

Maïwenn
LE BLANCHE
21415171



Module : Analyse de données - M. Forriez

RAPPORT D'ACTIVITÉ

Parcours débutants



source : Agence Niché, <https://agenceniche.com/tendance-actualite-evenementielle/analyse-donnees-optimiser-evenements/>

Master M1 GAED - EnviTERR
année universitaire 2025-2026

Sommaire :

Choix du parcours.....	3
Rapport d'activité - Séance 02.....	4-7
I/ Questions.....	
II/ Résultats.....	
III/ Les difficultés rencontrées et les constatations.....	
Rapport d'activité - Séance 03.....	8-10
I/ Questions.....	
II/ Résultats.....	
III/ Les difficultés rencontrées et les constatations.....	
Rapport d'activité - Séance 04.....	11-16
I/ Questions.....	
II/ Résultats.....	
III/ Les difficultés rencontrées et les constatations.....	
Rapport d'activité - Séance 05.....	17-19
I/ Questions.....	
II/ Résultats.....	
III/ Les difficultés rencontrées et les constatations.....	
Rapport d'activité - Séance 06.....	20-22
I/ Questions.....	
II/ Résultats.....	
III/ Les difficultés rencontrées et les constatations.....	
Retour personnel sur le module.....	23
Réflexion personnelle sur les humanités numériques.....	23-24

Choix du parcours :

Pour le module d'analyse de données, j'ai décidé de choisir le parcours débutants. Au départ, je souhaitais prendre "intermédiaires" puisque j'ai déjà eu par le passé des cours d'analyse univariée et bivariée. Cependant, ayant pris peur à cause de la programmation Python, j'ai choisi "débutants", en pensant cela plus accessible, puisque je n'ai que très peu d'expérience dans la programmation (j'ai autrefois fait un peu de R).

En finalité, je reconnais que ce n'était pas nécessairement le meilleur choix puisque les codes les plus accessibles se trouvaient dans les parcours plus élevés, ce que je n'avais pas compris au début du semestre. C'est un choix que j'ai tout de même décidé d'assumer.

Rapport d'activité - Séance 02

I/ Questions :

La géographie entretient une relation historiquement ambivalente avec la statistique. Elle s'est méfiée pendant longtemps de la statistique perçue comme extérieure à son champ disciplinaire alors même qu'elle produit des données massives nécessitant précisément ces outils pour être analysées. Cette tension conduit souvent à une sous-utilisation ou à un mauvais usage des méthodes statistiques par les géographes. Pourtant, la statistique constitue aujourd'hui un passage obligé pour faire de la géographie.

Il est impossible de prévoir le détail des comportements individuels ou des événements locaux, mais il est possible d'identifier des régularités globales et des tendances statistiques. La géographie adopte ainsi une posture intermédiaire entre nécessité et contingence. Le hasard n'empêche pas la mise en évidence de structures spatiales, dès lors que l'on raisonne à des échelles appropriées et que l'on distingue hasard bénin et hasard sauvage.

L'information géographique se divise en deux grands types. Le premier concerne les attributs des territoires, c'est-à-dire les caractéristiques humaines (population, activités économiques, structures sociales) ou physiques (température, précipitations, débits). Le second type porte sur la morphologie et la géométrie des ensembles spatiaux eux-mêmes. Dans un système d'information géographique, ces deux dimensions correspondent respectivement aux données attributaires et aux données géométriques.

La géographie a besoin de l'analyse de données pour comprendre la structure interne de jeux de données souvent volumineux et complexes. Cette analyse permet de résumer l'information, de dégager des tendances, d'identifier des relations entre variables et d'évaluer la fiabilité des résultats. Elle constitue une étape indispensable entre la production des données et l'interprétation géographique qui doit toujours confronter les résultats statistiques aux conditions de collecte et aux connaissances du terrain.

La statistique descriptive a pour objectif de décrire et de résumer les données observées à l'aide de paramètres numériques, de tableaux et de représentations graphiques, afin de fournir une image simplifiée de la réalité. Elle constitue une étape préalable indispensable. La statistique explicative vise quant à elle à relier des variables entre elles pour expliquer un phénomène ou prévoir des scénarios possibles en ajustant un modèle à partir des distributions de probabilité identifiées lors de la phase descriptive.

Les visualisations dépendent directement de la nature des variables étudiées. Les variables qualitatives sont représentées par des diagrammes en secteurs, tandis que les variables quantitatives continues sont visualisées par des histogrammes, des boîtes à moustaches ou des courbes cumulatives. Le choix d'une visualisation repose donc sur le type de variable et sur l'objectif de l'analyse, qu'il s'agisse de décrire une distribution, de comparer des modalités etc.

On distingue trois grandes familles de méthodes d'analyse de données. Les méthodes descriptives visent à visualiser et classer les données. Les méthodes explicatives cherchent à relier une variable à expliquer à des variables explicatives à l'aide de modèles statistiques. Enfin, les méthodes de prévision portent sur les séries chronologiques et visent à modéliser l'évolution d'un phénomène dans le temps.

La population statistique est l'ensemble des unités étudiées, tandis que l'individu statistique est un élément de cette population. Les caractères statistiques sont les propriétés observées sur chaque individu et leurs modalités sont les valeurs possibles de ces caractères qui doivent être exhaustives et exclusives. Les caractères se répartissent en variables qualitatives et quantitatives, elles-mêmes subdivisées en nominales, ordinaires, discrètes et continues.

L'amplitude correspond à la largeur d'une classe statistique et se calcule comme la différence entre sa borne supérieure et sa borne inférieure. Elle concerne toujours une classe et non l'ensemble de la série. La densité est définie comme le rapport entre l'effectif d'une classe et son amplitude. Elle permet de comparer des classes de tailles différentes et constitue un élément essentiel de la représentation par histogramme lorsque les classes ne sont pas d'amplitude égale.

Les formules de Sturges et de Yule servent à déterminer un nombre approximatif de classes lors de la discréttisation d'une variable quantitative continue. Elles visent à éviter deux écueils majeurs : un découpage trop fin, qui complique la lecture, et un découpage trop grossier, qui entraîne une perte d'information.

L'effectif correspond au nombre d'occurrences d'une modalité ou d'une classe dans la population étudiée. La fréquence est le rapport entre cet effectif et l'effectif total, ce qui permet de raisonner en proportions. La fréquence cumulée est obtenue en additionnant les fréquences des modalités inférieures ou égales à une valeur donnée. L'ensemble des effectifs ou des fréquences associés aux modalités d'un caractère constitue une distribution statistique, qui permet d'identifier empiriquement la forme générale de la répartition des données.

