (valeurs mesurables dans un intervalle avec décimales).

A travers le texte il s'exprime une hiérarchie: les caractères quantitatifs sont plus riches analytiquement parce qu'ils permettent des opérations mathématiques, alors que les caractères qualitatifs ont besoin des traitements adaptés pour leur analyse. Néanmoins, il est important à dire que tous les types de caractères sont complémentaires, à cause de que la géographie s'intéresse autant aux structures sociales et symboliques (analyse qualitative) qu'aux grandeurs mesurables. (analyse quantitative).

9. Comment mesurer une amplitude et une densité ?

L'amplitude et la densité sont deux indicateurs pour comprendre la répartition des données quantitatives dans un territoire ou un échantillon.

- L'amplitude est la différence entre la valeur la plus élevée et la valeur la plus basse d'un caractère mesuré ; à travers ça elle exprime l'étendue de la variation observée.
- La densité est une mesure de fréquence relative, une quantité observée à une surface, une population ou un intervalle.

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

Les formules de Sturges et de Yule servent à choisir un nombre de classes adapté lorsqu'on regroupe une variable quantitative continue en classes ou catégories.

Le nombre de classes ne doit être ni petit, ni très grand parce que si jamais on ne choisit pas le bon nombre il est possible d'avoir une très grande perte d'informations. Ces formules sont des règles pratiques pour trouver des divisions raisonnables entre précision et lisibilité dans la représentation des données continues.

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

L'effectif d'une modalité est le nombre de fois où cette modalité apparaît dans la population.

- Pour calculer une fréquence, ce que l'on doit faire est diviser l'effectif de la modalité considérée par l'effectif total, ce qui donne la fréquence.
- Pour calculer une fréquence cumulée, il est nécessaire d'additionner les effectifs (ou les fréquences) de toutes les modalités dont la valeur est inférieure ou égale à la modalité considérée, afin d'obtenir l'effectif ou la fréquence cumulée.

On définit une distribution statistique par l'ensemble des modalités d'un caractère et les fréquences associées, ce qui constitue une distribution statistique empirique à partir de laquelle on peut étudier le type de loi de probabilité.

Apprentissages Mise en Oeuvre Codage

- Manipuler un fichier C.S.V.
- Faire des sorties graphiques
- Utiliser les bibliothèques Pandas (données) et Matplotlib (graphiques)
- Calculer des effectifs
- Calculer des fréquences
- Faire des graphiques (diagrammes en bâton et circulaires, et histogrammes)

Réflexion de la séance

Lors de cette séance, j'ai appris à importer un fichier CSV avec Pandas et à explorer sa structure (dimensions, types de variables, noms des colonnes). J'ai ensuite manipulé les données en sélectionnant des colonnes et en calculant des effectifs (ex. total des inscrits) et des sommes sur les variables quantitatives. Enfin, j'ai utilisé Matplotlib pour produire des sorties graphiques : un diagramme en barres (comparaison inscrits/votants par département), un diagramme circulaire (répartition blancs/nuls/exprimés/abstention), et un histogramme (distribution du nombre d'inscrits).

Cette séance m'a aidé à comprendre le lien entre données brutes → calculs → visualisations, et l'importance d'organiser les fichiers de sortie (dossiers d'images) pour automatiser proprement les résultats.

Graphique: Inscrits / Votants / Département

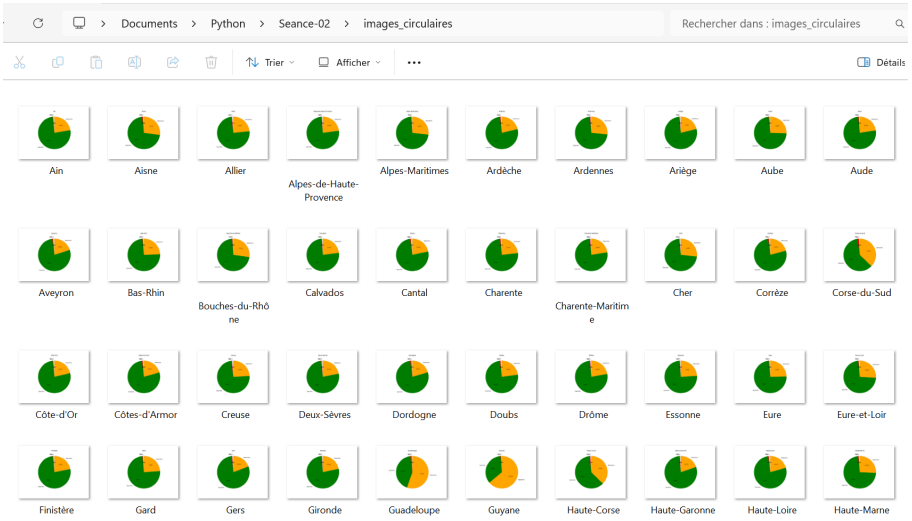
```
#Question 11 : Diagramme en barre nombre inscrits et votants / département (boucle)

for i in range(len(contenu)): #=>en haut = la boucle
    dept= contenu.loc[i, "Libellé du département"]
    inscrits = contenu.loc[i, "Inscrits"]
    votants = contenu.loc[i, "Votants"]
    plt.figure(figsize=(8,6)) #Le diagramme
    plt.bar(["Inscrits","Votants"], [inscrits, votants], color=['blue', 'red'])
    plt.title(f"{dept}")
    plt.ylabel("Nombre de personnes")
    plt.ticklabel_format(style='plain', axis='y')
    #Avoir les noms et pas les valeurs de matplotlib
    plt.savefig(f"{dept}.png")
    plt.close()
```

