Emilie Zekas

Questions de cours GEANDO

Séance 2. Les principes généraux de la statistique

1. La géographie entretient une relation tendue et paradoxale avec la statistique.
Elle produit des données massives, mais a longtemps méprisé les définitions mathématiques de la statistique, les considérant hors de son champ disciplinaire.
Pourtant, seul l'outil statistique permet d'étudier rigoureusement l'information géographique. La statistique est donc indispensable à la géographie, même si elle ne remplace pas le raisonnement explicatif.

2. Le hasard est avant tout une question philosophique.
En statistique (et donc en géographie) : on ne peut pas prévoir le détail des réalisations individuelles ; mais on peut dégager une tendance globale, une vraisemblance.
En géographie humaine, on ne peut pas prédire chaque action individuelle, mais on peut identifier des régularités statistiques. Le hasard n'interdit pas l'existence de tendances, il introduit seulement de la variabilité.

3. L'information géographique se divise en deux grandes séries statistiques :
- Les informations attributaires, qui comprennent la géographie humaine (population, caractéristiques économiques et sociales…) et la géographie physique (température, précipitations…)
- Les informations géométriques, qui représentent la forme, la surface ou la géométrie des objets spatiaux

4. La géographie a besoin de l'analyse de données pour structurer et résumer l'information massive, identifier des tendances globales, comparer des territoires ou des phénomènes, mesurer la fiabilité de l'information et dégager une structure dans l'aléatoire.

5. La statistique descriptive décrit et résume les données et met de l'ordre dans les distributions. Elle utilise des tableaux, des graphiques et des indicateurs pour préparer les comparaisons et les prédictions. La statistique explicative cherche à relier des variables en modélisant une variable à expliquer à partir d'autres variables explicatives par le biais de modèles tels que l'ANOVA et la régression linéaire.

6. Pour visualiser les variables quantitatives continues, on utilise le plus souvent des histogrammes et pour des variables qualitatives, c'est plutôt une représentation sectorielle. Le choix de représentation dépend du nombre de variables, du type de variable et de l'objectif de la représentation.

7. Il y a les méthodes descriptives (ACP, AFC, classification ascendante hiérarchique, nuées dynamiques), les méthodes explicatives (régression simple et multiple, ANOVA, régression

logistique, analyse discriminante) et les méthodes de prévision qui analyse des séries chronologiques.

8. a) Une population statistique correspond à un ensemble, au sens mathématique du terme. C'est l'ensemble des individus étudiés.
b) Un individu statistique (unité statistique ou éléments du niveau inférieur) correspond à un élément de la population statistique.
c) Les caractères statistiques sont les caractéristiques (c'est-à-dire les particularités) de l'individu pris parmi la population statistique sur laquelle l'analyse statistique porte.
d) Les modalités statistiques correspondent aux valeurs prises par un caractère. Elles sont exhaustives et disjointes.
Les types de caractères sont quantitatifs discrets, quantitatifs continus, qualitatifs ordinaux et qualitatifs nominaux. Il n'y a pas de hiérarchie stricte entre ces types de caractères mais une différence d'usage statistique.

9. Une aptitude de classe se mesure en faisant la différence entre la valeur maximum et minimum de la classe. La densité se mesure en prenant l'effectif de la classe rapportée à son amplitude.

10. Les formules de Sturges et de Yule servent à déterminer le nombre de classes lors de la discrétisation et évitent un découpage mal adapté, trop fin ou trop grossier.

11. Un effectif correspond au nombre d'apparition d'une variable dans une population donnée. La fréquence est le rapport entre l'effectif de la classe divisé par l'effectif total de cette classe. La fréquence cumulée est la somme des effectifs associés aux valeurs du caractère qui sont inférieures ou égales à cette même modalité. La distribution statistique représente l'ensemble des modalités ou classes d'un caractère avec leurs effectifs ou fréquences, décrivant la répartition des données.

Séance 3. Les paramètres statistiques élémentaires

1. Le caractère qualitatif est le plus général parce qu'il décrit une catégorie ou une modalité sans mesure numérique directe (type d'activité, catégorie socio-professionnelle…). Tout caractère quantitatif peut être vu comme un qualitatif ordonné, mais l'inverse n'est pas vrai.

