Rapport d'activité – Séance 6
La statistique d'ordre des variables qualitatives

Honoka OYAMADA

Introduction

Ce rapport synthétise les réponses aux questions de cours et les résultats produits par le script Python de la séance 6. L'objectif est de comprendre comment des statistiques d'ordre (rangs) permettent d'analyser des variables qualitatives ordinales (classements) et de comparer des classements entre eux (corrélation des rangs et concordance).

1. Questions de cours

1.1 Qu'est-ce qu'une statistique ordinale ?

Une statistique ordinale est une statistique construite à partir d'un ordre entre modalités : on ne mesure pas d'abord une distance numérique entre valeurs, mais un classement (rang 1, rang 2, etc.). Elle s'oppose principalement aux statistiques nominales (catégorielles sans ordre), où les catégories sont seulement différentes mais non hiérarchisées. Elle s'applique à des variables ordinales (p. ex. « faible / moyen / fort », ou un classement de pays par population). En géographie, ce type d'approche matérialise facilement une hiérarchie spatiale : par exemple une hiérarchie urbaine ou une hiérarchie d'États selon la population, la densité, la richesse, etc.

1.2 Quel ordre privilégier dans les classifications ?

Dans une classification, l'ordre à privilégier est celui qui est explicitement justifié par la question de recherche et par le sens géographique : en pratique, on privilégie un ordre monotone et interprétable (souvent décroissant pour faire apparaître les « plus grands / plus denses / plus influents » en tête). Quand l'objectif est de comparer des distributions très dissymétriques, un ordre décroissant et une visualisation rang-taille (souvent en log-log) sont particulièrement adaptés.

1.3 Corrélation des rangs vs concordance de classements

La corrélation des rangs (ex. Spearman) mesure une relation monotone entre deux séries de rangs : si une variable augmente, l'autre tend-elle à augmenter (en termes de rang) ? La concordance (ex. Kendall) s'intéresse davantage à la cohérence paire-à-paire des ordres : parmi toutes les paires d'objets, combien sont ordonnées dans le même sens dans les deux classements (paires concordantes) ou dans le sens opposé (paires discordantes) ? Les deux approches répondent à des questions proches mais pas identiques.

1.4 Différence entre Spearman et Kendall

Spearman (rho) est, en pratique, une corrélation calculée sur les rangs ; il est sensible à la structure globale des rangs et peut être vu comme une corrélation de Pearson appliquée aux rangs. Kendall (tau) s'appuie sur les paires concordantes/discordantes : il est souvent plus robuste et plus

directement interprétable comme un « taux de concordance » (à un coefficient près), au prix d'une valeur numérique souvent plus faible que Spearman pour une même association.

1.5 Goodman–Kruskal et Yule

Les coefficients de Goodman–Kruskal (par ex. gamma, lambda, tau selon le contexte) et ceux de Yule (Q, Y) sont des mesures d'association pour variables catégorielles, notamment dans des tableaux de contingence. Ils servent à quantifier la force (et parfois le sens) d'une association entre catégories, en particulier quand on ne souhaite pas supposer une échelle numérique. Ils sont utiles pour analyser des relations entre variables qualitatives (nominales ou ordinales selon le coefficient).

2. Mise en œuvre avec Python

2.1 Fonctions locales et logique du script

Le script repose sur des fonctions locales (ouvrirUnFichier, conversionLog, ordreDecroissant, ordrePopulation, classementPays) afin de découper le problème en tâches simples : lecture des données, transformation, tri, construction de classements, puis comparaison de classements. Cette factorisation rend le code plus lisible, réutilisable et testable.

2.2 Partie A – Îles et loi rang-taille

Le script charge un fichier « island-index.csv » puis isole la colonne de surface (km²), à laquelle il ajoute les surfaces des continents. La liste est triée par ordre décroissant et on trace la loi rang-taille (rang en abscisse, surface en ordonnée), d'abord en linéaire puis en log-log (conversionLog).

Remarque importante : dans l'environnement de rendu de ce rapport, le fichier « island-index.csv » n'était pas disponible ; je ne peux donc pas reproduire numériquement les figures de la partie A ici. En revanche, la logique méthodologique est la suivante : (1) construire une distribution très dissymétrique (quelques très grandes surfaces, beaucoup de petites), (2) visualiser en rang-taille, puis (3) passer en log-log pour rendre lisible la queue de distribution et évaluer une éventuelle structure de type loi de puissance.

