RAPPORT D'ACTIVITÉ

Analyse de données - Parcours Débutant

par GAUTHIER Youri
Master 1 GAED, parcours Géo Suds - Sociétés, Territoires et développement

SOMMAIRE

Séance 2

Les principes généraux de la statistique

Question 1

La discipline de la géographie met en perspective des espaces à plusieurs échelles, des hommes, des milieux naturels mais surtout émet des données. La mesure d'une population ou d'un territoire constitue de fait une statistique, ce qui peut interroger le rapport de la géographie aux statistiques.
D'abord, il convient de dire que la statistique désigne une science en tant que branche des mathématiques. Il s'agit d'un ensemble de données qui révèlent des faits qui sont quantifiés. Dans le cas de la géographie, l'analyse se réalise autour du domaine de l'information géographique, ou les géographes peuvent classer les statistiques en fonction des objets trouvés.
Néanmoins, l'approche fréquente humaine et sociale des géographes fait qu'ils rejettent en partie les statistiques, vues comme un bloc uni, malgré les atouts qu'elle peut apporter. Il existe une sous-estimation de l'analyse statistique en général, malgré certaines exceptions.

Question 2

Il existe un large débat dans la communauté scientifique autour de plusieurs aspects du hasard. La majorité des non statisticiens et des géographes défendent notamment l'idée selon laquelle le hasard serait à l'origine de toute chose. Il s'agit en ce sens, d'un phénomène opposé au déterminisme et à la rationalisation, propre aux méthodes des sciences dures, ce qui tend à écarter la discipline de la géographie de la science.
Or, l'école de l'analyse spatiale réfute cette thèse en optant une position qui oscille entre nécessité et contingence. D'abord, la nécessité désigne le fait qu'un phénomène doit se produire de manière inévitable, tandis que la contingence offre la possibilité qu'un phénomène puisse se produire ou non.
Cette dernière permet de produire un paramètre scientifique au sein de la géographie, avec l'élaboration de tendances (démographiques notamment).
De fait, il n'est pas possible d'anticiper l'action de chaque acteur sur un territoire donné mais des certitudes peuvent être établies avec les effets de contingence.

Question 3

L’information géographique se divise en deux grandes séries statistiques qui correspondent aux deux manières de “mesurer” l’espace.
D’abord, nous pouvons parler de l’information attributaire qui caractérise un lieu ou un territoire déjà délimité. Dans un Système d’Information Géographique (SIG), on appelle cela la base attributaire.
Cette information peut relever aussi bien de la géographie humaine (populations, indicateurs socio-économiques) ou la géographie physique (données climatiques). Cette information répond à la question “Qu’est-ce qu’il y a à cet endroit ?”, notamment.
D’un autre côté, l’information géométrique étudie la morphologie de la surface elle-même et non son contenu. Cette information pratique différentes échelles, en étudiant les contours, les distances et l’organisation spatiale des objets.

Question 4

Avec l’usage d’outils d’analyse plus performants et la massification de l’information géographique, de nouveaux enjeux dirigent la pratique de la géographie. Les besoins de la géographie en matière d’analyse de données sont dictés par une nécessité de transition, soit transiter d’une “méthode descriptive” à une “science des échelles” plus complètes.
Il convient d’adopter l’outil statistique pour structurer, résumer et interpréter l’abondance d’informations, ce qui permet à la géographie de comprendre plus en profondeur les dynamiques spatiales complexes (démographie, inégalités) pour permettre d’orienter les politiques adaptées.

Question 5

D’abord, nous pouvons dire que la différence fondamentale entre ces modèles réside dans l’objectif de l’analyse : la statistique descriptive cherche à simplifier et résumer l’information, tandis que la statistique explicative vise à comprendre les causes et les relations de dépendance.
Le caractère de la statistique descriptive à “mettre de l’ordre” peut se comprendre par l’objectif de résumer l’information avec le minimum de paramètres et de graphiques. Cela permet de dégager une structure principale et parlante.
La statistique explicative cherche plutôt à modeler des relations entre les phénomènes. Pour se faire, une variable “à expliquer” est mise en relation avec une ou plusieurs autres variables pour comprendre les variations et les interdépendances.

Question 6

On retrouve plusieurs types de visualisation de données en géographie : les représentations sectorielles, l'histogramme, le diagramme en bâtons, le nuage de points, la boîte à moustache ou le polygone de fréquence entre autres.
Le choix de la représentation cartographique s'effectue selon le type de variable statistique analysée. Ainsi, il convient d'utiliser un graphique en secteur pour des données nominales, tandis que pour des données qualitatives ordinales, nous utilisons un histogramme.

Question 7

Les méthodes d'analyse de données sont les méthodes descriptives, les méthodes explicatives et les méthodes de prévention.
D'abord, les méthodes descriptives servent à visualiser et à classer les données sans chercher de lien de causalité immédiat. Toutes les variables jouent un rôle équivalent. Ces méthodes descriptives permettent de synthétiser le flux d'informations et se divisent en Analyse en Composantes Principales (ACP), qui rendent visible les grandes structures d'un phénomène, puis en Analyse Factorielle des Correspondance (AFC), qui étudie les proximités dans un tableau de contingence. Enfin, on compte l'Analyse des Correspondance Multiples (ACM) qui s'appliquent pour l'analyse de plus de deux variables qualitatives.
Les méthodes explicatives cherchent à relier une variable à expliquer à des variables explicatives, en utilisant plusieurs procédés dont l'Analyse de la variance, qui étudie les implications, ou l'Analyse discriminante ou la segmentation.
Les méthodes de prévision concernent, elles, l'analyse des séries chronologiques. L'objectif est de construire un modèle reliant les données présentes à celles du passé pour anticiper le futur.

Question 8

D'abord, une population statistique peut être considérée comme un ensemble d'éléments au sens mathématiques, qui est quantifiable. C'est un ensemble d'éléments similaires sur lequel une étude statistique est réalisée.
Un individu statistique correspond à un élément de la population statistique, dont on observe les caractéristiques pour des analyses statistiques.
Un individu est localisable et cartographiable (unités spatiales). Les individus statistiques sont aussi composés d'un ensemble d'attributs.

Le caractère statistique est le caractère de l'individu pris dans la population statistique. Autrement dit, il s'agit de caractéristiques quantitatives (dont la nationalité, le sexe, la couleur des yeux).
In fine, la modalité statistique correspond aux différentes valeurs ou catégories qu'une variable peut prendre dans une étude statistique.

Les caractères sont notamment qualitatifs et quantitatifs. Les caractères qualitatifs sont nominal ou ordinal, tandis que les caractères quantitatifs sont discrets ou continus.
Il existe une hiérarchie entre les caractères qui est déterminée par la quantité d'information et de données dont ils disposent.

Question 9

Le calcul de la mesure de l'amplitude s'obtient en soustrayant la valeur de sa borne minimale à celle de sa borne maximale, tandis que la densité s'obtient par la formule : d=ni/b-a. C'est un ratio qui rapporte l'effectif d'une classe à son amplitude. L'usage de ces deux mesures permet la discrétisation de caractères quantitatifs (classer les valeurs).

Question 10

Les formules de Sturge et de Yule doivent répondre à la question du nombre de classes qui divisent une variable quantitative continue pour construire un histogramme pertinent.
D'abord, la formule de Sturges repose sur l'idée que la distribution des données se rapproche d'une "Loi Normale" et donne une valeur approximative du nombre de classes.