2. Les valeurs du quantitatif discret sont finies ou dénombrables (nombre d'élèves, nombre de logements) et les modalités sont distinctes et séparées. Les valeurs du quantitatif continu appartiennent à un intervalle réel (la température, les revenus, l'altitude) et théoriquement infini. Il faut les distinguer parce que les formules changent, somme pour le discret, intégrale pour le continu. Les représentations graphiques diffèrent aussi de diagrammes à histogrammes et enfin les méthodes de calcul des paramètres ne sont pas identiques (moyenne, médiane, quantiles).

3. Paramètres de position

Il existe plusieurs types de moyenne parce que les variables n'ont pas la même nature (discrète, continue…), les situations d'analyse sont différentes (vitesse, surfaces, taux…) et certaines moyennes sont plus adaptées à certains phénomènes.
La médiane n'est pas influencée par les valeurs extrêmes. Elle permet de mieux résumer les distributions dissymétriques et partage la population en deux parties égales.
Un mode est calculable lorsqu'une modalité possède un effectif maximal (discret) ou une densité maximale (continu).

4. Paramètres de concentration
La médiale partage la masse totale du caractère en deux parts égales et met en évidence la concentration de la variable. L'indice C de Gini mesure l'écart entre médiane et médiale, qualifie le degré d'inégalité dans la distribution et est visualisé par la courbe de Lorenz.

5. Paramètres de dispersion
— Pourquoi calculer une variance à la place de l'écart à la moyenne? Pourquoi la remplacer par l'écart type? L'écart à la moyenne peut être positif ou négatif, donc sa somme est nulle. La variance utilise la carré des écarts, ce qui élimine les signes et possède de bonnes propriétés mathématiques. L'écart type est utilisé car il est exprimé dans la même unité que la moyenne et il est plus interprétable.
L'étendue donne une première idée de la dispersion et est simple à calculer.
Un quantile sert à découper une série ordonnée en parts égales et à analyser une répartition interne des données. Les plus utilisés sont les quartiles, les déciles et les centiles.
Une boîte de dispersion sert à résumer graphiquement une distribution, à comparer plusieurs séries et à repérer la dispersion et la dissymétrie. La boîte correspond aux quartiles Q1 à Q3, le trait central correspond à la médiane et les moustaches sont les valeurs extrêmes.

6. Paramètres de forme
Les moments centrés sont calculés par rapport à la moyenne et les moments absolus sont calculés par rapport à l'origine. Les moments centrés sont plus informatifs : ordre 2, la variance, ordre 3, la dissymétrie et ordre 4, l'aplatissement. Ils permettent de caractériser la forme de la distribution.
Il faut vérifier la symétrie de la distribution pour savoir si la moyenne est représentative et pour comprendre la structure de la distribution. Pour cela, il faut comparer la moyenne, la médiane et le mode et calculer le coefficient de Pearson ou Fisher. Si le résultat est supérieur à 0, alors il y a une dissymétrie positive, si c'est en-dessous de 0, alors elle est négative et si elle est égale à 0, alors la distribution est symétrique.

Séance 4. Les distributions statistiques

1. Le choix dépend principalement de la nature du phénomène étudié et de la précision de l'information. On doit d'une part prendre en compte la nature du caractère. Pour une variable discrète, il faut qu'elle soit dénombrable ou finie comme le nombre d'habitants, d'accidents ou d'hôpitaux. D'autre part, la variable continue peut appartenir à un intervalle réel comme un

revenu, la température, la densité. Le mode de mesure est aussi différent que ce soit le comptage pour une variable discrète et la mesure ou une estimation de la variable continue. Le volume de la donnée est aussi un critère de choix, avec les valeurs continues ayant souvent beaucoup plus de données, devant être regroupées en classe alors que les valeurs discrètes peuvent restées telles quelles. Enfin, l'objectif de l'analyse diffère. Si on veut faire une analyse simple, descriptive, alors on utilise du discret mais si on veut analyser des tendances, de la modélisation ou des lois théoriques, on va préférer le continu.