Question 7 (test sur les rangs) : oui, on peut tester l'association entre deux classements via une corrélation des rangs (Spearman) et/ou une concordance (Kendall), par exemple si l'on compare le classement des surfaces au classement des traits de côte.

2.3 Partie B – Populations et densités des États (2007–2025)

Le script charge le fichier « Le-Monde-HS-Etats-du-monde-2007-2025.csv », isole les colonnes d'État, de population et de densité (2007 et 2025), puis construit quatre classements décroissants grâce à ordrePopulation(). La fonction classementPays() permet ensuite d'aligner deux classements sur les mêmes États afin de comparer les rangs population/densité.

Top 10 – Population (2007)

Rang	État	Pop (hab.)
1	Chine	1 311 400 000

2	Inde	1 121 800 000
3	États-Unis	299 100 000
4	Indonésie	225 500 000
5	Brésil	186 800 000
6	Pakistan	165 800 000
7	Bangladesh	146 600 000
8	Russie	142 300 000
9	Nigeria	134 500 000
10	Japon	127 800 000

Top 10 – Population (2025)

Rang	État	Pop (hab.)
1	Inde	1 450 000 000
2	Chine	1 400 000 000
3	États-Unis	345 400 000
4	Indonésie	283 400 000
5	Pakistan	251 200 000
6	Nigeria	232 600 000
7	Brésil	211 900 000
8	Bangladesh	173 500 000
9	Russie	144 800 000
10	Éthiopie	132 000 000

Top 10 – Densité (2007)

Rang	État	Densité (hab./km²)
1	Singapour	7500.00

2	Malte	1333.00
3	Bangladesh	1018.00
4	Maldives	1000.00
5	Bahreïn	1000.00
6	La Barbade	750.00
7	Maurice	650.00
8	Taïwan	633.00
9	Corée du Sud	490.00
10	Pays-Bas	400.00

Top 10 – Densité (2025)

Rang	État	Densité (hab./km²)
1	Monaco	19150.00
2	Singapour	9666.67
3	Bahreïn	2285.71
4	Malte	1800.00
5	Maldives	1760.00
6	Bangladesh	1204.86
7	La Barbade	705.00
8	Taïwan	644.44
9	Maurice	600.00
10	Liban	580.00

2.4 Corrélation des rangs population vs densité

Pour 2007, la corrélation de Spearman entre les rangs de population et de densité vaut rho = 0.0928 (p = 0.224). La concordance de Kendall vaut tau = 0.0668 (p = 0.192).

Pour 2025, Spearman vaut rho = -0.0269 (p = 0.709) et Kendall vaut tau = -0.0075 (p = 0.877).

Interprétation : sur ces données, il n'y a pas d'association monotone forte entre « être très peuplé » et « être très dense » au niveau mondial. Des États peuvent être très peuplés mais relativement peu denses (grande superficie) et inversement, des États très denses peuvent être de petite taille démographique.

