Rapport : Analyse de données (niveau débutant)

Victor CATHALA - Master 1 GAED Géopolitique – GEOINT

Dans le cadre du cours de Monsieur Maxime Forriez

Séance n°2

Question n°1 : Quel est le positionnement de la géographie par rapport aux statistiques ?

La géographie entretient historiquement une relation avec la statistique. Elle produit et mobilise des volumes importants de données empiriques, (données de population, d'économie, de climat ou de flux…). Cette discipline a longtemps manifesté une réticence vis-à-vis de la formalisation mathématique, souvent perçue comme étrangère à son identité disciplinaire. Cette situation amène à discuter d'un paradoxe. La géographie repose sur des données massives, elle a pourtant sous-estimé l'apport des outils statistiques, ou les a utilisés de manière intuitive ou peu rigoureuse. Cette critique a déjà été formulée par plusieurs géographes quantitativistes, comme Marchand (1972), Béguin (1979) ou Chadule (1997), qui ont montré que la statistique constitue un outil central pour structurer, comparer et interpréter les phénomènes spatiaux. Dans cette perspective, la statistique n'est pas une discipline auxiliaire, mais un instrument fondamental dans l'aspect scientifique puisqu'elle permet en effet à la géographie de dépasser le simple descriptif pour accéder à l'analyse poussée, à la modélisation et à la généralisation.

Question n°2 : Le hasard existe-t-il en géographie ?

La question du hasard relève d'abord d'un débat philosophique ancien, opposant déterminisme et contingence. En géographie, cette question est centrale, car les phénomènes étudiés (dynamiques territoriales, comportements humains, processus physiques) sont complexes, multifactoriels et fortement contextualisés. En effet, il est alors impossible de prévoir le détail des réalisations individuelles, mais il est toutefois possible de dégager une certitude globale soit une tendance. Cette idée se retrouve pleinement dans l'approche multiscalaire, fondamentale en géographie. A grande échelle, on trouve des régularités statistiques (par exemple des structures de peuplement ou de mobilité) à une échelle large, tandis que l'on peut constater des variations à l'échelle locale.

Le hasard n'empêche pas donc pas la scientificité de la géographie, la statistique permet de détecter des structures dans l'aléatoire, rejoignant les travaux fondateurs de Korčák (1940) ou Fréchet (1941) sur les lois de distribution géographiques

Question n°3 : Quels sont les types d'information géographique ?

On distingue d'abord les informations attributaires, soit celles qui décrivent les caractéristiques d'un territoire ou d'un objet spatial (population, revenus, types d'activités, températures, précipitations…). Ces informations constituent la base attributaire d'un SIG et relèvent directement de l'analyse statistique. Ensuite, il y'a les informations géométriques (forme, localisation, morphologie des objets géographiques comme les points, lignes, surfaces). Ainsi, l'analyse statistique porte prioritairement sur les attributs, tandis que l'analyse spatiale intègre ensuite la dimension géométrique.

Question n°4 : Quels sont les besoins de la géographie au niveau de l'analyse de données ?

Pour produire une analyse géographique, il est nécessaire de disposer de données produites ou collectées selon une nomenclature explicite, des concepts définis. Les données doivent être complétées par des méta-données qui permettent un examen critique des sources (conditions de collecte, dates, échelles spatiales, exhaustivité ou sondage, fiabilité des mesures). Ces méta-données sont indispensables pour éviter les interprétations abusives, comme l'ont souligné Dumolard et al. (2003). L'analyse de données emprunte surtout aux mathématiques, puisqu'il s'agit d'étudier la structure des données, de dégager des régularités, puis de confronter les résultats aux conditions de production des données et à la connaissance du phénomène étudié. La statistique ne remplace donc pas l'interprétation géographique, mais elle permet de la rendre plus rigoureuse et précise.

Question n°5 : Différences entre statistique descriptive et statistique explicative

La statistique descriptive décrit les données (population/échantillon), résume les distributions, produit des comparaisons et des prédictions. Elle permet de proposer des paramètres, tableaux et visualisations (ex. histogrammes pour continues, secteurs pour qualitatives).

La statistique explicative permet, elle, de relier une variable à expliquer **Y** à des variables explicatives **X1…Xk**, en ajustant un modèle selon la nature de Y (numérique : $Y = f(X) + aléa$; qualitative : probabilité conditionnelle).

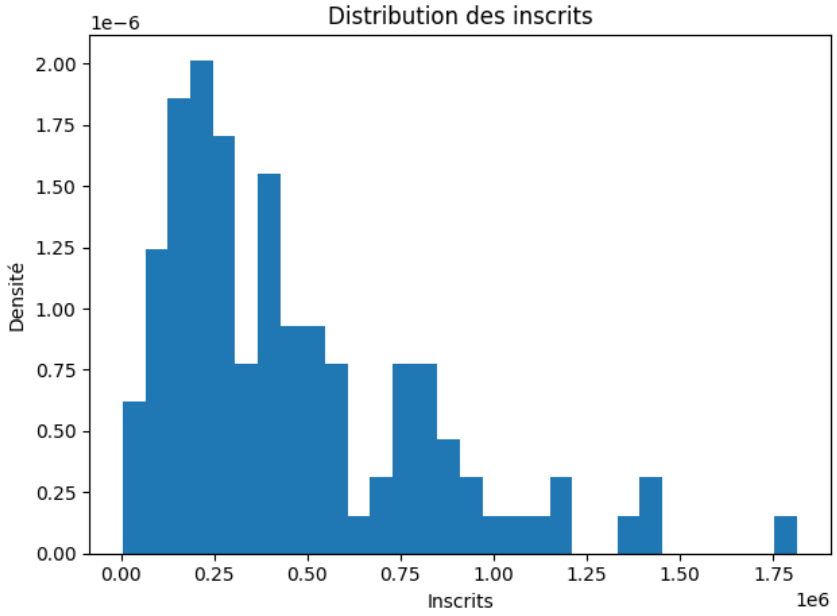
Pour résumer clairement, Tenenhaus (2007), propose que la statistique descriptive répond à la question « que montrent les données ? », tandis que la statistique explicative cherche à comprendre « pourquoi et comment ces phénomènes sont liés »

Question n°6 : Quels sont les types de visualisation de données en géographie ? Comment les choisir ?

