SORBONNE UNIVERSITÉ

**Lou Guilbot**

# Portfolio
# Cours d'analyse de données en géographie
# Master 1 - GEOINT
# Niveau Intermédiaire

## Séance 4. Les distributions statistiques

Lorsqu'on se penche sur le choix entre une distribution statistique basée sur des variables discrètes ou alors une distribution basée sur des variables continues, différents critères doivent être pris en considération afin de bien déterminer laquelle est la plus appropriée pour analyser un phénomène donné. En premier lieu, il faut considérer la nature des données elles-mêmes. Les variables discrètes se présentent comme des valeurs distinctes puis finies, souvent on les compte ou on les catégorise. On peut citer, par exemple, au nombre de maisons dans une commune, d'étudiants dans une école ou d'espèces observées dans un écosystème. Ces variables ne peuvent accepter que des valeurs spécifiques. On ne peut fractionner ces variables, qui sont généralement des entiers. Inversement, les variables continues figurent véritablement des mesures qui peuvent obtenir une multitude de valeurs dans un intervalle spécifique, comme par exemple la température, l'altitude, la distance ou aussi la densité de population. Elles sont bien adaptées pour des situations où la précision ainsi que la granularité sont importantes, car elles permettent de représenter certains changements infinitésimaux. Ensuite, il faut évaluer quel niveau de détail on souhaite analyser : soit on s'intéresse aux catégories distinctes ou si l'on compte précisément, alors la distribution discrète est privilégiée, soit l'on cherche à modéliser les phénomènes qui varient de manière fluide ou progressive, et donc la distribution continue offre une représentation exacte. Le type de représentation graphique et d'analyses statistiques prévues influence également le choix : les histogrammes et diagrammes en barres conviennent mieux aux variables discrètes, alors que les courbes de densité et les fonctions de répartition sont adaptées aux variables continues. Enfin, des critères pratiques comme la taille de l'échantillon, la précision des mesures et la facilité de collecte des données peuvent aussi guider la décision : des données difficiles à mesurer avec exactitude peuvent nécessiter un regroupement en classes pour utiliser des distributions continues de manière efficace.

En ce qui concerne les lois statistiques les plus utilisées en géographie, il est essentiel de considérer le type de phénomènes géographiques que l'on souhaite étudier et modéliser. Les géographes font fréquemment appel à la loi normale ou loi de Gauss, car de nombreux phénomènes naturels et humains suivent une distribution approximativement symétrique autour d'une moyenne : par exemple, la répartition des températures, des précipitations ou même certaines caractéristiques démographiques comme la taille des ménages ou le revenu moyen dans une population. Cette loi est extrêmement pratique car elle permet d'appliquer de nombreux outils statistiques, comme le calcul de probabilités ou d'écarts-types, pour décrire la variabilité des phénomènes. Ensuite, les lois exponentielles ou de Poisson sont souvent utilisées pour modéliser des événements rares ou discrets, comme le nombre de séismes dans une région donnée, les accidents de la route dans une ville ou l'occurrence de certaines espèces dans un territoire précis. La loi de Pareto ou loi des puissances est également très utilisée, notamment pour analyser des phénomènes géographiques liés à la concentration ou à l'inégalité, comme la distribution de la population dans les villes (où quelques grandes villes concentrent la majorité des habitants) ou la répartition des revenus et des ressources. Enfin, les géographes peuvent recourir à des lois spécifiques aux séries spatiales, comme les modèles de corrélation spatiale et les lois de distance, pour comprendre comment les événements ou les caractéristiques se dispersent ou se concentrent sur un territoire. Dans tous

les cas, le choix de la loi dépend fortement de l'objectif de l'étude : est-ce pour décrire un phénomène, pour prédire son évolution ou pour comparer plusieurs territoires ? Chaque loi apporte un cadre théorique qui permet de rendre les données compréhensibles et d'en tirer des conclusions fiables, ce qui est fondamental pour la géographie quantitative et pour la planification territoriale.