Graphiques – dispersion des rangs

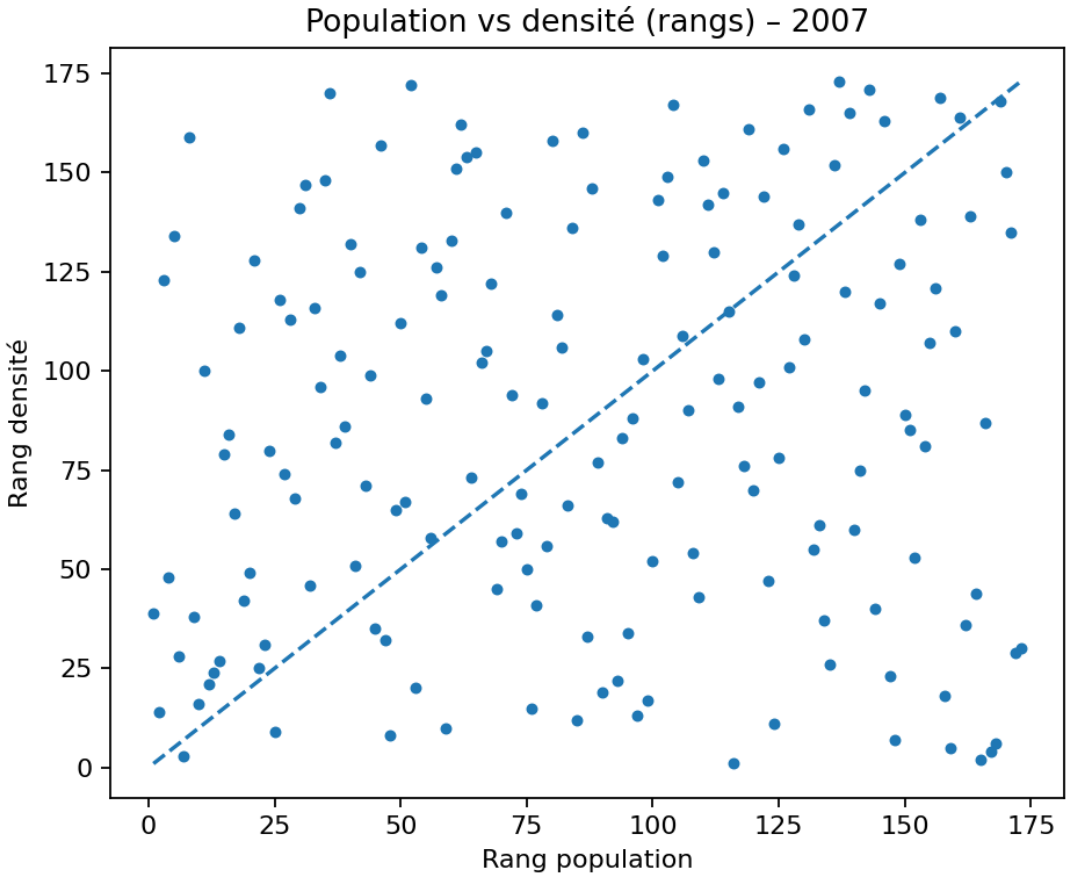

Figure 1 – Rangs population vs densité (2007)

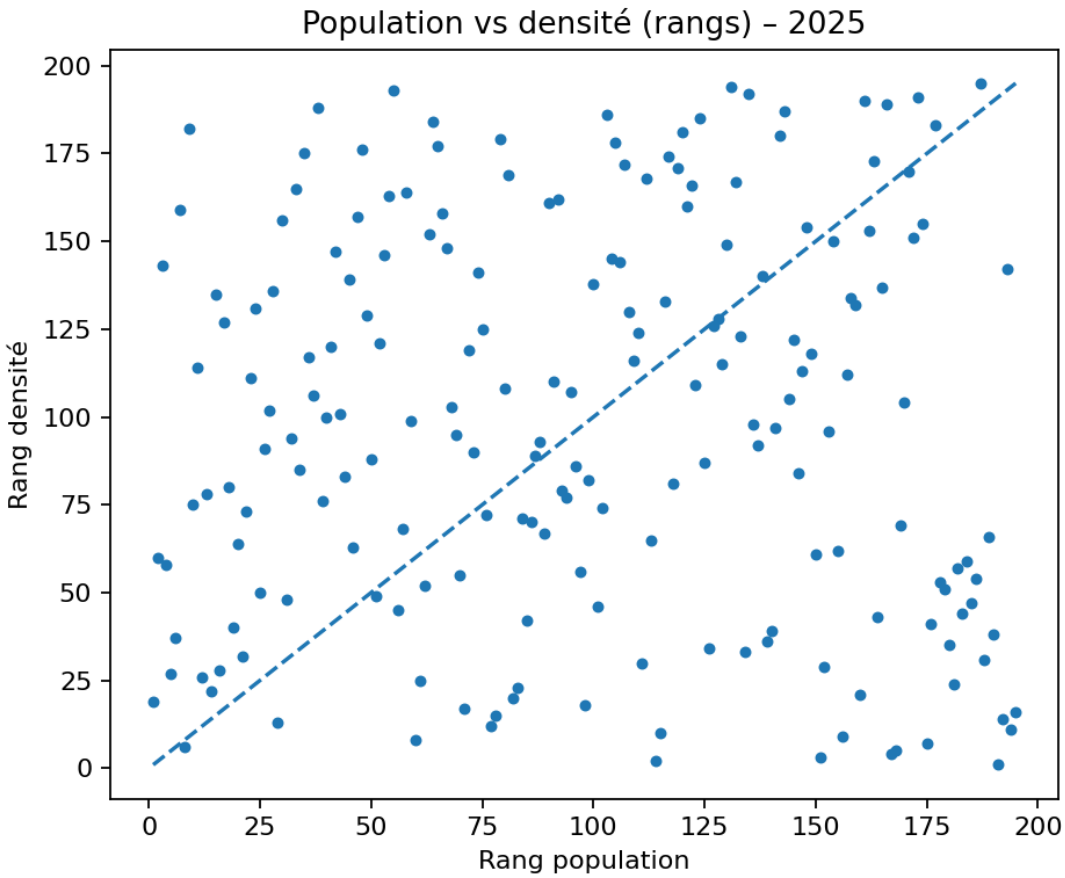

Figure 2 – Rangs population vs densité (2025)

2.5 Concordance des classements annuels (bonus)

Le script bonus factorise le calcul des coefficients (Spearman/Kendall) puis mesure, pour chaque année, la stabilité des classements par rapport à 2007. On compare donc le rang de chaque État en 2007 avec son rang dans une année donnée (sur l'ensemble des États communs).

