Séance 3
Les paramètres statistiques élémentaires

Honoka OYAMADA

Introduction

La séance 3 du cours d'analyse de données en géographie est consacrée aux paramètres statistiques élémentaires. Elle a pour objectif d'approfondir l'analyse descriptive des données en mobilisant les notions de paramètres de position, de dispersion et de concentration, ainsi que leur représentation graphique à l'aide des boîtes à moustaches.

La séance s'appuie sur des données réelles issues des résultats du premier tour de l'élection présidentielle française de 2022, agrégées par département, et met en œuvre le langage Python ainsi que les bibliothèques Pandas, NumPy et Matplotlib.

1. Questions de cours

1.1 Caractères qualitatifs et quantitatifs

Les caractères quantitatifs sont plus généraux que les caractères qualitatifs, car ils permettent l'application d'un plus grand nombre de traitements statistiques. Les caractères qualitatifs décrivent des catégories, tandis que les caractères quantitatifs mesurent des grandeurs et autorisent le calcul de paramètres de position, de dispersion et de forme.

1.2 Caractères quantitatifs discrets et continus

Les caractères quantitatifs discrets prennent des valeurs distinctes issues d'un comptage, comme le nombre de votants. Les caractères quantitatifs continus peuvent prendre une infinité de valeurs dans un intervalle donné, comme une surface. Cette distinction est essentielle pour le choix des méthodes statistiques et des représentations graphiques.

1.3 Paramètres de position

Il existe plusieurs types de moyenne car chacune possède des propriétés spécifiques et répond à des objectifs différents. La médiane est utilisée pour décrire une valeur centrale robuste aux valeurs extrêmes. Le mode peut être calculé lorsque certaines valeurs apparaissent plus fréquemment que d'autres, notamment pour des variables discrètes.

1.4 Paramètres de concentration

La médiale et l'indice de concentration de C. Gini permettent de mesurer le degré d'inégalité dans la répartition d'une variable au sein d'une population. Ces indicateurs sont utiles pour analyser des distributions fortement concentrées ou déséquilibrées.

1.5 Paramètres de dispersion

La variance est calculée afin d'éviter les compensations entre écarts positifs et négatifs. L'écart type, racine carrée de la variance, est plus facilement interprétable car il s'exprime dans la même unité que la variable étudiée. L'étendue mesure l'amplitude globale d'une distribution. Les quantiles, en particulier les quartiles et les déciles, permettent de décrire la dispersion

interne des données. Les boîtes à moustaches synthétisent ces informations de manière graphique.

1.6 Paramètres de forme

Les moments centrés décrivent la dispersion et la forme autour de la moyenne, tandis que les moments absolus mesurent les écarts indépendamment du signe. L'analyse de la symétrie d'une distribution permet d'identifier d'éventuelles asymétries et d'adapter l'interprétation statistique.

2. Mise en œuvre avec Python

2.1 Objectifs de l'exercice

L'exercice pratique vise à calculer les principaux paramètres statistiques descriptifs à l'aide de la bibliothèque Pandas, à mesurer la dispersion des variables quantitatives et à représenter graphiquement les distributions par des boîtes à moustaches.

2.2 Données et préparation

Le fichier CSV resultats-elections-presidentielles-2022-1er-tour.csv, contenant les résultats du premier tour de l'élection présidentielle française de 2022 par département, est chargé à l'aide de la méthode read_csv de la bibliothèque Pandas.

Les colonnes correspondant à des caractères quantitatifs sont ensuite sélectionnées afin d'exclure les variables qualitatives du calcul des paramètres statistiques.

2.3 Calcul des paramètres statistiques

Pour chaque colonne quantitative, les paramètres suivants sont calculés : moyenne, médiane, mode, écart type, écart absolu à la moyenne et étendue. Les valeurs sont arrondies à deux décimales afin d'assurer une présentation homogène des résultats.

Les statistiques obtenues sont regroupées dans un tableau récapitulatif et exportées aux formats CSV et Excel, ce qui garantit la traçabilité et la réutilisation des résultats.

2.4 Distances interquartile et interdécile

Les distances interquartile et interdécile sont calculées à l'aide des quantiles. Ces indicateurs permettent de mesurer la dispersion centrale des distributions en limitant l'influence des valeurs extrêmes.

2.5 Boîtes à moustaches

Des boîtes à moustaches sont produites pour chaque variable quantitative à l'aide de Matplotlib. Ces graphiques représentent la médiane, les quartiles, l'étendue et les éventuelles valeurs atypiques. Ils permettent de comparer rapidement la structure et la dispersion des distributions statistiques.

3. Catégorisation des surfaces d'îles

Dans le cadre des questions 9 et 10, le fichier island-index.csv est utilisé lorsque celui-ci est disponible. La colonne correspondant à la surface des îles (« Surface (km²) ») est sélectionnée et convertie en variable quantitative exploitable.

Cette variable continue est ensuite discrétisée en classes de surface définies par des intervalles, conformément aux bornes imposées dans l'énoncé. Cette catégorisation permet de transformer une variable quantitative continue en variable quantitative discrète et de dénombrer le nombre d'îles par classe de surface.

4. Difficultés rencontrées et apprentissages

Les principales difficultés ont concerné l'identification correcte des colonnes quantitatives et le calcul des paramètres statistiques sur des variables hétérogènes. Ces difficultés ont été résolues par un filtrage des variables selon leur type et par l'utilisation des méthodes adaptées de Pandas.

La séance a permis de mieux comprendre l'intérêt des paramètres de dispersion et l'utilité des boîtes à moustaches pour comparer des distributions statistiques.

5. Réflexion personnelle

Cette séance met en évidence le rôle central de l'analyse descriptive dans le traitement des données géographiques. Les paramètres statistiques élémentaires constituent une étape préalable indispensable avant toute analyse explicative ou spatiale plus avancée.

L'utilisation d'outils numériques comme Python renforce la rigueur et la reproductibilité des analyses, tout en nécessitant une attention particulière à la nature des variables et aux choix méthodologiques effectués.

Conclusion

La séance 3 a permis d'approfondir les bases de la statistique descriptive appliquée à la géographie. La combinaison des calculs statistiques et des représentations graphiques constitue un socle méthodologique essentiel pour les analyses quantitatives et spatiales qui seront développées dans les séances suivantes.