Sur chaque diagramme, on observe que la barre “Inscrits” est toujours plus élevée que “Votants”, ce qui est logique : tous les inscrits ne votent pas. L’écart visuel entre les deux barres donne une idée immédiate de la participation (écart faible = participation plus forte, écart fort = participation plus faible). On remarque aussi que les départements n’ont pas la même “taille électorale” : certains ont beaucoup plus d’inscrits que d’autres.

Pertinence:
Ce graphique est pertinent pour une comparaison rapide entre départements, mais surtout pour comprendre une relation simple : corps électoral → participation. C’est un bon premier graphique “diagnostic” car il rend la donnée lisible en un coup d’œil, avant de passer à des indicateurs plus fins (taux de participation, etc.).

Graphique diagramme circulaire votes blancs, nuls, votants, abstentions

```python
#Question 12 : diagramme circulaire votes blancs, nuls, votants, abstentions (boucle)

import os
os.makedirs("images_circulaires", exist_ok=True)

for i in range(len(contenu)): #la boucle pour département
    dept = contenu.loc[i, "Libellé du département"]
    blancs = contenu.loc[i, "Blancs"]
    nuls = contenu.loc[i, "Nuls"]
    votants = contenu.loc[i, "Votants"]
    abstention = contenu.loc[i, "Abstentions"]

    # les diff votes exprimés(donc, votants, blanc et les votes nuls)

    exprimés = votants - blancs - nuls

    valeurs = [blancs, nuls, exprimés, abstention]
    labels = ["Blancs", "Nuls", "Exprimés", "Abstention"]
    couleurs = ["lightgrey", "red", "green", "orange"]

    plt.figure(figsize=(8,6))
    plt.pie(valeurs, labels=labels, colors=couleurs, autopct='%1.1f%%', startangle=90) #Les % à afficher les % dans le graphique circulair
    plt.title(f"{dept}")

    #Sauvegarder
    plt.savefig(f"images_circulaires/ {dept}.png")
    plt.close()
```

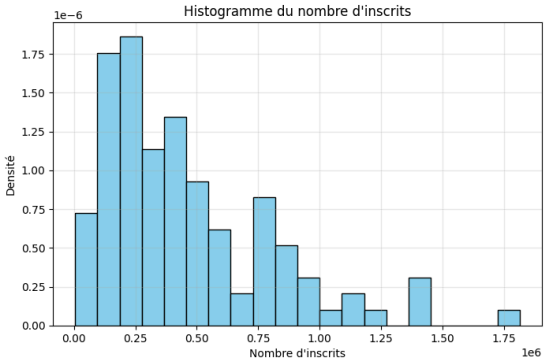


Commentaire sur le graphique obtenu :
Le graphique circulaire met en évidence la structure du scrutin : une part correspond à l'abstention, et le reste regroupe les comportements parmi les votants (votes blancs, nuls et exprimés). Visuellement, la part "Exprimés" est généralement la plus importante parmi les votes, tandis que "Blancs" et "Nuls" sont des parts plus petites, mais intéressantes car elles signalent des formes de participation "non exprimée".

Pertinence / ce que ça apporte :
Ce graphique est utile car il transforme des effectifs en proportions (%), donc il permet de comparer des départements même s'ils n'ont pas le même nombre d'inscrits. Il est pertinent pour discuter de la qualité de la participation : voter ne veut pas toujours dire "exprimer un choix", et l'abstention a un poids politique important.

Histogramme du nombre d'inscrits

```python
#Question 13 : Histogramme

plt.figure(figsize=(8,5))
plt.hist(contenu ["Inscrits"], bins=20, density=True, edgecolor='black', color='skyblue')
plt.title("Histogramme du nombre d'inscrits")
plt.xlabel("Nombre d'inscrits")
plt.ylabel("Densité")
plt.grid(alpha=0.3)
plt.show()
```



Commentaire sur le graphique obtenu :
L'histogramme montre que le nombre d'inscrits par département est très inégal : la majorité des départements se situe autour de quelques centaines de milliers d'inscrits, tandis qu'une minorité de départements atteint plus d'un million, formant une longue queue vers la droite. Cela justifie de comparer ensuite les territoires avec des pourcentages (fréquences) plutôt qu'uniquement avec des effectifs bruts.

Pertinence:
L'histogramme est pertinent car il permet de visualiser la distribution des inscrits et de constater une forte inégalité de taille entre départements. Il sert de diagnostic : comme les effectifs bruts dépendent fortement de la population électorale, l'analyse comparative doit privilégier les taux (fréquences) plutôt que les nombres absolus.