## Séance 5. Les statistiques inférentielles

L'échantillonnage est un concept central en statistique et en sciences quantitatives, qui permet de tirer des conclusions sur une population entière à partir d'un sous-ensemble représentatif de cette population. On peut définir l'échantillonnage comme la méthode consistant à sélectionner un certain nombre d'individus, d'objets ou d'unités statistiques, de manière systématique ou aléatoire, afin de recueillir des données permettant d'estimer des caractéristiques de la population totale. L'un des principaux arguments pour ne pas utiliser l'ensemble de la population réside dans des contraintes pratiques et logistiques. Dans la plupart des études, interroger chaque individu de la population serait extrêmement coûteux, chronophage, voire impossible : par exemple, mesurer les habitudes alimentaires de tous les habitants d'un pays demanderait des ressources considérables et un temps énorme. De plus, dans certaines situations, les populations sont en constante évolution, comme les flux migratoires ou la croissance démographique, ce qui rend la collecte exhaustive peu fiable ou rapidement obsolète. L'échantillonnage permet donc d'obtenir des informations représentatives avec un investissement beaucoup plus raisonnable.

Les méthodes d'échantillonnage sont diverses et leur choix dépend de la nature de la population, de la précision désirée et des objectifs de l'étude. L'échantillonnage aléatoire simple consiste à sélectionner au hasard des individus de la population, de manière à ce que chaque élément ait la même probabilité d'être choisi. Cette méthode est simple et intuitive, mais peut parfois produire des échantillons peu équilibrés si la population est hétérogène. L'échantillonnage stratifié, en revanche, divise la population en sous-groupes homogènes appelés strates, puis effectue un échantillonnage au sein de chaque strate selon un proportionnalité définie. Cette méthode permet de mieux représenter des sous-populations importantes, par exemple pour analyser les différences de revenus entre zones urbaines et rurales. L'échantillonnage systématique sélectionne des individus selon un intervalle fixe, par exemple tous les 10ème habitants d'une liste ordonnée. Enfin, l'échantillonnage par grappes choisit des groupes entiers (comme des écoles, quartiers ou entreprises) et inclut tous les individus de ces grappes. Le choix entre ces méthodes repose sur la connaissance de la population, la variabilité des données et la faisabilité logistique.

Une fois l'échantillon constitué, on utilise des estimateurs pour évaluer les paramètres inconnus de la population. Un estimateur est une statistique calculée à partir de l'échantillon qui fournit une approximation du paramètre, tandis que l'estimation correspond à la valeur numérique obtenue. Par exemple, si l'on souhaite connaître la moyenne de revenu d'une population, la moyenne des revenus de l'échantillon est l'estimateur, et le chiffre calculé

constitue l'estimation. L'intérêt de cette distinction réside dans le fait que l'estimateur est une variable aléatoire : sa valeur peut varier d'un échantillon à un autre, alors que l'estimation est une valeur concrète issue d'un échantillon précis. La qualité de l'estimateur est jugée selon plusieurs critères : son biais, sa variance, son efficacité et sa robustesse. Le biais mesure l'écart systématique entre l'estimateur et le paramètre réel ; un estimateur non biaisé fournit en moyenne la vraie valeur, tandis qu'un estimateur biaisé surestime ou sous-estime systématiquement le paramètre. La variance quantifie la dispersion des estimations autour de la moyenne de l'estimateur, et une variance faible indique une grande fiabilité. L'efficacité combine le biais et la variance pour juger la précision globale de l'estimateur, et la robustesse mesure la résistance aux valeurs aberrantes ou aux erreurs de mesure.

Pour compléter ces notions, il est important de comprendre la différence entre intervalle de fluctuation et intervalle de confiance. L'intervalle de fluctuation est une notion qui décrit la variabilité naturelle d'un échantillon autour d'une valeur théorique connue, souvent utilisée pour prévoir la fréquence à laquelle des valeurs similaires apparaissent dans différents échantillons. L'intervalle de confiance, quant à lui, est un outil d'inférence statistique qui indique, avec un certain niveau de probabilité (par exemple 95 %), l'intervalle dans lequel le paramètre inconnu de la population se situe. Ainsi, l'intervalle de fluctuation se concentre sur l'échantillon et ses variations, tandis que l'intervalle de confiance se concentre sur le paramètre et l'incertitude liée à son estimation.

Un autre point clé dans la théorie de l'estimation est celui du biais, qui correspond à une erreur systématique introduite par la méthode d'échantillonnage ou l'estimateur choisi. Par exemple, si l'échantillon est constitué uniquement de volontaires très motivés, certaines caractéristiques de la population peuvent être surreprésentées, ce qui conduit à une estimation biaisée. À l'inverse, lorsqu'une statistique est calculée sur la population entière, on parle de paramètre : il s'agit d'une valeur exacte qui décrit la population, comme la vraie moyenne ou la vraie variance. Avec l'émergence des données massives ou big data, il devient parfois possible de mesurer directement la population entière, mais même dans ce cas, des méthodes statistiques restent nécessaires pour traiter les anomalies, la qualité des données et les volumes gigantesques d'information.

