Thomas
Dupuy
GéoSuds

19/12/2025

Analyse de données - M. Forriez

Analyse de données - 1e Semestre

Parcours débutant

Séance 2 à 6

Thomas Dupuy

Séance 2 : Principe généraux de la statistique

Question 1 : La géographie est une discipline qui a énormément évolué depuis qu'elle existe et qu'elle est considérée comme scientifique. Les mathématiques, et plus particulièrement les statistiques, constituent un outil d'une très grande importance pour le géographe qui peut l'utiliser pour traiter une quantité importante de données. Une étude géographique d'un phénomène ou d'un espace dégage une masse de données que souvent seules les statistiques peuvent simplifier, expliciter, et transformer en une matière scientifique exploitable pour celui qui l'étudie. Néanmoins, il en reste que les mathématiques ne sont pas encore communément acceptées par tous les géographes, beaucoup restent réticents à son utilisation (réticence que l'on peut imputer à la formation de ces géographes souvent plus tournée vers les lettres que les mathématiques), et son caractère nouveau dans la discipline, qui entraîne de fait une vague de conservatisme.

Question 2 : Effectivement, il existe un hasard en géographie, comme il existe un hasard dans presque toutes les disciplines du monde. Question de positionnement philosophique dans un second temps, il semble que les scientifiques s'accordent sur le fait que le hasard existe et qu'il peut plus ou moins être limité, borné, pour établir une certitude globale. En géographie, possibilisme vidalien et déterminisme s'affrontent sur le positionnement global de la discipline quant au hasard, et il est admis aujourd'hui que c'est celle de Vidal de la Blache qui l'a remporté. Toutefois, malgré tout cela, la géographie française reste assez ambiguë dans la mesure où elle n'a pas pleinement introduit les mathématiques dans sa formation,formation, ce qui la différencie des autres nationalités, et surtout bride sa vision complète du hasard en géographie.

Question 3 : Il existe deux séries de statistiques qui servent à étudier soit la forme géométrique des attributs d'un espace, ou bien les attributs en eux-mêmes, leur présence sur cet espace et leurs caractéristiques. Pour traiter ces données, le géographe peut mettre en place une nomenclature de celles-ci, afin d'y gagner en lisibilité et les higéographede lesérarchiser (ou non). Les métadonnées qui en ressortent agissent comme vérificateur de sources, c'est-à-dire qu'elles regroupent en leur en leur sein d'autres données et nomenclatures, permettant ainsi de dézoomer la masse de données à une échelle bien plus compréhensible pour le cerveau humain.

Question 4 : Cfcf.Réponses Questions 1, 2 et 3 cf.3.

Question 5 : Les statistiques descriptives ont pour but de ranger, de classer, la quantité massive de données afin de décrire une situation, la réalité d'un espace, la distribution des attributs au sein d'un territoire. Elle représentera donc un phénomène ou une situation sous forme de graphiques ou d'indicateurs sans forcément les mettre en relation avec une variable. Les statistiques explicatives, quant à elle , rendent compte d'un phénomène en fonction d'un réel, d'une variable. Ici, on cherche à comprendre les liens d'inter-dépendance entre une variable dite dépendante et une autre dite indépendante. Ces statistiques sont beaucoup plus utilisée interdépendance utilisées pour rentrer en profondeur dans l'interrogation et l'explication d'un phénomène géographique inutilisées en fonction des caractéristiques du lieu d'occurence.

Question 6 : En géographie, les types de visualisation dépendent de la nature des données étudiées. Les variables quantitatives continues (comme l'altitude, le revenu ou le nombre d'habitants) sont souvent représentées par des histogrammes, ou des courbes cumulatives. Les variables qualitatives (comme le type d'usage du sol ou le parti politique en tête) sont mieux représentées par des diagrammes en barres ou des camemberts.