II/ Résultats :

étape 5	Afficher sur le terminal exécutant le conteneur la variable contenu	
étape 6	Calculer le nombre de lignes et de colonnes du tableau de données	
étape 7	Faire le point sur la nature statistique des variables en utilisant le lien vers les métadonnées fournies en commentaire. Faire une liste sur le type de chaque colonne (int, float, str ou bool).	
étape 8	Afficher sur le terminal le nom des colonnes	
étape 9	Afficher le nombre des inscrits	
étape 10	Calculer les effectifs de chaque colonne et les placer dans une liste	

étape 11	Faire des diagrammes en barres avec le nombre des inscrits et des votants pour chaque département.	<p>The bar chart displays two categories: 'Inscrits' (Registered) and 'Votants' (Voters). The y-axis represents the 'Nombre de personnes' (Number of people) from 0 to 400,000. The x-axis is labeled 'Catégories' (Categories). The 'Inscrits' bar reaches approximately 420,000, while the 'Votants' bar reaches approximately 340,000.</p> <table border="1"> <thead> <tr> <th>Catégorie</th> <th>Nombre de personnes</th> </tr> </thead> <tbody> <tr> <td>Inscrits</td> <td>~420,000</td> </tr> <tr> <td>Votants</td> <td>~340,000</td> </tr> </tbody> </table>	Catégorie	Nombre de personnes	Inscrits	~420,000	Votants	~340,000												
Catégorie	Nombre de personnes																			
Inscrits	~420,000																			
Votants	~340,000																			
étape 12	Faire des diagrammes circulaires avec les votes blancs, nuls, exprimés et l'abstention pour chaque département.	<p>The pie chart shows the distribution of votes for department 65. The segments are: 'Exprimés' (76.0%, green), 'Abstention' (22.3%, red), and 'Blancs' (0.7%, blue).</p> <table border="1"> <thead> <tr> <th>Type de vote</th> <th>Pourcentage</th> </tr> </thead> <tbody> <tr> <td>Exprimés</td> <td>76.0%</td> </tr> <tr> <td>Abstention</td> <td>22.3%</td> </tr> <tr> <td>Blancs</td> <td>0.7%</td> </tr> </tbody> </table>	Type de vote	Pourcentage	Exprimés	76.0%	Abstention	22.3%	Blancs	0.7%										
Type de vote	Pourcentage																			
Exprimés	76.0%																			
Abstention	22.3%																			
Blancs	0.7%																			
étape 13	Faire l'histogramme de la distribution des inscrits	<p>The histogram displays the normalized distribution of registered voters ('Nombre d'inscrits') on the x-axis (from 0.00 to 1.75e6) against density on the y-axis (from 0.0 to 1.4). The distribution is skewed right, with the highest frequency in the first bin (0.00 to 0.25).</p> <table border="1"> <thead> <tr> <th>Nombre d'inscrits (bin)</th> <th>Densité</th> </tr> </thead> <tbody> <tr> <td>[0.00, 0.25]</td> <td>~1.25</td> </tr> <tr> <td>[0.25, 0.50]</td> <td>~1.45</td> </tr> <tr> <td>[0.50, 0.75]</td> <td>~1.10</td> </tr> <tr> <td>[0.75, 1.00]</td> <td>~0.40</td> </tr> <tr> <td>[1.00, 1.25]</td> <td>~0.20</td> </tr> <tr> <td>[1.25, 1.50]</td> <td>~0.15</td> </tr> <tr> <td>[1.50, 1.75]</td> <td>~0.10</td> </tr> <tr> <td>[1.75, 2.00]</td> <td>~0.05</td> </tr> </tbody> </table>	Nombre d'inscrits (bin)	Densité	[0.00, 0.25]	~1.25	[0.25, 0.50]	~1.45	[0.50, 0.75]	~1.10	[0.75, 1.00]	~0.40	[1.00, 1.25]	~0.20	[1.25, 1.50]	~0.15	[1.50, 1.75]	~0.10	[1.75, 2.00]	~0.05
Nombre d'inscrits (bin)	Densité																			
[0.00, 0.25]	~1.25																			
[0.25, 0.50]	~1.45																			
[0.50, 0.75]	~1.10																			
[0.75, 1.00]	~0.40																			
[1.00, 1.25]	~0.20																			
[1.25, 1.50]	~0.15																			
[1.50, 1.75]	~0.10																			
[1.75, 2.00]	~0.05																			

⇒ Paragraphe de synthèse sur les résultats obtenus :

Lors de cette séance, nous avons pu réaliser une analyse statistique descriptive du fichier des résultats du premier tour de l'élection présidentielle de 2022 par département.

Les sommes calculées sur les variables quantitatives (inscrits, votants, abstentions, votes exprimés, blancs et nuls) fournissent une première vision agrégée des volumes électoraux.

Ensuite, les diagrammes en barres permettent de comparer, département par département, les inscrits et les votants, tandis que les diagrammes circulaires représentent la structure interne du vote pour chaque territoire. Par exemple, on peut remarquer dans le diagramme en barre de l'Ain (01), le nombre de votants est inférieur au nombre d'inscrits.

Cette différence est significative puisqu'elle montre qu' $\frac{1}{4}$ des inscrits ne participent pas au vote.

Enfin, l'histogramme des inscrits met en évidence la distribution statistique de la population électorale départementale, conformément aux principes vus dans le cours sur les distributions.

Tout ceci permet de décrire des phénomènes géographiques à partir de données électorales.

III/ Les difficultés rencontrées et les constatations :

J'ai eu des difficultés à afficher correctement les diagrammes. En effet, j'ai souhaité rendre plus lisible les diagrammes circulaires, puisque certaines données, du fait de leur faible proportion, s'entremêlent visuellement. J'ai essayé de modifier le code pour corriger cela mais je ne suis pas parvenue à un rendu satisfaisant.

Rapport d'activité - Séance 03

I/ Questions :

L'analyse statistique repose d'abord sur la nature du caractère étudié. Entre caractère quantitatif et qualitatif, le plus général est le caractère qualitatif, car il permet de décrire toute population, y compris lorsque la mesure numérique est impossible. Le caractère quantitatif constitue alors un cas particulier. En effet, celui-ci renvoie à des valeurs numériques ordonnées, tandis que les caractères qualitatifs regroupent des modalités ou catégories plus larges.

Parmi les caractères quantitatifs, il faut distinguer les variables discrètes et les variables continues. Les premières prennent des valeurs isolées, alors que les secondes s'inscrivent dans un intervalle continu. Cette distinction est essentielle car elle conditionne les méthodes de calcul, que cela soit somme ou intégrale, fréquence ou densité.

Il existe plusieurs types de moyennes (arithmétique, harmonique, géométrique, quadratique). Chacune de ces moyennes répond à un usage particulier. Leur coexistence découle des propriétés mathématiques recherchées, comme la sensibilité aux valeurs extrêmes ou la prise en compte des proportions. Par exemple, la moyenne arithmétique mesure la tendance centrale simple, tandis que la moyenne harmonique est adaptée aux vitesses.

La médiane est définie comme la valeur qui partage la distribution en deux parties égales. Elle est utilisée car elle n'est pas influencée par les valeurs extrêmes et convient mieux aux distributions asymétriques. Son existence dépend du classement des données et de la structure discrète ou continue de la variable.

Le mode correspond à la valeur la plus fréquente ou de densité maximale. On ne peut le calculer que si une fréquence maximale est identifiable. Il peut ne pas exister (distribution uniforme) ou être multiple (distribution bimodale), ce qui renvoie souvent à la présence de sous-populations.

La médiale et l'indice de Gini permettent d'évaluer la répartition interne d'un caractère. La médiale divise la masse totale d'un caractère en deux parts égales et son écart avec la médiane mesure le degré de concentration. L'indice de Gini issu de la courbe de Lorenz formalise cette inégalité. Ces outils sont pertinents lorsque l'homogénéité ou la dispersion ne peuvent être saisies uniquement par les paramètres de position.