Séance 3

Objectif Séance

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

Questions

1. **Quel caractère est le plus général : le caractère quantitatif ou le caractère qualitatif ? Justifier pourquoi.**
   Après mon interprétation, le caractère qualitatif est le plus général, car il peut correspondre à des données de natures très variées, tandis que le caractère quantitatif repose uniquement sur des valeurs numériques spécifiques.

2. **Que sont les caractères quantitatifs discrets et caractères quantitatifs continus ? Pourquoi les distinguer ?**
   Les caractères quantitatifs discrets ne prennent en compte que des valeurs entières, alors que les caractères quantitatifs continus concernent les nombres réels (entiers, décimaux, fractions). En Python, il faut les distinguer car un entier (type int) et un nombre décimal (type float) ne se traitent pas de la même manière : pour effectuer certains calculs, ils doivent être du même type. Il est néanmoins possible de convertir une valeur d'un type à l'autre dans Python.

3. **Paramètres de position**

   - **Pourquoi existe-t-il plusieurs types de moyenne ?**
     Il existe plusieurs moyennes parce qu'elles donnent des lectures différentes des données et ne traduisent pas les mêmes tendances. Le choix dépend de l'information recherchée et type de réponse cherchée, puisque les données ne sont pas mises en relation de la même façon selon le type de moyenne.

   - **Pourquoi calculer une médiane ?**
     La médiane coupe un ensemble de données en deux parties égales : une moitié inférieure et une moitié supérieure. Elle décrit mieux la série que la moyenne, 'car elle n'est pas affectée par les valeurs extrêmes.
   - **Quand est-il possible de calculer un mode ?**
     On peut le calculer lorsqu'au moins une valeur de la série statistique apparaît plusieurs fois. Il met en évidence la valeur la plus fréquente.

4. Paramètres de concentration

- **Quel est l'intérêt de la médiale et de l'indice de C. Gini ?**
  Son intérêt est de permettre de déterminer un indice de concentration dans une série statistique.

L'indice de Gini sert à décrire la répartition d'une variable au sein d'un groupe et à mesurer la concentration dans une population statistique. Ses valeurs vont de 0 à 1 : 0 correspond à une égalité parfaite et 1 à une situation d'inégalité maximale. L'intérêt est ainsi d'obtenir un indicateur hiérarchisé qui met en évidence la 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 ?**

Dans certains cas, elle est plus utile car elle montre la relation entre les valeurs d'un ensemble de données entre elles. À l'inverse, l'écart à la moyenne met surtout en évidence la distance des valeurs par rapport à la moyenne, par exemple pour repérer des valeurs aberrantes. La variance, elle, prend l'ensemble des données en compte.

- **Pourquoi calculer l'étendue ?**

L'étendue correspond à la différence entre la valeur minimale et la valeur maximale d'une série. La calculer est utile lorsqu'on veut se focaliser sur les extrêmes, car elle ne dépend pas des autres valeurs.

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

Un quantile permet de repérer les valeurs situées au centre d'une distribution, ainsi que celles autour du centre. Il sert à découper une série statistique en plusieurs parts. Les quantiles les plus courants sont les quartiles (en 4 parties), 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) offre une lecture visuelle synthétique d'une série statistique : elle met en évidence la médiane, les quartiles et les valeurs extrêmes. La médiane apparaît sous forme d'un trait dans la boîte ; les extrêmes se lisent aux deux extrémités (minimum à gauche, maximum à droite). La boîte représente l'intervalle interquartile, qui regroupe 50 % des observations, tandis que les points isolés indiquent des valeurs aberrantes. Enfin, la taille de la boîte reflète la dispersion des valeurs centrales, ce qui en fait un graphique très utile pour comprendre rapidement une distribution.

6. Paramètres de forme

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

Les moments centrés décrivent la distribution autour de la moyenne en prenant en compte le signe des écarts (valeurs au-dessus ou au-dessous de la moyenne). Les moments absolus, eux, mesurent uniquement l'ampleur de l'écart à un point de référence, sans considérer le signe. On les utilise pour mieux caractériser la dispersion et la forme d'une série statistique (variabilité, asymétrie, etc.).

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

Vérifier la symétrie d'une distribution permet de mieux interpréter les données et de choisir des indicateurs adaptés (moyenne vs médiane, écart-type vs IQR, etc.), car une distribution très asymétrique peut "tirer" certains indicateurs. Pour le faire, on peut analyser l'asymétrie (skewness) : si elle est proche de 0, la distribution est globalement symétrique ; si elle est très différente de 0, la distribution est asymétrique (à droite si l'asymétrie est positive, à gauche si elle est négative). On peut

aussi compléter par une vérification graphique (histogramme ou boîte à moustaches) pour confirmer visuellement.