En cartographie, les données peuvent aussi être visualisées à travers des symboles proportionnels ou des cartes de flux. Le choix de la visualisation dépend donc à la fois du type de variable (qualitative ou quantitative), de la question posée (comparer, répartir, localiser) et de l'échelle géographique d'analyse.

Question 7 : On distingue plusieurs grandes familles de méthodes. Les méthodes descriptives regroupent les techniques de réduction et de visualisation de données comme l'analyse en composantes principales (ACP) pour les variables quantitatives, ou l'analyse factorielle des correspondances (AFC) pour les variables qualitatives. Les méthodes explicatives s'appuient sur des modèles statistiques permettant de relier une variable à expliquer à d'autres variables explicatives, par exemple à travers une régression linéaire, une analyse de variance ou une régression logistique. Enfin, les méthodes de prévision concernent l'étude de séries chronologiques, où l'on cherche à prédire une valeur future à partir des données passées.

Question 8 : (a) Une population statistique peut être définie en occurrence définie comme un ensemble d'éléments ou d'attributs qui partage au moins les mêmes caractéristiques, à entendre comme un ensemble mathématique.

(b) Aussi appelé unité statistique, l'individu statistique correspond à un seul élément de la population statistique.

(partagent(c ()) On appelle caractère statistique toutes les singularités propres à chacune)chacune des unités et sur lesquelles peuvent se reposer l'étude statistique.

(d) La modalité est une condition que l'on pose pour trier les données selon si oui ou non elles présentent un (ou des) caractères spécifiques.

Il existe deux types de caractères : les variables qualitatives ou quantitatives. Toutefois, il n'existe pas de hiérarchie claire entre elles..

Question 9 : L'amplitude en statistique se calcule en soustrayant à la valeur la plus forte de la population la valeur la plus basse. Ainsi, sur la population d'un club de foot, on prendra l'âge du licencié le plus vieux et du plus jeune, leur soustraction nous indiquera l'amplitude d'âge dans le club. La densité, c'est l'effectif divisé par l'amplitude, elle permet de calculer des classes d'amplitude.

Question 10 : Les formules de Sturges et de Yule servent à déterminer le nombre de classes optimal pour construire un histogramme et gagner le plus possible en lisibilité. Elles permettent d'éviter un découpage trop fin (qui rend le graphique illisible) ou trop grossier (qui fait perdre de l'information). Elle se base sur le nombre de classes et l'amplitude des données recueillies.

Question 11 : L'effectif se définit comme le nombre d'occurrence d'une variable dans l'ensemble d'une classe. Dans la classe régions françaises avec plus de 5 millions d'habitants, s'il y a 7 régions avec plus de 5 5 M d'habitants, alors la fréquence absolue est de 7.

Elle se calcule de la sorte : fi=ni

La fréquence cumulée additionne les fréquences jusqu'à une certaine modalité ou classe, ce qui permet de connaître la part d'observations inférieures ou égales à une valeur donnée. Elle se calcule de la sorte : fi = $\sum i=1k$ ni ≤ k.

La fréquence permet d'établir une distribution statistique, qui elle-même se définit comme un tableau qui répertorie les classes de valeurs obtenues.

Séance 3 : statistiques univariées (1). Paramètres statistiques élémentaires

Questions de cour :

Question 1 : Les caractères dits qualitatifs sont plus généraux que les caractères dits quantitatifs. Les variables qualitatives recensent une qualité, une propriété d'une population, comme par exemple l'appartenance politique. Les variables quantitatives mesurent une quantité, donc quelque chose que l'on peut mesurer, par exemple, la taille (en cm). Toutefois, avec ce même exemple, on peut la transformer en variable qualitative si l'on divise les catégories en "petit", "moyen" et "grand". En ça, les caractères qualitatifs sont plus généraux.