Le choix d'un estimateur et des méthodes d'estimation est stratégique. Parmi les méthodes d'estimation les plus utilisées, on trouve l'estimation par moments, qui utilise les moments statistiques (moyenne, variance, etc.) pour construire l'estimateur, et l'estimation par maximum de vraisemblance, qui cherche le paramètre le plus probable étant donné les données observées. L'estimation bayésienne intègre des connaissances préalables pour affiner l'estimation. Le choix de la méthode dépend du type de données, de la distribution sous-jacente et du niveau de précision souhaité.

Les tests statistiques sont ensuite utilisés pour vérifier des hypothèses à partir des échantillons. On distingue différents tests : le test de Student pour comparer des moyennes, le test du chi² pour les distributions de fréquence, le test de corrélation pour évaluer la relation entre variables, et de nombreux autres tests spécialisés. La création d'un test implique la formulation d'une hypothèse nulle et d'une hypothèse alternative, le choix d'un niveau de

signification (souvent 5 %) et le calcul d'une statistique pour décider si l'hypothèse nulle peut être rejetée ou non.

Enfin, la statistique inférentielle fait l'objet de critiques, notamment parce qu'elle repose sur des hypothèses souvent simplifiées, des échantillons qui peuvent être non représentatifs, et sur l'idée que l'incertitude peut être quantifiée par des probabilités. Certains reprochent à la statistique inférentielle de donner une impression de certitude excessive ou de masquer des biais cachés. Pourtant, elle reste un outil indispensable pour comprendre des populations à partir d'échantillons et pour guider des décisions éclairées, à condition de respecter ses limites, de vérifier la qualité des données et de choisir judicieusement les méthodes d'échantillonnage, d'estimation et de test.

**Interprétation des résultats:** Dans un premier temps, les moyennes et fréquences des opinions (« Pour », « Contre », « Sans opinion ») sont calculées pour l'échantillon et pour la population mère, et l'on constate qu'elles sont très proches, ce qui suggère un échantillon représentatif. Les intervalles de fluctuation à 95 % confirment cette représentativité, puisque les fréquences observées de l'échantillon se situent toutes à l'intérieur des intervalles théoriques. Le calcul des intervalles de confiance permet ensuite d'estimer les proportions réelles dans la population avec un niveau de confiance donné, illustrant concrètement la logique de la statistique inférentielle. Enfin, les tests de normalité de Shapiro-Wilk appliqués aux jeux de données montrent que l'hypothèse de normalité est rejetée, ce qui indique que les distributions étudiées ne suivent pas une loi normale et justifie le recours à des méthodes adaptées. L'ensemble des résultats valide donc le bon fonctionnement du programme et met en évidence une démarche statistique rigoureuse, structurée et interprétable.

## Séance 6. La statistique d'ordre des variables qualitatives

Une statistique ordinale est une statistique qui repose sur des données pouvant être classées ou ordonnées selon un critère précis, mais pour lesquelles les écarts entre les catégories n'ont pas nécessairement de signification quantitative précise. En d'autres termes, elle permet de hiérarchiser ou de ranger des éléments, mais ne renseigne pas sur la distance exacte qui sépare chaque rang. Par exemple, on peut classer des villes selon le niveau de pollution de faible à élevé, ou des quartiers selon le niveau de satisfaction des habitants, mais l'écart entre le rang 1 et le rang 2 n'est pas forcément identique à celui entre le rang 2 et le rang 3. La statistique ordinale s'oppose directement à la statistique nominale, qui regroupe des données catégorielles sans ordre particulier, comme le type de sol, le code postal ou la couleur d'un bâtiment. Tandis que la statistique nominale se limite à identifier et regrouper, la statistique ordinale permet d'introduire une notion de hiérarchie. Dans un contexte spatial, cela devient particulièrement utile pour matérialiser des hiérarchies entre territoires, par exemple en classant des régions selon leur densité urbaine, leur niveau de développement économique ou leur vulnérabilité environnementale. Cette hiérarchisation permet ensuite de visualiser et de comparer les territoires selon des critères de rang, et de prendre des décisions éclairées en matière d'aménagement ou de politiques publiques.