La formule de Yule constitue une alternative, souvent utilisée pour des distributions plus courtes, en donnant souvent un nombre de classes légèrement plus élevé que Sturges, ce qui permet de constituer une structure plus importante de la distribution.

Question 11

Un effectif correspond au nombre de fois ou une modalité ou une valeur apparaît au sein de la population statistique analysée.
La fréquence exprime la part d'une valeur par rapport au total, en divisant l'effectif de la modalité par l'effectif total de la population statistique.
La fréquence obtenue est donc comprise entre 0 et 1.

Les fréquences cumulées servent à répondre à des questions de seuil.
On additionne les fréquences au fur et à mesure que l'on progresse dans les valeurs. Il s'agit de classer les données par ordre croissant, ce qui fait que la première fréquence cumulée est égale à la première fréquence simple, tandis que pour les suivantes, on ajoute la fréquence de la ligne actuelle à la somme des précédentes.
Une distribution statistique constitue la vision d'ensemble des données et révèle la manière dont les données sont réparties. En réunissant l'ensemble des couples (Valeur ; Effectif) ou (Valeur ; Fréquence), cette distribution montre si les données sont concentrées ou très dispersées.

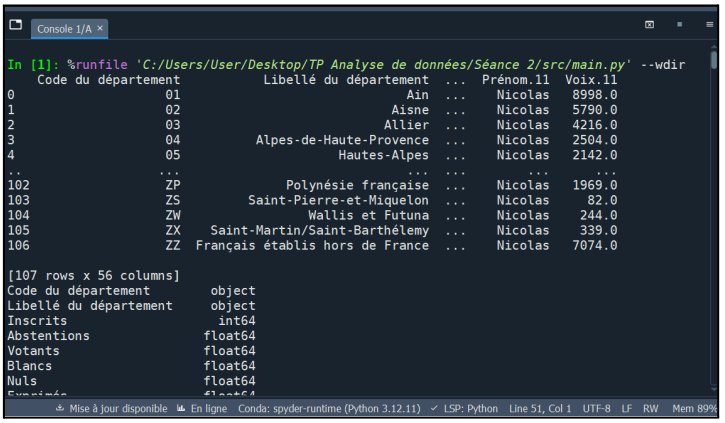
Manipulations

Affichage : print

Type de colonnes : print(contenu.dtypes)
types_colonnes = {}

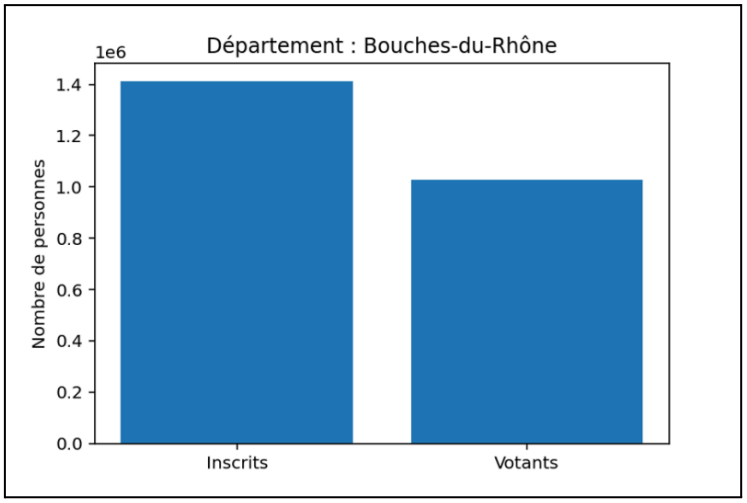
Lignes et colonnes : nb_lignes =len(contenu)
nb_colonnes= len(contenu.columns)

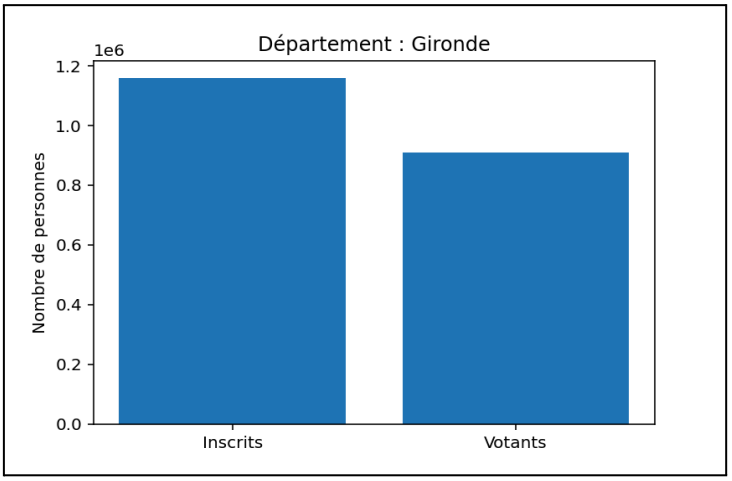
Sommes quantitatives : sommes_quantitatives = []



Commentaire de résultat : Lors de cette première séance, j'ai pu me familiariser aux commandes Python. Le plus dur résidait dans l'organisation des documents au sein des fichiers et pour y avoir accès sur Python.

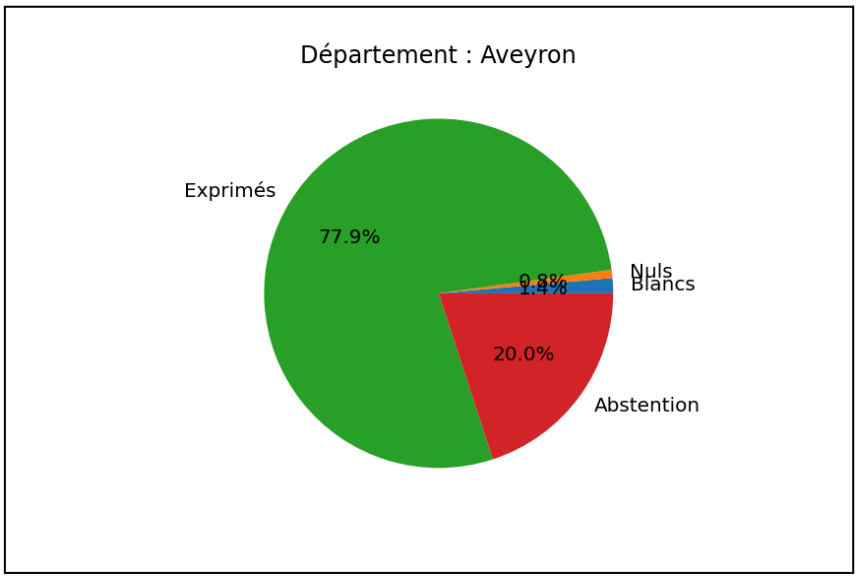
Grâce à des formules écrites, j'ai pu obtenir les sommes quantitatives du document Excel et pu faire apparaître les colonnes.
En utilisant la librairie Matplotlib, j'ai calculé le nombre de colonnes, de lignes et les types de variables. J'ai aussi appris qu'en écrivant du code et en l'enregistrant, on pouvait retrouver ce code dans notre dossier, si le chemin du fichier est bon.



