RAPPORT SEMESTRIEL: ANALYSE DE DONNÉES STATISTIQUES EN GÉOGRAPHIE (Parcours débutant)

M. FORRIEZ

PERCEVAL Emma
Master 1 GAED
Parcours Sociétés, Cultures, Territoires
Année 2025 - 2026

* Certaines manipulations *python* sont appuyées de l'aide d'une intelligence artificielle.

SÉANCE 2 - Principes généraux de la statistique

---

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

Les statistiques ne sont généralement pas valorisées comme essentielles en géographie, car celles-ci sont vues comme appartenant au monde des sciences dures, à l'inverse de la géographie qui s'est de plus en plus affiliée aux sciences humaines. Cependant, la géographie produit beaucoup de données statistiques et nécessite donc de prendre en compte la discipline de la statistique comme une de ses prérogatives.

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

Comme toute sciences ou disciplines, y compris les sciences dures, la géographie appréhende le hasard différemment selon les courants, les époques, les échelles. Si l'existence semble pouvoir empêcher à la géographie de produire des données scientifiques et des lois sur la relation de l'homme à l'espace, certains géographes prennent en compte l'importance de la statistique pour faire émerger des modèles généraux, des probabilités, des tendances. Dire qu'il n'y a pas de hasard en géographie, c'est se rapprocher du danger de sombrer dans le déterminisme, mais dire qu'elle n'est faite que de hasard décrédibilise la géographie en tant que sciences.

3. Quels sont les types d'information géographique ?

Il y a deux série statistiques possibles en information géographique. Premièrement, les informations qui caractérisent l'ensemble délimité en géographie humaine ou physique. Deuxièmement, les informations qui permettent d'étudier la morphologie même des ensembles délimités.

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

L'analyse de données statistique en géographie permet de rendre comparable les objets en créant de l'information systématique, de traiter cette information en la mettant en forme (exemple avec la cartographie numérique), de connaitre la fiabilité de cette information et enfin d'appliquer cette information au réel (exemple en géographie avec l'aménagement du territoire). Cependant, l'analyse de données n'a pas de caractère explicatif des phénomènes.

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

La statistique descriptive se base sur une distribution théorique connue pour en faire ressortir des propriétés remarquables. La statistique descriptive essentialise les données pour faciliter leur lecture. Cette lecture et ce « tri » des données permet de schématiser leur organisation et facilite le travail de la statistique explicative.
En effet, la statistique mathématique prend pour base le travail de la statistique descriptive et y applique des lois de probabilités pour établir des *scenarii* possibles.

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

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

Il y a d'abord les méthodes descriptives qui permettent de visualiser et classer des données. Ces deux méthodes s'appliquent différemment sur les données quantitatives et qualitatives, et incluent des analyses factorielles et multi-factorielles. Elles permettent d'identifier et de hiérarchiser des données, ainsi que de mettre en valeur des relations entre elles (ce qui permet aussi de les classer). L'analyse factorielle de données mixtes, par exemple, permet de traiter plusieurs types de données en même temps (qualitatif et quantitatif).

Les méthodes explicatives, comme leur noms l'indique, permettent d'expliquer des variables Y par des variables explicatives tels que X1…Xk. Cette étape passe par une modélisation de ce qu'est Y selon X1…Xk.

Les méthodes de prévision se construisent en rapport avec une temporalité (Xt) et prennent en compte l'aléa, afin de prévoir la suite d'une série de variables.

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?

En statistiques, on parle d'une « population » pour définir l'ensemble des données, le tout. Quand on isole un élément de cette population statistique, on parle d'individu statistique (ou unité statistique). Ce dernier peut apparaitre différemment dans la table attributaire: Soit il est un et cartographiable à lui tout seul (les unités spatiales); soit il comprend d'autres informations intérieures qui lui sont propres.

Les caractères statistiques sont définis en fonction de leurs modalité. Le but propres des caractères est d'être singuliers à chacune des variables pour les distinguer les unes des autres ou faire apparaitre des similarités. Les caractères peuvent autant être de nature qualitatifs que quantitatifs.

L'appellation des modalités diffère selon si elle concerne des variables qualitatives ou quantitatives. On parlera de catégories pour les variables qualitatives et de valeurs pour les variables quantitatives qui peuvent elles être mesurées.