Question 2 : Les variables quantitatives discrètes ne peuvent prendre qu'une certaine valeur, souvent entière, comprise dans un intervalle connu. Par exemple, si l'on considère un groupe d'individus où la personne avec le plus d'enfants au sein de celui-ci aurait 4 enfants, le probable nombre d'enfants des autres individus se situera forcément dans l'intervalle [0 ; 4], soit 5 possibilités. Le caractère quantitatif continu s'inscrit aussi dans un intervalle mais peut avoir une infinité de valeurs possibles, notamment avec les nombres décimaux. La grande différence entre ces deux caractères, c'est qu'un sert à compter/décompter tandis que l'autre cherche plus à mesurer.

Question 3 : Il existe plusieurs types de moyennes car chacune rend compte plus ou moins précisément d'un caractère que l'on souhaite mettre en valeur dans nos calculs, et toutes comportent leur lot de qualités et de défauts. Par exemple, la moyenne dite arithmétique est beaucoup plus sensible aux valeurs extrêmes.

La médiane a l'avantage de classer la population en deux catégories comprenant la même quantité de part et d'autre d'une valeur qu'on a fixée. Typiquement, le salaire médian permet de comprendre à combien de rémunération on doit être pour être plus riche que la moitié des Français. Elle n'est pas influencée par les valeurs extrêmes.

Le mode correspond à la valeur qui a le plus de chances de tomber dans une série statistique, il s'agit d'une moyenne de fréquence. Il peut être calculé par tous les types de variables, nominale, ordinale, discrète mais pas continue. Et si toutes les valeurs n'apparaissent qu'une seule fois, alors la série statistique ne peut avoir de mode.

Question 4 : L'intérêt principal de la médiale est qu'elle correspond plus ou moins à la médiane, sauf qu'elle en règle un des plus gros défauts, elle prend en compte les valeurs globales relatives de la population statistique. Ainsi, au lieu de seulement diviser la population en 2 selon une variable, il prend en compte le nombre total de la population et la divise en 2 pour que chaque partie ait une valeur égale. L'indice de Gini est un peu comme la mise en application graphique de la médiane

puisqu'elle a pour but de représenter dans un repère orthonormé l'effet de concentration d'une population. En lisant la courbe de l'indice de Gini, on comprend mieux si la concentration au sein d'une population est grande ou pas, et donc les effets que cela peut avoir sur la série statistique.

Question 5 : Pour mesurer la dispersion, l'écart à la moyenne n'est pas un bon moyen car il renvoie toujours à 0. Ainsi, en élevant toutes les valeurs au carré, on obtient un résultat positif qui nous permet de mieux interpréter les résultats même si les valeurs au carré ne nous aident pas. Alors, l'écart-type se retrouve en mettant au carré les valeurs obtenues avec la variance. De fait, on retombe dans l'unité statistique de base et on comprend dès lors beaucoup mieux l'écart à la moyenne.

L'étendue sert à mesurer l'amplitude totale des valeurs observées,

c'est-à-dire l'écart global entre les extrêmes d'un ensemble de données. Elle indique jusqu'où les données s'étalent.

Un quantile sert à découper une série de données ordonnées en parties égales afin d'analyser leur répartition. Les quantiles les plus utilisés sont les centiles (10 groupes créés), le quartile (4 groupes créés) et les centiles (100 groupes créés).

Une boîte de dispersion est une représentation graphique d'un traitement statistique permettant d'observer visuellement la répartition, la concentration et la variabilité d'une série de données. Pour l'interpréter, il faut prendre en compte 5 données importantes : le minimum, le premier quartile, la médiane, le second quartile et le maximum. Avec toutes ces données en tête, la boîte à moustaches en dit beaucoup plus sur la répartition statistique et s'impose comme le meilleur des graphiques pour cette catégorie.

Question 6 : Vérifier la symétrie d'une distribution, c'est voir si les données sont réparties de façon équilibrée autour de la moyenne ou de la médiane. On cherche donc à savoir si la distribution des valeurs est centrée ou décentrée. Et justement, l'asymétrie ou la symétrie des données joue un rôle sur l'interprétation de celle-ci et le choix des outils graphiques que l'on va utiliser pour les représenter. Ce sont les coefficients $\beta 1$ et $\beta 2$ de Pearson et de Fisher qui aident à calculer la symétrie ou l'asymétrie des données.

