PORTIER Marielle
M1 GAED Géopolitique-Geoint

Rapport analyse de données : *parcours intermédiaire*

Note : Les codes qui répondent aux manipulations demandées sont généralement dans des fichiers nommés "test" ou "ex" et non dans les "main.py" fournis pour l'exercice afin de ne pas le modifier directement.

Séance 4

Réponses aux questions de cours :

Pour choisir entre une distribution statistique avec des variables discrètes et une distribution statistique avec des variables continues, il existe différents critères. En effet, le choix dépend tout d'abord de la nature du phénomène étudié. Ce critère permet de choisir entre une loi discrète ou continue. Ensuite, il faut tenir compte de la forme de la distribution empirique, mais aussi des caractéristiques statistiques de l'ensemble de données (espérance, médiane, variance, écart-type, asymétrie, etc…). Enfin, le choix peut également dépendre du nombre de paramètres nécessaires à l'ajustement de la loi.
En somme, on utilise une loi discrète lorsque la variable ne peut prendre qu'un nombre fini ou dénombrable de valeurs (par exemple : résultats de sondages, tirages, comptages, jeux de hasard, événements rares).
A l'inverse, on utilise une loi continue lorsque la variable peut prendre n'importe quelle valeur réelle dans un intervalle (par ex : mesures physiques, revenus, altitudes, températures).
Selon moi, les lois les plus utilisées en géographie sont la loi normale, la loi log-normale et la loi de Zipf, en sachant que cela dépend du type de phénomène étudié (naturel, économique ou urbain). En effet, la loi normale est la plus fréquente pour modéliser les phénomènes naturels continus tels que l'altitude, la température, les précipitations ou encore les erreurs de mesure etc… Quant à la loi log-normale, elle est utilisée pour les phénomènes de croissance multiplicative, ce qui correspond à la distribution des revenus, la taille des villes, les surfaces agricoles etc… La loi de Zipf est également souvent utilisée en géographie urbaine pour les lois rang-taille, qui correspond à la relation entre le rang d'une ville et sa population. Enfin, nous pouvons aussi citer la loi de Poisson, qui permet de modéliser des catastrophes naturelles, ce qui correspond à la géographie des risques naturels (séismes, tempêtes, inondations…).

Mise en oeuvre avec Python :