Il est possible de hiérarchiser certaines variables mais pas toutes. Les données qualitatives ordinales peuvent être soumises à un classement par exemple. La discrétisation, qui consiste à créer des classes parmi une population de données quantitatives, peut aussi apparaitre comme une méthode de hiérarchisation.

9. Comment mesurer une amplitude et une densité?

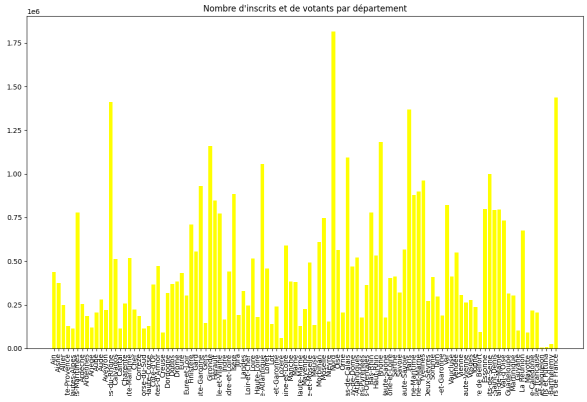
L'amplitude et la densité sont les deux paramètres pris en compte pour justifier la discrétisation, soit le regroupement de variables pour en faire des classes.

L'amplitude se calcule sur l'ensemble d'une classe formée, en soustrayant la valeur maximale de la classe à la valeur minimale.

La densité se calcule en connaissance de l'amplitude d'une classe, puisqu'il s'agit d'une division entre l'effectif (le nombre de variables) de la classe concernée et son amplitude.

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

Les formules de Sturges et de Yule permettent d'envisager le nombre de classes idéal sur un ensemble de variables. La première formule utilise les tables logarithmiques, la deuxième la racine carrée.
Grace à ces formules on peut créer des histogrammes. Exemple dans l'application de la séance 2.



11. Comment définir un effectif ? Comment calculer une fréquence et une fréquence cumulée? Qu'est-ce qu'une distribution statistique ?

L'effectif (fréquence absolue) noté « *n i* » tend à définir combien de fois une valeur notée « *x* » apparait dans une série de variables. La fréquence relative traite les probabilités sur une échelle plus large que l'effectif absolue, puisqu'elle compare « *n i* » à l'ensemble *« n »*.
A partir du calcul de la fréquence, la distribution statistique donne la probabilité d'apparition des caractères, et utilise une loi de probabilité pour cela.

SÉANCE 3 - Paramètres statistiques élémentaires

1. Quel caractère est le plus général : le caractère quantitatif ou le caractère qualitatif? Justifier pourquoi.

Les caractères qualitatifs sont plus généraux que les caractères quantitatifs car ils ne sont pas forcément doublés de modalités numériques. En effet, les caractères qualitatifs mesurent un état, une qualité alors que les caractères quantitatifs sont l'expression d'une quantité produite, précise, par une mesure ou un dénombrement.

2. Que sont les caractères quantitatifs discrets et caractères quantitatifs continus? Pourquoi les distinguer ?

Les valeurs discrètes sont par exemple le nombre d'habitants dans une commune, le nombre de réponses justes à un examen etc.. Elles sont le résultat d'un comptage et pas d'une mesure. Les

valeurs continues sont par contre le résultats d'une mesure et sont ininterrompues au sein d'une intervalle. C'est par exemple le poids d'une personne, le nombre de litres dans une citerne etc…

3. Paramètres de position
— Pourquoi existe-t-il plusieurs types de moyenne?

Il existe plusieurs types de moyennes tels que: la moyenne arithmétique, quadratique, harmonique, géométrique, mobile, fonctionnelle. Elles sont toutes adaptées à des situations où il est préférable d'obtenir certains résultats, et adaptées aux données que l'on rencontre.

Par exemple, la moyenne géométrique permet d'obtenir le milieu entre deux ordres de grandeurs, ce qui est utile avec les puissances, quand les ordres de grandeurs sont proches. Exemple: entre 1000 et 1 milliards il est plus naturel de dire que la moyenne est 10^6 (obtenu grâce à la moyenne géométrique) plutôt que de dire que c'est 1 million.

— Pourquoi calculer une médiane ?

La médiane identifie le milieu d'une distribution statistique et elle permet de rendre compte d'une régularité, qui n'apparait pas forcément en calculant une moyenne.

— Quand est-il possible de calculer un mode?

Le mode est une moyenne de fréquence qui indique la valeur qui revient le plus dans une série statistique. On peut le calculer plutôt avec des données discrètes et il peut y en avoir plusieurs à trouver.