Séance 4. Les distributions statistiques

Questions de cour :

Question 1. Je dégage 3 critères pour choisir entre une distribution statistique avec variables discrètes ou avec variables continues :

- Il faut d'abord regarder la nature du phénomène étudié. Si celui-ci est dénombrable et distinct, alors on choisira la distribution statistique avec des variables discrètes.
- En interprétant les principales caractéristiques des données fournies. Veut-on lisser les données (variables continues) ou bien veut-on montrer une fréquence absolue (variables discrètes) ?
- À la quantité de paramètres des lois ayant forcément une influence sur celle que l'on va choisir pour notre distribution statistique.

Question 2. Selon moi, on retrouve deux grandes lois mathématiques très fréquemment utilisées en géographie pour aider au traitement statistique à grande échelle. La première est la loi de Zipf Mandelbrot car, en mettant en relation le classement des unités et leur décroissement au fil de ce rang, cette loi nous permet de modéliser des systèmes régionaux tout en tenant compte de la diversité des villes. La loi simple de Zipf permet quant à elle de détecter les cas de croissance macrocéphale et donc d'y palier si la croissance est trop concentrée.

Deuxièmement, la loi normale ou courbe de Gauss est aussi très utilisée en géographie car énormément de phénomènes suivent une distribution normale ou quasi-normale. Elle permet donc d'analyser la distribution spatiale, elle mesure la probabilité d'un événement, ou encore d'établir des classes ou intervalles statistiques.

Séance 5 : Les statistiques inférentielles

Questions de cour

Question 1 : Un échantillon est un modèle réduit de la population étudiée.udi chantillon est dit représentatif si sa structure (sociale, économique, géographique) reflète celle de la population de rérencer On ne peut souvent pas étudier toute la population, car cela coûterait trop cher, prendrait trop de temps ou serait matériellement impossible (ex : recenser toute une ville chaque semaine).

Question 2 : Un estimateur est une formule ou une règle statistique qui permet d'estimer un paramètre inconnu (comme la moyenne réelle). L'estimation est la valeur numérique concrète obtenue à partir de cet estimateur appliqué à l'échantillon.

Question 3 : L'intervalle de fluctuation décrit l'intervalle de valeur compris entre une borne supérieure et une borne inférieure de valeurs que peut prendre une proportion ou une moyenne d'échantillon dès lors que l'on lequel plusieurs fois. L'intervalle de confiance indique dans quelle interval de valeurs se trouve le vrai paramètre de la population avec une certaine probabilité de 95%

Question 4 : Un biais est une erreur systématique qui fait qu'un estimateur ne donne pas, en moyenne, la vraie valeur du paramètre. Il peut venir d'un mauvais échantillonnage, ou d'un modèle mal adapté.

Question 9 : La statistique inférentielle est une démarche statistique consistant à extrapoler à une population entière les propriétés mises en évidence sur un ensemble d'individus enquetés, appelé échantillon. De fait les critiques qui lui sont adressées sont plutôt adressées sont clairement légitimes. Elles remettent en cause l'extrapolation de résultat qui sont résultats échantillon justes mais qui ne peuvent l'être pleinement juste en élargissant la population étudiée

Théorie de l'échantillonnage

Question 1 La moyenne du nombre de personnes ayant répondu “Pour” est de 391. Le nombre de personnes ayant répondu “Contre” est de 416 et enfin le nombre de personne ayant répondu opinion est de 193

Question 2

Pour les fréquences cf Excel

Quest, cf. 3

Idem pour l’intervalle de fluctuation cf Excel

Quest, cf. 4

L’intervalle de fluctuation que nous avons calculé avec les moyennes et les effectifs de les échantillons a des de vérifier la représentativité des échantillon si on les échantillons l'ensemble de la population mère. Si la fréquence de l’échantillon se trouve dans l’intervalle de fluctuation, alors on peut en conclure que l'échantillon est bel et bien représentatif et donc que notre raisonnement est cohérent. Le cas échéant, il nous faudrait reconsidérer la population étudiée. En clair, il nous permet de valider ou d’infirmer le tirage effectué initialement.