La variance mesure l'écart quadratique moyen à la moyenne. On utilise le carré des écarts plutôt que la simple distance car le carré possède des propriétés algébriques utiles. L'écart type est ensuite la racine carrée de la variance, ce qui permet de revenir à l'unité de la variable tout en conservant l'information de dispersion.

L'étendue est la différence entre la valeur maximale et la valeur minimale. Elle donne une première idée de l'amplitude des données, mais elle dépend uniquement des valeurs extrêmes et perd rapidement sa pertinence lorsque les effectifs augmentent.

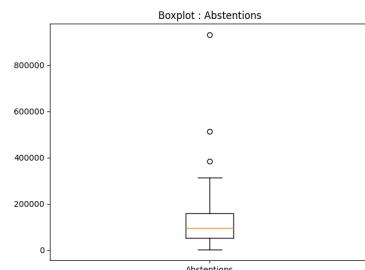
Les quantiles divisent une distribution en parts égales. Leur utilité repose sur leur capacité à décrire la répartition. Les quantiles les plus utilisés sont les quartiles, notamment pour la construction de l'écart interquartile et des boîtes de dispersion.

La boîte à moustaches ou box-plot représente graphiquement les quartiles, la médiane et les valeurs extrêmes. Elle permet une lecture rapide de la distribution, de sa symétrie, de ses valeurs aberrantes et de la comparaison entre séries différentes.

Les moments centrés mesurent l'écart des valeurs à la moyenne en tenant compte du signe (sauf pour les moments d'ordre pair). Les moments absolus, eux, s'affranchissent du signe dans le calcul et donnent des mesures plus robustes dans certains cas. Ils servent à caractériser la forme générale des distributions.

La symétrie se vérifie notamment par le moment centré d'ordre 3 (asymétrie ou skewness). Une distribution est symétrique si ce moment est nul.

II/ Résultats :

étape 5	Calculer : les moyennes de chaque colonne les médianes de chaque colonne les modes de chaque colonne l'écart type de chaque colonne l'écart absolu à la moyenne de chaque colonne l'étendue de chaque colonne	
étape 6	Afficher la liste des paramètres sur le terminal	
étape 7	Calculer la distance interquartile et interdécile de chaque colonne quantitative	
étape 8	faire des boîtes à moustache de chaque colonne quantitative	
étape 10	Concevoir un organigramme	

⇒ Paragraphe de synthèse sur les résultats obtenus :

Lors de cette séance, nous avons pu approfondir l'analyse statistique descriptive des données électorales en mobilisant des indicateurs de tendance centrale et de dispersion calculés sur les variables quantitatives.

Le calcul des moyennes, médianes etc. permet de caractériser les valeurs centrales des effectifs électoraux, tandis que l'écart type, l'écart absolu à la moyenne et l'étendue renseignent sur la dispersion des données autour de ces valeurs. Le recours aux quantiles (quartiles et déciles) nous permet de compléter l'analyse avec une distribution interne des variables, indépendante des valeurs extrêmes.

Les boîtes à moustaches produites synthétisent les informations en rendant visibles la médiane, la dispersion interquartile et les éventuelles asymétries des distributions.

L'ensemble des résultats obtenus s'inscrit dans une démarche de statistique descriptive.

III/ Les difficultés rencontrées et les constatations :

J'ai eu des problèmes avec des conteneurs orphelins qui ont posé problème. J'ai retrouvé la formule "*docker-compose down –remove-orphans*" pour résoudre le problème et retirer les conteneurs inutiles.

Rapport d'activité - Séance 04

I/ Questions :

Le premier critère essentiel pour choisir entre une distribution statistique discrète ou continue est la nature du phénomène étudié. Ce choix dépend avant tout des caractéristiques intrinsèques de la variable. Certaines réalités sont dénombrables (nombre d'événements, individus, objets) tandis que d'autres sont mesurables sur un continuum (distances, températures, durées). Le choix de la loi repose sur la nature du phénomène étudié afin de choisir entre loi discrète et loi continue. Lorsqu'un phénomène ne peut prendre que des valeurs isolées, on a nécessairement affaire à une variable discrète. À l'inverse, lorsqu'une variable se définit sur un ensemble continu de valeurs, aucune probabilité n'est assignée à un point particulier mais à un intervalle. C'est le principe même d'une variable continue, pour laquelle " $\Pr(X = x) = 0$ ".

Un second critère déterminant est la forme empirique de la distribution observée. La distribution réelle doit être examinée pour orienter le choix du modèle. La forme de la distribution empirique intervient directement dans la sélection de la loi adéquate. Une distribution en escalier suggère un processus discret, tandis qu'une densité lisse et continue orientera vers un modèle continu. Ce critère est essentiel en pratique car il permet d'adapter la loi théorique à la réalité observée.

Un troisième critère concerne les paramètres descriptifs disponibles, comme l'espérance, la variance, la médiane, l'asymétrie ou encore l'aplatissement. Il est nécessaire d'interpréter ces caractéristiques afin de choisir entre les lois. Par exemple, une moyenne identique à la variance peut orienter vers une loi de Poisson, tandis qu'un phénomène symétrique avec une variance définie peut correspondre à une loi normale. Le comportement statistique de la variable oriente ainsi le choix du modèle, notamment dans les situations où plusieurs lois sont a priori envisageables.

Enfin, le nombre de paramètres constitue un critère méthodologique important. Une loi comportant davantage de paramètres est capable de mieux épouser la forme d'une distribution empirique complexe. Une loi dépendant de plusieurs paramètres peut s'adapter plus facilement à une distribution. C'est particulièrement vrai lorsque l'on compare des lois simples et des lois paramétriques plus flexibles. Le choix final doit donc équilibrer entre simplicité, conformité empirique et pertinence théorique.

On retrouve plusieurs lois utilisées en géographie. La première d'entre elles est la loi de Zipf, présentée comme un outil directement appliquée au système urbain. Cette loi figure parmi les modèles typiques utilisés en analyse territoriale. En géographie, cette loi se rencontre dans les lois rang-taille confrontant, au sein d'un territoire, le nombre d'habitants d'une ville avec son rang. Ce modèle permet d'analyser la hiérarchie des villes, les mécanismes de concentration urbaine et la structuration des réseaux de centralités. Sa généralisation par Zipf-Mandelbrot permet de décrire des systèmes urbains plus complexes ou déviants du schéma idéal.

Une autre famille de lois particulièrement mobilisée est celle des lois normales. Elles ont une grande importance statistique générale puisqu'elles constituent la distribution statistique la plus fréquente pour de nombreux phénomènes aléatoires. En géographie, elles interviennent dans l'analyse de grandeurs continues comme les altitudes, les températures, ou encore les résidus de modèles statistiques. Elles servent également à l'établissement

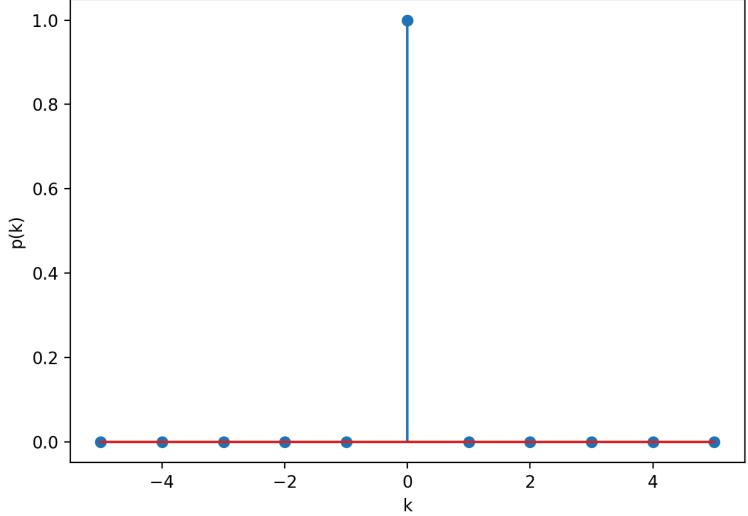
d'intervalles de confiance, à la modélisation des erreurs de mesure et à l'évaluation de la variabilité d'un phénomène spatialement distribué.