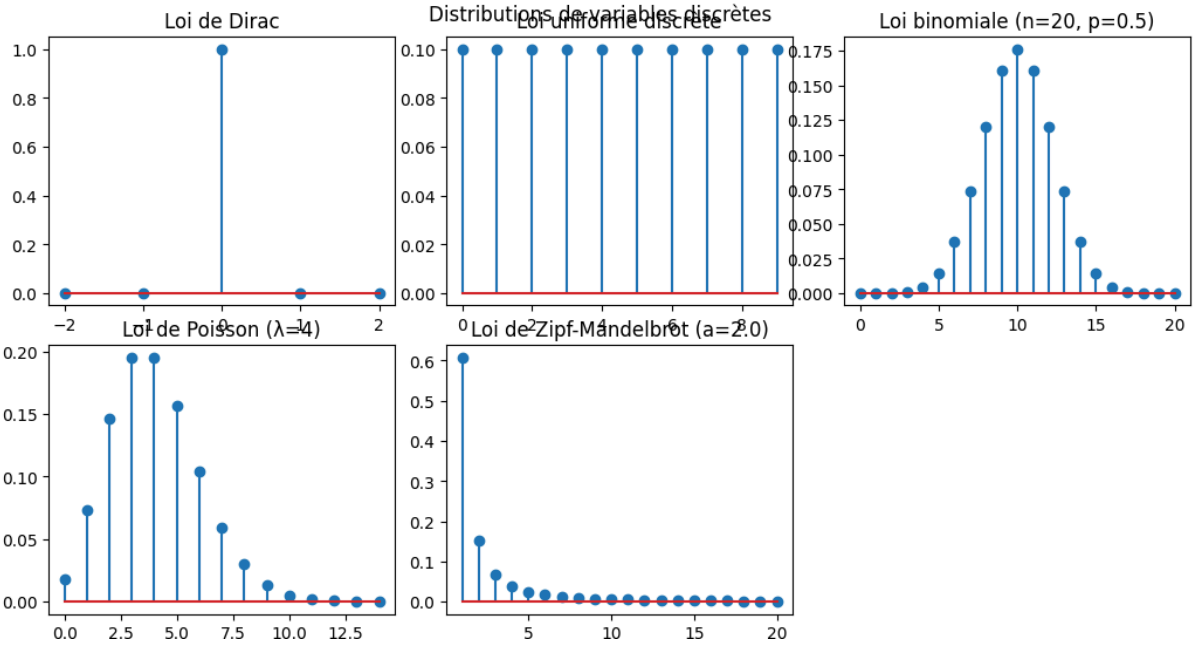

Figure 1 : Distributions des lois discrètes

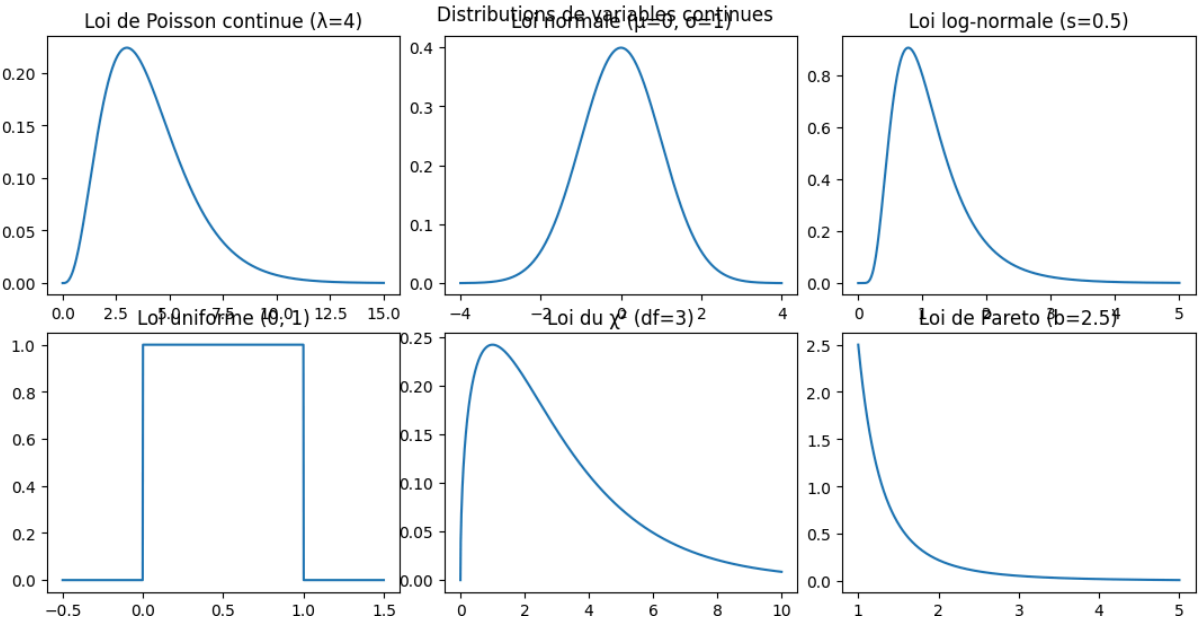

Figure 2 : Distributions de variables continues

Séance 5

Réponses aux questions de cours :