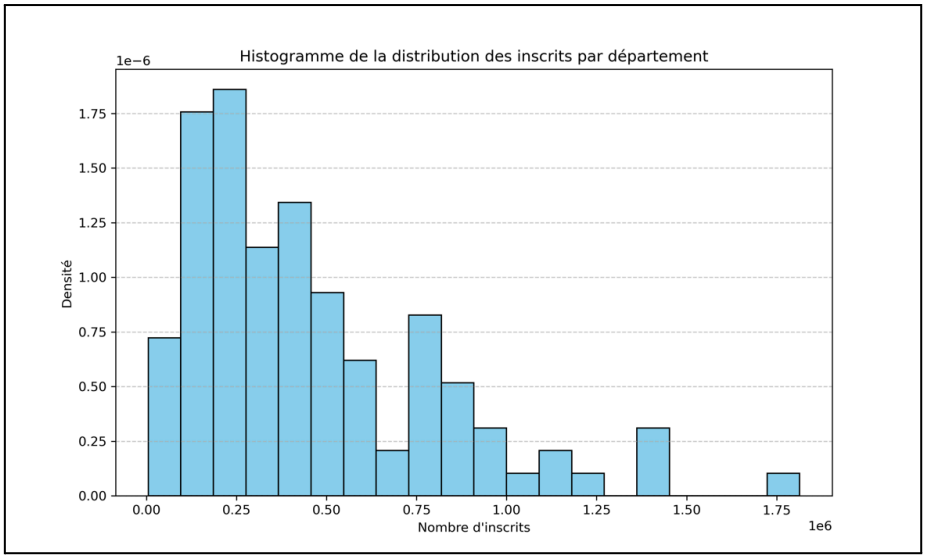


<u>Commentaire de résultat :</u> On obtient des diagrammes qui synthétisent les informations du département en question en deux colonnes distinctes, une "inscrits" et une autre "votants". Ils s'obtiennent en demandant au code de les figurer. On observe que le nombre d'inscrits dans le département des Bouches-du-Rhône atteint le total d'1,4 million de personnes, bien que la participation soit plus faible, située ici autour d'1 million de personnes.

Nous pouvons constater que près de 30% des inscrits n'ont pas voté dans ce département.
Ce phénomène se retrouve dans une moindre mesure dans le département de la Gironde. L'effectif total d'inscrits est similaire, avec 1,2 million d'habitants, tandis que le nombre de votants s'élève à près de 900.000 personnes. Ainsi, environ 25% des inscrits n'ont pas voté dans le département de la Gironde.



<u>Commentaire de résultat :</u> Ce diagramme circulaire représente la répartition en détail des votes lors du premier tour de l'élection présidentielle de 2022 dans le département de l'Aveyron. Nous pouvons observer une structure plus large et détaillée de l'information, avec la part des votes "Nuls" et "Blancs".
J'ai obtenu ce diagramme grâce à la fonction "os.*makedirs*" dans le code.
Ainsi, la part majoritaire des votes ont été exprimés (quasiment 78%).
Le taux d'abstention est la deuxième variable importante et s'élève à 20%. Enfin, les variables de votes "Blancs" et "Nuls" sont minimes et comptent pour 2,2% cumulées.



Commentaire de résultats : L'histogramme obtenu, qui représente la distribution du nombre d'inscrits par département lors du premier tour de l'élection présidentielle 2022, permet de comprendre plusieurs choses.
Il démontre notamment la dispersion et l'asymétrie de la distribution du nombre d'inscrits, qui peut révéler d'inégales répartitions dans le territoire français.

Séance 3

Les paramètres statistiques élémentaires

Question 1

On considère généralement que le caractère qualitatif est le plus abondant car il définit une précision supplémentaire, sous forme de mesure numérique.
Ce caractère permet de rendre compte de types très variés de donnés, tandis que le caractère quantitatif sert de base de la classification.

Question 2

Les caractères quantitatifs discrets prennent en compte des des valeurs isolées, généralement des nombres entiers. Entre les deux valeurs, il ne peut pas y avoir

d’intermédiaires. Le caractère quantitatif continu est défini par sa capacité à prendre n’importe quelle valeur dans un intervalle donné.
Entre deux données, il existe une quantité de valeurs possibles, dont les nombres décimaux.
La distinction de ces caractères permet la méthode d’analyse statistique afin d’éviter les confusions liées à des taux (démographie) et sur Python, la fonction “int” correspond au caractère discret, tandis que la fonction “float” s’applique au caractère continu.

Question 3

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

Il existe plusieurs types de moyenne dans l’objectif d’avoir une représentation des données différentes qui n'expriment pas les mêmes tendances.
La moyenne est un résumé statistique et selon la nature de ce que l’on mesure, la logique change et les données sont mises en rapport différemment.

— Pourquoi calculer une médiane ?

Il convient de calculer une médiane afin d’obtenir une véritable répartition des distributions car elle permet de diviser la série statistique en deux, entre d’un côté les valeurs inférieures à la médiane et les valeurs supérieures.
Elle permet de rendre compte de situations de distribution et d’asymétrie importante, dans le cas de l’étude géographique des inégalités socio-économiques, par exemple.

— Quand est-il possible de calculer un mode ?

Le mode correspond à une modalité liée à l’effectif maximal et peut être calculé sur tous les types de données.

Nous pouvons calculer le mode dans l’étude des caractères qualitatifs, en raison du caractère nominal notamment. Le mode constitue l’élément dominant dans la statistique et peut être considéré comme une mesure de tendance globale.

Question 4

L'intérêt majeur de la médiale est de démasquer la quantité de données en servant de "révélateur" d'une distribution inégale et de la captation d'une partie des ressources par une minorité, en général.
On observe sur les données que les ressources sont proches (bien réparties) ou éloignées (mal réparties).
L'intérêt du coefficient de Gini est de rendre compte de ces inégalités par une étude qui quantifie les valeurs entre 0 et 1. Plus les valeurs se rapprochent de 0, plus les ressources sont réparties, tandis que plus les valeurs se rapprochent de 1, plus les ressources sont détenues par une minorité.
L'intérêt est d'avoir un indice hiérarchisé qui permet de démontrer une concentration des données.

Question 5

— Pourquoi calculer une variance à la place de l'écart à la moyenne ? Pourquoi la remplacer par l'écart type ?

D'abord, la variance constitue la moyenne de la somme des écarts par rapport à la moyenne arithmétique. Calculer la variance peut être utile pour démontrer l'absence de grandes valeurs aberrantes, ce qui permet d'avoir une représentation plus juste des données.
L'écart à la moyenne permet de montrer seulement la distance et le lien des valeurs par rapport à la moyenne et représente notamment les grandes valeurs aberrantes. Cependant, il convient de privilégier l'utilisation de l'écart-type, la racine carrée de la variance, qui permet de comparer la dispersion à la moyenne.

— Pourquoi calculer l'étendue ?

L'étendue correspond à la différence entre la valeur maximale et la valeur minimale. Elle donne une idée immédiate de l'amplitude totale des données. C'est le paramètre le plus synthétique pour vérifier les bornes d'un phénomène, comme l'amplitude thermique d'une surface.

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

Un quantile permet de découper une série de données en parts égales pour situer un individu par rapport au reste du groupe.
En individualisant une donnée, cela permet de savoir ou se place cette donnée sans être pollué par la moyenne.
Pour cela, plusieurs types de quantile sont utilisés, selon l'objectif de représentation d'un phénomène. D'abord, il y a la médiane, qui se divise en deux classes, qui est le plus connu mais limite l'analyse.
Les quartiles sont aussi utilisés et permettent de diviser la série statistique en 4 classes. Les déciles sont notamment utilisées pour mesurer les inégalités socioprofessionnelles et découpent la série statistique en 10 classes.
Enfin, les centiles structurent la série statistique en 100 classe, ce qui s'applique pour les grandes statistiques.