Lorsqu'il s'agit de classifications, l'ordre à privilégier dépend directement de l'objectif de l'analyse et de la nature des variables. En général, on recommande de classer les éléments selon un ordre croissant ou décroissant du critère d'intérêt, ce qui permet de repérer facilement les extrêmes et de détecter les tendances. Par exemple, dans une étude géographique sur la richesse par département, un classement décroissant permet d'identifier immédiatement les zones les plus riches, tandis qu'un classement croissant met en évidence les zones les plus défavorisées. Cet ordre structuré facilite aussi la lecture et la communication des résultats, en particulier lorsque les classements sont visualisés sur des cartes ou des graphiques.

Pour ce qui est des relations entre variables ordinales, il est important de distinguer la corrélation des rangs de la concordance de classements. La corrélation des rangs mesure l'intensité et la direction de la relation monotone entre deux variables ordonnées, c'est-à-dire si des valeurs élevées d'une variable correspondent généralement à des valeurs élevées de l'autre. La concordance de classements, en revanche, se concentre sur l'accord ou le désaccord global entre deux classements : elle examine si les éléments occupent des positions similaires dans deux classements différents, sans nécessairement quantifier la force de la relation. Ces notions sont complémentaires : la corrélation des rangs permet d'évaluer l'existence et la direction d'un lien, tandis que la concordance mesure la cohérence globale des classements.

Deux des tests les plus utilisés pour analyser les données ordinales sont Spearman et Kendall. Le test de Spearman, ou coefficient rho, évalue la corrélation monotone entre deux variables ordinales ou continues transformées en rangs. Il se base sur la différence entre les rangs correspondants et fournit une mesure de l'association qui varie entre -1 et 1. Le test de Kendall, ou tau, évalue également la corrélation entre deux rangs, mais il se concentre sur les paires concordantes et discordantes pour calculer un coefficient qui reflète la proportion d'accord global. La différence principale réside donc dans le calcul : Spearman se base sur la différence des rangs, tandis que Kendall examine la proportion de paires correctement ordonnées. En pratique, Kendall est considéré comme plus robuste pour les petits échantillons ou lorsqu'il y a beaucoup d'ex-æquo dans les rangs.

Enfin, certains coefficients permettent de mesurer des relations spécifiques entre variables catégorielles ou ordinales, même dans des tableaux de contingence. Le coefficient de Goodman-Kruskal sert à évaluer la force de l'association entre une variable indépendante et une variable dépendante ordinales ou catégorielles, et indique dans quelle mesure la connaissance de la première permet de prédire la seconde. Le coefficient de Yule, quant à lui, est utilisé pour mesurer l'association entre deux variables dichotomiques et permet de quantifier la tendance d'occurrence simultanée ou inverse des deux caractéristiques. Ces coefficients sont particulièrement utiles en géographie pour analyser des relations entre caractéristiques spatiales, comme la relation entre le niveau de développement économique et le type d'occupation du sol, ou pour détecter des tendances et corrélations dans des données sociales et environnementales.

En résumé, les statistiques ordinales sont des outils puissants pour organiser, hiérarchiser et comparer des territoires ou des phénomènes géographiques, offrant à la fois une compréhension qualitative et quantitative des données. Leur utilisation implique des choix méthodologiques précis, tant dans l'ordre des classements que dans les tests statistiques appliqués, et permet de transformer des observations disparates en analyses cohérentes, facilitant la prise de décision et la communication des résultats.

**Interprétation des résultats:** Ce résultat met en évidence une analyse comparative détaillée des populations et des densités de population à l'échelle mondiale entre 2007 et 2025. Les classements montrent d'abord des évolutions nettes parmi les pays les plus peuplés : l'Inde dépasse la Chine en 2025, tandis que des pays comme le Pakistan et le Nigeria gagnent plusieurs places, ce qui traduit une croissance démographique plus rapide que celle d'autres grandes puissances comme le Brésil ou le Japon, ce dernier reculant nettement dans le classement. Du côté des densités de population, les premières places restent occupées par des États de petite superficie ou des cités-États, mais on observe des changements notables, comme l'apparition de Monaco en tête en 2025 et le recul de certains pays européens comme les Pays-Bas. La comparaison des rangs entre 2007 et 2025 montre que les classements de population et de densité évoluent différemment : des pays très peuplés ne sont pas nécessairement très denses, et inversement. Cette dissociation est confirmée par les résultats des corrélations de rangs, qui indiquent une relation faible et non significative entre population totale et densité, aussi bien en 2007 qu'en 2025. Autrement dit, le volume de population d'un pays n'explique pas directement son niveau de densité. L'ensemble des données souligne donc la complexité des dynamiques démographiques mondiales, où la taille des États, leur superficie et leur évolution démographique jouent un rôle déterminant dans la structuration des classements.