2. Pour moi, la loi normale est la plus utilisée car elle permet de modéliser des phénomènes résultant de nombreux facteurs indépendants. De plus, elle est symétrique autour de la moyenne, qui est aussi égale à la médiane et au mode. Elle permet l'utilisation de l'écart type et des intervalles de confiance. Pour une localisation ponctuelle des phénomènes naturels, en géographie physique, on peut aussi utiliser la loi de Poisson pour des évènements rares et discrets. Elle porte sur le nombre d'occurrences dans un espace ou un temps donné. La loi binomiale modélise également à deux issues possibles et est utilisée lorsque la probabilité est connue notamment dans le cas de la présence ou l'absence d'un équipement ou la réussite/échec de celui-ci. Enfin, en géographie, les lois de concentration sont aussi très utilisées pour décrire des distributions très dissymétriques surtout en géographie humaine comme pour la répartition des revenus, la taille des villes ou la concentration foncière. Elles sont liées à l'étude des inégalités spatiales.

Séance 5. Les statistiques inférentielles

1. L'échantillonnage consiste à prélever une partie (un échantillon) d'individus dans une population mère afin d'obtenir des informations sur les paramètres inconnus de cette population. Cette méthode de statistique inférentielle permet de tirer des conclusions générales à partir d'un nombre limité d'observations. L'utilisation de la population entière est souvent impossible ou irréaliste en raison de sa taille, de son coût ou du temps nécessaire à son étude. Les méthodes d'échantillonnage peuvent être aléatoires (tirages équiprobables, avec ou sans remise), non aléatoires (échantillonnage systémique ou par quotas), ou par les méthodes de Monte Carlo, reposant sur des tirages aléatoires répétés. Le choix d'une méthode dépend de la représentativité recherchée, des contraintes pratiques et du degré de précision attendu.

2. Un estimateur est une variable aléatoire, construite à partir des observations, destinée à approcher un paramètre inconnu de la population mère. Il dépend donc du hasard lié à l'échantillonnage. Une estimation est la valeur numérique obtenue lorsque l'on calcule cet estimateur sur un échantillon donné.

3. L'intervalle de fluctuation suppose que le paramètre de la population est connu et décrit la variabilité possible de l'échantillon autour de ce paramètre. L'intervalle de confiance encadre lui, un paramètre inconnu de la population à partir d'un estimation calculée sur l'échantillon, avec un risque a.

4. Un biais est une erreur systématique liée à un estimateur. Il correspond à la différence entre l'espérance mathématique de l'estimateur et la valeur réelle du paramètre. Un estimateur est dit sans biais lorsque son espérance est égale au paramètre à estimer.

5. Une statistique qui contient toute l'information sur le paramètre est appelée statistique exhaustive. Quand on travaille sur la population totale (ou de très grandes bases de données), il faut réduire l'information. Cela fait écho aux données massives, où l'on cherche à résumer efficacement des bases très volumineuses sans perte d'information essentielle.

6. Les choix autour d'un estimateur visent à minimiser l'erreur quadratique moyenne (ERQM), à obtenir un estimateur sans biais, de variance minimale (précision) et à assurer la convergence lorsque la taille de l'échantillon augmente tout en conservant le maximum de l'information.

7. On peut estimer un paramètre sans biais de variance minimale avec les méthodes de l'utilisation de statistiques exhaustives ou grâce à l'analyse de l'information de Fisher qui permet de mesurer la précision intrinsèque d'un estimateur et de fixer une borne inférieure sur sa variance via l'inégalité de Cramer. Lorsque les données contiennent des valeurs aberrantes, on peut utiliser des estimations plus ponctuelles (moyenne, variance, proportion), des estimations par intervalle de confiance, la méthode de vraisemblance, qui choisit la valeur du paramètre, rendant les observations plus probables ou enfin par la méthode du bootstrap, fondée sur le rééchantillonnage de l'échantillon principal. Le choix dépend du modèle statistique , de la taille de l'échantillon et des propriétés souhaitées.

8. Les tests statistiques servent à prendre une décision sur une hypothèse concernant une population à partir d'un échantillon, avec un risque mesuré. Les tests statistiques existants sont le test de Student, le test du Khi2, le test de Fisher-Snedecor, les tests sur les proportions (normale ou de Poisson). Il y a aussi des tests non paramétriques comme celui de Mann-Whitney, de Wilcoxon ou le test sur les signes.
Pour créer un test, il faut formuler une hypothèse, choisir une statistique de test, connaître sa loi sous H0, fixer un risque d'erreur a, comparer la valeur observée à la zone critique des tables statistiques puis enfin conclure par acceptation ou rejet de H0.