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

La boîte de dispersion permet d'offrir un résumé visuel complet de la distribution d'une série statistique, quand un chiffre ou une moyenne peuvent être trompeurs. Cette boîte permet de diviser l'échantillon en 4 zones contenant 25% des effectifs. On constate la répartition de la majorité de la population et la symétrie ou non des valeurs. Les points isolés sont les valeurs aberrantes et la taille de la boîte indique la dispersion des valeurs centrales. La représentation en seul graphique est utile et efficace pour l'étude statistique d'un phénomène.

Question 6

— 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, lui mesure la distance à un point de référence sans tenir compte du signe.
Ce sont des outils performants pour l'analyse de la dispersion et la forme d'une série statistique.

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

Il convient de vérifier la symétrie pour choisir les bons indicateurs de résumé de l’information, d’abord. Dans le cas d’une distribution symétrique, nous pouvons utiliser la moyenne comme indicateur, tandis que dans le cas d’une distribution asymétrique, la médiane est plus pertinente. Vérifier la symétrie d’une distribution aide aussi à comprendre le phénomène étudié, par une asymétrie positive ou négative, soit la concentration des valeurs d’un côté ou de l'autre.
Si la distribution est proche de 0, elle est symétrique mais si elle est éloignée de 0, elle est asymétrique.

Manipulations

Calcul des paramètres de position

Moyenne : df[quant_cols]
Médiane : round(series.median(), 2)
Mode : round(series.mode().iloc[0], 2) if not series.mode().empty else np.nan

Calcul des paramètres de dispersion

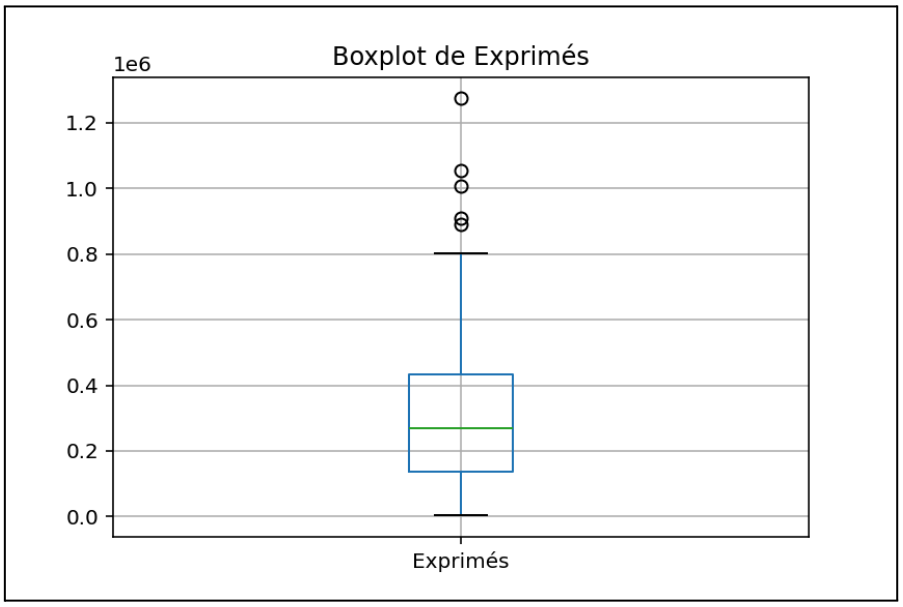
L’écart-type : round(series.std(), 2)
L’écart absolu à la moyenne : round((abs(series - series.mean()))).mean(), 2)
L’étendue : round(series.max() - series.min(), 2)

Quantiles et distances interquantiles

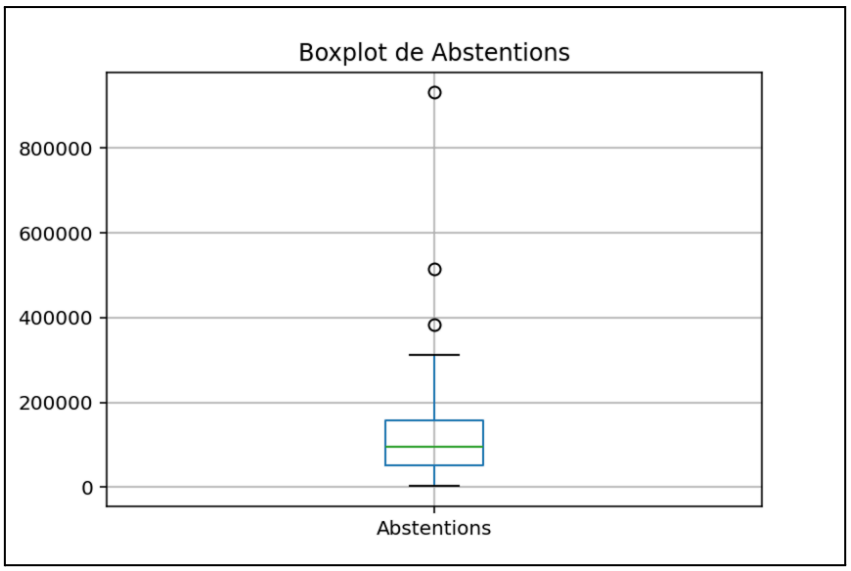
Distance interquantille : round(series.quantile(0.75) - series.quantile(0.25), 2)
Distance interdécile : round(series.quantile(0.9) - series.quantile(0.1), 2)

Boîte de dispersion : plt.figure(figsize=(12,8))
quant_df.boxplot()
plt.title("Boîtes de dispersion des colonnes quantitatives")
plt.xticks(rotation=45)
plt.tight_layout()
plt.savefig("boites_dispersion.png")
plt.show()

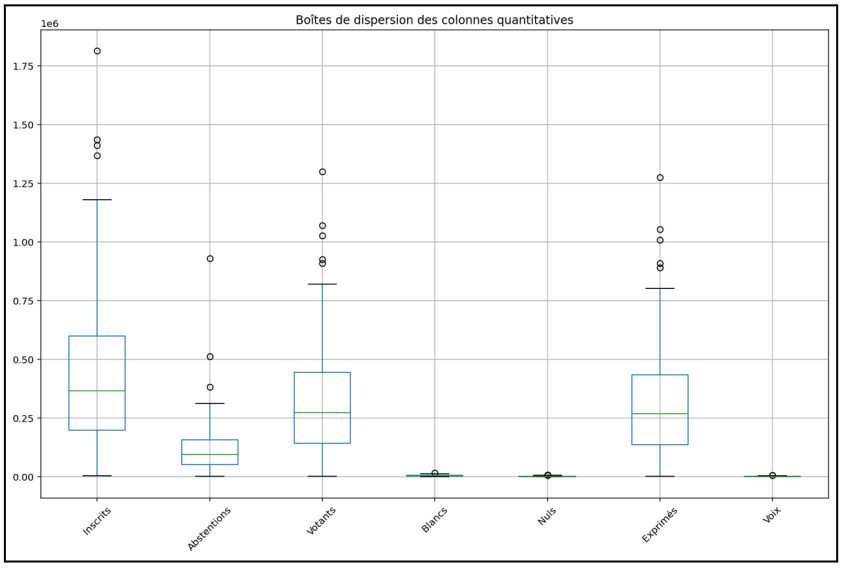
Variable quantitative continue : df["Classe_Surface"] = liste_surface
liste_comptage = df["Classe_Surface"].value_counts()



Commentaire de résultat : Ce graphique représente la boîte de dispersion du nombre de votes exprimés lors du premier tour de l'élection présidentielle de 2022.
On observe d'abord une forte asymétrie de la distribution. En effet, la médiane (en vert) est située plus bas que le milieu exacte de la boîte bleue, ce qui indique qu'une grande partie des observations se concentre sur des valeurs faibles.
La partie supérieure est plus longue que la partie inférieure, ce qui nous fait dire que les données s'étirent vers le haut.
Surtout, on retrouve la présence de valeurs aberrantes qui se distinguent de la masse de données générale, comprise entre 200.000 et 400.000 votes.
Nous pouvons interpréter cela par les cas exceptionnels des grandes villes qui ont un nombre d'inscrits plus important.