Exercice codage

Les colonnes à caractères quantitatifs

Pour cette partie d'affichage; il se sélectionne automatiquement les colonnes quantitatives du fichier, puis supprime les valeurs manquantes avant de faire les calculs. Pour chaque variable, il calcule des indicateurs de position (moyenne, médiane, mode) et de dispersion (écart-type, écart absolu, étendue). Enfin, il regroupe tous les résultats dans un tableau récapitulatif (stats) arrondi à 2 décimales pour une lecture plus claire.

```
5.Calcul des paramètres statistiques pour les colonnes quantitatives :
        Colonne    Moyenne   Médiane     Mode  Ecart-type  Ecart absolu moy    Etendue
0      Inscrits  455587.63  366859.0   5045.0   351003.78         272240.72  1808861.0
1   Abstentions  119852.05   95369.0   2272.0   117017.80          74959.07   929183.0
2       Votants  335735.58  274372.0   2773.0   258393.81         201517.17  1297100.0
3        Blancs    5080.46    4001.0   4577.0     3492.52           2817.95    17389.0
4          Nuls    2309.82    2039.0     17.0     1501.38           1131.99     8236.0
5      Exprimés  328345.30  268568.0   2701.0   253758.58         197762.20  1272080.0
6          Voix    1842.00    1627.0   1203.0     1268.37            977.36     7651.0
7        Voix.1    7499.27    5968.0     19.0     6501.29           4474.96    45883.0
8        Voix.2   91430.45   67831.0    534.0    77226.14          59929.14   372286.0
9        Voix.3   10293.34    8944.0  17010.0     7464.32           5140.37    48168.0
10       Voix.4   76017.08   64543.0    459.0    60278.10          42514.72   372668.0
11       Voix.5   23226.41   16885.0   9657.0    20760.60          15278.36   108537.0
12       Voix.6   72079.63   51556.0    501.0    66210.68          49157.01   316871.0
13       Voix.7    5761.48    4881.0     75.0     4581.79           3333.34    22826.0
14       Voix.8   15213.58    9561.0     72.0    14807.62          11136.57    80196.0
15       Voix.9   15691.60   11918.0     51.0    13027.13           9432.01    69513.0
16      Voix.10    2513.12    2118.0   3663.0     1781.41           1404.50     8686.0
17      Voix.11    6777.35    6152.0   7271.0     4636.02           3689.50    20535.0
```

6. Liste des paramètres statistiques calculés :

A partir du code, il s'affiche simplement la liste des paramètres statistiques présents dans le tableau stats. Il permet de vérifier rapidement quelles mesures ont été calculées et enregistrées (ex. moyenne, médiane, écart-type, etc.). C'est une étape de contrôle utile avant de poursuivre l'analyse.

```
6.Liste des paramètres statistiques calculés :
['Colonne', 'Moyenne', 'Médiane', 'Mode', 'Ecart-type', 'Ecart absolu moy', 'Etendue']
```

7. Distance quartile et interdécile

Pour chaque colonne quantitative, deux mesures de dispersion robustes : l'IQR (Q3-Q1) et l'IDR (D9 - D1). Il utilise *dropna()* pour ignorer les valeurs manquantes, puis récupère les quantiles (quartiles et déciles) avec *quantile()*. Enfin, il ajoute ces résultats au tableau stats et affiche une version arrondie pour faciliter la lecture.

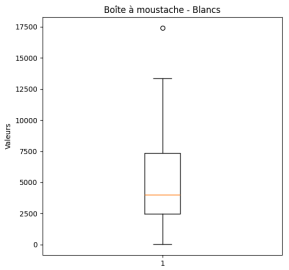
7.Calcul de l'IQR et de l'IDR :

	Colonne	Moyenne	Médiane	Mode	Ecart-type	Ecart absolu moy	Etendue	IQR	IDR
0	Inscrits	455587.63	366859.0	5045.0	351003.78	272240.72	1808861.0	401050.0	793988.8
1	Abstentions	119852.05	95369.0	2272.0	117017.80	74959.07	929183.0	106489.0	193676.2
2	Votants	335735.58	274372.0	2773.0	258393.81	201517.17	1297100.0	301770.5	602687.2
3	Blancs	5080.46	4001.0	4577.0	3492.52	2817.95	17389.0	4852.5	8845.8
4	Nuls	2309.82	2039.0	17.0	1501.38	1131.99	8236.0	1917.0	3240.6
5	Exprimés	328345.30	268568.0	2701.0	253758.58	197762.20	1272080.0	296870.5	590169.2
6	Voix	1842.00	1627.0	1203.0	1268.37	977.36	7651.0	1517.5	3015.6
7	Voix.1	7499.27	5968.0	19.0	6501.29	4474.96	45883.0	6264.5	13104.2
8	Voix.2	91430.45	67831.0	534.0	77226.14	59929.14	372286.0	101317.0	177340.2
9	Voix.3	10293.34	8944.0	17010.0	7464.32	5140.37	48168.0	7999.5	13813.0
10	Voix.4	76017.08	64543.0	459.0	60278.10	42514.72	372668.0	63342.0	130094.6
11	Voix.5	23226.41	16885.0	9657.0	20760.60	15278.36	108537.0	20638.5	43668.8
12	Voix.6	72079.63	51556.0	501.0	66210.68	49157.01	316871.0	60743.5	159421.2
13	Voix.7	5761.48	4881.0	75.0	4581.79	3333.34	22826.0	4779.0	10712.2
14	Voix.8	15213.58	9561.0	72.0	14807.62	11136.57	80196.0	14833.5	38190.8
15	Voix.9	15691.60	11918.0	51.0	13027.13	9432.01	69513.0	13265.5	27686.8
16	Voix.10	2513.12	2118.0	3663.0	1781.41	1404.50	8686.0	2466.0	4266.6
17	Voix.11	6777.35	6152.0	7271.0	4636.02	3689.50	20535.0	6146.5	12311.0