4. Paramètres de concentration
— Quel est l'intérêt de la médiale et de l'indice de C. Gini?

La médiale est une sorte de médiane pondérée. En coupant en deux la distribution, on distingue deux groupes qui sont égaux en effectif et en importance (masse).

L'indice de Gini correspond à l'air entre deux courbes, c'est un calcul qui peut être obtenu grâce à une intégrale.

5. Paramètres de dispersion
— Pourquoi calculer une variance à la place de l'écart à la moyenne? Pourquoi la remplacer par l'écart type?

La variance fait partie des paramètres de dispersion. On calcule la variance en faisant la moyenne des carrés des écarts à la moyenne. La variance a des propriétés mathématiques plus intéressantes que l'écart à la moyenne.
Il faut d'abord calculer la variance pour avoir l'écart-type (c'est la racine de la variance), afin de savoir à quel point une variable de la distribution est loin de la moyenne.

— Pourquoi calculer l'étendue?

L'étendue est un calcul facile pour observer la dispersion des données de la plus petite à la plus grande. Elle simplifie la lecture de la distribution.

— À quoi sert-il de créer un quantile? Quel(s) est (sont) le(s) quantile(s) le(s) plus utilisé(s)?
— Pourquoi construire une boîte de dispersion ? Comment l'interpréter?

Les quantiles et les boites de dispersion permettre de lire une distribution de données de manières séquencée. Cela permet par exemple de créer des classes (par amplitude/ par effectifs notamment) et de s'en servir pour cartographier un phénomène. Par exemple, sur des données qui représentent les notes de 100 élèves d'une promotion, on peut créer les classes par rapport aux quantiles, ce sera alors des classes par rapport à l'effectif et non pas par rapport à l'amplitude (dans les 25 premiers élèves les notes pourraient aller de 1 à 9, ce qui est très vaste; et dans la deuxième classe de 9 à 12 pour les 25 autres élèves par exemple).

6. Paramètres de forme
— Quelle différence faites-vous entre les moments centrés et les moments absolus ? Pourquoi les utiliser?
— Pourquoi vérifier la symétrie d'une distribution et comment faire ?

SÉANCE 4 - Les distributions statistiques

1. Quels critères mettriez-vous en avant pour choisir entre une distribution statistique avec des variables discrètes et une distribution statistique avec des variables continues?

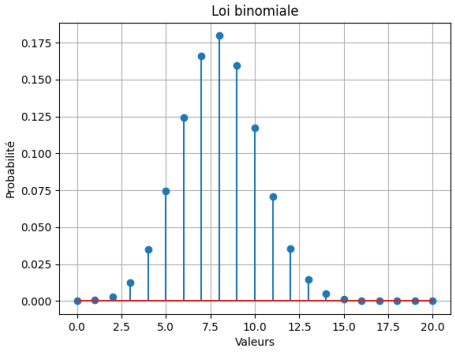
La nature du phénomène étudié, la forme de distribution empirique ainsi que la connaissance et l'interprétation de l'ensemble des données et nombre de paramètres des lois, peuvent etre des facteurs de choix.

2. Expliquez selon vous quelles sont les lois les plus utilisées en géographie?

La loi Bernouilli: elle permettait par exemple pour un sujet de géographie sur l'accessibilité des transports en commun aux personnes à mobilité réduite de modéliser la probabilité qu'une bouche de métro dans Paris soit équipée d'un ascenseur PMR ou non.

La loi Binomiale: Celle-ci est différente puisqu'elle analyse plutôt les situations ou la variable est dut au hasard et assez rare. Par exemple, en géographie on pourrait l'utiliser pour déterminer combien de fois le RER est en retard au cours de la semaine, on regarderait le résultat sur toute la semaine.

Sur cette image générée par la séance 4, on voit une illustration de la loi binomiale et on remarque en effet qu'elle n'est pas constante.



SÉANCE 5 - Les statistiques inférencielles

1. Comment définir l'échantillonnage. Pourquoi ne pas utiliser la population en entier? Quelles sont les méthodes d'échantillonnage? Comment les choisir ?

2. Comment définir un estimateur et une estimation?

3. Comment distingueriez-vous l'intervalle de fluctuation et l'intervalle de confiance ?

4. Qu'est-ce qu'un biais dans la théorie de l'estimation?

5. Comment appelle-t-on une statistique travaillant sur la population totale? Faites le lien avec la notion de données massives [1] ?