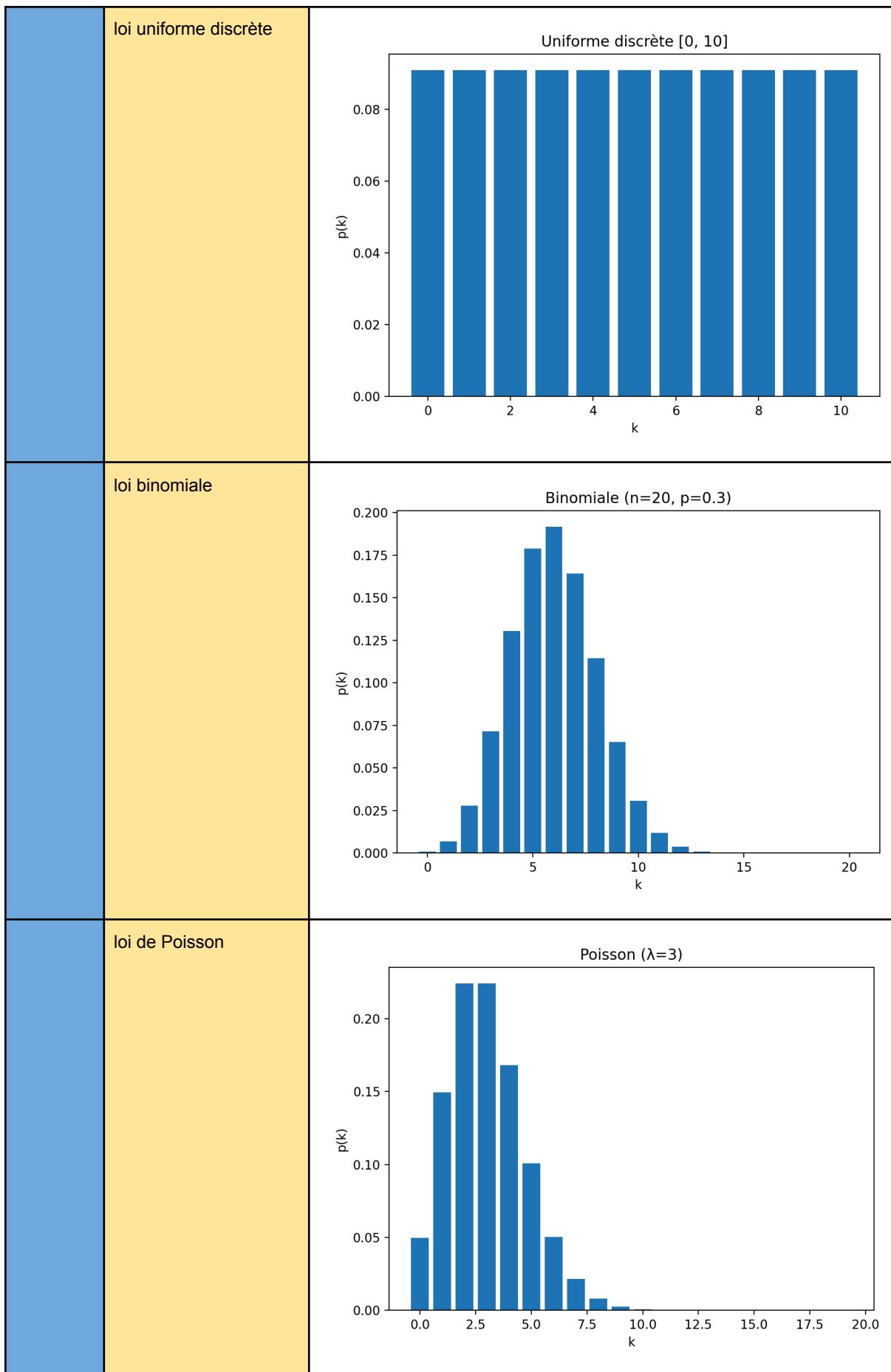
Il y a aussi la loi log-normale qui est essentielle pour modéliser les phénomènes multiplicatifs ou les distributions très asymétriques. En géographie, de nombreuses grandeurs physiques ou socio-économiques suivent ce type de distribution, comme les surfaces des bassins versants ou les tailles de populations, etc. Ce modèle permet de comprendre des dynamiques de croissance proportionnelle ou cumulative, fréquentes dans les systèmes géographiques.

Les lois de Poisson et les lois exponentielles se révèlent également utiles pour travailler sur la distribution d'événements dans l'espace ou dans le temps. La loi de Poisson est une loi des événements rares, ce qui en fait un modèle pertinent pour analyser les occurrences d'événements à faible probabilité mais à répétition possible. La loi exponentielle associée aux processus de Poisson modélise les durées entre événements selon un taux constant, comme le temps de retour ou bien le temps d'attente par exemple.

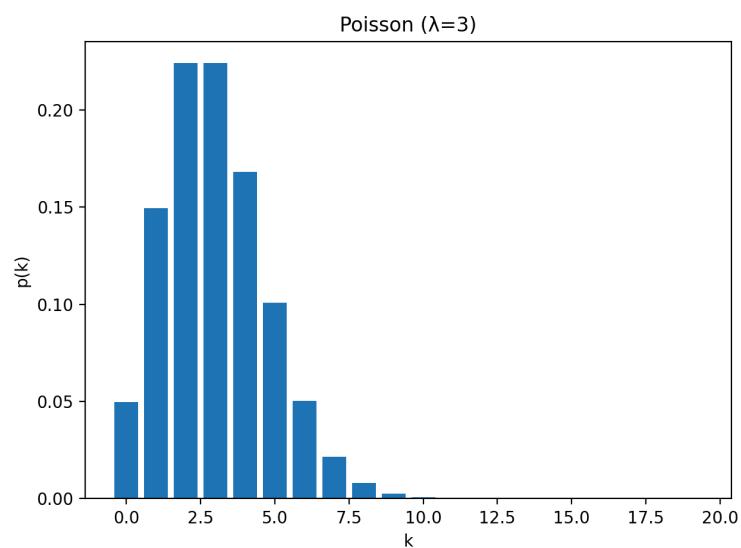
Enfin, la loi uniforme continue occupe une place particulière dans la modélisation théorique et dans la simulation. En effet, elle traduit l'hypothèse d'équirépartition, ce qui en fait un modèle de référence pour les espaces isotropes en géographie théorique ou pour la génération aléatoire de points dans des analyses spatiales. Elle est également utilisée dans les approches bayésiennes lorsque l'on modélise une ignorance totale ou une absence d'information préalable.