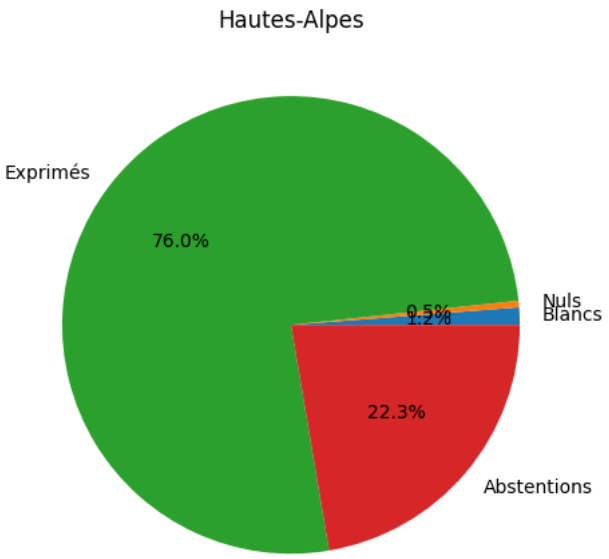
Le choix des visualisations dépend directement du type de variable étudiée. Pour les variables quantitatives continues, l'histogramme est le type de visualisation le plus répandu, permettant

de représenter une distribution statistique et d'en analyser la forme. Pour les variables qualitatives, la représentation sectorielle est plus adaptée, puisqu'elle permet de mettre en évidence les parts relatives de chaque modalité. Il existe ensuite une multitude de représentation, comme le polygone de fréquence, la courbe cumulative, ou encore la boîte à moustaches que nous avons pu maitriser dans le cadre du cours de « Méthode quantitative » avec le Professeur Huguenin-Richard, en sachant que chacun répond à des objectifs spécifiques d'analyse. On utilise ainsi une représentation selon le type d'analyse, afin de conditionner l'interprétation et doit surtout être justifié scientifiquement, car en effet, une mauvaise visualisation peut conduire à des conclusions fausses ou des mauvaises interprétations.

Exemple tiré du travail de codage de la séance n°3 : histogramme



Exemple tiré du travail de codage de la séance n°3 : diagramme circulaire



Question n°7 : Quelles sont les méthodes d'analyse de données possibles ?

Il existe trois grandes méthodes d'analyse de données. D'abord, les méthodes descriptives multidimensionnelles (ACP, AFC, ACM), permettant de réduire la dimension des données et de visualiser les structures principales. Les méthodes de classification (CAH, nuées dynamiques) servent à regrouper des individus ou des territoires homogènes. Ensuite, les méthodes explicatives permettent d'établir des relations entre des variables (régressions, analyse discriminante…). Enfin, les méthodes de prévision concernent l'analyse des séries chronologiques, très utilisées en géographie économique ou climatique, pour anticiper l'évolution d'un phénomène à partir de ses valeurs passées, pour faire de la prospective par exeemple.

Question n°8 : Définitions statistiques fondamentales et types de caractères

La population statistique est définie comme un ensemble au sens mathématique. L'individu statistique est un élément de cette population, souvent localisable et cartographiable en géographie. Les caractères statistiques sont les propriétés observées sur les individus, tandis que les modalités sont les valeurs prises par ces caractères. Elles doivent être incompatibles et exhaustives. On peut distinguer quatre types de caractères : qualitatifs nominaux, qualitatifs ordinaux, quantitatifs discrets et quantitatifs continus.

Question n°9 : Comment mesurer une amplitude et une densité ?

Lorsqu'une variable quantitative est discrétisée en classes, l'amplitude correspond à la largeur d'une classe ($b - a$). Elle concerne toujours une classe particulière. La densité, quant à elle, rapporte l'effectif d'une classe à son amplitude. Elle permet de comparer des classes de largeur différente et constitue la base de la construction correcte d'un histogramme statistique.

Question n°10 : À quoi servent les formules de Sturges et de Yule ?

Les formules de Sturges et de Yule servent à estimer le nombre optimal de classes lors de la discrétisation d'une variable quantitative. Elles permettent d'éviter un découpage trop fin (bruit) ou trop grossier (perte d'information). Ces formules fournissent une valeur indicative, qui doit toujours être interprétée et adaptée au contexte géographique et à la distribution observée.

Question n°11 : Effectifs, fréquences et distribution statistique

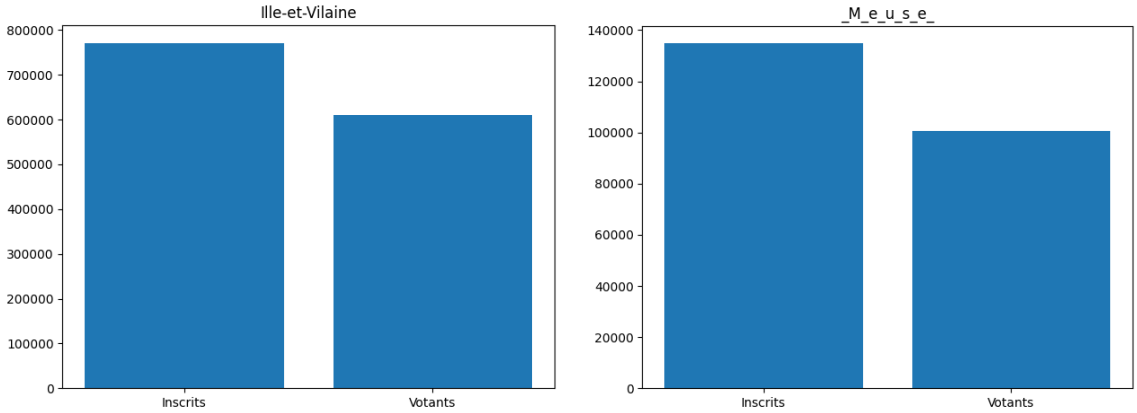
L'effectif correspond au nombre d'occurrences d'une modalité ou d'une classe. La fréquence est l'effectif rapporté à l'effectif total, tandis que la fréquence cumulée additionne les fréquences jusqu'à une valeur donnée.

La distribution statistique empirique est l'ensemble des fréquences observées. Elle constitue le point de départ de toute interprétation probabiliste et permet de formuler des hypothèses sur la loi de probabilité sous-jacente

Rapport de la séance n°2 :

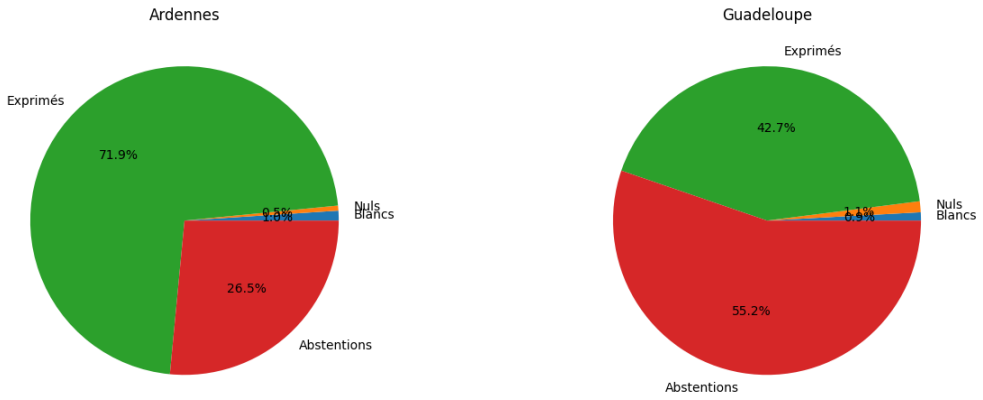
- Les **diagrammes en barres Inscrits / Votants** montrent de fortes disparités démographiques entre départements, avec une majorité de territoires aux effectifs modérés et quelques départements très peuplés concentrant des effectifs nettement plus élevés, ce qui traduit une distribution inégale de la population électorale.