Commentaire de résultat : Ce graphique représente la boîte de dispersion du nombre d'abstentions lors du premier tour de l'élection présidentielle de 2022. Nous pouvons constater que 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, en se situant autour de 100.000 abstentions. On retrouve aussi une forte asymétrie entre les données, ou la dispersion s'étend de quelques milliers à près de 800.000, pour la valeur la plus extrême.



Commentaire de résultat : Ce graphique représente les boîtes de dispersion des différentes compositions du vote lors du premier tour de l'élection présidentielle de 2022. Cette mise en comparaison des boîtes de dispersion rend compte de plusieurs choses. D'abord, la distribution des variables de masse, regroupant les "Inscrits", "Votants" et "Exprimés" est similaire, car elles sont en partie corrélées : le nombre de votes exprimés dépend directement du nombre d'inscrits.
La boîte des "Abstentions" est plus basse, avec une médiane autour 100.000 mais reste une boîte significative, alors que les boîtes des "Blancs" et "Nuls" sont résiduelles et s'avèrent être quasiment plates, ce qui révèle leur faible importance.
Le caractère majeur de distribution est l'asymétrie entre les valeurs, dans le cas des "Inscrits", notamment, en raison de territoires et de grandes villes plus peuplées que le reste du pays. Aussi, en comparant les boîtes des "Inscrits" et des "Exprimés", on peut observer un décalage, en raison de la présence de la boîte des "Inscrits" qui commence plus haut que celle des "Exprimés".
Cet écart entre "Inscrits" (environ 350.000) et celle des "Exprimés" (environ 250.000) exprime l'impact cumulé de l'abstention, notamment.

Séance 4

Les distributions statistiques

Question 1

Le choix entre une variable discrète et une variable continue est fondamental en statistique, car il détermine le type de modèles, de tests et de représentations graphiques. Pour trancher, il convient de voir si la variable résulte d'un processus de comptage ou de mesure.
Les critères qui permettent de choisir sont multiples et dépendent d'abord de l'ensemble que l'on souhaite étudier. Dans le cas d'une distribution discrète, on considère des variables qui ne peuvent prendre que des valeurs isolées et finies, souvent représentées par des nombres entiers. L'utilisation de cette variable est pertinente dans le cas d'une étude sur le nombre d'élèves, le nombre d'accidents ou de morts. A l'inverse, on utilise une distribution continue lorsque la variable peut théoriquement prendre n'importe quelle valeur au sein d'un intervalle réel, avec une précision infinie. La possibilité d'étude est plus élargie, comme l'étude de la croissance chez les hommes (taille, poids) ou des mesures environnementales et physiques (crues, tempête). Ainsi, il convient de se demander s'il est plus pertinent d'étudier avec des valeurs discrètes ou continues, pour les données à étudier.

Question 2

Je crois que les lois les plus utilisées sont celles qui représentent la répartition spatiale et sa hiérarchie. Il y a la Loi de Zipf ou Rang-Taille qui modélise la hiérarchie des villes au sein d'un système urbain. Cette loi permet de classer les villes en fonction du rang de leur population. La relation qui s'établit entre la population de chaque ville et son rang hiérarchique dans un classement par nombre d'habitants serait une constante
L'utilisation en parallèle du principe de Pareto doit aussi être important, car elle permet de relever les effets de concentration de population ou de richesse au sein d'une partie du territoire, notamment.

Manipulations

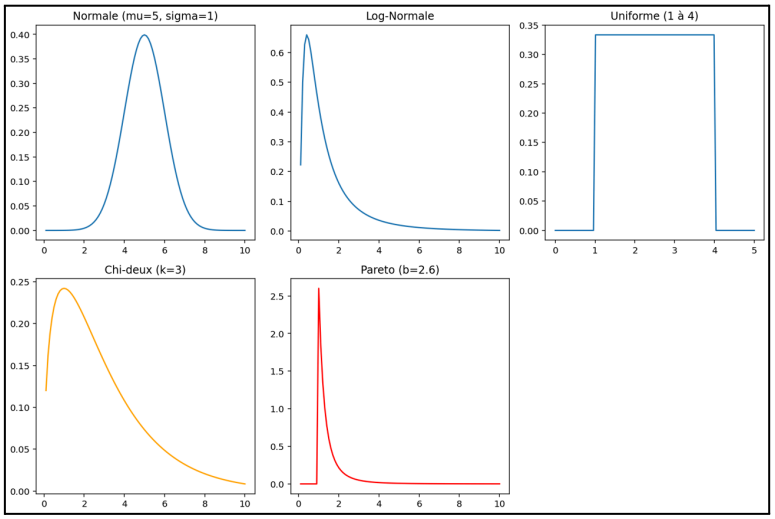
Loi de Dirac :
```
# 1. Dirac (centrée en 5)
plt.subplot(2, 3, 1)
x_dirac = np.arange(0, 11)
y_dirac = np.where(x_dirac == 5, 1, 0)
plt.stem(x_dirac, y_dirac)
plt.title("Loi de Dirac (x=5)")
```

Loi uniforme discrète :
```
# 2. Uniforme Discrète (Dé à 6 faces)
plt.subplot(2, 3, 2)
x_unif = np.arange(1, 7)
plt.bar(x_unif, randint.pmf(x_unif, 1, 7), alpha=0.7)
plt.title("Uniforme Discrète (1 à 6)")
```

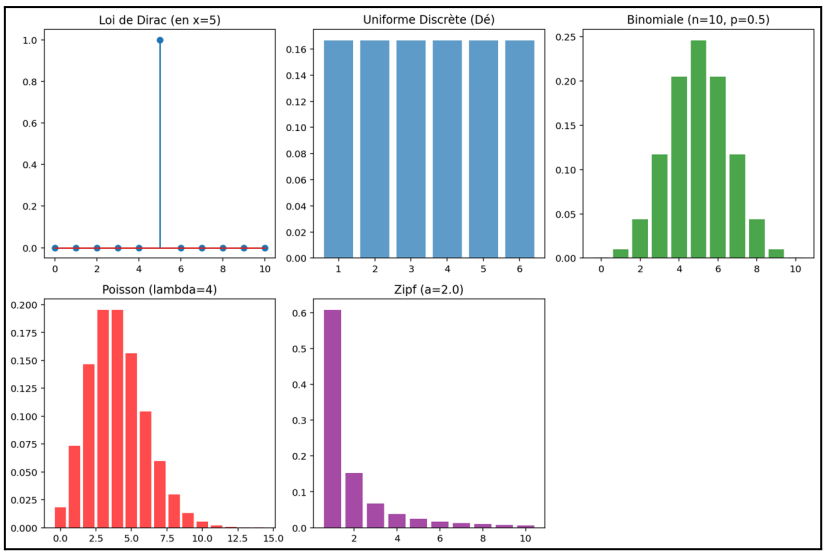
Loi binomiale :
```
plt.subplot(2, 3, 3)
n, p = 10, 0.5
x_binom = np.arange(0, n+1)
plt.bar(x_binom, binom.pmf(x_binom, n, p), color='g', alpha=0.7)
plt.title(f"Binomiale (n={n}, p={p})")
```