II/ Résultats :

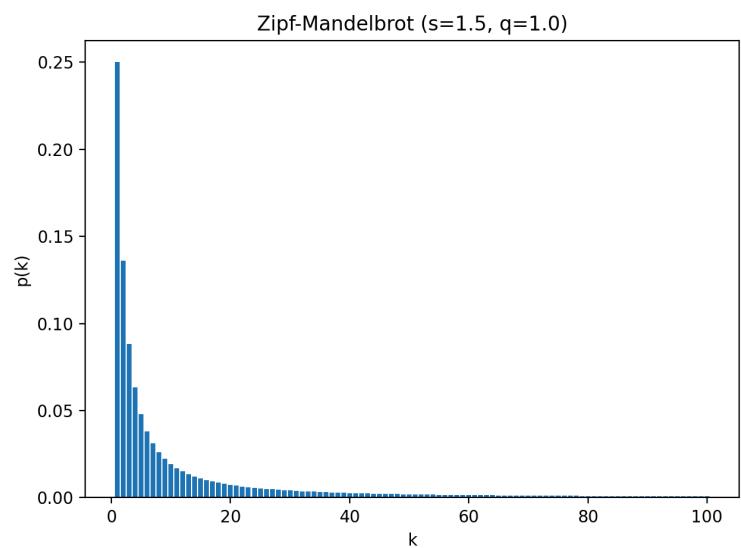
étape 1	Visualiser :																									
	loi de Dirac	 <p style="text-align: center;">Loi de Dirac ($k_0=0$)</p> <table border="1"> <caption>Data for Loi de Dirac ($k_0=0$)</caption> <thead> <tr> <th>k</th> <th>p(k)</th> </tr> </thead> <tbody> <tr><td>-5.0</td><td>0.00</td></tr> <tr><td>-4.0</td><td>0.00</td></tr> <tr><td>-3.0</td><td>0.00</td></tr> <tr><td>-2.0</td><td>0.00</td></tr> <tr><td>-1.0</td><td>0.00</td></tr> <tr><td>0.0</td><td>1.00</td></tr> <tr><td>1.0</td><td>0.00</td></tr> <tr><td>2.0</td><td>0.00</td></tr> <tr><td>3.0</td><td>0.00</td></tr> <tr><td>4.0</td><td>0.00</td></tr> <tr><td>5.0</td><td>0.00</td></tr> </tbody> </table>	k	p(k)	-5.0	0.00	-4.0	0.00	-3.0	0.00	-2.0	0.00	-1.0	0.00	0.0	1.00	1.0	0.00	2.0	0.00	3.0	0.00	4.0	0.00	5.0	0.00
k	p(k)																									
-5.0	0.00																									
-4.0	0.00																									
-3.0	0.00																									
-2.0	0.00																									
-1.0	0.00																									
0.0	1.00																									
1.0	0.00																									
2.0	0.00																									
3.0	0.00																									
4.0	0.00																									
5.0	0.00																									