L’échantillonnage peut être défini comme une procédure méthodique consistant à extraire une partie limitée d’une population afin d’en analyser les caractéristiques et d’en déduire des informations valables pour l’ensemble, une démarche rendue indispensable par l’impossibilité matérielle, financière ou temporelle d’observer tous les individus d’une population donnée. Étudier la population entière, bien que théoriquement idéal, est le plus souvent irréaliste, ce qui justifie le recours à des échantillons supposés représentatifs, c’est-à-dire suffisamment proches de la structure réelle de la population pour permettre une généralisation prudente des résultats. Il existe différentes méthodes d’échantillonnage, notamment les méthodes aléatoires qui reposent sur le tirage au sort et garantissent l’égalité des chances de sélection, et les méthodes non aléatoires, comme les quotas ou l’échantillonnage systématique, qui cherchent à reproduire certaines caractéristiques connues de la population ; le choix de la méthode dépend du contexte de l’étude, des contraintes pratiques et du niveau de précision recherché. À partir des données collectées, un estimateur désigne une fonction mathématique appliquée aux observations aléatoires, tandis que l’estimation correspond à la valeur concrète obtenue une fois les données observées, ces deux notions étant au cœur de l’inférence statistique. Il est alors essentiel de distinguer l’intervalle de fluctuation, qui décrit la variabilité attendue d’une statistique lorsque le paramètre théorique est supposé connu, de l’intervalle de confiance, qui fournit une plage de valeurs plausibles pour un paramètre inconnu à partir de l’échantillon, avec un niveau de confiance donné. Dans la théorie de l’estimation, un biais apparaît lorsque l’estimateur ne se centre pas, en moyenne, sur la vraie valeur du paramètre, produisant ainsi une erreur systématique qui nuit à la fiabilité des conclusions. À l’inverse, lorsqu’une statistique est calculée sur l’ensemble de la population, on parle de statistique exhaustive, situation aujourd’hui rapprochée du traitement des données massives, où l’abondance d’informations peut donner l’illusion d’une connaissance totale tout en posant de nouveaux défis liés à la qualité et à l’interprétation des données. Le choix d’un estimateur constitue donc un enjeu majeur, car il implique de trouver un équilibre entre absence de biais, faible variance, convergence et robustesse face aux données aberrantes. Les paramètres peuvent être estimés selon différentes approches, telles que la méthode des moments ou le principe de vraisemblance, la sélection de la méthode dépendant des hypothèses sur la loi sous-jacente, de la quantité d’information disponible et des propriétés souhaitées pour l’estimateur. Les tests statistiques prolongent cette logique en permettant de prendre des décisions à partir des données, en comparant une hypothèse nulle à une hypothèse alternative et en mesurant le risque d’erreur associé à cette décision ; leur construction repose sur le choix d’une statistique de test, d’une loi de référence et d’un seuil de signification. Enfin, bien que la statistique inférentielle fasse l’objet de critiques, notamment concernant l’abus de tests ou la rigidité de certaines hypothèses, elle demeure un outil fondamental pour analyser l’incertitude et structurer le raisonnement scientifique, à condition d’être utilisée avec discernement et esprit critique

Mise en oeuvre avec Python :

Le test de Shapiro-Wilk permet de déterminer si une distribution suit une loi normale. L'interprétation se fait comme suit :

1. H0 : La distribution suit une loi normale

2. H1 : La distribution ne suit pas une loi normale
3. Si p-value > 0.05 : On ne peut pas rejeter H0
4. Si p-value < 0.05 : On rejette H0

La distribution avec la p-value > 0.05 peut être considérée comme suivant une loi normale avec un niveau de confiance de 95%.

Séance 6

Réponses aux questions de cours :

La statistique ordinale s'applique à des variables qualitatives dont les modalités peuvent être mises dans un ordre logique, sans que l'on puisse mesurer précisément les écarts entre elles ; elle s'oppose ainsi à la statistique nominale, qui ne repose sur aucune hiérarchisation des catégories. Ce type de variables permet d'exprimer des classements, lesquels sont particulièrement pertinents pour rendre visibles des formes d'organisation ou de hiérarchie spatiale, par exemple entre villes, territoires ou régions. Dans ce cadre, l'ordre naturel, généralement croissant, est à privilégier car il facilite l'interprétation, la comparaison et la détection de valeurs extrêmes, même si certaines exceptions existent selon les problématiques étudiées. La corrélation des rangs vise à mesurer le degré de similarité entre deux classements établis sur les mêmes objets, tandis que la concordance de classements cherche plus spécifiquement à évaluer l'accord ou le désaccord global entre plusieurs ordonnancements, en s'appuyant sur le comptage des paires concordantes et discordantes. Les tests de Spearman et de Kendall poursuivent un objectif commun, mais se distinguent par leur logique de calcul : le premier repose sur les écarts entre rangs et s'apparente à une corrélation classique adaptée aux données ordinales, alors que le second s'appuie sur la comparaison systématique des paires pour apprécier l'ordre relatif, ce qui le rend plus robuste et plus facilement généralisable à plusieurs classements. Enfin, les coefficients de Goodman-Kruskal et de Yule servent à mesurer l'intensité de l'association entre variables ordinales ou dichotomiques en comparant le surplus de paires concordantes par rapport aux paires discordantes, le coefficient de Yule constituant un cas particulier adapté aux tableaux de contingence de dimension deux, permettant ainsi d'interpréter clairement le sens et la force d'une relation ordonnée.

