

Séance 2 questions - Analyse de données

1. Quel est le positionnement de la géographie par rapport aux statistiques ?

La géographie entretient une relation complexe avec les statistiques :

Elle produit des données massives que seul l'outil statistique permet d'étudier .

Historiquement, elle a souvent sous-estimé leur utilité, mais aucun géographe ne peut aujourd'hui s'en passer.

Les statistiques permettent la réduction de l'incertitude et l'analyse structurée des phénomènes géographiques.

Les statistiques sont donc un outil indispensable pour faire de la géographie une véritable science basée sur l'analyse rigoureuse.

2. Le hasard existe-t-il en géographie ?

Deux visions coexistent :

Le déterminisme (Laplace) : le hasard n'existe pas, tout a une cause.

Le hasard comme cause cachée, explorable grâce à la progression des connaissances.

La géographie utilise ces deux approches :

elle cherche des tendances globales, même si les détails individuels ne sont pas prévisibles. Le hasard existe sous forme de variabilité, que les statistiques permettent de modéliser (loi normale, loi de Pareto, etc.) .

En géographie, le hasard existe localement, mais des tendances globales sont identifiables.

3. Quels sont les types d'information géographique .

Il y a deux grands types d'information géographique : d'une part les données attributaires (caractéristiques des territoires) : la population, les variables sociales, économiques, climatiques... et d'autre part les données géométriques : les formes, contours, surfaces, réseaux et morphologies spatiales. Ces deux éléments sont aussi constitutifs de la base d'un SIG

4. Quels sont les besoins de la géographie au niveau de l'analyse de données?

La géographie nécessite la production et la collecte de données (nomenclature, métadonnées) mais aussi l'étude de la structure interne des données (matrice individus, variables) et l'usage d'outils statistiques pour résumer, modéliser, comparer et visualiser les phénomènes complexes. Le but est d'extraire des connaissances, de détecter des structures et de confronter théorie et réalité.

5. Quelles sont les différences entre la statistique descriptive et la statistique explicative?

La statistique descriptive, décrit et résume les données en utilisant les moyennes, quantiles, histogrammes, ACP, AFC, ACM... Son objectif est d'ordonner des données et de préparer des comparaisons et des prévisions .

La statistique explicative cherche quant à elle à comprendre une variable Y à partir de variables explicatives X1... Elle inclut : régression, analyse discriminante, régression logistique, ANOVA... et son objectif est d'expliquer ou prédire

6. Quelles sont les types de visualisation de données en géographie? Comment choisir celles-ci?

Les diagrammes sectoriels sont le type de visualisation adapté pour les variables qualitatives. Pour les variables quantitatives on peut utiliser des historiographes, despotés à moustaches, des courbes cumulatives ou encore des polygones de fréquences etc. Enfin pour des données multivariées : des cartes de proximités, des cartes thématiques (dans contexte SIG, évoqué indirectement), ACP, AFC, ACM

Le choix se fait donc selon le type de variable, la continuité ou non, la nécessité de comparer plusieurs variables ainsi que la volonté de réduire la dimension

7. Quelles sont les méthodes d'analyse de données possibles?

On retrouve trois grandes familles :

Les méthodes descriptives :

- ACP, AFC, ACM
- Classification (CAH, nuées dynamiques)
- Analyses de proximité .

Les méthodes explicatives :

- Régression simple/multiple
- Régression logistique
- Analyse discriminante
- Analyse de la variance
- Modèles linéaires généraux .

les méthodes de prévision :

- Analyse des séries chronologiques
- Modèles : $(X_t = f(X_{t-1}, X_{t-2}, \dots) + \text{aléa})$

8. Comment définiriez-vous : (a) population statistique? (b) individu statistique ? (c) caractères statistiques? (d) modalités statistiques? Quels sont les types de caractères? Existe-t-il une hiérarchie entre eux?

a. Population statistique
Ensemble des unités étudiées

Ex : toutes les villes d'une région

b. Individu statistique

Un élément de la population (unité spatiale)

Ex : une commune, un ménage

c. Caractères statistiques

Caractéristique mesurée sur chaque individu (âge, superficie, revenu...).

d. Modalités statistiques

Valeurs possibles du caractère, exclusives et exhaustives (ex : homme/femme, 0-77ans etc).

Types de caractères

Qualitatifs nominal / ordinal

Quantitatifs discret / continu

Existe-t-il une hiérarchie entre eux?

Oui, les variables quantitatives permettent plus d'opérations (tests paramétriques, distribution), les qualitatives moins

9. Comment mesurer une amplitude et une densité?

Amplitude d'une classe : ($A = b - a$)

Densité d'une classe : ($d = \frac{n_i}{b - a}$)

→ effectif de la classe divisé par son amplitude .

10. À quoi servent les formules de Sturges et de Yule?

Elles servent à déterminer le nombre optimal de classes lors de la discrétisation d'un caractère quantitatif.

Sturges :

($k \approx 1 + 3.222 \times \log_{10}(n)$)

Yule :

($k \approx 2.5 \sqrt[4]{n}$)

→ Elles évitent un découpage trop fin ou trop grossier.

11. Comment définir un effectif ? Comment calculer une fréquence et une fréquence cumulée?

Qu'est-ce qu'une distribution statistique ?

Effectif : nombre d'apparitions d'une modalité dans la population .

Fréquence (f_i) :

Calcul : ($f_i = \frac{n_i}{n}$)

Proportion de la modalité

Fréquence cumulée

Somme des fréquences des modalités $<$ une valeur donnée :

$$(F_k = \sum_{i=1}^k f_i).$$

La distribution statistique correspond à l'ensemble des fréquences observées pour les différentes modalités, elle sert à identifier la loi de probabilité sous-jacente

Comment s'est passée la séance 2 :

Difficile car c'est le début des séances, et longue.