8.Boîte à moustache

à Partir le code, ce bloc génère une boîte à moustaches pour chaque variable quantitative afin de visualiser rapidement la médiane, la dispersion et les valeurs atypiques. Les données sont d'abord nettoyées avec dropna(), puis chaque graphique est enregistré automatiquement dans le dossier img sous un nom différent (boxplot_nomColonne.png). plt.close() ferme la figure pour éviter d'empiler trop de graphiques en mémoire.

```
#Question 8 : boite à moustache

for col in colonnes_quantitatives:
    data = contenu[col].dropna()
    plt.figure(figsize=(6,6))
    plt.boxplot(data)
    plt.title(f"Boite à moustache - {col}")
    plt.ylabel("Valeurs")
    plt.savefig(f"img/boxplot_{col}.png") #choisir l'emplacement
    plt.close()
```



9. Introduire le dossier data le fichier island-index.csv

Avec bloc codage charge le fichier island-index.csv depuis le dossier data afin de pouvoir l'analyser ensuite. L'option low_memory=False évite des erreurs liées à des colonnes dont le type peut varier lors de la lecture. Le script affiche le contenu pour vérifier que l'importation s'est bien déroulée.

```
#Question 9
print("Lecture du fichier island-index.csv :")

with open ("data/island-index.csv", encoding= "utf-8") as fichier2:
    contenu = pd.read_csv(fichier2, low_memory=False)
print(contenu)
```

```
Lecture du fichier island-index.csv :
```

10. Nombre d'îles et superficies et organigramme

```python
#question 10 : nombres d'iles/superficie
print("Nombre d'îles par catégorie de superficie :")

contenu["Surface (km²)"] = pd.to_numeric(contenu["Surface (km²)"], errors='coerce')#
surface = contenu["Surface (km²)"].dropna() #sélection de la colonne
bins = [0, 10, 25, 50, 100, 2500, 5000, 10000, float('inf')]
labels = ("0-10", "10-25","25-50","50-100", "100-2500", "2500-5000","5000-10000", "10000+")
categories = pd.cut(surface, bins=bins, labels=labels, right=True, include_lowest=True)
compte = categories.value_counts().sort_index()
print(compte)
```

```
Nombre d'îles par catégorie de superficie :
Surface (km²)
0-10          78423
10-25          2327
25-50          1164
50-100          788
100-2500       1346
2500-5000        60
5000-10000       40
10000+           71
Name: count, dtype: int64
```

```python
#organigramme
#DEBUT
#Convertir "Surface (km²)" en numérique (erreurs → NaN)
#Supprimer les valeurs manquantes (dropna)
#Définir les classes de superficie (bins) + étiquettes (labels)
#Classer chaque île dans une classe (pd.cut)
#Compter le nombre d'îles par classe (value_counts) + trier (sort_index)
#Afficher les effectifs
# FIN
```

Séance 4

Objectif Séance

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

Questions

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 ?

Pour choisir entre une distribution statistique avec des variables discrètes ou variables continues, il est nécessaire de déterminer la nature du phénomène à étudier. Si les données à analyser pour ce phénomène sont des événements ou des objets comptabilisés à nombre entiers, il est recommandé de choisir des lois discrètes. Autre part, si les données sont des données avec des décimales ou nombre continus, il est à utiliser les lois continues. Aussi, comme critère donné par le texte, il est nécessaire, de manière empirique, de considérer la forme de la distribution dans les données et ses caractéristiques statistiques. En conclusion, je considère que le choix de la ne dépend pas seulement de la nature du phénomène (discret ou continu), mais aussi de la forme de la distribution empirique et de la connaissance et de l'interprétation des principales caractéristiques de l'ensemble de données.

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

Je considère que les lois les plus utilisées en géographie sont la loi de Poisson, la loi normale et les lois de type Pareto ou Zipf, parce qu'elles sont présentées comme adaptées à des situations communes à des études géographiques.

D’abord, la loi de Poisson est essentielle pour modéliser des événements rares que l’on compte dans l’espace ou dans le temps , comme dit dans le texte: *“loi binomiale converge souvent vers une loi de S. D. Poisson*” et que “*la loi de S. D. Poisson correspond à la loi des événements rares (ou la loi des petites probabilités)*”. Et elle est utilisée pour des événements *peu probables dans une succession d’épreuves très nombreuses*. Aussi, la distribution de Poisson est qualifiée de “*fondamentale*”, parce qu’ elle *décrit beaucoup de processus dont la probabilité est petite et constante*. Cela la rend très utile pour modéliser de nombreux phénomènes géographiques peu courants. La loi normale est une loi très utilisée quand on étudie des variables continues et qu’on veut décrire la façon dont les valeurs sont réparties autour d’une moyenne. Dans le texte, il est expliqué qu’un grand nombre de phénomènes peuvent être approchés par une loi normale, notamment quand on additionne beaucoup de petites influences indépendantes.