Mise en oeuvre avec Python :

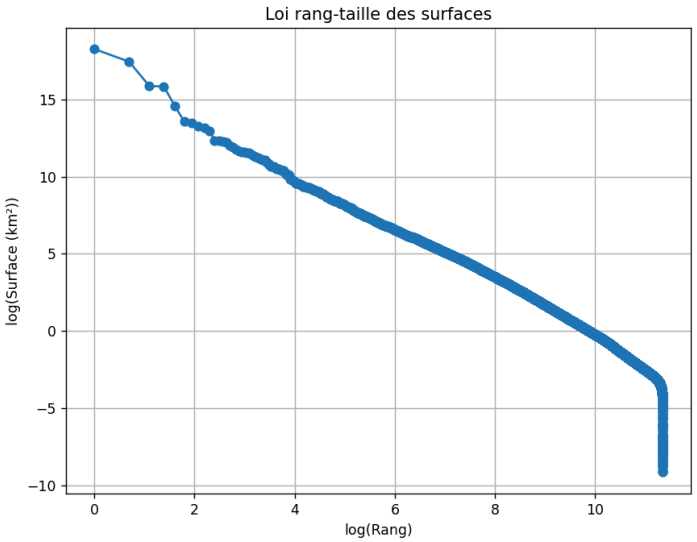

Figure 3 : Loi rang-taille des surfaces

Séance 7

Réponses aux questions de cours :

Passer de l'analyse univariée à la bivariée permet de ne plus se limiter à la description isolée d'un phénomène, mais d'examiner les liens éventuels entre deux variables et de mieux comprendre leurs interactions. Dans ce cadre, la corrélation vise à mesurer l'intensité et le sens d'une relation statistique entre deux variables quantitatives, tandis que les correspondances concernent plutôt l'étude des relations entre modalités de variables qualitatives ; le rapport de corrélation, quant à lui, évalue dans quelle mesure une variable peut être expliquée par une autre, sans supposer nécessairement une relation linéaire. Les valeurs marginales décrivent les distributions globales de chaque variable prise séparément, alors que les valeurs conditionnelles rendent compte de la répartition d'une variable en tenant compte d'une modalité précise de l'autre, distinction essentielle pour analyser la dépendance. De même, la variance mesure la dispersion d'une variable autour de sa moyenne, tandis que la covariance indique comment deux variables évoluent conjointement. Mesurer la corrélation ou tester l'indépendance permet ainsi de déterminer si une relation existe réellement et d'en apprécier la force. La méthode des moindres carrés repose sur la minimisation des écarts entre les valeurs observées et celles prévues par un modèle, afin d'obtenir la droite d'ajustement la plus pertinente. La théorie de la corrélation simple s'intéresse précisément à la quantification du lien linéaire entre deux variables et à son interprétation statistique. Toutefois, l'autocorrélation constitue un piège lorsque des

observations successives ne sont pas indépendantes, ce qui fausse les résultats. La régression linéaire consiste alors à modéliser la relation entre une variable expliquée et une variable explicative par une droite, afin de prédire ou d'expliquer. Le coefficient de corrélation indique la force et le sens du lien, tandis que le coefficient de détermination précise la part de la variabilité expliquée par le modèle. Enfin, tester les deux droites de régression est nécessaire car la relation n'est pas symétrique : expliquer X par Y n'équivaut pas à expliquer Y par X.