Loi Poisson :
```
plt.subplot(2, 3, 4)
mu = 4
x_pois = np.arange(0, 15)
plt.bar(x_pois, poisson.pmf(x_pois, mu), color='r', alpha=0.7)
plt.title(f"Poisson (lambda={mu})")
```

…



Commentaire de résultat : Ces graphiques illustrent plusieurs types de distributions statistiques qui sont générées comparées par le code. D'abord, la Loi Normale est parfaitement symétrique autour de sa moyenne (5).
Ensuite, Log-Normale monte rapidement et descend avec une longue "queue" vers les valeurs élevées. La Loi de Chi-deux se trouve très asymétrique, en s'aplatissant au fur et à mesure que les valeurs augmentent. La Loi Uniforme illustre une dominante entre 1 et 4, ce qui nous fait dire que la probabilité d'apparition de ces valeurs est très forte.
Enfin, le principe de Pareto représenté ici constitue la distribution la plus extrême. Elle montre une très forte concentration de valeurs au début, pour rapidement descendre, ce qui peut illustrer des structures d'inégalités spatiales.



Commentaire de résultat : Ces graphiques révèlent un caractère hétérogène des données, avec une très forte dispersion. Le graphe le plus parlant est celui de la Loi Zipf, ou les colonnes sont tirées par des valeurs importantes et qui représentent la domination de grandes villes sur le reste des villes.

Séance 5

Les statistiques inférentielles

Question 1

L'échantillonnage peut se définir comme une méthode qui prélève au hasard un petit groupe d'individus statistique au sein d'une population délimitée. Cette étude limitée vise à former une tendance, pour déduire des informations sur toute la population délimitée et former une tendance générale, ce qui implique, déjà, de penser à réaliser un échantillonnage représentatif. Il faut remplir cette condition, dans la mesure où il n'est pas possible d'étudier une population entière quand elle est trop grande.
Pour réaliser ces études, on distingue deux méthodes d'échantillonnage.
D'abord, nous pouvons parler des échantillons non biaisés qui désigne un tirage au hasard d'individus, où chacun à autant de chances de tomber dans l'échantillon.
Les échantillons biaisés, eux, correspondent à des sélections préalables.

Le choix de la méthode d'échantillonnage dépend d'un compromis entre précision et moyens. L'échantillonnage aléatoire et non biaisé s'applique dans le cadre ou il n'y a pas de liste complète de tous les individus de la population. Le risque d'erreur doit être minimisé, dans le cadre d'études médicales ou politiques.
De l'autre côté, l'échantillonnage non-aléatoire s'applique sur une population déjà définie.

Question 2

L'estimation statistique est le processus qui permet de déduire les caractéristiques d'une population entière à partir de l'observation d'un échantillon réduit. Ainsi, il est primordial de distinguer l'estimateur de l'estimation. Un estimateur se définit comme une règle de calcul, une formule mathématique ou une fonction que l'on applique aux données de l'échantillon pour évaluer un paramètre inconnu de la population, tel que la moyenne ou la variance. Sur le plan théorique, l'estimateur est considéré comme une variable aléatoire : sa valeur numérique est susceptible de changer à chaque fois que l'on sélectionne un nouvel échantillon. On recherche généralement des estimateurs possédant des qualités de convergence et d'absence de biais afin de garantir que, sur un grand nombre d'échantillons, le résultat moyen soit le plus proche possible de la réalité.À l'inverse, l'estimation représente le résultat numérique concret obtenu une fois que la formule de l'estimateur a été appliquée aux données réelles d'un échantillon précis. Contrairement à l'estimateur, l'estimation est une valeur fixe et unique pour un jeu de données donné. On distingue d'ailleurs l'estimation ponctuelle, qui fournit un chiffre unique, de l'estimation par intervalle (ou intervalle de confiance), qui propose une fourchette de valeurs à l'intérieur de laquelle le paramètre recherché a une forte probabilité de se situer.

Question 3

L'intervalle de fluctuation est utilisé lorsque l'on connaît ou suppose les caractéristiques d'une population globale. Son objectif est de définir une zone dans laquelle les résultats d'un futur échantillon ont de fortes chances de tomber. Cet intervalle est essentiellement utilisé pour valider une hypothèse.
D'un autre côté, l'intervalle de confiance intervient lorsque l'on ignore tout de la population globale et que l'on ne dispose que des données d'un seul échantillon.
Il convient donc d'utiliser l'estimation calculée sur cet échantillon pour déduire une fourchette de valeurs. C'est un outil d'estimation qui va du particulier vers le général.

Question 4

Le biais est un indicateur qui mesure l'erreur systématique d'un estimateur.
Il représente la différence entre l'espérance mathématique de l'estimateur et la véritable valeur du paramètre que l'on cherche à estimer dans la population.
Autrement dit, le biais précise si la l'expérience d'échantillonnage réalisé a tendance en moyenne à être correcte, ou non, en surestimer ou sous-estimer la réalité.

Question 5

La statistique qui travaille sur la population totale s'appelle un recensement.
Ce dénombrement réalisé dans un cadre officiel vise l'exhaustivité en collectant des données dans chaque unité statistique de la population.
Le lien avec la notion de données massives (*Big Data*) réside dans la nature de collecte moderne d'informations. Ainsi, les données massives permettent d'atteindre la quasi-totalité d'une population rendant l'échantillonnage moins important pour certaines analyses.

Question 6

L'intérêt d'un estimateur est qu'il soit le plus proche possible de la valeur du paramètre. Il convient de trouver la méthode de calcul la plus fiable pour traduire les données d'un échantillon en une réalité globale. En fournissant une information partielle, l'estimateur doit pouvoir minimiser les erreurs et ce en ne contenant pas de biais. Il faut également qu'il y ait une variance minimale parmi tous les estimateurs qui ne sont pas biaisés. L'estimateur doit aussi être robuste et convergent.
Le choix de l'estimateur ne se fait pas au hasard et est entouré de plusieurs conditions.

Question 7

L'estimation d'un paramètre est un processus statistique consistant à utiliser les informations d'un échantillon pour déduire les caractéristiques inconnues d'une population globale. Nous pouvons parler de deux méthodes principales qui sont utilisées : l'estimation ponctuelle et l'estimation par intervalle.
L'estimation ponctuelle fournit une valeur numérique unique pour le paramètre recherché, comme l'utilisation de la moyenne d'un échantillon pour estimer celle de la population. Bien que directe, cette méthode ne permet pas de mesurer l'incertitude inhérente au tirage de l'échantillon.