Pour comprendre et lire cette grande quantité de données, la loi normale sert de référence : on transforme (centre et réduit) les données pour pouvoir les comparer à cette loi et voir si la distribution observée lui ressemble. Néanmoins, il est important de dire que “ce n’est pas parce que la loi est dite normale qu’elle est toujours valable”. Cependant, cela reste une loi de base pour analyser de nombreuses distributions continues en géographie.

Enfin, les lois de type Pareto et Zipf sont particulièrement importantes pour décrire des structures hiérarchiques et des distributions très inégalitaires, omniprésentes en géographie (systèmes de villes, tailles d’unités, intensités extrêmes, etc.). Ces lois servent surtout à analyser la hiérarchie des villes et la structure des systèmes urbains Pour ça, je considère que ces trois familles sont les lois les plus utilisées en géographie, car elles correspondent chacune à des formes de phénomènes géographiques que l’on rencontre très fréquemment.

Commentaire de Séance

Pour cette séance, la partie la plus difficile pour moi a été de comprendre comment afficher correctement les graphiques et surtout comment les enregistrer au bon endroit. Au début, je faisais les calculs et les simulations, mais je n’arrivais pas à “voir” le résultat final, donc je ne savais pas si mon code fonctionnait vraiment. Avec l’aide de Zara Huston, j’ai réussi à corriger cela en organisant mieux la sortie : création automatique d’un dossier (img_finale) et utilisation de plt.savefig() pour conserver chaque figure. Ensuite, j’ai pu comparer plus facilement les lois discrètes et continues (barres vs histogrammes + densité théorique) et comprendre ce que représentent les courbes. À la fin, j’ai eu un script complet et plus clair.

Graphiques Obtenus

Séance 5

Objectif Séance

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

Questions

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

L'échantillonnage est la façon dont on prélève un morceau de population d'une population mère. A partir des nécessités de l'étude et constitution, on choisit un taille spécifique pour la population du l'échantillonnage à partir de la population mère (le tout). On n'utilise pas la population entière car souvent, il est impossible d'effectuer (toute un population, un pays, un continent, etc).

Les méthodes d'échantillonnage peuvent être aléatoires (tirage au sort, par exemple) ou non aléatoires (systématique; des quotas, par exemple). On choisit la méthode à utiliser à partir des données qu'on a avant faire l'échantillonnage: avons - nous une base de sondage complète (parfait pour les méthodes aléatoires)? Ou si notre base n'est pas parfaite mais on connaît sa structure, on choisit des méthodes non aléatoires pour récupérer des informations.

2. Comment définir un estimateur et une estimation ?
Un estimateur est, dans le vocabulaire de l'analyse des données, une statistique. C'est à dire une fonction des données aléatoire choisie pour approcher un paramètre inconnu. Il est présent dans les variables aléatoires. Il est un objet théorique ou règle du calcul.

L'estimation à sa part est une valeur numérique prise, le nombre concret, une fois qu'on a observé effectivement les données à partir de l'échantillon.

3. Comment distingueriez-vous l'intervalle de fluctuation et l'intervalle de confiance ?
Ces deux éléments sont des notions rélevantes et liées mais de logique inverse. Pour le distinguer il est important de reconnaître qu'ils ne répondent pas à la même question. D'un côté l'intervalle de fluctuation prend une donnée connue à partir de l'échantillon et arrive à calculer l'intervalle de la fréquence observée dans l'échantillon. Lorsque l'intervalle de confiance est utilisé quand on ne connaît pas cette quantité. Il s'agit d'une estimation du paramètre.

4. Qu'est-ce qu'un biais dans la théorie de l'estimation ?
Un biais dans la théorie de l'estimation est la différence entre ce qui est vu par un estimateur et la réalité du paramètre des résultats des données. Ceci peut être résultat des anticipations d'un estimateur ou de ce qu'il aperçoit qui peut être différent à ce qui est vraiment prouvé par la statistique.

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

On appelle ça un recensement: « *le recensement est une enquête exhaustive* », un échantillon est juste une partie. Le lien avec les données massives / big data vient du fait que, quand le taux de sondage n/Nn/Nn/N est très élevé, « s'il est fort, il rend les estimations plus précises. On parle aussi d'un *échantillonnage exhaustif* » . Dans le big data, on est souvent dans cette situation : on n'a plus un petit échantillon, mais des données sur une énorme partie de la population (transactions, traces numériques, etc.), ce qui se rapproche d'un recensement ou d'un échantillonnage exhaustif même si, en pratique, ces données peuvent rester incomplètes ou biaisées en comparaison à la population totale.