Mise en oeuvre avec Python :

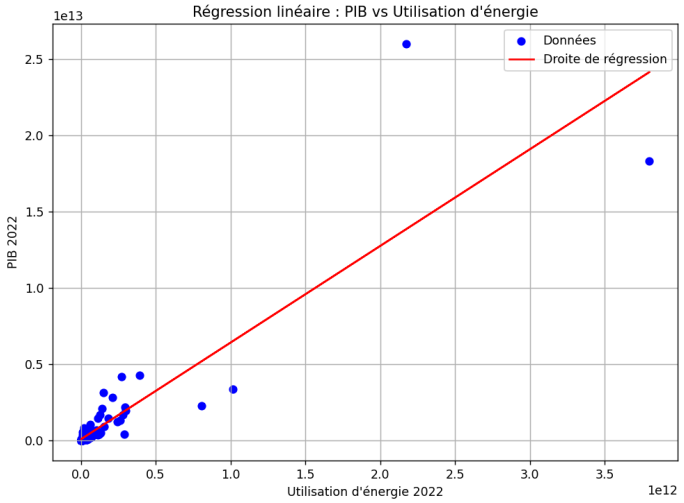

Figure 4 : Régression linéaire mettant en parallèle le PIB face à l'utilisation de l'énergie

Pente	6,34
Ordonnée à l'origine	73103136147
Coefficient de corrélation (r-value)	0,88654
P-value	1,03E-48
Corrélation simple entre les deux colonnes	0,8865430088

Figure 5 : Données du résultat de la régression linéaire

Séance 8

Réponses aux questions de cours :

La notion de corrélation, au sens strict, ne s'applique pas directement aux variables qualitatives, car elle suppose des valeurs numériques ordonnées et des écarts mesurables ; en revanche, on peut étudier l'existence de relations entre variables qualitatives à l'aide d'outils

adaptés, comme les tableaux de contingence. C'est précisément pour vérifier si deux caractères qualitatifs sont liés ou non que l'on utilise le test d'indépendance du khi-deux, qui permet de comparer les effectifs observés à ceux attendus en cas d'absence de relation. L'analyse de la variance à simple entrée, quant à elle, vise à déterminer si les moyennes d'une variable quantitative diffèrent significativement selon les modalités d'un seul facteur explicatif, en séparant la variabilité due aux groupes de celle liée aux fluctuations internes. Le rapport de corrélation mesure la part de la dispersion d'une variable expliquée par une autre, sans imposer une relation linéaire, alors que la correspondance renvoie à l'étude des associations entre catégories de variables qualitatives. L'analyse factorielle constitue une famille de méthodes dont l'objectif est de résumer l'information contenue dans un ensemble de variables en un nombre réduit de dimensions synthétiques, facilitant ainsi l'interprétation des structures sous-jacentes. Dans cette logique, l'analyse factorielle des correspondances applique ces principes aux tableaux croisant des variables qualitatives, afin de représenter graphiquement les relations entre modalités et de mettre en évidence les proximités ou oppositions entre elles.

Mise en oeuvre avec Python :

Tableau des effectifs attendus :

190.61490774	176.38509226
1015.92032574	940.07967426
3472.61927295	3213.38072705
3858.52356847	3570.47643153
3940.0672756	3645.9327244
3028.54367044	2802.45632956
172.95576098	160.04423902
12574.87065772	11636.12934228
63.88456036	59.11543964

Statistique du chi2	4812,419369
P-value	0
Degrés de liberté	8

Colonne 1	Agriculteurs exploitants	Artisans, commerçants	Cadres et professions	Professions intermédiaires	Employés	Ouvriers	Chômeurs n'ayant j	Inactifs	Non classés
Femmes	94	661	2889	3918	5770	1193	167	13566	60
Hommes	273	1295	3797	3511	1816	4638	166	10645	63

Figure 6 : Tri croisé entre les valeurs liées au genre et celles liées aux catégories socio-professionnelles

---

Réflexion sur les humanités numériques

Par définition, les humanités numériques désignent un champ de réflexion et de recherche qui combine les sciences humaines et les technologies numériques. De ce fait, elles interrogent plus largement la manière dont le numérique transforme nos façons de produire, d'analyser et de transmettre le savoir. D'un point de vue méthodologique, les humanités numériques ouvrent de nouvelles possibilités. L'analyse de corpus massifs de textes, la visualisation de données ou encore la cartographie interactive permettent de renouveler les approches classiques. Ces outils facilitent l'exploration de phénomènes à grande échelle, tout en rendant visibles des tendances, des réseaux ou des évolutions qui échappaient auparavant à l'analyse humaine seule. Toutefois, cette puissance technique ne remplace pas l'interprétation mais elle exige au contraire une vigilance accrue quant aux choix méthodologiques, aux biais algorithmiques et à la contextualisation des résultats.