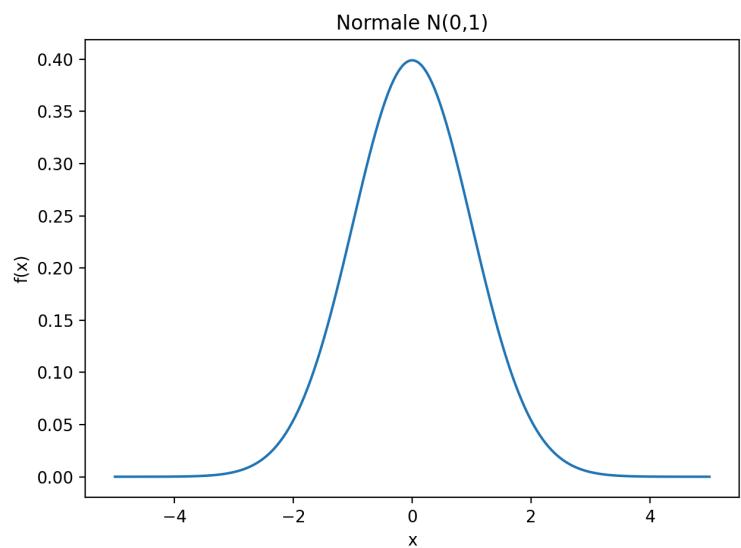
loi de Poisson continue



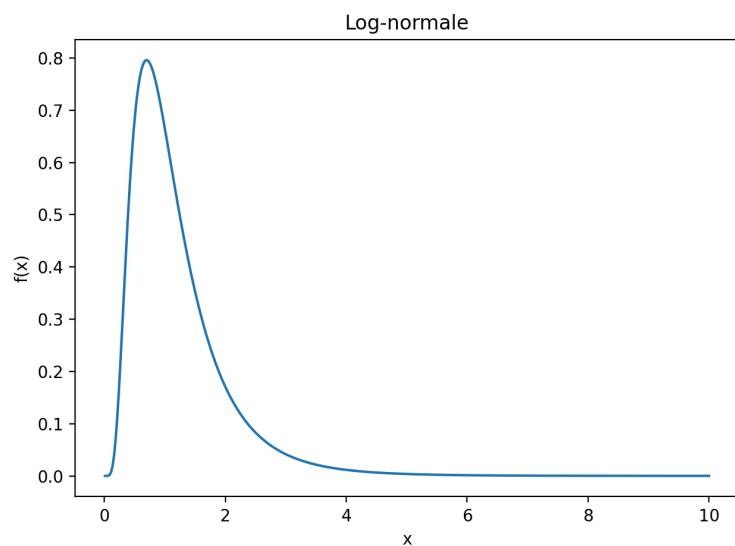
loi de Zipf-Mandelbrot



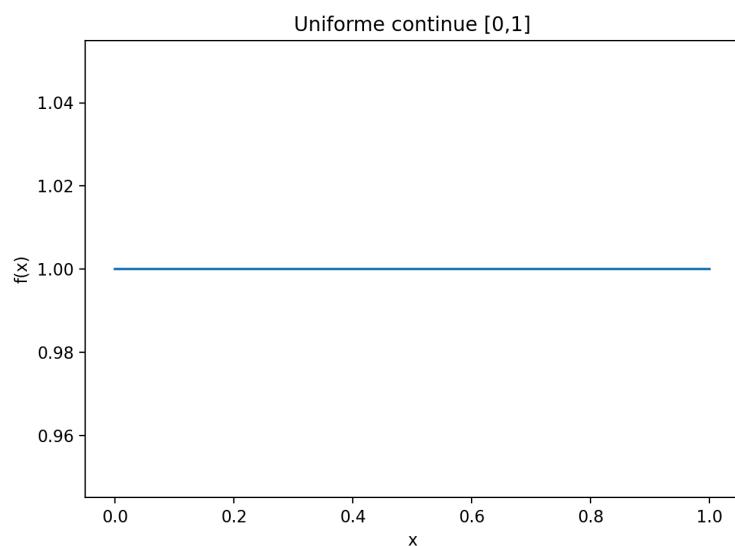
loi normale



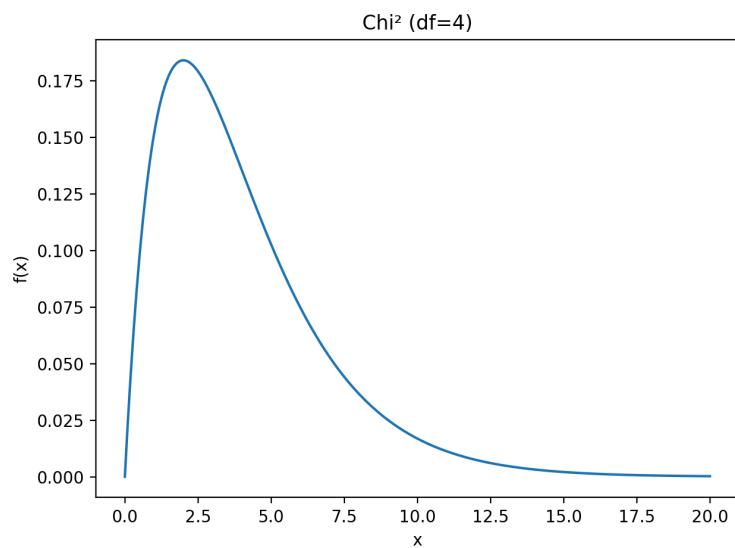
loi log-normale

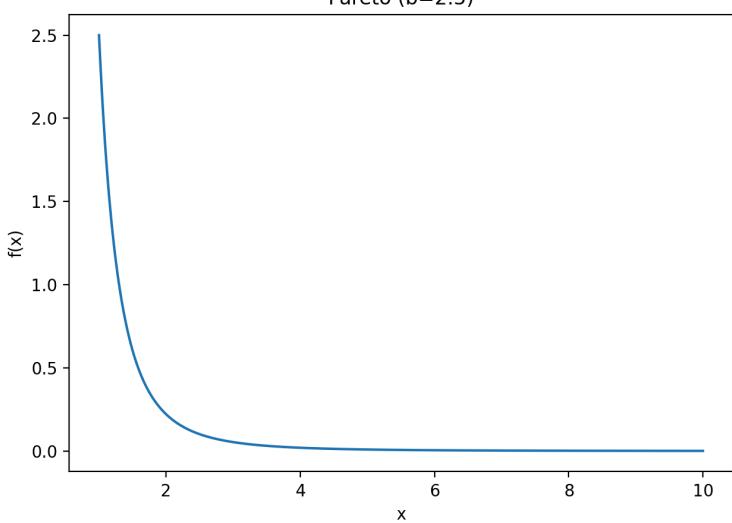


loi uniforme



loi du χ^2



	loi de Pareto	
étape 2	Calculer la moyenne, l'écart-type	

⇒ **Paragraphe de synthèse sur les résultats obtenus :**

Dans cette séance, on génère la visualisation de distributions statistiques théoriques pour un ensemble de lois discrètes ; loi de Dirac, loi uniforme discrète, loi binomiale, loi de Poisson, loi de Zipf-Mandelbrot, et continues ; loi de Poisson traitée en continu, loi normale, loi log-normale, loi uniforme, loi du χ^2 , loi de Pareto. Cela nous permet de visualiser les caractéristiques de chaque loi.

Nous avons ensuite calculé la moyenne et l'écart-type de ces distributions simulées. Cela permet de faciliter la différenciation entre les lois.

III/ Les difficultés rencontrées et les constatations :

Grâce aux conseils de mes camarades de classe, j'ai réalisé pas à pas chaque loi, et n'ai donc pas rencontré de problème spécifique lors de cette séance.

Rapport d'activité - Séance 05

I/ Questions :

Les statistiques inférentielles reposent sur l'idée qu'il est généralement impossible, trop coûteux ou inutile d'observer l'intégralité d'une population. Dans de nombreuses situations comme l'étude d'intentions de vote de plusieurs dizaines de millions d'individus, on privilégie l'analyse d'un échantillon limité mais représentatif car il fournit une approximation suffisamment fiable des caractéristiques de la population mère tout en réduisant considérablement les coûts de collecte et de traitement des données. L'échantillonnage consiste alors à prélever un sous-ensemble tiré au hasard dans la population, ce qui permet d'étendre les résultats obtenus à l'ensemble étudié en tenant compte de la fluctuation d'échantillonnage.

Plusieurs méthodes d'échantillonnage existent. Les méthodes aléatoires (avec ou sans remise) garantissent l'équiprobabilité des tirages tandis que les méthodes non aléatoires, telles que l'échantillonnage systématique ou la méthode des quotas, reproduisent la structure de la population sans recours au tirage pur. Le choix dépend de l'accessibilité à une base de sondage, du coût des tirages répétés, et de la capacité à garantir la représentativité de l'échantillon.

Sur la base de ces échantillons, la statistique inférentielle vise à estimer des paramètres inconnus de la population. Pour cela, elle mobilise des estimateurs, définis comme des fonctions des observations issues de l'échantillon. Une fois calculé sur des données observées, l'estimateur fournit une estimation. La qualité d'un estimateur repose sur des propriétés théoriques comme l'absence de biais (écart nul entre l'espérance de l'estimateur et la valeur réelle du paramètre), la faible variance, la convergence lorsque la taille de l'échantillon augmente et l'efficacité au sens de la borne de Cramér-Rao. Par exemple, la moyenne empirique constitue un estimateur sans biais et convergent de la moyenne réelle, tandis que la variance empirique non corrigée est biaisée d'où la nécessité d'appliquer le facteur correctif $n/(n-1)$ pour obtenir un estimateur sans biais.

L'inférence statistique distingue également deux outils fondamentaux : l'intervalle de fluctuation et l'intervalle de confiance. Le premier s'applique lorsque la vraie valeur du paramètre est connue. Il décrit l'ensemble des valeurs possibles que peut prendre une fréquence observée lors d'un tirage aléatoire de taille n . Il permet donc d'évaluer si une observation est compatible avec le paramètre théorique au seuil considéré. L'intervalle de confiance, lui, s'utilise lorsque le paramètre est inconnu. Construit à partir d'un estimateur et de son erreur standard, il encadre avec un niveau de confiance donné la valeur probable du paramètre réel. La logique s'inverse donc. Le premier vérifie la compatibilité d'une observation avec un paramètre connu. Le second encadre un paramètre inconnu à partir d'une observation.

Les statistiques inférentielles s'intéressent aussi à la question des statistiques portant sur la totalité de la population. Lorsqu'une statistique résume sans perte l'ensemble de l'information contenue dans l'échantillon, on parle de statistique exhaustive. Dans un contexte de données massives, cette idée prend une importance particulière. Lorsque la population entière est observée, l'inférence perd sa fonction traditionnelle, car il n'est plus nécessaire d'estimer les paramètres. Ceux-ci peuvent être calculés directement, et la variabilité due à l'échantillonnage disparaît.

Le choix d'un estimateur soulève plusieurs enjeux : garantir la précision (variance minimale), l'absence de biais, la convergence vers la valeur réelle, mais aussi la robustesse. Les méthodes classiques sont sensibles aux valeurs aberrantes. Les estimateurs robustes sont importants car capables de limiter l'influence de ces observations atypiques. À côté de ces méthodes analytiques, des approches numériques comme la méthode de Monte-Carlo ou le bootstrap permettent de simuler des distributions d'estimateurs dans des situations complexes ou mathématiquement intraitables.