6. Quels sont les enjeux autour du choix d'un estimateur ?

7. Quelles sont les méthodes d'estimation d'un paramètre? Comment en sélectionner une ?

8. Quels sont les tests statistiques existants ? À quoi servent-ils ? Comment créer un test ?

9. Que pensez-vous des critiques de la statistique inférentielle ?

SÉANCE 6 - Statistique d'ordre des variables qualitatives

1. Qu'est-ce qu'une statistique ordinale? À quel autre statistique catégorielle s'oppose-t-elle? Quel type de variables utilise-t-elle ? En quoi cela peut matérialiser une hiérarchie spatiale?

La statistique ordinale est une branche de la statistique qui permet d'ordonner les données de nature qualitatives, en faisant notamment ressortir les plus grandes valeurs (l'ordre croissant est donc préférable pour l'ordination ici), les plus remarquables par rapport à une distribution etc. Ainsi, ces ordinations peuvent donner suite à des interprétations, notamment sur les hiérarchies qui prennent place dans l'espace. En effet, grâce aux ordinations de données, on peut réaliser des cartes où apparaissent les différences d'échelles entre une villes et une métropole par exemple (particulièrement en géographie humaine), sans pour autant que leur relation métrique soit clairement explicitée. La statistique de moment s'oppose à la statistique d'ordre car elle s'intéresse plus aux valeurs numériques des données qu'à leur classement.

2. Quel ordre est à privilégier dans les classifications?

La logique de classification des données doit rester propre à chaque phénomène étudié mais de manière générale c'est l'ordre croissant qui doit être privilégié dans l'ordination car il permet de mettre en valeur les positions dominantes.

3. Quelle est la différence entre une corrélation des rangs et une concordance de classements?

Les deux systèmes permettent de comparer les classements qui sont en partie le fruit de l'arbitraire, mais ils le font différemment.

La corrélation de rangs est plus générale car elle mesure un degré de relation entre les rangs de deux classements différents. Par exemple: si on prend 5 pays et qu'on les classe dans un tableau par densité et dans l'autre par population (dans un ordre décroissant), avec la logique de la corrélation des rangs on regarde ligne par ligne si les pays garde globalement le même rang d'un tableau à l'autre. Si oui dans ce cas-là on peut dire « Il y a une forte corrélation entre la forte population d'un pays et sa forte densité ».

La concordance de classements est plus précise car elle vise à observer un accord structurel entre deux classements, notamment en établissant des comparaisons de structure par pairs. Pour le même exemple, ici on va prendre les deux tableaux et essayer de distinguer des pays qui suivent soit la même trajectoire (ne bougent pas, baissent d'un rang etc) ou alors on observe des inversions etc . Le résultat attendu est de distinguer des pairs concordantes et discordantes.

4. Quelle est la différence entre les tests de Spearman et de Kendal?

Le test de Spearman suit la logique de corrélation de rangs alors que le test de Kendal suit la logique de la concordance de classements pour Kendal.
Le test de Superman applique la uniforme discrète et permet d'établir deux hypothèses.

5. À quoi servent les coefficients de Goodman-Kursdal et de Yule?

Ces deux coefficients permettent de mesurer l'association entre variables qualitatives.

Le coefficient de Goodman-Kursdal permet la réduction de l'erreur de prédiction d'une variable à l'autre. Ainsi, on peut répondre, à partir de données, à une question telle que: « avoir des informations sur le taux de signalisation d'un espace permet-il de prédire le taux d'accidents? ». Le coefficient a une capacité prédictive.

Le coefficient de Yule permet lui de mesurer l'association entre deux variables dichotomiques à partir d'un tableau de contingence (met en relation deux variables qualitatives). Il doit y avoir deux variables pour chaque proposition par exemple: « taux de signalisation » / « taux accidents ». Pour ce cas-là, le coefficient distingue les pairs discordantes et concordantes soit là ou la valeur est la plus grande et là ou les valeurs sont les plus petites. Ce travail permet d'interpréter et de dire « corrélation entre faible taux de signalisation et taux élevé d'accidents. Le coefficient met en valeur la force de l'association.

LES HUMANITÉS NUMÉRIQUES

Les humanités numériques font la liaisons entre les matières dites littéraires et les matière dites de sciences dures. Dans le cas de la géographie, elles permettent de donner à la discipline une pleine légitimité et d'appuyer les analyse de données traitées et utilisables.