Théorie de l’estimation

Question 1 : La somme de la ligne est égale à 1000 et les fréquences sont à consulter directement sur l’excel

Question 2 : Pour les fréquences cf Excel

Question 3 : Idem pour l’intervalle de fluctuation cf Excel

Séance 6. La statistique d'ordre des variables qualitatives

Question de cours

Question 1 Une statistique ordinale est une statistique qui s'appuie sur des données ordonnées, c'est-à-dire des données que l'on peut classer selon un ordre, sans que les écarts entre les valeurs aient nécessairement un sens quantitatif précis.

Question 2 Dans les classifications statistiques et spatiales, l'ordre décroissant est à privilégier.L'ordre décroissant permet de placer les éléments les plus importants en premier.

Il facilite la lecture des hiérarchies en mettant en évidence les valeurs dominantes. Il est particulièrement adapté aux analyses de type rang-taille, où l'on étudie la relation entre la taille d'un objet spatial et son rang.

Question 3. La corrélation des rangs mesure l'intensité et le sens de la relation entre deux classements alors que la concordance de classements évalue le degré d'accord global entre plusieurs classements.

Question 4. Les tests de Spearman et de Kendall sont deux tests de corrélation non paramétriques, basés sur les rangs. Ils servent à mesurer l'existence d'une relation monotone entre deux variables ordinales ou quantitatives non normales. L'un mesure la corrélation entre deux séries de rangs et compare les différences de rang entre observations et l'autre repose sur le nombre de paires concordantes et discordantes ainsi que mesure le degré d'accord entre deux classements.

Question 5. Les coefficients de Goodman-Kruskal et de Yule servent à mesurer l'intensité de l'association entre deux variables catégorielles, à partir de tableaux de contingence. Le coefficient de Goodman-Kruskal évalue la réduction d'erreur de prédiction d'une variable par une autre, tandis que le coefficient de Yule mesure la force et le sens de l'association entre deux variables dichotomiques. Ils sont utilisés lorsque les données ne sont ni numériques ni ordinales, mais purement qualitatives.

Commentaire sur les humanités numériques

Il semble bien qu'aujourd'hui les humanités numériques prennent de plus en plus de place dans la recherche scientifique. Dans un monde où tout semble de plus en plus connecté et où l'industrie de la tech s'impose comme le futur de l'humanité, cette discipline a tout intérêt à se développer au rythme des innovations technologiques qui ne cessent de faire évoluer notre quotidien. Elles sont tant d'outils précieux pour le chercheur qui, s'il ne s'adapte pas aux caractéristiques de son époque, ne peut la comprendre pleinement. Plus encore, ces outils numériques tels que nous avons appris à les manipuler dans le cours doivent servir de support à la recherche scientifique. À l'échelle du travail universitaire, cela ne fait que quelques dizaines d'années que le traitement statistique peut être réalisé dans de telles proportions. Les ordinateurs augmentant chaque année leur puissance de calcul, peut-on vraiment imaginer qu'un jour ils s'arrêtent ? Dans l'intérêt du chercheur, il ne le faudrait pas. Car plus l'ordinateur est puissant, plus le chercheur va pouvoir réaliser des études s'étendant sur une population toujours plus grande. Ainsi, les humanités numériques sont le prolongement de ce constat. Si dans un premier temps on peut trouver le terme oxymorique, humanité et numérique ne s'inscrivant pas du tout dans la même réalité, il semble bien qu'un croisement est possible. Et c'est le domaine de la recherche qui en est le théâtre car seul le chercheur peut également employer au maximum cet outil. Finalement, l'humanité numérique ne peut que se développer dans les prochaines années/décennies, et plus encore, elle peut constituer une forme d'aboutissement dans la recherche avec des traitements statistiques toujours plus importants et brassant une population toujours plus large. L'humanité, si elle cherche sa place dans le futur, la trouvera surement dans le numérique.