Enfin, les tests statistiques constituent un outil essentiel de l'inférence. Ils permettent de juger si les données observées sont compatibles avec une hypothèse en contrôlant un risque d'erreur prédéfini. Leur construction s'appuie sur les lois de probabilité des estimateurs et sur les intervalles de fluctuation ou de confiance, selon que le paramètre théorique est connu ou non. Ils interviennent notamment dans les tests de signification qui sont omniprésents dans les applications pratiques de la statistique.

Des critiques de la statistique inférentielle, bien que non formulées émergent alors : dépendance à des hypothèses souvent idéalisées, risque de biais si l'échantillon n'est pas représentatif, sensibilité excessive de certains estimateurs aux données aberrantes, ou encore obsolescence relative dans les contextes où la population totale devient observable. Ces limites rappellent que l'inférence est un ensemble d'outils puissants, mais qui nécessitent une utilisation prudente et éclairée, en tenant compte des fondements théoriques de chaque méthode.

III/ Résultats :

étape 1	Calculer la moyenne pour chaque colonne	
	Calculer la somme des 3 moyennes obtenues, diviser le résultat par l'ensemble des moyennes. Faire de même avec les fréquences de la population mère	
	Calculer intervalle de fluctuation et les valeurs réelles de la population mère. Que pouvez-vous en conclure ?	Intervalle de fluctuation à 95 % : {'Pour': (0.36, 0.42), 'Contre': (0.39, 0.45), 'Sans opinion': (0.17, 0.21)} Ce résultat signifie que l'échantillon est représentatif de la population mère pour la variable étudiée. On peut donc utiliser les résultats de l'échantillon pour décrire la population, dans la limite de l'incertitude statistique associée au niveau de confiance de 95 %.
étape 2	Calculer la somme de la ligne et les fréquences	Fréquences du premier échantillon : [0.4, 0.4, 0.21]
	Calculer l'intervalle pour chaque opinion	Intervalle de confiance du premier échantillon à 95 % : [(0.37, 0.43), (0.37, 0.43), (0.18, 0.24)]
	Interpréter le résultat obtenu et comparer avec le résultat	Les intervalles de confiance à 95 % contiennent les proportions réelles de la population mère. Comme précédemment avec les intervalles de fluctuation, les écarts observés sont attribuables au hasard d'échantillonnage. L'échantillon est donc représentatif de la population.

	précédent	
étape 3	Laquelle est une distribution normale ?	Test de normalité Shapiro-Wilk : Fichier Test 1 -> Stat=0.964, p=0.000 -> Non normale Fichier Test 2 -> Stat=0.261, p=0.000 -> Non normale Il n'y a pas de distribution normale.
	Pourquoi c'est une distribution normale ?	Il n'y a pas de distribution normale.

⇒ **Paragraphe de synthèse sur les résultats obtenus :**

Les résultats de cette séance montrent la cohérence globale des outils mobilisés pour relier un échantillon à sa population mère. Le calcul des moyennes par colonne à partir des 100 échantillons permet d'obtenir une estimation globale des effectifs “Pour”, “Contre” et “Sans opinion”, qui, une fois transformée en fréquences, présente des écarts avec les valeurs réelles de la population mère. Toutefois, le calcul des intervalles de fluctuation à 95 % montre que les fréquences réelles de la population mère sont bien contenues dans ces intervalles, ce qui indique que les écarts observés sont compatibles avec le hasard d'échantillonnage et que l'échantillon est représentatif. Cette conclusion est confirmée par la théorie de l'estimation. Pour le premier échantillon, les intervalles de confiance à 95 % calculés pour chaque opinion englobent également les proportions réelles de la population mère, conduisant à une interprétation similaire. Enfin, la théorie de la décision, mobilisée à travers le test de normalité de Shapiro-Wilk, montre que les deux distributions testées ne suivent pas une loi normale ce qui souligne l'importance de vérifier les hypothèses statistiques avant d'appliquer certains tests.

III/ Les difficultés rencontrées et les constatations :

Du fait de la précipitation, j'ai commencé ma programmation sans télécharger les fichiers csv. De ce fait, un message d'erreur s'est affiché dans le terminal. Il m'a suffi de revenir sur votre GitHub et de télécharger les données pour résoudre le problème.

Rapport d'activité - Séance 06

I/ Questions :

La statistique ordinale occupe une place centrale dans l'analyse géographique car elle repose sur l'établissement de classements permettant d'ordonner des objets, des individus ou des territoires. Une statistique ordinale utilise des variables ordinales. Elle s'oppose ainsi aux statistiques nominales qui portent sur des catégories sans ordre. Grâce à cet ordre, les statistiques ordinaires permettent de matérialiser et analyser des hiérarchies spatiales car classer des villes, des régions ou des phénomènes revient à représenter leur position relative dans un espace social ou géographique. C'est pourquoi la statistique ordinale constitue le "cœur de la géographie humaine", où classements urbains, hiérarchies de centralités, ou positions socio-économiques sont omniprésents.

La première règle méthodologique de l'ordination consiste à privilégier l'ordre croissant. Cet ordre facilite l'identification d'anomalies comme les valeurs extrêmes et permet d'étudier des phénomènes importants tels que la loi rang-taille en géographie urbaine ou l'analyse des maxima en géographie physique. À partir d'une série ordonnée, les statistiques d'ordre reposent sur les valeurs de rang, notées $X(1)$, $X(2) \dots X(n)$, qui structurent toute une famille de lois utiles pour étudier la distribution des phénomènes.

Lorsqu'il s'agit de comparer plusieurs classements, deux approches principales apparaissent : la corrélation des rangs et la concordance des classements. La corrélation des rangs vise à mesurer la similitude entre deux classements en étudiant la relation monotone entre leurs positions. La concordance, quant à elle, se concentre sur le nombre de paires d'objets classées dans le même ordre ou dans l'ordre inverse. Ainsi, la corrélation examine la force et le sens de la relation tandis que la concordance s'intéresse à la manière dont les couples respectent l'ordre d'un classement à l'autre.

Ces différences se retrouvent dans les deux tests classiques, Spearman et Kendall. Le test de Spearman repose sur un coefficient obtenu en appliquant la formule de corrélation de Bravais-Pearson aux rangs, après les avoir standardisés. Il s'agit d'un test non paramétrique utilisant les différences de rangs $(ui - vi)^2$, qui varie entre -1 (classements inverses) et $+1$ (identité parfaite). À l'inverse, le test de Kendall repose uniquement sur le comptage des paires concordantes et discordantes. Son coefficient T , compris entre -1 et $+1$, exprime directement le surplus de concordances. Kendall se distingue par sa capacité à s'étendre naturellement au cas de p classements, ce qui en fait un outil puissant pour analyser plusieurs critères de hiérarchisation simultanément.

Au-delà de ces deux tests fondamentaux, nous retrouvons aussi deux autres outils essentiels : le coefficient Γ de Goodman-Kruskal et le coefficient Q de Yule. Le coefficient Γ est défini comme T par le rapport entre la différence et la somme des paires concordantes et discordantes : $\Gamma = (Na - Nd) / (Na + Nd)$. Il se situe également entre -1 et $+1$ et mesure la force d'association entre deux classements ou deux variables ordinaires. Toutefois, Γ diffère de Kendall par l'interprétation et l'usage. Il représente une proportion et peut être nul même en présence d'une relation non aléatoire lorsque le nombre de concordances et de discordances s'équilibrent. Enfin, le coefficient Q de Yule, cas particulier du coefficient de Goodman-Kruskal, s'applique exclusivement aux tables de contingence 2×2 . Il évalue l'association entre deux variables ordinaires binaires selon la formule $Q = (ad - bc) / (ad + bc)$, variant là aussi entre -1 (association négative parfaite) et $+1$ (association positive parfaite).