Sur le plan épistémologique, les humanités numériques remettent en question certaines frontières disciplinaires. Elles encouragent le travail collaboratif entre chercheurs, ingénieurs, bibliothécaires et designers, bouleversant l'image traditionnelle du chercheur isolé. Ainsi, le savoir devient plus collectif, plus ouvert, parfois plus expérimental. Cette dynamique pose également la question de l'évaluation scientifique : selon quels critères peut-on reconnaître la valeur scientifique d'un projet numérique, d'une base de données ou d'un outil, au même titre qu'un article ou un ouvrage académique ?

Ensuite, les humanités numériques soulèvent des enjeux éthiques et politiques majeurs. L'accès aux données, la préservation des archives numériques, la dépendance aux plateformes privées ou encore l'exclusion de certaines populations face aux technologies interrogent la promesse d'un savoir plus démocratique. Le numérique peut à la fois favoriser l'ouverture et renforcer des inégalités existantes. Il appartient donc aux humanités numériques de développer une réflexion critique sur leurs propres pratiques, afin de ne pas confondre innovation technologique et progrès intellectuel ou social.

En définitive, les humanités numériques ne constituent pas une rupture totale avec les humanités traditionnelles, mais plutôt une reconfiguration. Elles invitent à repenser les outils, les méthodes et les valeurs qui fondent les sciences humaines. En combinant l'esprit critique et l'innovation technologique, elles offrent un espace privilégié pour interroger notre rapport contemporain au savoir, à la culture et à la mémoire dans une société de plus en plus façonnée par le numérique.

Les humanités numériques appliquées à la géographie

Dans un contexte de digitalisation massive des données géographiques, la maîtrise des outils liés aux humanités numériques est essentielle. Dans la discipline qu'est la géographie, les humanités numériques ont profondément renouvelé les pratiques de recherche et d'analyse de l'espace. En effet, l'usage des systèmes d'information géographique (SIG), de la cartographie numérique et des outils de visualisation interactive permet de croiser des données quantitatives et qualitatives, issues aussi bien d'archives que d'enquêtes de terrain, ou de sources numériques contemporaines. Ces différentes approches facilitent l'étude des dynamiques territoriales, des mobilités, des paysages ou des représentations de l'espace, à différentes échelles. En intégrant des méthodes computationnelles tout en conservant une lecture critique des données et des cartes produites, les humanités numériques en géographie interrogent également le pouvoir de la cartographie, les biais liés aux sources et les enjeux éthiques de la géolocalisation. Elles contribuent ainsi à une géographie plus collaborative, réflexive et ouverte, mais aussi attentive aux interactions entre les espaces, les sociétés et les technologies numériques.

Problèmes rencontrés et difficultés d'apprentissage

- Difficultés à mettre en route Python avec Docker malgré toutes les installations nécessaires. De ce fait, j'ai installé Python en brut et suis parvenue à mettre en œuvre les codes.

- N'ayant pas de bases en Python, il a été difficile de comprendre le fonctionnement de ce langage. J'ai donc regardé des vidéos explicatives et me suis aidée de l'intelligence artificielle de Visual Studio Code pour certaines manœuvres.

- Les cours de chaque séances étant très riches en informations, il est parfois difficile de tout assimiler d'une semaine à l'autre, surtout dans le cadre d'un enseignement utilisant la pédagogie inversée.

- En termes de méthode d'apprentissage, selon moi, plusieurs séances d'initiation au codage pour se familiariser davantage avec le langage Python seraient utiles afin de ne pas se sentir perdu face à ces divers nouveaux outils (VS Code, Git, GitHub, Docker etc…).