## Séance 7. Régression et corrélation statistique de deux variables

Passer des statistiques univariées aux statistiques bivariées représente une étape essentielle dans l'analyse des données, car cela permet de dépasser l'étude d'une seule variable isolée pour explorer les relations entre deux variables. Alors que les statistiques univariées se concentrent sur des mesures comme la moyenne, la médiane, la variance ou l'écart type pour une variable donnée, les statistiques bivariées permettent de comprendre comment deux variables interagissent, s'influencent mutuellement, ou présentent des associations spatiales ou temporelles. Par exemple, en géographie, on peut étudier séparément la population d'une ville et son revenu moyen (univariées), mais c'est en les comparant (bivariées) que l'on peut identifier des relations telles que les corrélations entre densité urbaine et niveau de vie, ou encore analyser l'impact des infrastructures sur la distribution de la population. Cette approche fournit une dimension explicative et prédictive, fondamentale pour la modélisation et la prise de décision.

Dans ce cadre, il est crucial de distinguer corrélation et correspondances. La corrélation mesure l'intensité et la direction d'une relation linéaire entre deux variables quantitatives,

tandis que les correspondances sont souvent utilisées pour explorer des relations dans des tableaux de contingence ou entre variables qualitatives. Le rapport de corrélation, quant à lui, est un coefficient, souvent noté rrr, compris entre -1 et 1, qui indique la force et la direction de la relation : un rrr proche de 1 signifie une corrélation positive forte, un rrr proche de -1 une corrélation négative forte, et un rrr proche de 0 l'absence de relation linéaire.

Les notions de valeurs marginales et de valeurs conditionnelles sont également fondamentales. Les valeurs marginales correspondent aux totaux ou moyennes calculés sur une seule variable, indépendamment des autres, tandis que les valeurs conditionnelles tiennent compte de la valeur d'une autre variable. Par exemple, dans un tableau représentant le nombre d'habitants selon la classe d'âge et le niveau de revenu, les valeurs marginales indiquent le total par âge ou par revenu, alors que les valeurs conditionnelles montrent, par exemple, la distribution des revenus pour chaque tranche d'âge. Distinguer les deux permet de mieux comprendre les dépendances ou l'indépendance entre variables et d'éviter les interprétations erronées liées à la globalisation des données.

Pour mesurer la variabilité et les relations, on utilise variance et covariance. La variance décrit la dispersion des valeurs autour de la moyenne d'une seule variable, tandis que la covariance mesure la façon dont deux variables varient simultanément. Une covariance positive indique que les variables augmentent ou diminuent ensemble, alors qu'une covariance négative signifie que lorsque l'une augmente, l'autre diminue. Mesurer la corrélation ou l'indépendance permet donc d'identifier si les variables sont liées et dans quelle direction, ce qui est crucial pour l'analyse prédictive et la planification.

Le principe de la méthode des moindres carrés est au cœur de la régression linéaire. Il consiste à trouver la droite qui minimise la somme des carrés des écarts entre les valeurs observées et les valeurs prédites par le modèle. Cette méthode sert à ajuster un modèle linéaire aux données afin de prédire la valeur d'une variable dépendante à partir d'une variable indépendante. La théorie de la corrélation simple s'inscrit dans cette logique : elle étudie la relation linéaire entre deux variables quantitatives et cherche à quantifier la force et la direction de cette relation à travers le coefficient de corrélation.