Ainsi, les statistiques d'ordre permettent d'examiner non seulement les hiérarchies spatiales et sociales, mais aussi la cohérence entre plusieurs systèmes de classement. En géographie comme dans d'autres sciences sociales, elles constituent des outils essentiels pour comprendre, comparer et interpréter les structures ordonnées, qu'il s'agisse de populations, de territoires, d'activités économiques ou de dynamiques urbaines.

II/ Résultats :

étape 3	Isoler la colonne « Surface (km ²) »	
étape 4	Ordonner la liste obtenue	
étape 5	Visualiser la loi rang-taille	<p style="text-align: center;">Conversion logarithmique des surfaces des îles et continents</p> <p>The figure is a scatter plot titled "Conversion logarithmique des surfaces des îles et continents". The x-axis is labeled "Log(Rang)" and ranges from 0 to 12. The y-axis is labeled "Log(Surface (km²))" and ranges from -10 to 18. A series of blue data points follows a downward-sloping curve, indicating a negative correlation. The points are densely clustered between Log(Rang) values of 0 and 4, and more sparsely distributed for higher values.</p>
étape 7	Est-il possible de faire un test sur les rangs ?	<i>Voir dans le main.py</i>
étape 10	Isoler les colonnes « État », « Pop 2007 », « Pop 2025 », « Densité 2007 » et « Densité 2025 »	
étape 11	ordonner de manière décroissante les listes « Pop 2007 », « Pop 2025 », « Densité 2007 » et « Densité 2025 »	
étape 12	obtenir une liste avec deux colonnes ordonnées par rapport au classement de 2007	

étape 13	Isoler les deux colonnes	
étape 14	calculer le coefficient de corrélation des rangs et la concordance des rangs. Commenter les résultats	<p>Corrélation de Spearman : SpearmanResult(correlation=0.1226419321593404, pvalue=0.10898725808133632) Concordance de Kendall : KendalltauResult(correlation=0.0898952808377533, pvalue=0.08001081132586232)</p> <p>Le coefficient de Spearman $p = 0,123$ indique une association positive très faible entre les rangs des deux variables. En pratique, cela signifie que lorsque l'une des variables augmente, l'autre tend légèrement à augmenter aussi, mais de manière très peu marquée.</p> <p>Le coefficient de Kendall $\tau = 0,090$ indique également une association positive très faible, encore plus faible que celle mesurée par Spearman. Kendall τ est généralement plus conservateur et plus robuste aux ex æquo.</p>

⇒ **Paragraphe de synthèse sur les résultats obtenus :**

Les manipulations réalisées au cours de la séance mettent en évidence deux résultats principaux. D'une part, l'analyse rang–taille des surfaces des îles et des continents montre qu'après ordonnancement décroissant et passage en logarithme, la distribution devient lisible et révèle une forte hiérarchisation des surfaces. D'autre part, la comparaison des classements des États selon la population et la densité (2007–2025), à l'aide des tests de corrélation des rangs, indique une concordance très faible entre ces deux dimensions. Les coefficients de Spearman ($\rho = 0,123$) et de Kendall ($\tau = 0,090$) traduisent une association positive mais très peu marquée, suggérant que les pays les plus peuplés ne sont pas nécessairement ceux qui présentent les plus fortes densités. Ces résultats confirment l'intérêt des tests sur les rangs pour comparer des hiérarchies, tout en soulignant les limites de la corrélation lorsque les phénomènes observés obéissent à des logiques spatiales et démographiques distinctes.

III/ Les difficultés rencontrées et les constatations :

Ma programmation en Python n'a pas su récupérer les données dans les fichiers csv du fait des noms des colonnes, notamment avec la colonne "Surface (km2)". De ce fait, j'ai modifié ma programmation afin que les différences de titres etc ne soient plus un problème dans la lecture des données.

Retour personnel sur le module :

Le début du semestre dans ce module fut assez compliqué. En effet, possédant un Macbook, j'ai eu de grandes difficultés à installer les logiciels nécessaires pour réaliser les séances. C'est alors que j'ai fait le choix de recommencer et de faire la demande d'un prêt d'ordinateur à la faculté, pour pouvoir travailler sur un ordinateur Windows. Cette décision a largement facilité mon travail. Je n'ai eu aucune difficulté à installer les logiciels une fois sur ce nouvel ordinateur, avec le tutoriel que vous avez mis sur votre GitHub.

De plus, le retard que j'ai pris lors du semestre à cause de mon ordinateur m'a été bénéfique. En effet, bon nombre de mes camarades avaient déjà une ou plusieurs séances d'avance sur moi. De ce fait, j'ai évité un grand nombre d'erreurs et de problèmes dans le codage puisqu'ils m'ont aidé à réussir les séances. C'est un des points positifs majeurs sur la méthode de pédagogie inversée, puisque nous étions libres pendant les séances d'échanger avec nos camarades pour s'entraider.

Les limites de ce cours ont été, pour moi, avant tout le matériel. En effet, quelques-uns de mes camarades n'avaient pas à disposition un ordinateur suffisamment puissant pour réaliser avec aisance les séances. Je pense à mon camarade qui a dû acheter un nouvel ordinateur au milieu du semestre. Je pense aussi à mon autre camarade qui possédait un Macbook et qui a décidé, comme moi, d'emprunter un ordinateur de la faculté, mais qui s'est retrouvé avec un autre type d'ordinateur que le mien, bien pire que son Macbook.

Réflexion personnelle sur les humanités numériques :

Les humanités numériques sont une transformation des manières de produire, d'analyser et de discuter le savoir. Elles constituent avant tout une culture issue de l'environnement numérique, qui modifie les pratiques intellectuelles, les formes d'autorité scientifique et les modes de transmission des connaissances. De ce fait, le numérique n'est pas un simple support technique, mais un cadre épistémologique qui influe sur les questions posées autant que sur les méthodes employées.

Les humanités numériques favorisent un déplacement du regard, en rendant possibles des analyses à grande échelle tout en posant des enjeux nouveaux de réflexivité critique. En effet, l'usage d'algorithmes et de modèles computationnels implique des choix implicites qui doivent être interrogés au même titre que les sources traditionnelles. Le numérique reconfigure l'interprétation et nous oblige, nous chercheurs et étudiants, à expliciter les hypothèses et les chaînes de traitement.

De plus, les humanités numériques soulèvent d'importantes questions politiques et éthiques centrales, comme l'accès aux données, la reproductibilité des résultats, ou bien encore la dépendance aux infrastructures techniques et aux plateformes. Elles invitent à repenser le rôle du chercheur non seulement comme analyste mais aussi comme médiateur.

De ce fait, les humanités numériques ne remplacent pas les humanités classiques. Celles-ci prolongent leurs interrogations fondamentales sur le sens, le pouvoir et la production du savoir dans un monde numérisé.