Pour ce qui est du cours qu'il nous a été donné de suivre ce semestre, je l'ai trouvé particulièrement difficile à comprendre sans véritable accompagnement Vous ne pouvez pas demander à des étudiants en master d'investir autant de temps et de ressources dans un domaine que très peu, et j'insiste sur le très peu, de personnes n'avaient pratiqué avant. Alors vous avez surement dû créer des vocations chez des personnes qui n'avaient jamais pratiqué Python avant et qui ont adoré le faire pendant vos cours. Pour autant, l'écrasante majorité des personnes ne sont pas dans ce cas. J'entends l'ambition que vous aviez de faire découvrir un langage informatique, aussi important soit-il, à des étudiants en master de géographie. Un outil qui par ailleurs devrait se développer grandement dans les prochaines années. Pour autant, quand on connaît le profil des géographes en France, c'est-à-dire très littéraire, où la plupart d'entre nous ont arrêté les mathématiques au lycée, certains dès la Seconde, on ne va pas reprendre tous les programmes de mathématiques

depuis le lycée pour comprendre une seule matière du semestre. Je n'ai assisté à aucun cours, je n'ai posé aucune question et pourtant j'ai réussi à rendre toutes les séances, ce qui montre que ce n'est pas impossible, même pour quelqu'un qui a arrêté les mathématiques en 3e. Pour autant, je ne vous cache pas mon énervement (le mot est faible), pendant les dizaines d'heures que j'ai passées à ce compte rendu. Des dizaines d'heures que j'ai passées sans même savoir si ce que je faisais était bien puisque je n'ai jamais fait de Python. C'est peut-être le pire dans cette histoire, c'est d'y avoir mis autant de temps pour probablement avoir une note médiocre car je suis bien conscient que ce que j'ai rendu l'est. Des heures passées à comprendre votre cours, littéralement des heures, avant que je sois résigné et que j'utilise l'intelligence artificielle pour m'aider car c'était le meilleur outil que j'avais pour avancer. Je pensais bien m'y connaitre en informatique, et c'est ce qui m'a donné une légère avance sur mes camarades, mais je n'imagine pas ceux qui ont de base des difficultés avec les ordinateurs. Vous nous avez donné votre cours de 600 pages et puis on était censés avoir toutes les clés pour réussir. Encore une fois, je ne remets pas en cause votre bonne volonté de nous faire réussir dans votre matière, la preuve vous avez rédigé un cours de 600 pages. Néanmoins, il faudra pour les années suivantes beaucoup plus d'accompagnement pour les élèves, revenir à la base et prendre le temps de tout expliquer. J'ai beaucoup appris, mais dans la souffrance et je pense que le cerveau est fait de telle façon que j'ai associé ces connaissances à un mauvais moment et donc je serai amené à tout oublier (ou en partie). Voilà comme je finirai ce commentaire sur le cours en lui-même, un mauvais moment. Votre bonne volonté est indéniable, maintenant à vous de revoir l'approche avec les étudiants en prenant en considération que la plupart n'y connaissent rien en informatique, et que personne ne peut investir autant de temps pour une matière qui nous a été imposée.

Voilà, j'ai peut-être été dur mais j'ai cherché à être constructive tout en vous montrant la frustration que j'ai eu pendant ce semestre.

Je n'ai pas pu intégrer les images générés par mes codages python car les ayant stockés sur mon environnement Linux de mon chromebook, au moment de rédiger leur commentaire, mon ordinateur les a supprimées de l'espace Linux. Je ne peux pas non plus ouvrir VS Code pour les générer de nouveaux.

https://github.com/Thomasdpy/Analyse-de-don-e---Rendu-Final---Thomas-Dupuy.git