Ici l'exemple de l'Ille-et-Vilaine (Bretagne) avec 800 000 inscrits et la Meuse (Grand-Est) et ses 140 000.



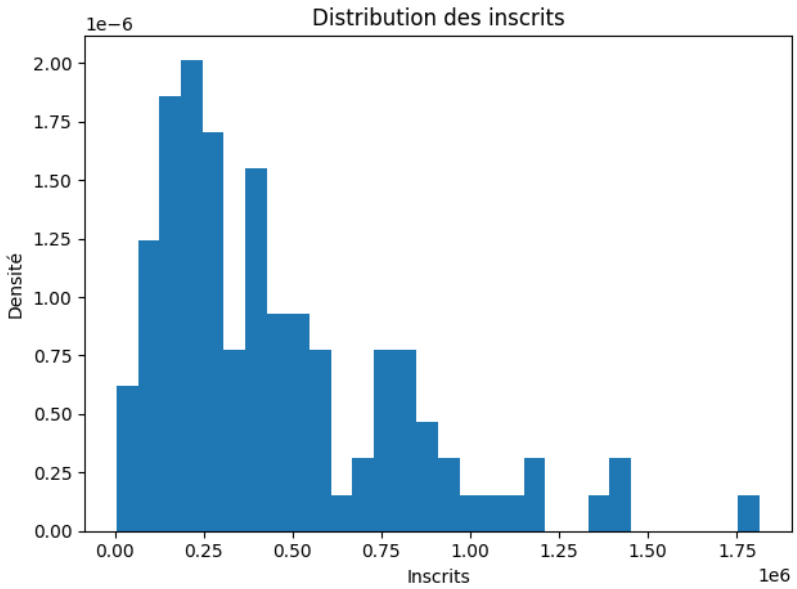
- Ces mêmes **diagrammes en barres Inscrits / Votants** mettent également en évidence un écart systématique entre inscrits et votants, révélant une abstention présente partout et particulièrement marquée dans certains départements. En Ille-et-Vilaine, on obtient un taux d'abstention d'environ 20%, que l'on retrouve dans la Meuse.

- Les diagrammes circulaires Blancs / Nuls / Exprimés / Abstentions montrent que les votes exprimés restent majoritaires, mais que l'abstention représente une part importante du total, tandis que les votes blancs et nuls, bien que minoritaires, apparaissent relativement plus élevés dans certains départements, ce qui constitue des profils atypiques.



Nous avons ici l'exemple parfait de la Guadeloupe, qui montre un taux d'abstention de 55,2%, tandis qu'elle est plus généralement entre 15 et 30% en France métropolitaine.

- Enfin, l'histogramme de la distribution des inscrits met en évidence une distribution fortement dissymétrique, caractérisée par une concentration de départements dans les classes de faibles et moyens effectifs et par la présence de quelques valeurs extrêmes correspondant aux départements les plus peuplés, confirmant l'existence de contrastes territoriaux marqués.



Séance n°3 :

Question n°1 : Quel caractère est le plus général : quantitatif ou qualitatif ? Pourquoi ?

Le caractère le plus général est le caractère qualitatif car toute variable quantitative peut être transformée en catégorie, donc devenir qualitative, tandis que l'inverse n'est pas forcément possible.

Question n°2 : Que sont les caractères quantitatifs discrets et continus ? Pourquoi les distinguer ?

Les caractères quantitatifs discrets prennent un nombre fini ou dénombrable de valeurs (le nombre d'enfants par exemple). Les caractères quantitatifs continus peuvent prendre une infinité de valeurs dans un intervalle (comme l'âge ou un salaire). On les distingue car ils ne se traitent pas de la même manière statistiquement (calculs, graphiques, lois de probabilité).

Question n°3 : Les paramètres de position

a) Pourquoi existe-t-il plusieurs types de moyenne ?

Il existe plusieurs moyennes car elles ne répondent pas toutes aux mêmes situations. Selon la nature des données (valeurs positives, vitesses, évolutions), certaines moyennes sont plus adaptées que la moyenne arithmétique.

b) Pourquoi calculer une médiane ?

La médiane permet de représenter la valeur centrale d'une distribution sans être influencée par les valeurs extrêmes. Elle est particulièrement utile lorsque la distribution est dissymétrique (comme l'analyse des salaires en France).

c) Quand est-il possible de calculer un mode ?

Le mode peut être calculé lorsqu'une valeur ou une classe est plus fréquente que les autres. Il n'existe pas toujours ou alors il peut être multiple.

Question n°4 : Les paramètres de concentration

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

La médiale permet de mesurer comment une masse totale (revenus, salaires…) est répartie dans la population. L'indice de Gini sert à mesurer le degré de concentration ou d'inégalité d'une distribution.

Question n°5 : Les paramètres de dispersion

a) Pourquoi calculer une variance plutôt que l'écart à la moyenne ? Pourquoi l'écart type ?

On calcule la variance car les écarts simples à la moyenne se compensent. Le carré des écarts évite ce problème. L'écart type est utilisé car il est plus lisible puisqu'il s'exprime dans la même unité que la variable.

b) Pourquoi calculer l'étendue ?

L'étendue permet de donner une idée simple de la dispersion en mesurant l'écart entre la valeur minimale et la valeur maximale en restant cependant très sensible aux valeurs extrêmes.

c) À quoi sert un quantile ? Lesquels sont les plus utilisés ?

Les quantiles servent à découper une distribution en parts égales, les plus utilisés sont les quartiles, les déciles et les centiles.

d) Pourquoi construire une boîte de dispersion ? Comment l'interpréter ?

La boîte de dispersion permet de résumer visuellement une distribution et de comparer plusieurs séries, elle permet de montrer la médiane, la dispersion centrale et les valeurs extrêmes.