9. La statistique inférentielle est critiquable puisqu'elle repose sur des hypothèses probabilistes, des fluctuations d'échantillonnage, des résultats qui peuvent être fortement biaisés par des valeurs aberrantes et les intervalles de confiance qui ne garantissent pas la certitude absolue. Par contre, il ne faut pas oublier que cela repose sur des marges d'erreur et des biais possibles. On peut aussi améliorer la fiabilité en augmentant la taille de l'échantillon mais il y a toujours des limites de représentativité qui sont néanmoins quantifiées avec le niveau de confiance et le risque a. La statistique inférentielle reste par ailleurs indispensable quand il est impossible d'avoir accès à la population totale.

Séance 6. La statistique d'ordre des variables qualitatives

1. La statistique ordinale est un statistique qui repose sur le classement d'objets, d'individus ou d'entités géographiques selon un ordre. Elle ne s'intéresse pas directement aux valeurs elles-mêmes, mais à leur rang dans une série ordonnée. Elle s'oppose à la statistique nominale, qui classe les individus en catégories sans ordre intrinsèque. Elle utilise des variables qualitatives ordinales. Cela peut créer une hiérarchie spatiale car les classements révèlent des positions de dominantes ou subordonnées, des dynamiques de montée, stagnation ou déclassement mais aussi des structures hiérarchiques entre les villes, régions ou territoires.

2. L'ordre croissant (aussi ordre naturel) est à privilégier dans les classifications car il facilite la lecture des classements, l'identification des valeurs aberrantes et les valeurs extrêmes.

3. La corrélation des rangs vise à mesurer le degré de dépendance entre deux variables ordinales, chacune exprimée sous forme de rangs. Elle mesure la relation statistique entre deux classements. La concordance de classements mesure le degré de similarité globale entre plusieurs classements, en s'appuyant sur le nombre de paires concordantes et discordantes.

4. Le test de Spearman mesure une corrélation basée sur les rangs qui repose sur le coefficient de corrélation des rangs et s'appuie sur les différences entre les rangs, particulièrement adapté à la comparaison de deux classements. Le test de Kendall mesure une concordance fondée sur les relations d'ordre entre le nombre de paires concordantes et discordantes et peut être généralisée à plusieurs classements.

5. Le coefficient de Goodman-Kruskal sert à mesurer l'association entre deux variables qualitatives ordinales, en comparant le nombre de paires concordantes et discordantes. Il montre le surplus de concordances par rapport aux discordances et varie entre -1 et 1. Le coefficient de Yule est un cas particulier du coefficient Goodman-Kruskal car il s'applique uniquement aux tableaux de contingence 2x2 et mesure l'intensité et le sens de l'association entre deux variables qualitatives binaires.

Réflexion sur les sciences des données et les humanités numériques

Le cours m'a permis de prendre en main le langage Python et certaines bibliothèques essentielles comme Pandas. J'ai appris à manipuler des jeux de données, à calculer des statistiques descriptives, à produire des visualisation et à analyser des données. Ces exercices ont montré que l'analyse de données repose autant sur la compréhension des données que sur les outils utilisés.

Néanmoins, plusieurs difficultés ont été rencontrées au cours du parcours. Sur le plan technique, la syntaxe de python et la gestion des erreurs ont constitué un premier obstacle. Ces problèmes ont été progressivement surmontés avec la pratique et l'analyse des messages d'erreur. Le démarrage a aussi été compliqué avec l'installation de tous les outils nécessaires à la réalisation de ce cours et aux complications engendrés par nos machines personnelles. La manipulation des données a également posé problème, en particulier le nettoyage des valeurs manquantes ou incohérentes. D'autre part, l'interprétation des résultats était importante afin de distinguer une corrélation d'une causalité.

D’un point de vue plus général, ce cours m’a amené à réfléchir au lien entre les sciences des données et des humanités numériques. Cela permet de traiter de grands ensembles d’informations afin de faire émerger des structures, des hiérarchies ou des tendances dans des phénomènes sociaux, culturels ou géographiques. J’ai aussi compris l’importance de la corréler avec l’analyse qualitative et la complémentarité des deux. En effet, l’interprétation humaine reste indispensable pour donner du sens aux résultats.

Finalement, ce cours m’a permis d’acquérir des connaissances sur le plan méthodologique et technique et de mieux analyser les enjeux des données numériques et leur apport aux humanités, notamment à la géographie, en gardant une posture critique face aux résultats produits.