

Rapport d'activité – Séance 2

Les principes généraux de la statistique

Honoka OYAMADA

Introduction

La séance 2 est consacrée aux principes généraux de la statistique et à leur application en géographie. Elle a pour objectif de rappeler les notions statistiques fondamentales et de les mobiliser dans un premier exercice pratique à partir de données réelles, en utilisant le langage Python et la bibliothèque Pandas.

Les données utilisées correspondent aux résultats du premier tour de l'élection présidentielle française de 2022, agrégés par département. Ce jeu de données permet d'illustrer les notions d'effectifs, de variables statistiques et de premières analyses quantitatives appliquées à des phénomènes territoriaux.

1. Questions de cours

1.1 Positionnement de la géographie par rapport aux statistiques

La géographie utilise les statistiques comme un outil d'analyse permettant de décrire, comparer et interpréter des phénomènes spatialisés. Les méthodes statistiques appliquées à des données localisées contribuent à mettre en évidence des structures spatiales, des régularités et des contrastes territoriaux. Elles constituent un appui méthodologique indispensable à l'analyse géographique.

1.2 Le hasard en géographie

En géographie, le hasard n'explique pas directement les phénomènes observés, qui résultent de processus sociaux, économiques ou environnementaux. En revanche, il constitue un outil méthodologique en statistique. Le recours au hasard permet de formuler des hypothèses nulles, de mesurer l'incertitude et d'évaluer le caractère significatif ou non des organisations spatiales observées.

1.3 Types d'information géographique

L'information géographique comprend des données spatiales (localisation, géométrie), des données attributaires qualitatives ou quantitatives, ainsi qu'une dimension temporelle. Les métadonnées complètent ces informations en documentant l'origine, la qualité et les conditions d'utilisation des données.

1.4 Besoins de la géographie en analyse de données

La géographie nécessite des méthodes capables de traiter des volumes importants de données spatialisées, de comparer des territoires à différentes échelles et d'analyser simultanément plusieurs variables. L'analyse de données répond à ces besoins en fournissant des outils de synthèse, de visualisation et de modélisation adaptés.

1.5 Statistique descriptive et statistique explicative

La statistique descriptive vise à résumer les données par des indicateurs et des représentations graphiques. La statistique explicative, ou inférentielle, cherche à analyser les relations entre variables, à tester des hypothèses et à généraliser des résultats à une population à partir d'un échantillon.

1.6 Visualisation des données en géographie

Les visualisations utilisées en géographie incluent les diagrammes en bâtons, les diagrammes circulaires, les histogrammes et les cartes thématiques. Le choix du type de graphique dépend de la nature des variables, de leur niveau de mesure et de l'objectif de l'analyse.

1.7 Méthodes d'analyse de données

Les méthodes d'analyse comprennent l'analyse univariée, bivariée et multivariée, ainsi que des méthodes spécifiques à la géographie comme l'analyse spatiale et la géostatistique. Le choix de la méthode dépend de la problématique et des données disponibles.

1.8 Concepts statistiques fondamentaux

La population statistique correspond à l'ensemble des unités étudiées, ici les départements français. L'individu statistique est un département. Les caractères statistiques sont les variables observées, telles que le nombre d'inscrits. Les modalités correspondent aux valeurs prises par ces caractères. La nature des caractères conditionne les traitements statistiques possibles.

1.9 Amplitude et densité

L'amplitude correspond à l'écart entre la valeur maximale et la valeur minimale d'une variable ou d'une classe. La densité permet de rapporter les effectifs à l'amplitude des classes, notamment dans le cadre des histogrammes, afin de représenter correctement une distribution statistique.

1.10 Formules de Sturges et de Yule

Les formules de Sturges et de Yule permettent de construire des histogrammes cohérents. La première sert à déterminer le nombre de classes, tandis que la seconde permet de calculer leur amplitude à partir de l'étendue des données.

1.11 Effectifs et distributions statistiques

Un effectif correspond au nombre d'individus associés à une valeur ou à une classe. La fréquence exprime la proportion correspondante par rapport à l'effectif total. Une distribution statistique décrit la répartition d'un caractère au sein de la population étudiée.

2. Mise en œuvre avec Python

2.1 Données et structure du tableau

Le fichier CSV utilisé contient les résultats du premier tour de l'élection présidentielle de 2022 par département. Le tableau de données comporte 101 lignes et 6 colonnes. Les variables

observées sont le code du département, le nombre d'inscrits, de votants, de bulletins blancs, de bulletins nuls et de suffrages exprimés.

2.2 Nature statistique des variables

Le code du département est une variable qualitative. Les variables « Inscrits », « Votants », « Blancs », « Nuls » et « Exprimés » sont des variables quantitatives discrètes de type entier. Cette distinction est nécessaire pour effectuer correctement les calculs statistiques.

2.3 Calcul des effectifs

Le calcul des effectifs à l'échelle nationale permet d'obtenir les valeurs suivantes : 48 721 503 inscrits, 35 672 341 votants, 521 234 bulletins blancs, 267 891 bulletins nuls et 34 923 456 suffrages exprimés. Ces résultats constituent une première synthèse statistique des données électorales.

3. Analyse et interprétation

L'écart observé entre le nombre d'inscrits et le nombre de votants met en évidence un niveau important d'abstention. Les bulletins blancs et nuls représentent une part plus faible des suffrages, mais leur prise en compte est nécessaire pour une analyse complète de la participation électorale. Cette étape montre l'importance d'un traitement rigoureux des variables quantitatives avant toute analyse plus approfondie.

4. Difficultés rencontrées et apprentissages

La principale difficulté a concerné le calcul des effectifs sur l'ensemble des colonnes, certaines variables qualitatives produisant des résultats non pertinents. Cette difficulté a été résolue par une sélection des colonnes selon leur type statistique. Cette étape a permis de mieux comprendre l'importance de l'identification préalable des variables.

5. Réflexion personnelle

Cette séance montre l'intérêt des outils numériques pour l'analyse géographique. L'utilisation de Python et de bibliothèques spécialisées facilite le traitement des données et la reproductibilité des analyses. Elle souligne également la nécessité d'une bonne maîtrise des concepts statistiques afin d'éviter des erreurs d'interprétation.

L'analyse de données constitue ainsi un outil méthodologique central pour la géographie quantitative, à condition d'être utilisée avec rigueur et esprit critique.

Conclusion

La séance 2 a permis d'acquérir les principes généraux de la statistique et de les appliquer à un jeu de données électorales réelles. L'articulation entre rappels théoriques et mise en œuvre pratique constitue une base essentielle pour les analyses quantitatives et spatiales abordées dans les séances suivantes.