Question n°6 : Les paramètres de forme

a) Différence entre moments centrés et moments absolus ? Pourquoi les utiliser ?

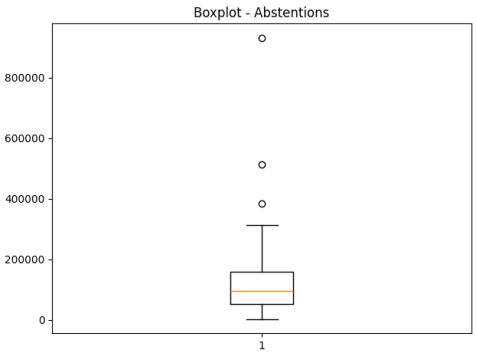
Les moments absolus décrivent une distribution par rapport à l'origine, alors que les moments centrés décrivent la distribution par rapport à la moyenne. Ils sont utilisés pour caractériser la forme d'une distribution (comme la dispersion, l'asymétrie, ou l'aplatissement).

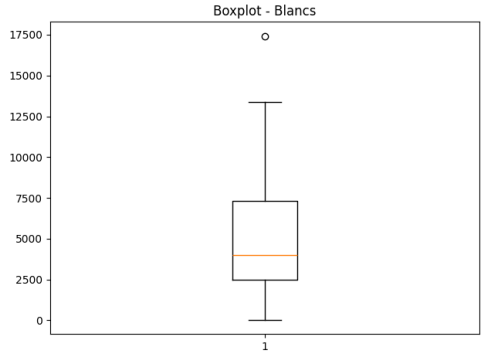
b) Pourquoi vérifier la symétrie d'une distribution et comment faire ?

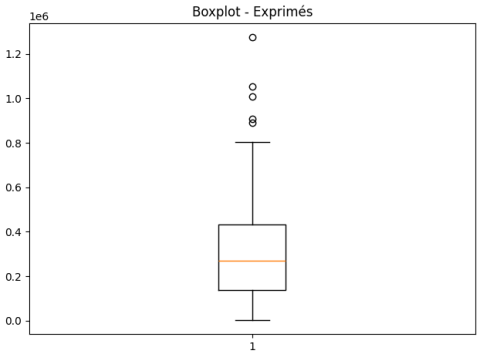
Vérifier la symétrie permet de savoir si une distribution est équilibrée ou dissymétrique. On peut le faire en comparant la moyenne, la médiane et le mode, ou à l'aide du coefficient d'asymétrie.

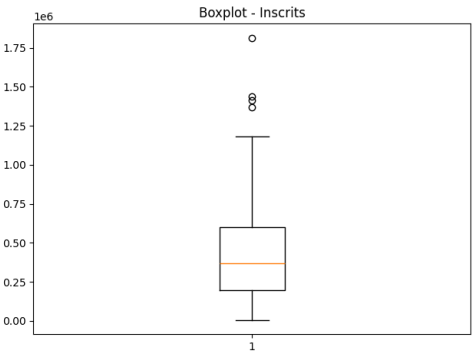
Rapport de la séance 3 :

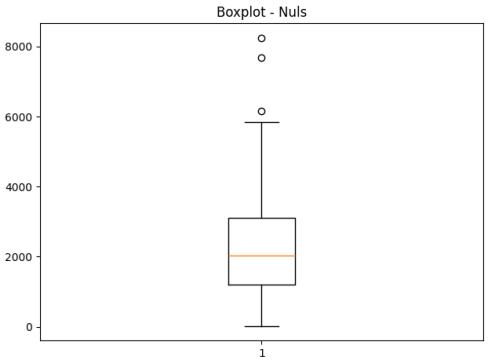
- Les statistiques descriptives montrent de fortes disparités entre départements pour l'ensemble des variables électorales. Les moyennes et médianes diffèrent nettement pour les effectifs d'inscrits, de votants et d'abstentions, ce qui indique des distributions asymétriques, confirmées par les histogrammes et les boîtes à moustaches.







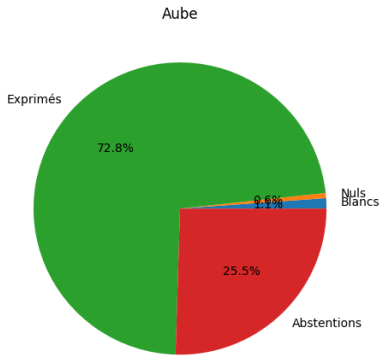


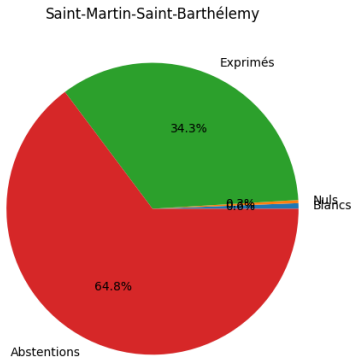


- Les écarts-types, les étendues et les distances interquartiles et interdéciles sont élevés, ce qui traduit une forte dispersion et la présence de valeurs extrêmes, correspondant aux départements les plus peuplés (exemple de l'Ile-de-France). Les diagrammes en barres Inscrits/Votants mettent en évidence un écart systématique entre inscrits et votants, révélant une abstention présente partout mais particulièrement marquée dans certains départements. Les diagrammes circulaires montrent que les votes exprimés restent majoritaires, mais que l'abstention représente une part importante du total et que les votes blancs et nuls, bien que minoritaires, sont relativement plus élevés dans certains territoires, ce qui constitue des profils atypiques. Enfin, les boxplots confirment que

certaines observations s'écartent fortement du reste de la distribution, illustrant des contrastes territoriaux marqués plutôt que de véritables anomalies statistiques.

L'exemple comparatif entre territoire national et outre-mer montre remarquablement cet écart pour le taux d'abstention.





Séance n°4 :

Question n°1 : Quels critères pour choisir entre une distribution discrète et une distribution continue ?

Le premier critère pour choisir entre une distribution discrète et une distribution continue est la nature de la variable étudiée. Si la variable correspond à un comptage (par exemple un nombre d'événements, d'accidents ou d'individus), on utilise plutôt une distribution discrète. En revanche, si la variable correspond à une mesure (comme un revenu, une distance ou une surface), on utilise une distribution continue.

Le deuxième critère important est le phénomène géographique étudié. Les distributions discrètes sont adaptées aux phénomènes ponctuels, alors que les distributions continues conviennent mieux aux phénomènes étalés dans l'espace ou le temps.