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

Le choix d'un estimateur est important parce qu'il détermine la qualité de l'information qu'on tire de l'échantillon: « un aspect fondamental de l'inférence statistique consiste à obtenir des estimations fiables des caractéristiques d'une population à partir d'un échantillon » et que ces estimations seront « inévitablement entachées d'erreurs que l'on doit minimiser autant que possible »

Un bon estimateur doit être *sans biais* (son espérance doit être égale au paramètre) et avoir une variance faible.Un estimateur *efficace* (vaincu ses enjeux): « un estimateur de θ est dit efficace s'il est sans biais et s'il est de variance minimale parmi tous les estimateurs sans biais de θ » . En pratique, choisir un estimateur, c'est donc trouver un compromis entre plusieurs critères : peu de biais, faible variance (bonne précision), robustesse aux valeurs extrêmes, et une méthode qui reste calculable et adaptée aux données que l'on a.

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

Il existe différentes façons d'estimer un paramètre. D'abord, la méthode des moindres carrés, souvent utilisée quand on analyse plusieurs variables, consiste à choisir les paramètres qui rendent minimale la somme des carrés des résidus. Ensuite, la méthode du maximum de vraisemblance est une démarche plus générale : elle compare les valeurs possibles du paramètre selon leur probabilité et retient celle qui maximise la fonction de vraisemblance. Enfin, on peut aussi recourir à des approches plus classiques, comme l'estimation ponctuelle (donner une valeur unique) ou l'estimation par intervalle de confiance (proposer un intervalle ayant une forte probabilité de contenir la vraie valeur du paramètre).

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

Un test statistique est « une méthode de calcul permettant de décider si une série d'observations est compatible avec une loi de probabilité entièrement spécifique » . Il y a plusieurs types de tests selon la question : tests de conformité (un échantillon vs une loi théorique), d'homogénéité (plusieurs échantillons entre eux), d'adéquation à une loi, ou d'indépendance de deux caractères.

On a aussi deux grandes familles : tests paramétriques, où la forme de la loi est supposée connue et l'on teste un paramètre (moyenne, variance, etc.), et tests non paramétriques, qui ne supposent pas de loi particulière et s'appliquent aussi bien à des variables quantitatives (Mann-Whitney, Wilcoxon…) qu'à des variables qualitatives (test du χ^2, test exact de Fisher)

Pour créer un test, la démarche générale est :
1. formuler une hypothèse nulle et une hypothèse alternative,
2. Choisir une statistique de test adaptée
3. fixer un risque α, c'est-à-dire la probabilité de rejeter
4. définir une région de rejet,
5. calculer la valeur observée de la statistique et décider si l'on rejette ou non.

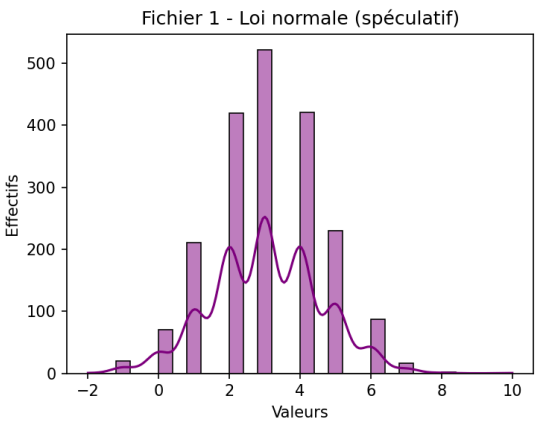
Un test sert à répondre à « l'effet que j'observe est-il compatible avec le simple hasard ? ».

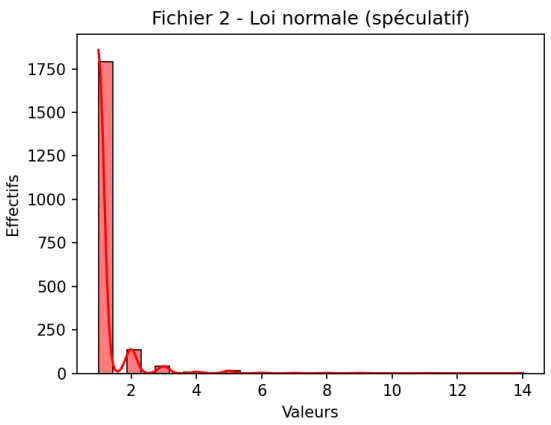
9. Que pensez-vous des critiques de la statistique inférentielle ?
Pour moi la controverse autour du choix de la statistique inferentielle veut dire : gardons les tests, mais toujours avec du contexte, des intervalles de confiance, la taille des effets et du bon sens.

Théorie de l'échantillonnage: Le lien entre l'intervalle de fluctuation et les valeurs réelles de la population mère. Que pouvez-vous en conclure par rapport aux échantillons utilisés pour le calcul ?
L'échantillonnage se base sur la répétitivité et on peut conclure que les échantillons utilisés sont représentatifs globalement, avec seulement de petites variations dues au hasard d'échantillonnage.