Pour pallier cette limite, on utilise l'estimation par intervalle, ou intervalle de confiance, qui définit une plage de valeurs à l'intérieur de laquelle le véritable paramètre a une forte probabilité (souvent 95 %) de se trouver.
La sélection d'une méthode d'estimation repose sur plusieurs enjeux de solidité, en cherchant un estimateur sans biais.
Enfin, le choix dépend de la convergence de l'estimateur, qui doit se rapprocher de la valeur réelle à mesure que la taille de l'échantillon augmente, et de sa robustesse. Dans certains cas, on sélectionne un estimateur robuste, comme la médiane plutôt que la moyenne, pour limiter l'impact des valeurs aberrantes qui pourraient fausser les résultats. Le choix final est donc un arbitrage entre la justesse, la précision et la nature des données observées.

Question 8

Les tests statistiques sont un ensemble d'outils analytiques qui permettent de prendre des décisions rigoureuses à partir de données échantillonnées.
Leur rôle est de définir si un effet observé est réellement significatif ou s'il est dû au hasard. On identifie différents tests. D'abord, on retrouve les tests paramétriques (t de Student, F de Fisher), non paramétriques (Mann-Whitney, Whitney, Wilcoxon) ou d'ajustement (Chi-2).
Il faut énoncer une hypothèse nulle et une alternative pour réaliser un test.
Avec un seuil de risque, on établit une dimension critique et on calcule et compare pour décider d'accepter ou de rejeter l'hypothèse.

Question 9

La statistique inférentielle peut être critiquée, notamment pour son usage rigide et parfois abusif de ses outils pour valider des découvertes scientifiques.
Je crois que ces critiques peuvent s'entendre, notamment sur le fait qu'il y ait des résultats produits qui ne sont pas fiables entièrement car il y a une marge d'erreur dans l'estimation. Cependant, la statistique inférentielle reste un pilier de la statistique de données et se trouve être plus efficace que les méthodes d'échantillonnage.

Manipulations

Fréquences des échantillons : # Fréquences des échantillons (arrondies à 2 décimales)
f_pour = round(moy_p / total_moyennes, 2)
f_contre = round(moy_c / total_moyennes, 2)
f_sans = round(moy_s / total_moyennes, 2)

Fréquences de la population mère : # Fréquences de la population mère (Réalité)
pop_mere_total = 2185
p_pour = round(852 / pop_mere_total, 2)
p_contre = round(911 / pop_mere_total, 2)
p_sans = round(422 / pop_mere_total, 2)

Test de Shapiro-Wilk : rint("\nThéorie de la décision (Test de Shapiro-Wilk)")

# 1. Chargement des deux fichiers de test)
test1 = ouvrirUnFichier("./data/Loi-normale-Test-1.csv")
test2 = ouvrirUnFichier("./data/Loi-normale-Test-2.csv")

2. Application du test de Shapiro-Wilk
shapiro() retourne deux valeurs : la statistique (W) et la p-value
stat1, p_val1 = scipy.stats.shapiro(test1)
stat2, p_val2 = scipy.stats.shapiro(test2)

Test 1 : print(interpreter_shapiro(p_val1, "Test 1"))
Test 2 : print(interpreter_shapiro(p_val2, "Test 2"))

Séance 6

La statistique d’ordre des variables qualitatives

Question 1

Une statistique ordinale désigne une catégorie de données qualitatives dont les modalités peuvent être classées ou hiérarchisées selon un ordre logique, sans que l'on puisse pour autant mesurer précisément l'écart numérique entre elles. Contrairement aux données quantitatives, on ne peut pas effectuer d'opérations arithmétiques (comme une addition) sur ces valeurs, mais on peut établir une relation de supériorité ou d'infériorité (par exemple : "satisfait" est supérieur à "peu satisfait"). Elle s'oppose principalement à la statistique nominale. Alors que la statistique ordinale impose un rang (comme le niveau d'études ou une mention à un examen), la statistique nominale se contente de classer les individus dans des catégories mutuellement exclusives sans aucun ordre de valeur (comme le sexe, la couleur des yeux ou le département de résidence). Dans une variable nominale, aucune catégorie n'est "supérieure" à une autre, elles sont simplement différentes. Dans le domaine de l'analyse de données, cette statistique peut matérialiser une hiérarchie spatiale en traduisant l'organisation et l'importance relative de différents lieux. Par exemple, au lieu de mesurer simplement la population exacte, on peut classer des zones géographiques selon leur rang administratif ou leur niveau de centralité.

Question 2

Dans le domaine de la statistique et de l'analyse de données, le choix de l'ordre dans une classification n'est jamais anodin car il conditionne la lecture et l'interprétation des résultats. Pour les variables de nature qualitative ordinale, il est impératif de respecter la hiérarchie intrinsèque des modalités afin de ne pas perdre le sens de la mesure. On privilégiera alors un ordre logique, qu'il soit croissant ou décroissant, comme c'est le cas pour des niveaux de satisfaction, des échelons de diplômes ou des rangs administratifs. Cette organisation permet de matérialiser visuellement une progression ou une hiérarchie, facilitant ainsi la compréhension de la structure des données sans avoir besoin de recourir à des chiffres complexes.
À l'inverse, lorsqu'on traite des variables qualitatives nominales, qui ne possèdent pas d'ordre naturel, la stratégie change pour se focaliser sur l'importance statistique. Dans ce contexte, l'ordre à privilégier est celui de la fréquence décroissante des effectifs.

En classant les catégories de la plus représentée à la moins représentée, l'analyste permet une identification immédiate des groupes dominants et des tendances majeures de l'échantillon. Cette méthode est particulièrement efficace dans les graphiques de type Pareto ou les diagrammes en bâtons, où la clarté visuelle dépend de la capacité à distinguer rapidement les masses principales des catégories marginales, souvent regroupées sous l'étiquette "Autres".
Enfin, pour les variables quantitatives, qu'elles soient discrètes ou continues et regroupées en classes, l'ordre arithmétique des valeurs est le seul choix possible. Briser cet ordre numérique empêcherait toute analyse de la forme de la distribution, rendant impossible l'observation de la symétrie, de l'asymétrie ou de l'aplatissement de la courbe. En somme, l'ordre idéal est celui qui transforme un simple inventaire de données en une information structurée et scannable, que ce soit par le respect d'une logique préexistante ou par la mise en avant des poids statistiques les plus significatifs.

Question 3

La distinction entre la corrélation des rangs et la concordance de classements repose principalement sur le nombre d'observateurs ou de critères impliqués et sur l'objectif de la mesure statistique. Bien que ces deux notions traitent de données ordinales, elles ne s'appliquent pas aux mêmes configurations d'analyse.

La corrélation des rangs, illustrée par des coefficients comme le Rho de Spearman ou le Tau de Kendall, est utilisée pour mesurer l'intensité et la direction de la relation entre deux variables ordonnées pour un même groupe d'individus. Elle cherche à déterminer si, lorsqu'un individu monte dans le classement d'une première variable, il a tendance à monter (corrélation positive) ou à descendre (corrélation négative) dans le classement d'une seconde variable. C'est l'outil privilégié pour comparer deux séries de mesures, par exemple pour savoir si les élèves les mieux classés en mathématiques sont aussi les mieux classés en physique.

À l'inverse, la concordance de classements, souvent mesurée par le W de Kendall, intervient lorsque l'on souhaite évaluer l'accord global entre plus de deux classements. Elle ne s'intéresse plus seulement à la liaison entre deux variables, mais à la cohérence d'un ensemble de juges ou de critères sur un même objet. Par exemple, si dix critiques de cinéma doivent classer les mêmes cinq films, la concordance mesurera à quel point leurs avis convergent vers un consensus. Contrairement à la corrélation qui peut être négative, la concordance varie généralement de 0 (désaccord total) à 1 (unanimité parfaite), car elle quantifie la similarité d'opinion au sein d'un groupe.