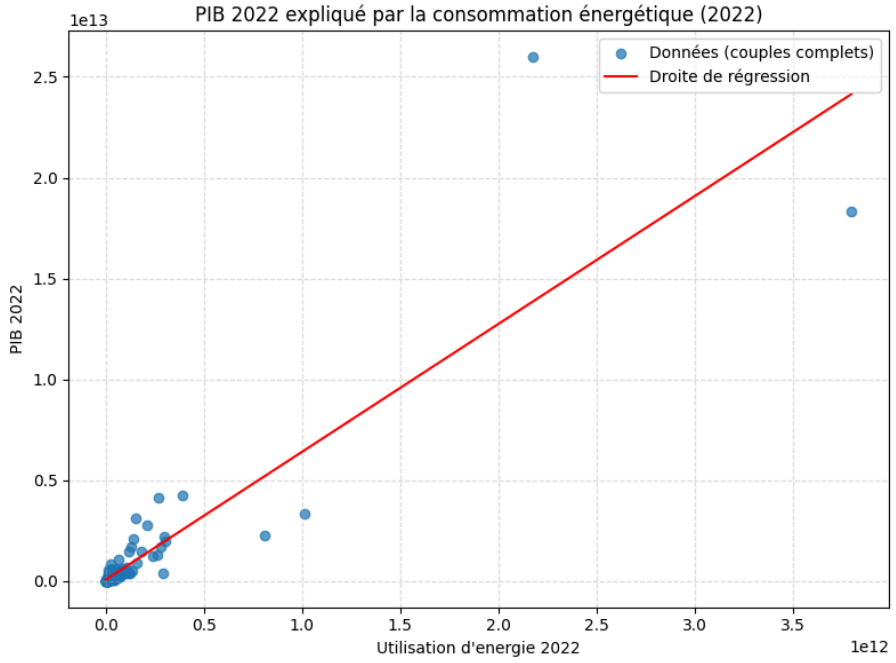
Cependant, il existe un piège classique, appelé autocorrélation, qui se produit lorsque les valeurs successives d'une variable ne sont pas indépendantes, comme dans des séries temporelles ou des données spatiales. L'autocorrélation peut biaiser les analyses et conduire à des estimations incorrectes si elle n'est pas prise en compte. La régression linéaire permet alors de modéliser la relation entre une variable indépendante et une variable dépendante par une droite d'ajustement, facilitant la prédiction et l'interprétation de tendances.

Une distinction importante à connaître est celle entre coefficient de corrélation et coefficient de détermination. Le premier, rrr, mesure la force et la direction de la relation entre deux variables, tandis que le second, R2R^2R2, indique la proportion de la variance de la variable dépendante expliquée par la variable indépendante dans le modèle de régression. Tester les deux droites de régression, c'est-à-dire en inversant le rôle des variables dépendante et indépendante, est essentiel pour vérifier la robustesse du modèle et comprendre la symétrie

ou l'asymétrie de la relation, surtout lorsque l'on soupçonne des influences mutuelles ou des variables cachées.

En résumé, passer des statistiques univariées aux statistiques bivariées permet de transformer une simple description des données en une analyse relationnelle, d'identifier des corrélations et dépendances, de construire des modèles prédictifs fiables et de détecter des pièges tels que l'autocorrélation. La combinaison de mesures comme la covariance, la corrélation, la régression et les tests de significativité constitue le socle des analyses quantitatives avancées, essentielles pour la géographie, l'économie, les sciences sociales et bien d'autres disciplines où comprendre les liens entre variables est crucial.

**Interprétation des résultats:** Ce résultat met clairement en évidence une relation très forte entre le niveau de PIB et la consommation d'énergie en 2022 pour les territoires étudiés. Sur les 217 territoires initiaux, seuls 142 disposent de données complètes pour ces deux variables, ce qui montre déjà une certaine inégalité dans la disponibilité statistique à l'échelle mondiale. L'examen des couples de valeurs confirme que les économies les plus riches sont aussi celles qui consomment le plus d'énergie : les PIB les plus élevés correspondent systématiquement à des niveaux de consommation énergétique très importants. La corrélation observée est très élevée, ce qui signifie que les pays ayant une forte utilisation d'énergie tendent presque toujours à afficher un PIB important. Le graphique de régression renforce cette lecture visuelle : les points sont globalement alignés selon une tendance croissante, avec peu de dispersion autour de la droite. Cela suggère que, à l'échelle mondiale, la richesse économique reste étroitement liée à l'intensité des activités productives et donc à la consommation d'énergie. En revanche, la présence d'une constante élevée et de quelques écarts autour de la tendance rappelle que tous les territoires ne transforment pas l'énergie en richesse de la même manière, traduisant des différences de structures économiques, de niveaux d'industrialisation ou d'efficacité énergétique.



Le graphique montre une relation clairement positive entre la consommation d'énergie et le PIB en 2022 pour les 142 pays disposant de données complètes. Le nuage de points est globalement aligné selon une tendance croissante, ce qui est confirmé par la droite de régression ascendante : plus un pays consomme d'énergie, plus son PIB est élevé. La majorité des pays se concentre dans la zone de faible consommation énergétique et de PIB modéré, tandis que quelques pays à très forte consommation et très fort PIB apparaissent comme des points extrêmes, étirant l'échelle du graphique. Malgré une certaine dispersion, notamment pour les niveaux élevés de consommation, l'alignement général des points autour de la droite indique une relation forte et cohérente entre les deux variables, suggérant que la consommation énergétique constitue un bon indicateur du niveau de richesse économique.