Stabilité des classements de population (comparaison à 2007 – années sélectionnées)

Année(colonne)	Spearman rho	p Spearman	Kendall tau	p Kendall	N pays
Pop 2008	0.9996	7.34e-267	0.9888	4.11e-83	173
Pop 2012	0.9975	9.52e-199	0.9668	1.57e-79	173
Pop 2016	0.9945	1.50e-169	0.9448	4.97e-76	173
Pop 2020	0.9914	3.81e-153	0.9263	3.63e-73	173

Pop 2025	0.9864	4.03e-136	0.9054	5.64e-70	173

Stabilité des classements de densité (comparaison à 2007 – années sélectionnées)

Année(colonne)	Spearman rho	p Spearman	Kendall tau	p Kendall	N pays
Densité 2008	0.9992	9.41e-240	0.9821	5.18e-82	173
Densité 2012	0.9932	2.02e-161	0.9460	3.21e-76	173
Densité 2016	0.9859	1.31e-134	0.9175	8.29e-72	173
Densité 2020	0.9796	3.69e-121	0.8943	2.51e-68	173
Densité 2025	0.9678	2.16e-104	0.8605	2.19e-63	173

Graphiques – évolution des coefficients vs 2007

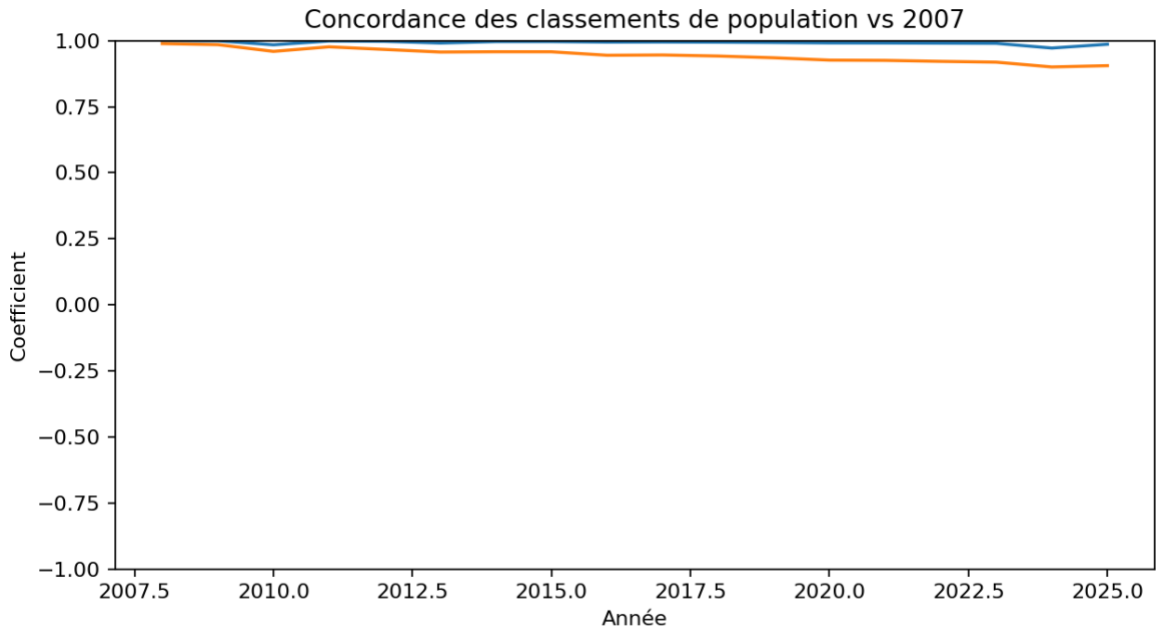

Figure 3 – Concordance population vs 2007 (2008–2025)