Enfin, le choix dépend aussi de la forme de la distribution observée et de l'objectif de l'analyse. Si les effectifs sont élevés, certaines variables discrètes peuvent être approchées par des lois continues. Le but est toujours de choisir la distribution la plus cohérente avec la réalité observée, par rapport à ce que l'on cherche à démontrer ou à modéliser.

Question n°2 : Quelles sont, selon vous, les lois les plus utilisées en géographie ?

En géographie, certaines lois statistiques sont plus utilisées car elles correspondent bien aux phénomènes spatiaux. La loi de Poisson est très fréquente car elle permet de modéliser des événements rares et indépendants, comme des accidents ou des occurrences ponctuelles dans l'espace ou dans le temps. Elle est particulièrement adaptée aux phénomènes de comptage.

La loi normale est aussi largement utilisée, notamment comme loi de référence. Elle sert à décrire des phénomènes influencés par de nombreux facteurs, mais le cours rappelle qu'elle ne s'applique pas à tous les phénomènes géographiques.

La loi log-normale est très importante en géographie économique et urbaine, car elle décrit bien des variables positives et très asymétriques, comme les revenus ou les tailles de villes.

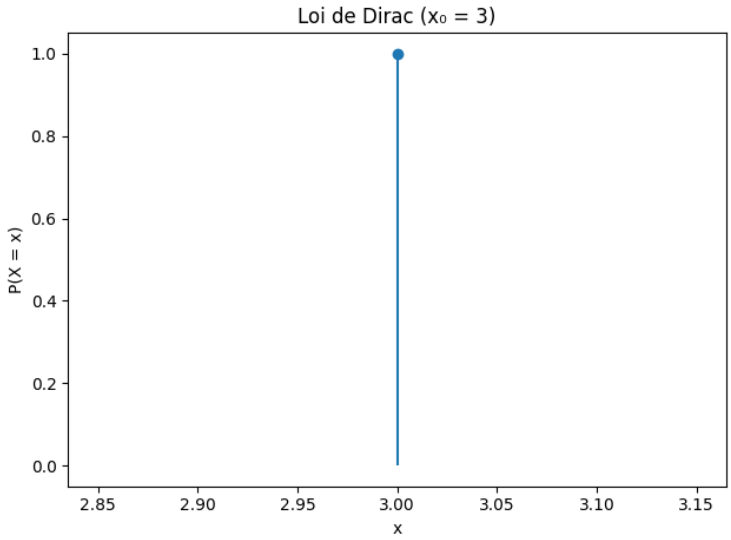
Enfin, la loi de Zipf-Mandelbrot est emblématique en géographie, notamment pour l'étude des systèmes urbains à travers les lois rang–taille. Elle met en évidence des régularités fortes dans l'organisation des territoires, malgré leur apparente complexité.

Rapport de la séance n°4 :

A travers cette séance, nous nous sommes penchés sur les différentes lois utiles en statistique. Les graphiques que nous avons produits illustrent les principales distributions statistiques utilisées en géographie, en distinguant clairement les lois discrètes et continues.

- La loi de Dirac montre un cas extrême où toute la probabilité est concentrée sur une seule valeur, ce qui sert surtout de référence théorique.

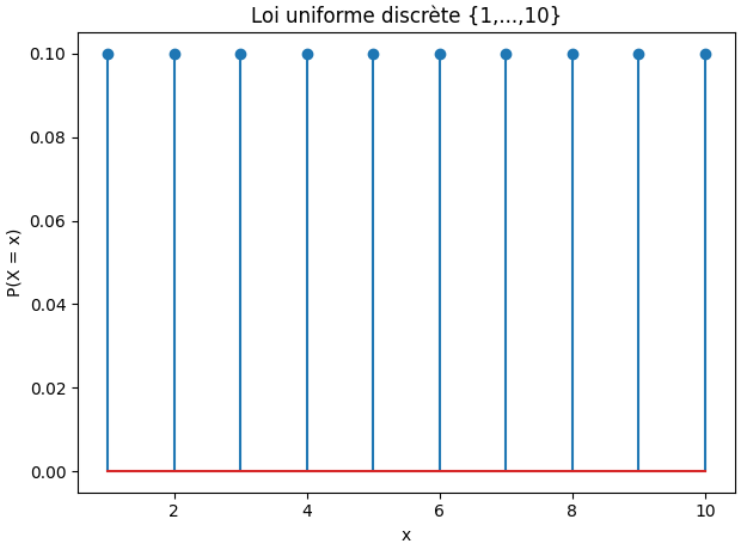
Ici, une illustration tirée du travail de codage de la loi de Dirac :



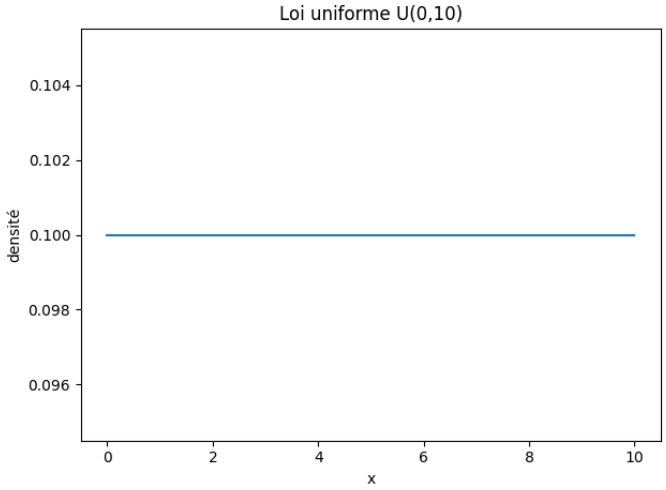
.

- La loi uniforme discrète met en évidence une situation sans préférence, où chaque valeur a la même probabilité.

Ici une illustration tirée du travail de codage de la loi uniforme discrète :

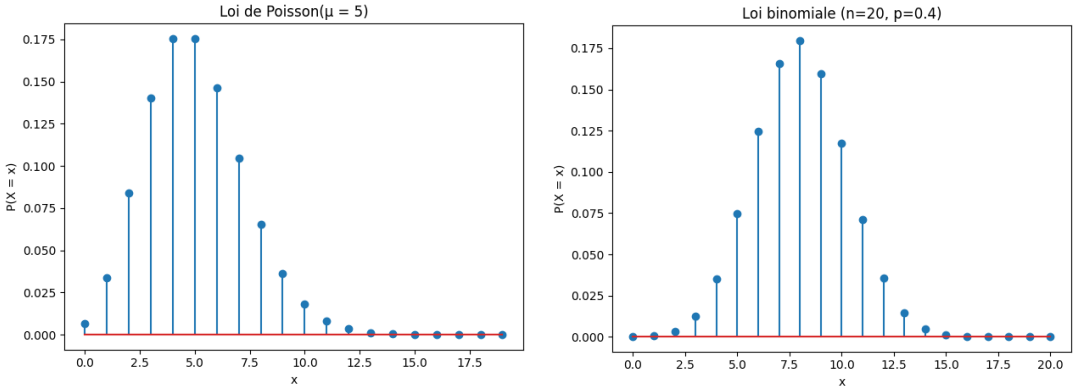


- La loi uniforme est correctement représentée par une courbe plate, ce qui signifie que toutes les valeurs de l'intervalle ont la même probabilité d'occurrence, sans valeur privilégiée. À l'inverse, la loi normale est représentée par une courbe en cloche, symétrique autour de la moyenne, ce qui traduit une forte concentration des valeurs autour de cette moyenne et une diminution progressive des probabilités vers les valeurs extrêmes. L'état des courbes correspond donc bien aux propriétés théoriques de ces deux lois.