## Séance 8. Étude de deux variables qualitatives

La question de savoir si la corrélation entre deux variables qualitatives a un sens mérite une attention particulière. En général, la corrélation classique (comme le coefficient de Pearson) est conçue pour des variables quantitatives continues et repose sur l'idée de mesurer la relation linéaire entre valeurs numériques. Pour deux variables qualitatives, surtout nominales, ce type de corrélation n'a donc pas de sens direct, car il n'existe pas de notion d'ordre ou de distance entre les catégories. Cependant, il est possible de mesurer une forme d'association ou de dépendance entre deux variables qualitatives en utilisant d'autres outils, comme le chi² de Pearson, les coefficients de Cramer ou les coefficients de Yule, qui quantifient la force de la relation sans supposer de linéarité ou de distances entre catégories.

Le test d'indépendance du $\chi^2$ est pratiqué précisément pour évaluer si deux variables qualitatives sont indépendantes ou liées. L'objectif est de comparer les fréquences observées dans un tableau de contingence avec celles qui seraient attendues si les variables étaient effectivement indépendantes. Si les écarts entre valeurs observées et attendues sont significatifs, on conclut que les variables sont associées. Ce test est fondamental dans les analyses géographiques, sociologiques ou marketing, car il permet de détecter des relations entre catégories, comme la relation entre type de logement et niveau de revenus, ou entre localisation géographique et préférences de consommation.

L'analyse de la variance à simple entrée (ANOVA) constitue une autre technique essentielle pour comparer plusieurs groupes. Elle permet d'évaluer si les moyennes d'une variable quantitative diffèrent significativement selon les modalités d'une variable qualitative. Par exemple, on peut tester si le revenu moyen diffère selon différentes régions ou si la densité urbaine varie selon les types de quartier. L'ANOVA compare la variance entre les groupes avec la variance au sein des groupes, et si la variance entre groupes est significativement plus grande, on conclut que les moyennes ne sont pas toutes égales. Cette méthode est simple mais puissante pour analyser l'effet d'une seule variable catégorielle sur une variable quantitative.

Le rapport de corrélation, comme indiqué précédemment, mesure l'intensité et la direction de la relation entre deux variables, tandis que la correspondance s'intéresse davantage à l'accord

global ou à l'association entre rangs ou catégories. La corrélation quantifie la force et la direction d'une relation, souvent avec un coefficient numérique, alors que la correspondance se concentre sur la cohérence ou l'association des positions ou classes, ce qui est plus pertinent pour des données ordinales ou qualitatives.

L'analyse factorielle est une technique d'exploration multivariée qui permet de résumer l'information contenue dans un grand nombre de variables en un nombre réduit de facteurs ou dimensions principales, tout en conservant autant que possible la variance et la structure originale des données. Elle est très utilisée pour identifier des motifs sous-jacents, des corrélations cachées et pour simplifier l'interprétation de jeux de données complexes. Par exemple, en géographie, elle peut être utilisée pour regrouper des indicateurs socio-économiques, environnementaux et démographiques afin de dégager des profils régionaux ou des typologies de territoires.

L'analyse factorielle des correspondances (AFC) est une variante adaptée aux tableaux de contingence ou aux variables qualitatives. Son objectif est de représenter graphiquement et de synthétiser l'association entre modalités de deux variables ou plus. Elle projette les modalités sur un plan factoriel, où la proximité entre points reflète l'association ou la similarité des catégories. Par exemple, dans l'étude des préférences culturelles selon différentes régions, l'AFC permet de visualiser quels types de préférences sont typiques de certaines régions et de détecter des groupes de modalités fortement associées. Elle constitue ainsi un outil extrêmement puissant pour interpréter des relations complexes dans les données qualitatives et pour guider la prise de décision ou l'aménagement du territoire.

En résumé, ces méthodes — du test du $\chi^2$ à l'ANOVA, en passant par la corrélation, la correspondance et les analyses factorielles — offrent un ensemble cohérent d'outils pour passer de la simple description de données à l'analyse des relations, associations et structures sous-jacentes. Elles permettent de comprendre non seulement si des variables sont liées, mais aussi comment elles se regroupent, se hiérarchisent ou se combinent pour révéler des patterns significatifs, ce qui est fondamental dans des disciplines comme la géographie, la sociologie et l'économie.