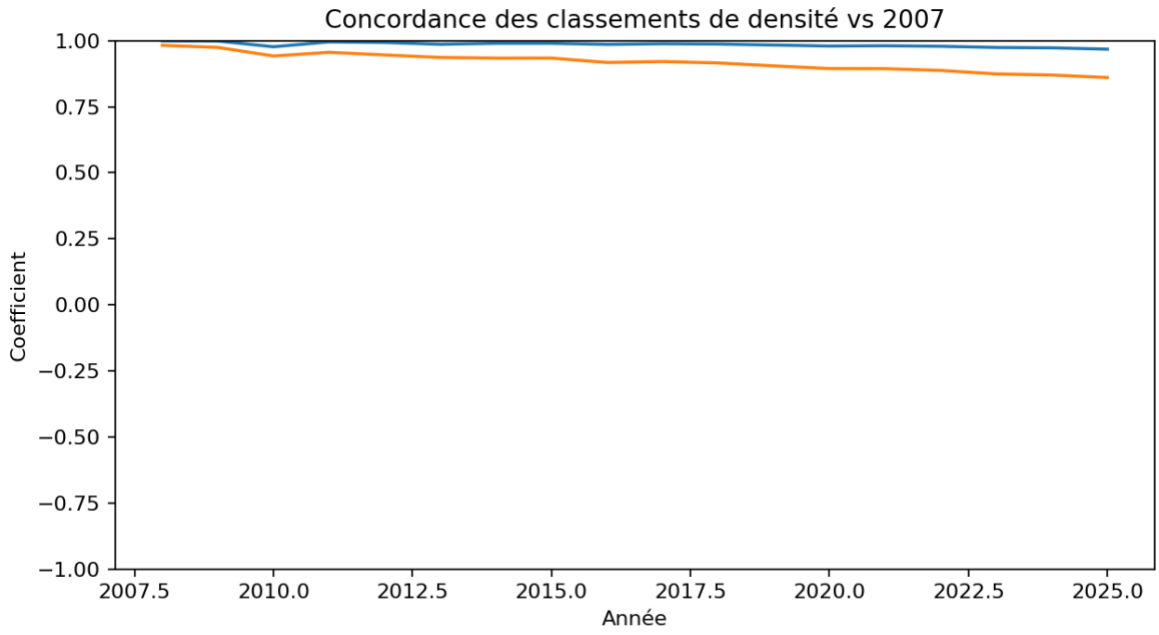

Figure 4 – Concordance densité vs 2007 (2008–2025)

Commentaire : les coefficients restent très élevés au fil du temps (proches de 1), ce qui indique que la hiérarchie mondiale de la population et celle de la densité sont globalement stables. Les variations observées proviennent surtout de changements relatifs entre pays proches en rang, plutôt que d'un renversement massif des hiérarchies.

3. Difficultés rencontrées et résolutions (FAQ)

• Typage : une difficulté fréquente est la présence de valeurs non numériques (ou de NaN). Le script force le typage via astype(float) et ignore les NaN lors de la construction des classements.
• Alignement des classements : comparer deux rangs nécessite d'aligner les mêmes États dans les deux listes ; classementPays() résout ce point en construisant des dictionnaires de rangs puis en ne conservant que l'intersection.
• Lisibilité des distributions dissymétriques : la loi rang-taille est souvent illisible en linéaire ; la transformation log-log (conversionLog) rend lisible la queue de distribution.

4. Réflexion personnelle – sciences des données et humanités numériques

Les exercices de cette séance montrent que des méthodes simples (tri, rangs, corrélations de rangs) suffisent à produire des analyses interprétables pour des objets géographiques (États, îles) sans supposer une relation linéaire entre variables. En humanités numériques, la logique est comparable : on classe des objets (textes, auteurs, thèmes, lieux), puis on compare des classements (au fil du temps, entre corpus, entre indicateurs). L'intérêt est double : (1) rendre visible une hiérarchie et (2) tester sa stabilité et ses déformations. Enfin, la factorisation du code (fonctions locales) est une compétence transversale : elle facilite la reproductibilité, dimension essentielle des sciences des données comme des humanités numériques.

Annexe mathématique (bonus) – équations comprises

• Transformation logarithmique : $y = \log(x)$, utilisée pour linéariser certaines relations rang-taille.
• Spearman : corrélation de Pearson appliquée aux rangs ($rho = corr(rank(X), rank(Y))$).
• Kendall : $tau = (C - D) / (C + D)$ (avec ajustements possibles), où C est le nombre de paires concordantes et D le nombre de paires discordantes.