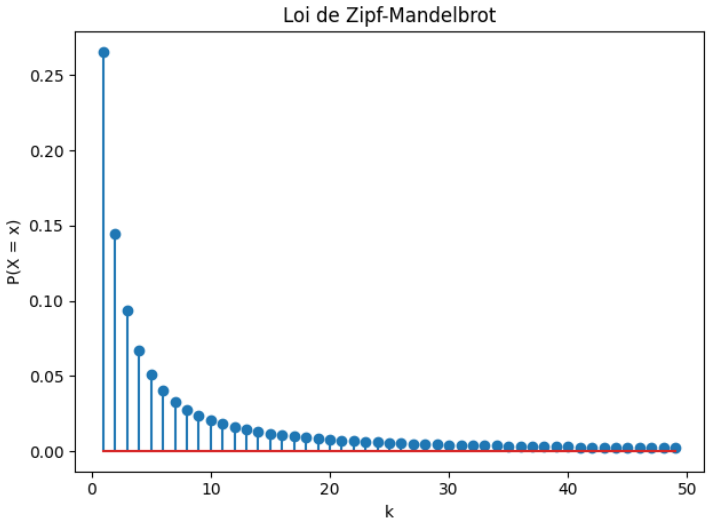
- La loi binomiale et la loi de Poisson présentent des distributions asymétriques, adaptées à la modélisation de phénomènes de comptage, la loi de Poisson étant particulièrement utilisée pour des événements rares.

Illustration de la loi de Poisson et de la loi binomiale tirée du travail de codage de la séance n°4 :



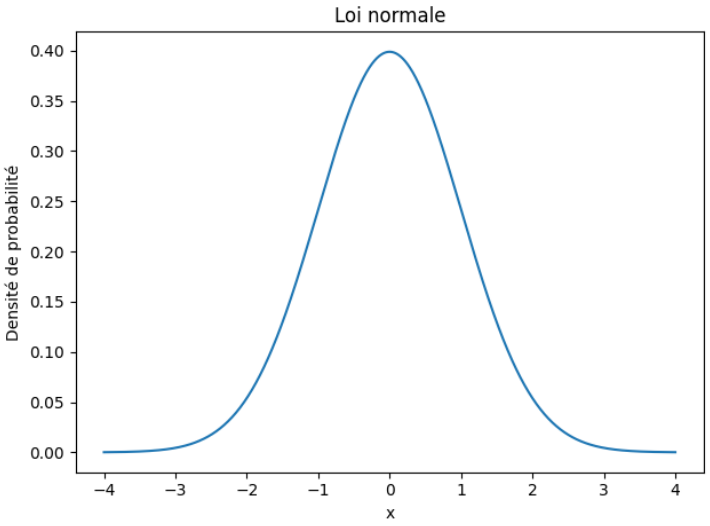
- La loi de Zipf-Mandelbrot se distingue par une forte décroissance des probabilités avec le rang, illustrant des phénomènes de hiérarchie et de concentration typiques des systèmes urbains.

Illustration de la loi de Zipf-Mandelbrot tirée du travail de codage de la séance n°4 :



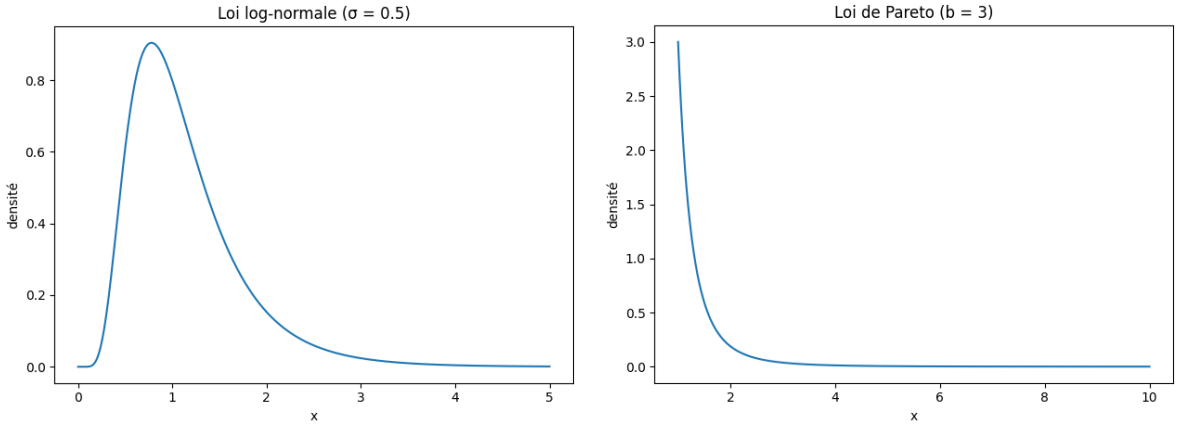
- Du côté des lois continues, la loi normale apparaît symétrique et centrée, servant souvent de loi de référence.

Illustration de la loi normale tirée du travail de codage de la séance n°4 :



Tandis que les lois log-normale, de Pareto et du χ^2 sont asymétriques et mettent en évidence des distributions à queue longue, fréquentes en géographie économique et sociale. Le calcul de la moyenne et de l'écart-type, notamment pour la loi de Poisson où ils sont proches, confirme la cohérence entre les propriétés théoriques des lois et leur représentation graphique.

Illustration de la loi log-normale et de Pareto tirée du travail de codage de la séance n°4 :



Séance n°5

Question n°1 : Comment définir l'échantillonnage ? Pourquoi ne pas utiliser toute la population ? Méthodes et choix

L'échantillonnage consiste à prélever une partie de la population (échantillon), afin d'en tirer des informations sur la population entière. On parle alors de population mère et d'échantillon. On n'utilise pas systématiquement la population entière car les données seraient trop longues à récolter, voire trop couteux (par exemple connaître l'opinion de toute la population française). L'échantillonnage permet donc d'obtenir des résultats fiables avec des moyens limités, en se basant seulement sur une frange plus réduite.