**Interprétation des résultats:** Ce résultat montre des données de 54 522 individus répartis par catégorie socioprofessionnelle et par sexe, avec des effectifs globalement comparables entre femmes (28 318) et hommes (26 204), mais une distribution très différente selon les catégories. Les femmes sont nettement surreprésentées parmi les employés (5 770 contre 1 816) et les professions intermédiaires, tandis que les hommes dominent largement chez les ouvriers (4 638 contre 1 193), les artisans-commerçants et les agriculteurs. La catégorie des cadres est plus équilibrée mais reste majoritairement masculine, alors que les chômeurs n'ayant jamais travaillé sont quasiment à parité. Les inactifs constituent de loin le groupe le plus nombreux pour les deux sexes, avec un écart marqué en faveur des femmes. Le test du chi-deux très élevé ($\chi^2 \approx 4\ 812$, ddl = 8) et la p-value nulle indiquent que ces écarts ne relèvent pas de fluctuations aléatoires : la catégorie socioprofessionnelle et le sexe sont statistiquement liés. La valeur de Cramér ($V \approx 0{,}30$) montre que cette liaison est d'intensité

modérée mais substantielle, traduisant une structuration sexuée marquée des positions professionnelles dans l'ensemble des données.

## Réflexion sur les sciences des données et les humanités numériques :

Au fil des différentes séances, ce parcours en sciences des données et en humanités numériques m'a permis de découvrir un domaine que je connaissais très peu au départ, et qui m'a souvent mis en difficulté, mais aussi fait progresser. Les premières séances ont posé les bases de la manipulation de données, de leur structuration et de leur interprétation, ce qui m'a fait prendre conscience que derrière des chiffres ou des tableaux apparemment simples se cachent en réalité des choix méthodologiques importants. J'ai compris que les humanités numériques ne se limitent pas à utiliser des outils informatiques, mais qu'elles consistent aussi à réfléchir à la manière dont les données représentent des réalités sociales, économiques ou culturelles.

À partir des séances plus techniques, notamment à partir de la séance 4, les difficultés ont augmenté de manière significative. J'ai rencontré beaucoup de problèmes liés à l'installation de Python, aux différentes versions du langage, ainsi qu'à la configuration de l'environnement de travail. Entre les conflits de versions, les chemins mal reconnus, les bibliothèques manquantes et les erreurs incompréhensibles au départ, il m'est arrivé de passer plus de temps à essayer de faire fonctionner l'outil qu'à travailler réellement sur les données. L'installation des extensions et des bibliothèques nécessaires (pandas, matplotlib, scipy, etc.) a été une source de frustration importante, car une petite erreur suffisait à bloquer toute la séance.

Les séances 4 à 8 ont été particulièrement marquantes sur ce point, et surtout la séance 8. Même si les exercices proposés étaient intéressants et formateurs, je n'avais quasiment aucune base préalable en programmation ou en analyse statistique avec Python, ce qui a rendu le début très compliqué. Comprendre les messages d'erreur, savoir s'ils venaient du code, des données ou de l'environnement Python lui-même n'était pas évident. J'ai souvent dû chercher des solutions par essais et erreurs, relire plusieurs fois le code, ou m'appuyer sur des ressources extérieures pour comprendre ce qui bloquait. Certaines difficultés ont été résolues progressivement, notamment en comprenant mieux la logique du langage et l'organisation d'un script, tandis que d'autres ont parfois été contournées sans être totalement maîtrisées.

Malgré ces obstacles, ces séances m'ont permis de mieux comprendre le lien entre données et sciences humaines. Les exercices sur le PIB, la consommation d'énergie ou encore les catégories socioprofessionnelles m'ont montré comment des outils quantitatifs peuvent éclairer des phénomènes sociaux réels, comme les inégalités économiques ou les différences de répartition entre les sexes. Même lorsque la technique prenait le dessus, j'ai réalisé que l'objectif restait l'analyse et l'interprétation des données, et non le code pour le code. Cela m'a aidé à donner du sens aux difficultés rencontrées.

En termes d’apprentissage, ce parcours m’a surtout appris la patience et l’autonomie. J’ai pris conscience que les sciences des données demandent du temps, de la rigueur et une capacité à accepter de ne pas tout comprendre immédiatement. Le fait de partir de presque rien a rendu la progression lente, mais aussi plus visible. Même si certaines notions restent encore floues, je me sens aujourd’hui plus à l’aise face à un jeu de données, un script Python ou un résultat statistique qu’au début du parcours. En ce sens, malgré les nombreuses difficultés techniques et le sentiment de découragement à certains moments, ces séances constituent une expérience formatrice qui m’a permis de mieux appréhender les enjeux et les méthodes des humanités numériques.