Question 4

Le coefficient Rho de Spearman repose sur une logique de distance. Il transforme les données en rangs, puis calcule la corrélation de Pearson sur ces rangs. En d'autres termes, il mesure la force de la relation monotone entre deux variables en observant l'écart entre les positions des individus dans chaque classement. C'est une méthode très intuitive et largement utilisée car elle est facile à interpréter comme une extension de la corrélation classique, mais elle peut se montrer sensible aux valeurs extrêmes au sein des rangs.
À l'inverse, le coefficient Tau de Kendall s'appuie sur une logique de probabilité et de concordance. Au lieu de regarder les distances de rangs, il examine toutes les paires d'observations possibles pour déterminer si elles sont concordantes (elles évoluent dans le même sens) ou discordantes (elles évoluent en sens inverse). Cette approche rend le test de Kendall plus robuste et plus fiable, notamment sur les petits échantillons ou lorsque les données comportent de nombreuses exæquo (valeurs identiques). De plus, le Tau de Kendall offre une interprétation plus fine de la probabilité de trouver des paires dans le même ordre.

Question 5

Les coefficients de Goodman-Kruskal et de Yule sont des outils statistiques essentiels pour mesurer l'intensité de l'association entre deux variables qualitatives, particulièrement lorsque l'on travaille avec des tableaux de contingence. Bien qu'ils partagent l'objectif commun de quantifier un lien, ils s'appliquent à des structures de données différentes.
Le coefficient de Yule (notamment le Q de Yule) est spécifiquement conçu pour les tableaux de petite taille, plus précisément les tableaux de type 2x2. Il permet de mesurer l'association entre deux variables binaires ou dichotomiques, comme la présence ou l'absence d'un caractère. Ce coefficient varie entre -1 et +1 : une valeur proche de +1 indique une association positive parfaite, tandis qu'une valeur proche de -1 indique une opposition totale. Son grand intérêt réside dans sa simplicité de calcul et sa capacité à révéler si deux catégories s'attirent ou se repoussent mutuellement de manière significative.
Les coefficients de Goodman et Kruskal ont une portée plus large. Ils servent à mesurer l'association entre deux variables qualitatives ordinales, c'est-à-dire des variables dont les catégories suivent un ordre logique, au sein de tableaux de contingence pouvant aller bien au-delà du format 2x2. Tout comme le Tau de Kendall, le Gamma repose sur l'analyse des paires concordantes et discordantes : il évalue si, de manière générale, une position élevée dans la première variable tend à correspondre à une position élevée dans la seconde.

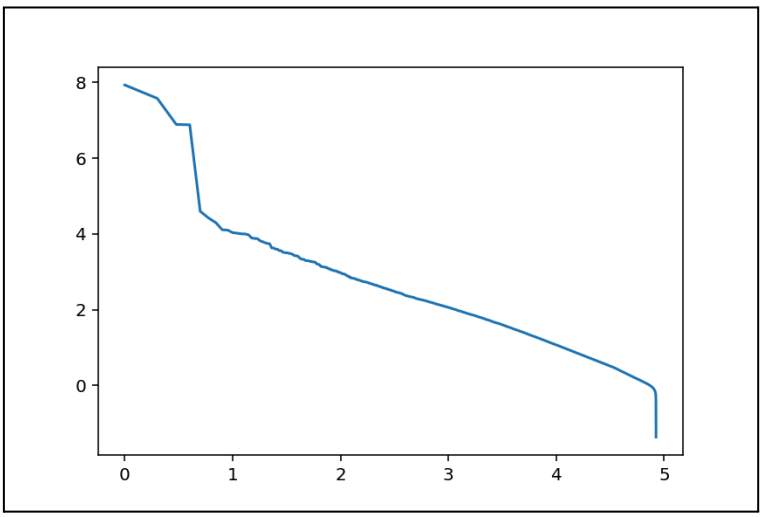
Dans la famille Goodman-Kruskal, on trouve également le Lambda (λ), qui est utilisé pour des variables nominales. Contrairement au Gamma, le Lambda mesure le pouvoir prédictif : il nous indique à quel point la connaissance de la catégorie d'une variable nous aide à prédire avec précision la catégorie de l'autre variable. Si le Lambda est égal à 0, cela signifie que la connaissance de la première variable n'apporte aucune aide pour deviner la seconde ; s'il est égal à 1, la prédiction est parfaite.

Manipulations

Ordre de la population : classe_pop2007 = ordrePopulation(pop2007, etats)
classe_pop2025 = ordrePopulation(pop2025, etats)
classe_den2007 = ordrePopulation(den2007, etats)
classe_den2025 = ordrePopulation(den2025, etats)

Classement des pays : comp2007 = classementPays(classe_pop2007, classe_den2007)
comp2025 = classementPays(classe_pop2025, classe_den2025)

Loi rang-taille (log-log)



Commentaire de résultat : On observe une courbe fortement décroissante produite par le code, avec les grandes surfaces qui sont présentes.
De fait, la Loi rang-taille possède une distribution asymétrique, avec un petit nombre de grandes valeurs et une majorité de petites valeurs.

Conclusion

Ce parcours démontre que la statistique n'est pas une fin en soi mais un langage et permet de mieux comprendre la géographie. Nous sommes partis des définitions épistémologiques du hasard et de la donnée (Séance 2), pour ensuite étudier les paramètres statistiques élémentaires (Séance 3). De plus, nous nous sommes intéressés aux distributions statistiques (Séance 4), avant d'analyser les statistiques inférentielles (Séance 5). *In fine*, nous avons abordé la question de la statistique d'ordre des variables qualitatives (Séance 6).
La statistique est un langage plutôt qu'un bloc visible de données, qui mérite d'être compris dans sa conception.
Cet outil est indispensable pour la géographie, avec le besoin de choisir les bons outils d'analyse essentiels pour une bonne pratique géographique.
Si la géographie est enseignée de manière littéraire, ces six séances ont été importantes pour changer de perspective et faire face à des problèmes nécessaires. L'étudiant de géographie se retrouve donc mieux armé pour mieux comprendre et pratiquer la géographie actuelle.

Remarques sur le cours

J’ai trouvé ce parcours d’Analyse de données difficile mais très intéressant pour comprendre les statistiques dans lesquelles j’ai été quelque peu initié en L3. Cependant, j’ai eu le sentiment de ne pas avoir su avancer efficacement dans ce semestre, la faute à un manque de compréhension des chemins des fichiers et d’un manque d’expertise en la matière. J’ai tenté de m’accrocher, bien que ça m’ait coûté du temps et de la fatigue, j’en ressort satisfait des concepts que j’ai appris et forgé, qui me seront précieux dans ma pratique de la géographie.
J’émet une réserve sur le fonctionnement de votre cours, bien que j’ai été séduit en début de semestre par cette organisation. Il aurait fallu à mon avis montrer les premières touches python et les chemins des fichiers pour que les personnes qui ont suivi un parcours littéraire comme moi s’y retrouvent (aucun déterminisme, bien sûr). Toutefois, j’ai beaucoup appris avec cette méthode de cours, en lisant les ressources sur votre GitHub, très complet et j’en suis satisfait.