Il existe plusieurs méthodes d'échantillonnage. Les méthodes aléatoires (comme le tirage au sort simple) sont les plus rigoureuses, car chaque individu a la même probabilité d'être choisi. Il existe aussi des méthodes non aléatoires, comme la méthode des quotas, qui cherchent à reproduire la structure de la population. Le choix dépend ainsi du contexte, des données disponibles, du coût, et surtout du besoin de représentativité. Un petit échantillon bien choisi est plus représentatif et produit des résultats plus rigoureux qu'un grand échantillon biaisé.

Question n°2 : Comment définir un estimateur et une estimation ?

Un estimateur est une formule ou une statistique calculée à partir d'un échantillon, qui sert à approcher un paramètre inconnu de la population (comme une moyenne ou une proportion). L'estimation est la valeur numérique obtenue quand on applique cet estimateur à un échantillon concret. Par exemple, la moyenne d'un échantillon est un estimateur de la moyenne de la population, et la valeur calculée est l'estimation. L'estimateur est théorique, l'estimation est pratique

Question n°3 : Différence entre intervalle de fluctuation et intervalle de confiance

L'intervalle de fluctuation est utilisé quand on connaît le paramètre de la population. Il sert à vérifier si une fréquence observée dans un échantillon est cohérente avec cette valeur théorique. On est donc dans une logique de contrôle. L'intervalle de confiance, au contraire, est utilisé quand on ne connaît pas le paramètre de la population. Il permet d'estimer une plage de valeurs dans laquelle ce paramètre a de fortes chances de se trouver.

Question n°4 : Qu'est-ce qu'un biais dans la théorie de l'estimation ?

Un biais correspond à une erreur systématique dans l'estimation. Un estimateur est biaisé si, en moyenne, il ne tombe pas sur la vraie valeur du paramètre. À l'inverse, un estimateur sans biais peut donner des valeurs trop grandes ou trop petites, mais ces erreurs se compensent sur le long terme.

Question n°5 : Comment appelle-t-on une statistique travaillant sur toute la population ? Lien avec les données massives

Une statistique calculée sur toute la population est appelée une statistique exhaustive. Elle correspond à un recensement plutôt qu'à une estimation. Avec le développement des données massives à l'ère du big data, certaines situations se rapprochent de cette logique exhaustive. Cependant, même avec beaucoup de données, des biais peuvent subsister, ce qui rappelle à nouveau que récolter beaucoup de données ne signifie pas qu'elles sont bonnes, sans biais.

Question n°6 : Quels sont les enjeux du choix d'un estimateur ?

On choisit un estimateur afin de faire un juste milieu entre la justesse et la précision. Pour cela, un bon estimateur doit pouvoir être sans biais, précis (donc faible en dispersion), et convergent (c'est-à-dire qu'il s'améliore quand la taille de l'échantillon augmente). Choisir un estimateur, c'est chercher un compromis entre justesse et précision. Un mauvais estimateur peut conduire à des conclusions fausses, même avec beaucoup de données. Le choix de l'estimateur est donc un enjeu central en statistique inférentielle pour obtenir une analyse rigoureuse.

Question n°7 : Quelles sont les méthodes d'estimation d'un paramètre ? Comment en choisir une ?

Il existe plusieurs méthodes d'estimation, comme la méthode des moments, la méthode du maximum de vraisemblance ou encore des méthodes par simulation (Monte-Carlo). Le choix dépend de la loi supposée, de la taille de l'échantillon, de la complexité du modèle et du niveau de précision recherché. On choisit donc souvent une méthode simple mais robuste, adaptée aux données disponibles.

Question n°8 : Quels sont les tests statistiques ? À quoi servent-ils ? Comment en construire un ?

Les tests statistiques servent à prendre une décision à partir de données, en tenant compte du hasard. Ils permettent par exemple de vérifier une hypothèse, de comparer des groupes ou de tester une loi de distribution. Il existe différents types de tests selon la nature des variables (quantitatives ou qualitatives) et les hypothèses formulées. Construire un test consiste à formuler une hypothèse nulle, choisir un seuil de risque, calculer une statistique de test et prendre une décision. D'abord, les tests de normalité. Ils servent à vérifier si une distribution suit une loi normale. Un exemple courant est le test de Shapiro-Wilk, pour décider si une série peut être considérée comme normale. Ensuite, les tests de comparaison. Ils permettent de comparer des moyennes ou des distributions entre groupes. Par exemple, on peut tester si deux échantillons ont la même moyenne ou si une moyenne observée est différente d'une valeur théorique. Le type de test dépend de la nature des variables et des hypothèses (normalité, indépendance, taille de l'échantillon). Il existe aussi les tests sur les proportions et les

fréquences. Ils servent à vérifier si une fréquence observée est compatible avec une proportion théorique connue. Enfin, on trouve les tests d'ajustement à une loi. Ils permettent de vérifier si une distribution observée correspond à une loi théorique (normale, uniforme, etc.). C'est le cas lorsqu'on cherche à valider un modèle statistique avant de faire de l'inférence.

Question n°9 : Que penser des critiques de la statistique inférentielle ?

Les critiques de la statistique inférentielle sont en partie légitimes. Les résultats dépendent fortement des hypothèses, de la qualité des données et du respect des conditions d'application. Cependant, ces limites ne remettent pas en cause son utilité. La statistique inférentielle reste un outil indispensable, à condition d'être utilisée avec esprit critique, transparence et rigueur méthodologique.

Rapport de la séance n°5 :

- Cette séance a pour objectif d'illustrer les principes de la statistique inférentielle, à travers l'échantillonnage, l'estimation et la prise de décision. Les moyennes calculées sur les 100 échantillons sont de 391 individus "Pour", 416 "Contre" et 193 "Sans opinion", soit des fréquences observées de 0,39, 0,42 et 0,19, exactement identiques aux fréquences de la population mère (0,39 ; 0,42 ; 0,19), ce qui montre que l'échantillonnage reproduit correctement la structure de la population.
- Les intervalles de fluctuation à 95 % confirment cette cohérence : [0,36 ; 0,42] pour *Pour*, [0,389 ; 0,451] pour *Contre* et [0,166 ; 0,214] pour *Sans opinion*, indiquant que les écarts observés relèvent du hasard d'échantillonnage. À partir d'un échantillon unique, les fréquences observées sont de 0,40, 0,40 et 0,21, avec des intervalles de confiance à 95 % respectivement égaux à [0,365 ; 0,425], [0,366 ; 0,426] et [0,184 ; 0,234], ce qui fournit une estimation encadrée des proportions réelles dans la population.
- Enfin, les tests de normalité de Shapiro-Wilk, réalisés sur deux jeux de données de taille 2000, donnent des p-values extrêmement faibles ($\approx 6{,}3\times10^{-22}$ et $\approx 7{,}0\times10^{-67}$), ce qui conduit à rejeter l'hypothèse de normalité dans les deux cas et illustre concrètement la logique de décision statistique.