Théorie de l'estimation: Interprétez le résultat obtenu et vous le comparerez avec le résultat précédent.
Avec le premier échantillon, les intervalles de confiance (ex. Pour [0,37 ; 0,43]) contiennent les vraies fréquences de la population mère (0,39 / 0,42 / 0,19). Donc cet échantillon donne une estimation "correcte" : les petites différences viennent juste du hasard. Par rapport à l'intervalle de fluctuation, on conclut pareil : les résultats sont cohérents avec la population.

Théorie de la décision: Laquelle est une distribution normale ? Vous expliquerez dans votre rapport pourquoi.
Le test de Shapiro–Wilk donne des p-values très faibles pour les deux fichiers ($p < 0,05$), ce qui conduit à rejeter l'hypothèse de normalité : statistiquement, les données ne suivent donc pas une loi normale. Cependant, ce test peut être très sensible et détecter de petites déviations, surtout quand la taille d'échantillon est importante, ce qui peut conduire à un rejet même si la distribution paraît "presque" normale. Pour compléter ce diagnostic, il est pertinent d'ajouter une vérification graphique (histogramme et idéalement QQ-plot) afin d'évaluer la forme globale de la distribution et l'importance réelle des écarts à la normalité.





Séance 6

Objectifs Séance

- Manipuler des fonctions locales et comprendre la nécessité de factoriser son code en une liste de fonctions ou de procédures exécutant une tâche unique
- Créer des fonctions locales spécifiques au traitement d'un problème
- Comprendre l'analyse de variables qualitatives ordinales

Questions

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 classe des données en catégories ordonnées (du plus faible au plus fort) sans mesurer l'écart entre les niveaux. Elle s'oppose à la statistique nominale (catégories sans ordre) et utilise des variables qualitatives ordinales. En géographie, elle montre une hiérarchie spatiale en classant des lieux (villes, régions) selon un critère.

2. Quel ordre est à privilégier dans les classifications ?

On privilégie en général l'ordre croissant, car il est plus clair et permet d'utiliser médiane et quantiles. Exception : certains classements comme la loi rang-taille utilisent l'ordre décroissant.

3. Quelle est la différence entre une corrélation des rangs et une concordance de classements ?
La corrélation des rangs compare deux classements. La concordance mesure l'accord entre plusieurs classements (souvent avec le W de Kendall).

4. Quelle est la différence entre les tests de Spearman et de Kendal ?
Spearman compare les rangs via leurs écarts. Kendall compare des paires concordantes/discordantes, et est souvent préféré quand on veut raisonner en accord.

5. À quoi servent les coefficients de Goodman-Kursdal et de Yule ?
Ils servent à mesurer l'association entre variables catégorielles. Yule est un cas particulier pour des variables binaires.

Retours sur le cours

Au début du semestre, j'ai vécu ce cours comme un choc assez fort: pas seulement parce qu'il fallait "faire des stats", mais parce que le cours demandait d'entrer dans une manière de penser différente une logique de procédure, d'étapes, d'erreurs à corriger, de résultats à justifier très différent à comment j'avais jamais appris python. J'avais eu une expérience plus libre et concentrée à choisir plus la manière dont on apprenait que le contrôle subis par cette méthode. Ainsi, la demande de la langue ajoutée au codage a été vraiment difficile pour moi, contrairement à comment j'avais vécu python avant.

Ma première impression a donc été la difficulté, et même une certaine intimidation. Dans mon cas, comme le français n'est pas ma langue maternelle, coder peut devenir un espace plus neutre, plus stable: les règles du langage sont les mêmes, où que l'on soit. J'ai eu beaucoup d'hésitation à apprendre à cause des limitation de la langue et la manière très spécifique de ce cours pour un langage universel. C'est pour ça qu'une grande partie du cours a été fait en compagnie de ma camarade Zara Huston, qui prenait le temps de m'expliquer le raisonnement du code, partageait ses conclusions et m'aidait à mettre les miennes et aussi m'aidait avec la langue (surtout car je trouvais une logique attendue très différente à celle j'avais eu expérience avec pour montrer mes connaissances en python). Elle prenait le temps de faire chaque pas à mes côtés et m'expliquait pour que je puisse avancer dans l'exercice. Elle prenait même des salles à la bibliothèque pour faire un pas à pas.

Enfin, sur le fonctionnement en présentiel, je pense que l'idée d'un espace de questions est bonne, mais qu'elle ne correspond pas toujours à la manière dont les difficultés se manifestent. Personnellement, je suis souvent quelqu'un qui comprend ses blocages après coup, en travaillant seule : sur le moment, je ne sais pas forcément formuler ce qui ne va pas, ou je ne repère pas immédiatement l'origine du problème. Résultat : j'ai parfois eu le sentiment de ne pas exploiter pleinement ces sessions, non par manque d'intérêt, mais parce que les questions apparaissent "trop tard". Et aussi sur discord, je ne suis pas trop à utiliser.

Je repars avec quelque chose de précieux : une initiation à Python, une compréhension plus claire de ce que signifie analyser des données, et l'envie d'intégrer cet outil dans mon parcours, notamment en géographie.