Séance n°6 :

Question n°1 : Qu'est-ce qu'une statistique ordinale ? À quoi s'oppose-t-elle ? Quel type de variables utilise-t-elle ? En quoi matérialise-t-elle une hiérarchie spatiale ?

La statistique ordinale est une statistique qui repose sur le classement des individus ou des objets, et non sur leurs valeurs exactes. Elle s'intéresse à l'ordre (premier, deuxième, troisième…) plutôt qu'aux écarts numériques. Elle s'oppose à la statistique nominale, qui classe sans ordre (par exemple des catégories sans hiérarchie). La statistique ordinale utilise donc des variables qualitatives ordinales, c'est-à-dire des variables pour lesquelles un ordre a du sens. En géographie, cela permet de matérialiser des hiérarchies spatiales, comme le classement des villes par population, des États par densité ou des territoires par niveau de développement. Même sans mesurer précisément les écarts, le classement révèle une organisation de l'espace.

Question n°2 : Quel ordre est à privilégier dans les classifications ?

L'ordre à privilégier est l'ordre croissant, aussi appelé ordre naturel. C'est l'ordre le plus simple et le plus lisible, car il va de la plus petite valeur à la plus grande. Il existe toutefois des exceptions en géographie, comme la loi rang/taille, où l'on privilégie un ordre décroissant pour mieux analyser les phénomènes de domination et de hiérarchie (comme par exemple les systèmes urbains).

Question n°3 : Quelle est la différence entre une corrélation des rangs et une concordance de classements ?

La corrélation des rangs mesure le lien entre deux classements pris globalement. Elle cherche à savoir si, quand un objet est bien classé dans un classement, il l'est aussi dans l'autre. La concordance de classements, elle, s'intéresse plus finement à la comparaison des classements, en examinant les paires concordantes et discordantes. Elle mesure le degré d'accord réel entre plusieurs classements. En résumé, la corrélation des rangs mesure une relation globale, tandis que la concordance mesure un accord précis entre classements.

Question n°4 : Quelle est la différence entre les tests de Spearman et de Kendall ?

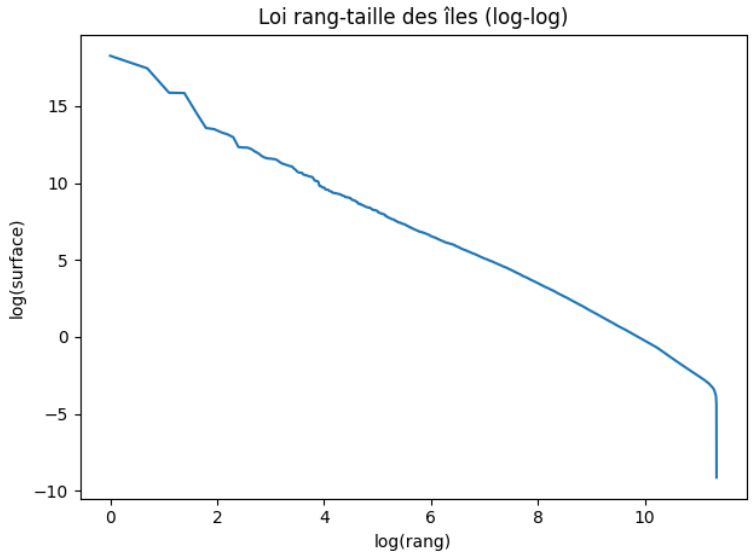
Le test de Spearman mesure une corrélation entre deux classements en comparant les différences de rang. Il est simple à utiliser et assez intuitif, mais il est sensible aux écarts importants. Le test de Kendall repose sur la comparaison des paires concordantes et discordantes. Il est souvent considéré comme plus robuste, notamment pour les petits échantillons, et il peut être généralisé à plus de deux classements. Les deux tests sont non paramétriques et adaptés aux variables ordinales, mais Kendall est souvent plus strict dans l'interprétation de l'accord entre classements.

Question n°5 : À quoi servent les coefficients de Goodman-Kruskal et de Yule ?

Les coefficients de Goodman-Kruskal servent à mesurer l'association entre des variables ordinales, à partir du nombre de paires concordantes et discordantes. Ils permettent de savoir si deux classements vont globalement dans le même sens. Le coefficient Q de Yule est un cas particulier, utilisé pour des tableaux de contingence 2×2. Il mesure la force et le sens de l'association entre deux variables binaires. Ces coefficients sont très utiles en géographie humaine pour analyser des relations ordinales, par exemple entre des catégories sociales, des niveaux de développement ou des classements territoriaux.

Rapport de la séance n°6 :

- La séance n°6 porte sur l'analyse des hiérarchies spatiales à l'aide des statistiques ordinales et des lois rang-taille. La loi rang-taille des îles met en évidence une très forte concentration des surfaces : les premières valeurs correspondent à des surfaces exceptionnellement élevées (par exemple ≈ 85 millions km², ≈ 38 millions km², ≈ 7,7 millions km² et ≈ 7,6 millions km² ajoutées au jeu de données), tandis que la majorité des îles présentent des surfaces beaucoup plus faibles. Cette inégalité apparaît clairement sur le graphique "Loi rang-taille des îles", où la courbe décroît rapidement avec le rang (graphique ci-dessous).



- Le passage en échelle logarithmique sur le graphique log–log transforme cette courbe en une relation quasi linéaire, ce qui est caractéristique d'une loi de type Zipf–Pareto et confirme une organisation hiérarchique stable.
  Dans un second temps, la comparaison des classements des États selon leur population en 2007 et leur densité en 2007 montre que ces deux hiérarchies ne se superposent pas parfaitement : certains États très peuplés ne sont pas nécessairement très denses, et

inversement. Cette relation est mesurée par les coefficients de corrélation des rangs de Spearman et de Kendall, qui indiquent une association positive mais imparfaite entre les deux classements. Ces résultats, visibles uniquement à travers des rangs et non des valeurs brutes, montrent l'intérêt des statistiques ordinales pour analyser les structures spatiales hiérarchisées, là où les tests statistiques classiques ne sont pas adaptés.